

The Proteomics Protocols Handbook

The Proteomics Protocols Handbook

Edited by

John M. Walker

University of Hertfordshire, Hatfield, UK

HUMANA PRESS  TOTOWA, NEW JERSEY

© 2005 Humana Press Inc.
999 Riverview Drive, Suite 208
Totowa, New Jersey 07512

www.humanapress.com

All rights reserved. No part of this book may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, microfilming, recording, or otherwise without written permission from the Publisher.

The content and opinions expressed in this book are the sole work of the authors and editors, who have warranted due diligence in the creation and issuance of their work. The publisher, editors, and authors are not responsible for errors or omissions or for any consequences arising from the information or opinions presented in this book and make no warranty, express or implied, with respect to its contents.

This publication is printed on acid-free paper. 
ANSI Z39.48-1984 (American Standards Institute)

Permanence of Paper for Printed Library Materials.

Production Editor: Tracy Catanese
Cover design by Patricia F. Cleary

For additional copies, pricing for bulk purchases, and/or information about other Humana titles, contact Humana at the above address or at any of the following numbers: Tel.: 973-256-1699; Fax: 973-256-8341; E-mail: humana@humanapr.com; or visit our Website: www.humanapress.com

Photocopy Authorization Policy:

Authorization to photocopy items for internal or personal use, or the internal or personal use of specific clients, is granted by Humana Press Inc., provided that the base fee of US \$30.00 per copy is paid directly to the Copyright Clearance Center at 222 Rosewood Drive, Danvers, MA 01923. For those organizations that have been granted a photocopy license from the CCC, a separate system of payment has been arranged and is acceptable to Humana Press Inc. The fee code for users of the Transactional Reporting Service is: [1-58829-343-2/05 \$30.00].

Printed in the United States of America. 10 9 8 7 6 5 4 3 2 1

e-ISBN: 1-59259-890-0

Library of Congress Cataloging in Publication Data

Proteomics protocols handbook / edited by John M. Walker.

p. ; cm.

Includes bibliographical references and index.

ISBN 1-58829-343-2 (hardcover : alk. paper) -- ISBN 1-58829-593-1 (pbk. : alk. paper)

1. Proteomics--Handbooks, manuals, etc.

[DNLM: 1. Proteomics--methods. QU 58.5 P9675 2005] I. Walker, John M., 1948-

QP551.P75675 2005

572'.6--dc22

2004016126

Preface

Recent developments in the field of proteomics have revolutionized the way that proteins, and their contribution to cellular functions, are studied. The subsequent increased understanding of the mechanisms of cellular function and malfunction will have particular impact in the area of medical research, where disease processes will be better understood, many new (protein) therapeutic targets identified, and novel therapeutic agents developed. At the basic research level, phenotype will be explained in terms of cellular mechanisms.

The completion of the sequences of an ever-widening range of genomes—not least of all, the human genome—has provided the molecular biologist with a wealth of data that needs to be analyzed and interpreted. For a variety of reasons (including alternative mRNA splicing, varying translational stop/start sites, frameshifting, and the inability to deduce posttranslational modifications), complete sequences of genomes are insufficient to elucidate the protein components of cells. The focus of attention has therefore turned to directly examining these protein components as the means of understanding cell function, as well as the cellular changes involved in disease states. However, the wealth of gene sequencing data now available has produced a glut of information that challenges the protein chemist to develop new tools to utilize this flood of genomic data.

From the beginning, the cornerstone of proteomics has been the use of two-dimensional gel electrophoresis to compare proteomes of different tissues (for example, normal and diseased tissue) with the subsequent identification of protein differences by the use of mass spectrometry and database searching. These still remain valuable techniques and receive appropriate coverage in this book. However, the term proteomics now encompasses a range of newly developed methodologies for determining the structure and function of a protein. I have therefore included in *The Proteomics Protocols Handbook* a number of novel mass spectrometry and LC-MS techniques, protein array technology, new bioinformatics tools, and the range of techniques central to structural and functional proteomics that are needed to deduce the function of newly discovered protein sequences. The use of these techniques, and no doubt further ones that will be developed in the coming years, will lead to achieving the ultimate goal of proteomics, namely to catalog the identity and function of all proteins in living organisms.

The Proteomics Protocols Handbook should prove a valuable resource for molecular biologists, protein chemists, clinical/medical researchers, structural chemists/biochemists, and microbiologists, as well as those involved in bioinformatics and structural/functional genomics.

John M. Walker

Contents

Preface	vii
Contributors	xiii
1 Extraction and Solubilization of Proteins for Proteomic Studies <i>Richard M. Leimgruber</i>	1
2 Preparation of Bacterial Samples for 2-D PAGE <i>Brian Berg Vandahl, Gunna Christiansen, and Svend Birkelund</i>	19
3 Preparation of Yeast Samples for 2-D PAGE <i>Joakim Norbeck</i>	27
4 Preparation of Mammalian Tissue Samples for Two-Dimensional Electrophoresis <i>Frank A. Witzmann</i>	31
5 Differential Detergent Fractionation of Eukaryotic Cells <i>Melinda L. Ramsby and Gregory S. Makowski</i>	37
6 Serum or Plasma Sample Preparation for Two-Dimensional Gel Electrophoresis <i>Anthony G. Sullivan, Stephen Russell, Henry Brzeski, Richard I. Somiari, and Craig D. Shriver</i>	49
7 Preparation of Plant Protein Samples for 2-D PAGE <i>David W. M. Leung</i>	55
8 Laser-Assisted Microdissection in Proteomic Analyses <i>Darrell L. Ellsworth, Stephen Russell, Brenda Deyarmin, Anthony G. Sullivan, Henry Brzeski, Richard I. Somiari, and Craig D. Shriver</i>	59
9 Purification of Cellular and Organelle Populations by Fluorescence-Activated Cell Sorting for Proteome Analysis <i>William L. Godfrey, Colette J. Rudd, Sujata Iyer, and Diether Recktenwald</i>	67
10 Purification of Nucleoli From Lymphoma Cells and Solubilization of Nucleolar Proteins for 2-DE Separation <i>Régis Dieckmann, Yohann Couté, Denis Hochstrasser, Jean-Jacques Diaz, and Jean-Charles Sanchez</i>	79
11 Prefractionation of Complex Protein Mixture for 2-D PAGE Using Reversed-Phase Liquid Chromatography <i>Volker Badock and Albrecht Otto</i>	87
12 Fractionation of Complex Proteomes by Microscale Solution Isoelectrofocusing Using ZOOM™ IEF Fractionators to Improve Protein Profiling <i>Xun Zuo, Ki-Boom Lee, and David W. Speicher</i>	97
13 Large-Format 2-D Polyacrylamide Gel Electrophoresis <i>Henry Brzeski, Stephen Russell, Anthony G. Sullivan, Richard I. Somiari, and Craig D. Shriver</i>	119
14 Analysis of Membrane Proteins by Two-Dimensional Gels <i>Michael Fountoulakis</i>	133

15	2-D PAGE of High-Molecular-Mass Proteins <i>Masamichi Oh-Ishi and Tadakazu Maeda</i>	145
16	Using Ultra-Zoom Gels for High-Resolution Two-Dimensional Polyacrylamide Gel Electrophoresis <i>Sjouke Hoving, Hans Voshol, and Jan van Oostrum</i>	151
17	NEpHGE and pI Strip Proteomic 2-D Gel Electrophoretic Mapping of Lipid-Rich Membranes <i>Steven E. Pfeiffer, Yoshihide Yamaguchi, Cecilia B. Marta, Rashmi Bansal, and Christopher M. Taylor</i>	167
18	Silver Staining of 2-D Gels <i>Julia Poland, Thierry Rabilloud, and Pranav Sinha</i>	177
19	Zn ²⁺ Reverse Staining Technique <i>Carlos Fernandez-Patron</i>	185
20	Multiplexed Proteomics Technology for the Fluorescence Detection of Glycoprotein Levels and Protein Expression Levels Using Pro-Q® Emerald and SYPRO® Ruby Dyes <i>Birte Schulenberg and Wayne F. Patton</i>	193
21	Multiplexed Proteomics Technology for the Fluorescence Detection of Phosphorylation and Protein Expression Levels Using Pro-Q® Diamond and SYPRO® Ruby Dyes <i>Birte Schulenberg, Terrie Goodman, Thomas H. Steinberg, and Wayne F. Patton</i>	201
22	Sensitive Quantitative Fluorescence Detection of Proteins in Gels Using SYPRO® Ruby Protein Gel Stain <i>Birte Schulenberg, Nancy Ahnert, and Wayne F. Patton</i>	209
23	Rapid, Sensitive Detection of Proteins in Minigels With Fluorescent Dyes: <i>Coomassie Fluor Orange, SYPRO® Orange, SYPRO Red, and SYPRO Tangerine Protein Gel Stains</i> <i>Thomas H. Steinberg, Courtenay R. Hart, and Wayne F. Patton</i>	215
24	Differential In-Gel Electrophoresis in a High-Throughput Environment <i>Richard I. Somiari, Stephen Russell, Stella B. Somiari, Anthony G. Sullivan, Darrell L. Ellsworth, Henry Brzeski, and Craig D. Shriver</i>	223
25	Statistical Analysis of 2-D Gel Patterns <i>Françoise Seillier-Moiseiwitsch</i>	239
26	2-DE Databases on the World Wide Web <i>Christine Hoogland, Khaled Mostaguir, and Ron D. Appel</i>	259
27	Computer Analysis of 2-D Images <i>Patricia M. Palagi, Daniel Walther, Gérard Bouchet, Sonja Voordijk, and Ron D. Appel</i>	267
28	Comparing 2-D Electrophoretic Gels Across Internet Databases: An Open Source Application <i>Peter F. Lemkin, Gregory C. Thornwall, and Jai Evans</i>	279
29	Sample Cleanup by Solid-Phase Extraction/Pipet-Tip Chromatography <i>Alastair Aitken</i>	307
30	Protein Identification by In-Gel Digestion and Mass Spectrometric Analysis <i>Michele Learmonth and Alastair Aitken</i>	311

31	Peptide Sequences of 2-D Gel-Separated Protein Spots by Nanoelectrospray Tandem Mass Spectrometry <i>Alastair Aitken</i>	315
32	Identification of Proteins by MALDI-TOF MS <i>Alastair Aitken</i>	319
33	Sequencing of Tryptic Peptides Using Chemically Assisted Fragmentation and MALDI-PSD <i>John Flensburg and Maria Liminga</i>	325
34	The <i>In Situ</i> Characterization of Membrane-Immobilized 2-D PAGE-Separated Proteins Using Ink-Jet Technology <i>Patrick W. Cooley, Janice L. Joss, Femia G. Hopwood, Nichole L. Wilson, and Andrew A. Gooley</i>	341
35	Protein Identification by Peptide Mass Fingerprinting <i>Alastair Aitken</i>	355
36	Analysis of the Proteomes in Human Tissues by In-Gel Isoelectric Focusing and Mass Spectrometry <i>Francesco Giorgianni and Sarka Beranova-Giorgianni</i>	367
37	Liquid Chromatography Coupled to MS for Proteome Analysis <i>Alastair Aitken</i>	375
38	Quantitative Analysis of Proteomes and Subproteomes by Isotope-Coded Affinity Tags and Solid-Phase Glycoprotein Capture <i>Eugene Yi, Hui Zhang, Kelly Cooke, Ruedi Aebersold, and David R. Goodlett</i>	385
39	Amino Acid-Coded Mass Tagging for Quantitative Profiling of Differentially Expressed Proteins and Modifications in Cells <i>Xian Chen</i>	393
40	Mass-Coded Abundance Tagging for Protein Identification and Relative Abundance Determination in Proteomic Experiments <i>Gerard Cagney and Andrew Emili</i>	407
41	Virtual 2-D Gel Electrophoresis by MALDI Mass Spectrometry <i>Angela K. Walker, Gary Rymar, and Philip C. Andrews</i>	417
42	Identification of Posttranslational Modification by Mass Spectrometry <i>Alastair Aitken</i>	431
43	Approaches to the O-Glycoproteome <i>Franz-Georg Hanisch and Stefan Müller</i>	439
44	Identification of Protein Phosphorylation Sites by Mass Spectrometry <i>Alastair Aitken</i>	459
45	Quantitative Analysis of Protein Phosphorylation Status and Protein Kinase Activity on Microassays Using Pro-Q™ Diamond Dye Technology <i>Karen Martin and Wayne F. Patton</i>	467
46	New Challenges and Strategies for Multiple Sequence Alignment in the Proteomics Era <i>Julie D. Thompson and Olivier Poch</i>	475
47	The Clustal Series of Programs for Multiple Sequence Alignment <i>Julie D. Thompson</i>	493

48	FASTA Servers for Sequence Similarity Search <i>Biju Issac and Gajendra P. S. Raghava</i>	503
49	Protein Sequence Analysis and Domain Identification <i>Chris P. Ponting and Ewan Birney</i>	527
50	Mammalian Genes and Evolutionary Genomics <i>Leo Goodstadt and Chris P. Ponting</i>	543
51	Computational Identification of Related Proteins: <i>BLAST, PSI-BLAST, and Other Tools</i> <i>Qunfeng Dong and Volker Brendel</i>	555
52	Protein Identification and Analysis Tools on the ExPASy Server <i>Elisabeth Gasteiger, Christine Hoogland, Alexandre Gattiker, Séverine Duvaud, Marc R. Wilkins, Ron D. Appel, and Amos Bairoch</i>	571
53	Protein Sequence Databases <i>Michele Magrane, Maria Jesus Martin, Claire O'Donovan, and Rolf Apweiler</i>	609
54	<i>In Silico</i> Characterization of Proteins: <i>InterPro and Proteome Analysis</i> <i>Nicola Jane Mulder, Manuela Pruess, and Rolf Apweiler</i>	619
55	Computational Prediction of Protein–Protein Interactions <i>Anton J. Enright, Lucy Skrabaneck, and Gary D. Bader</i>	629
56	The Yeast Two-Hybrid System for Detecting Interacting Proteins <i>Ilya G. Serebriiskii, Erica A. Golemis, and Peter Uetz</i>	653
57	Antibody–Affinity Purification to Detect Interacting Proteins <i>Sonia Navarro and Lucio Comai</i>	683
58	Biomolecular Interaction Analysis Coupled With Mass Spectrometry to Detect Interacting Proteins <i>Setsuko Hashimoto, Toshiaki Isobe, and Tohru Natsume</i>	689
59	Assessment of Antibody–Antigen Interaction Using SELDI Technology <i>Li-Shan Hsieh, Ramy Moharram, Emilia Caputo, and Brian M. Martin</i>	699
60	Protein and Peptide Microarray-Based Assay Technology <i>Scott T. Clarke</i>	709
61	Production of Protein Microarrays Using Robotic Pin Printing Technologies <i>Ye Fang, Ann M. Ferrie, and Fang Lai</i>	723
62	PCR-Directed Protein <i>In Situ</i> Arrays <i>Joe Boutell and Mingyue He</i>	735
63	Site-Specific Immobilization of Proteins in a Microarray <i>Yee-Peng R. Lue, Su-Yin D. Yeo, Lay-Pheng Tan, Grace Y. J. Chen, and Shao Q. Yao</i>	743
64	A Guide to Protein Interaction Databases <i>Tiffany B. Fischer, Melissa Paczkowski, Michael F. Zettel, and Jerry Tsai</i>	753
65	Deriving Function From Structure: Approaches and Limitations <i>Annabel E. Todd</i>	801
66	Comparative Protein Structure Modeling <i>M. S. Madhusudhan, Marc A. Marti-Renom, Narayanan Eswar, Bino John, Ursula Pieper, Rachel Karchin, Min-Yi Shen, and Andrej Sali</i>	831
67	Classification of Protein Sequences and Structures <i>S. Rackovsky</i>	861

68	How to Use Protein 1-D Structure Predicted by PROFphd <i>Burkhard Rost</i>	875
69	Classification of Protein Folds <i>Robert B. Russell</i>	903
70	Protein Threading <i>Andrew E. Torda</i>	921
71	High-Throughput Crystallography for Structural Proteomics <i>Jeff Yon, Mladen Vinković, and Harren Jhoti</i>	939
72	Automated High-Throughput Protein Crystallization <i>Arezou Azarani</i>	955
73	NMR-Based Structure Determination of Proteins in Solution <i>Andrzej Ejchart and Igor Zhukov</i>	967
	Index	983

Contributors

RUEDI AEBERSOLD • *Institute for Systems Biology, Seattle, WA*

NANCY AHNERT • *Molecular Probes Inc., Eugene, OR*

ALASTAIR AITKEN • *School of Biomedical and Clinical Laboratory Sciences, University of Edinburgh, UK*

PHILIP C. ANDREWS • *Michigan Proteome Consortium, University of Michigan, Ann Arbor, MI*

RON D. APPEL • *Swiss Institute of Bioinformatics, University and Geneva University Hospital, Geneva, Switzerland*

ROLF APWEILER • *European Bioinformatics Institute, Cambridge, UK*

AREZOU AZARANI • *Protagen Consulting, San Jose, CA*

GARY D. BADER • *Computational Biology Center, Memorial Sloan-Kettering Cancer Center, New York, NY*

VOLKER BADOCK • *Max Delbrueck Center for Molecular Medicine, Berlin, Germany*

AMOS BAIROCH • *Swiss Institute of Bioinformatics, University of Geneva, Geneva, Switzerland*

RASHMI BANSAL • *Department of Neuroscience, University of Connecticut Medical School, Farmington, CT*

SARKA BERANOVA-GIORGIANNI • *Department of Pharmaceutical Sciences, University of Tennessee Health Science Center, Memphis, TN*

SVEND BIRKELUND • *Department of Medical Microbiology and Immunology, University of Aarhus, Denmark, Loke Diagnostics ApS, Denmark*

EWAN BIRNEY • *European Bioinformatics Institute, Cambridge, UK*

GÉRARD BOUCHET • *Swiss Institute of Bioinformatics, Geneva, Switzerland*

JOE BOUTELL • *Discerna Ltd., Cambridge, UK*

VOLKER BRENDL • *Department of Genetics, Development and Cell Biology, Department of Statistics, Iowa State University, Ames, IA*

HENRY BRZESKI • *Functional Genomics and Proteomics Unit, Windber Research Institute, Windber, PA*

GERARD CAGNEY • *Department of Medical Research, University of Toronto, Canada*

EMILIA CAPUTO • *Institute of Genetics and Biophysics, Naples, Italy*

GRACE Y. J. CHEN • *Department of Biological Sciences, Department of Chemistry, National University of Singapore, Republic of Singapore*

XIAN CHEN • *Biosciences Division, Los Alamos National Laboratory, Los Alamos, NM*

GUNNA CHRISTIANSEN • *Department of Medical Microbiology and Immunology, University of Aarhus, Denmark*

SCOTT T. CLARKE • *Molecular Probes Inc., Eugene, OR*

LUCIO COMAI • *Department of Molecular Microbiology and Immunology, Keck School of Medicine, University of Southern California, Los Angeles, CA*

KELLY COOKE • *Institute for Systems Biology, Seattle, WA*

PATRICK W. COOLEY • *MicroFab Technologies Inc., Plano, TX*

YOHANN COUTÉ • *Centre de Génétique Moléculaire et Cellulaire, France*

- BRENDA DEYARMIN • *Clinical Breast Care Project, Windber Research Institute, Windber, PA*
- JEAN-JACQUES DIAZ • *Centre de Genetique Moléculaire et Cellulaire, France*
- RÉGIS DIECKMANN • *Biomedical Proteomics Research Group, Central Clinical Chemistry Laboratory, Geneva University Hospital, Geneva, Switzerland*
- QUNFENG DONG • *Department of Genetics, Development and Cell Biology, Iowa State University, Ames, IA*
- SÉVERINE DUVAUD • *Swiss Institute of Bioinformatics, Geneva, Switzerland*
- ANDRZEJ EJCHART • *Institute of Biochemistry and Biophysics, Poland*
- DARRELL L. ELLSWORTH • *Cardiovascular Disease Research Program and Clinical Breast Care Project, Windber Research Institute, Windber, PA*
- ANDREW EMILI • *Banting and Best Departments of Medical Research, University of Toronto, Canada*
- ANTON J. ENRIGHT • *Computational Biology Center, Memorial Sloan-Kettering Cancer Center, New York, NY*
- NARAYANAN ESWAR • *Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, California Institute for Quantitative Biomedical Research, University of California, San Francisco, CA*
- JAI EVANS • *LECB, CIT, NIH, Frederick, MD*
- YE FANG • *Biochemical Technologies, Science and Technology Division, Corning Incorporated, Corning, NY*
- CARLOS FERNANDEZ-PATRON • *Heart and Stroke Foundation New Investigator, Department of Biochemistry, University of Alberta, Canada*
- ANN M. FERRIE • *Biochemical Technologies, Science and Technology Division, Corning Incorporated, Corning, NY*
- TIFFANY B. FISCHER • *Department of Biochemistry and Biophysics, Texas A&M University, TX*
- JOHN FLENSBURG • *Amersham Biosciences AB, GE Healthcare, Uppsala, Sweden*
- MICHAEL FOUNTOULAKIS • *F. Hoffman-LaRoche Ltd., Center for Medical Genomics, Basel, Switzerland, Foundation for Biomedical Research of the Academy of Athens, Greece*
- ELISABETH GASTEIGER • *Swiss Institute of Bioinformatics, Geneva, Switzerland*
- ALEXANDRE GATTIKER • *Swiss Institute of Bioinformatics, Geneva, Switzerland*
- FRANCESCO GIORGIANNI • *Charles B. Stout Neuroscience Mass Spectrometry Laboratory, University of Tennessee Health Science Center, Memphis, TN*
- WILLIAM L. GODFREY • *Molecular Probes, Inc., Eugene, OR*
- ERICA A. GOLEMIS • *Division of Basic Science, Fox Chase Cancer Center, Philadelphia, PA*
- DAVID R. GOODLETT • *Institute for Systems Biology, Seattle, WA*
- TERRIE GOODMAN • *Molecular Probes Inc., Eugene, OR*
- LEO GOODSTADT • *MRC Functional Genetics Unit, University of Oxford, Department of Human Anatomy and Genetics, Oxford, UK*
- ANDREW A. GOOLEY • *Proteome Systems Ltd., Sydney, Australia*
- FRANZ-GEORG HANISCH • *Institute of Biochemistry II and Center for Molecular Medicine Cologne, Cologne, Germany*
- COURTENAY R. HART • *Molecular Probes Inc., Eugene, OR*

- SETSUKO HASHIMOTO • *Biacore K.K., Tokyo, Japan*
- MINGYUE HE • *Discerna Ltd., Cambridge, UK*
- DENIS HOCHSTRASSER • *Biomedical Proteomics Research Group, Central Clinical Chemistry Laboratory, Geneva University Hospital, Geneva, Switzerland*
- CHRISTINE HOOGLAND • *Swiss Institute of Bioinformatics, Geneva, Switzerland*
- FEMIA G. HOPWOOD • *Proteome Systems Ltd., Australia*
- SJOUKE HOVING • *Novartis Institutes for Biomedical Research, Functional Genomics/Proteome Sciences, Basel, Switzerland*
- LI-SHAN HSIEH • *Division of New Drug Chemistry 1, Center for Drug Evaluation and Research, Food and Drug Administration, Rockville, MD*
- TOSHIAKI ISOBE • *Department of Chemistry, Graduate School of Science, Tokyo Metropolitan University, Tokyo, Japan*
- BIJU ISSAC • *Institute of Microbial Technology, Chandigarh, India*
- SUJATA IYER • *BD Biosciences, San Jose, CA*
- HARREN JHOTI • *Astex Technology Ltd., Cambridge, UK*
- BINO JOHN • *Computational Biology Center, Memorial Sloan-Kettering Cancer Center, New York, NY*
- JANICE L. JOSS • *Proteome Systems Ltd., Sydney, Australia*
- RACHEL KARCHIN • *Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, California Institute for Quantitative Biomedical Research, University of California, San Francisco, CA*
- FANG LAI • *Biochemical Technologies, Science and Technology Division, Corning Incorporated, Corning, NY*
- MICHELE LEARMONTH • *School of Biomedical and Clinical Laboratory Sciences, University of Edinburgh, UK*
- KI-BOOM LEE • *The Wistar Institute, Philadelphia, PA*
- RICHARD M. LEIMGRUBER • *Pfizer, Inc., PGRD-World Wide Safety Sciences, St. Louis, MO*
- PETER F. LEMKIN • *LECB, NCI-Frederick, Frederick, MD*
- DAVID W. M. LEUNG • *School of Biological Sciences, University of Canterbury, New Zealand*
- MARIA LIMINGA • *Amersham Biosciences AB, GE Healthcare, Uppsala, Sweden*
- YEE-PENG R. LUE • *Department of Biological Sciences, National University of Singapore, Republic of Singapore*
- M. S. MADHUSUDHAN • *Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, California Institute for Quantitative Biomedical Research, University of California, San Francisco, CA*
- TADAKAZU MAEDA • *Kitasato University School of Science, Japan*
- MICHELE MAGRANE • *European Bioinformatics Institute, Cambridge, UK*
- GREGORY S. MAKOWSKI • *Department of Laboratory Medicine, School of Medicine, University of Connecticut Health Center, Farmington, CT*
- CECILIA B. MARTA • *Department of Neuroscience, University of Connecticut Medical School, Farmington, CT*
- MARC A. MARTI RENOM • *Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, California Institute for Quantitative Biomedical Research, University of California, San Francisco, CA*

- BRIAN M. MARTIN • *Laboratory of Neurotoxicology, NIMH, National Institutes of Health, Bethesda, MD*
- KAREN MARTIN • *Molecular Probes Inc., Eugene, OR*
- MARIA JESUS MARTIN • *European Bioinformatics Institute, Cambridge, UK*
- RAMY MOHARRAM • *Laboratory of Neurotoxicology, NIMH, National Institutes of Health, Bethesda, MD*
- KHALED MOSTAGUIR • *Swiss Institute of Bioinformatics, Geneva, Switzerland*
- NICOLA JANE MULDER • *European Bioinformatics Institute, Cambridge, UK*
- STEFAN MÜLLER • *Institute of Biochemistry II and Center for Molecular Medicine Cologne, Cologne, Germany*
- TOHRU NATSUME • *National Institute of Advances Industrial Science and Technology, Biological Information Research Center, Tokyo, Japan*
- SONIA NAVARRO • *Department of Molecular Microbiology and Immunology, Keck School of Medicine, University of Southern California, Los Angeles, CA*
- JOAKIM NORBECK • *Chalmers Technical University, Göteborg, Sweden*
- CLAIRE O'DONOVAN • *European Bioinformatics Institute, Cambridge, UK*
- MASAMICHI OH-ISHI • *Kitasato University School of Science, Japan*
- JAN VAN OOSTRUM • *Novartis Institutes for Biomedical Research, Functional Genomics/Proteome Sciences, Basel, Switzerland*
- ALBRECHT OTTO • *Department of Neuroproteomics, Max Delbrueck Center of Molecular Medicine, Berlin, Germany*
- MELISSA PACZKOWSKI • *Department of Biochemistry and Biophysics, Texas A&M University, TX*
- PATRICIA M. PALAGI • *Swiss Institute of Bioinformatics, Geneva University Hospital, Geneva, Switzerland*
- WAYNE F. PATTON • *Molecular Probes Inc., Eugene, OR*
- STEVEN E. PFEIFFER • *Department of Neuroscience, University of Connecticut Medical School, Farmington, CT*
- URSULA PIEPER • *Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, California Institute for Quantitative Biomedical Research, University of California, San Francisco, CA*
- OLIVIER POCH • *Laboratoire de Biologie et Genomique Structurales, Institut de Génétique et de Biologie Moléculaire et Cellulaire, France*
- JULIA POLAND • *Institut für Medizinische und Chemische Labordiagnostik, Landeskrankenhaus Klagenfurt, Klagenfurt, Austria*
- CHRIS P. PONTING • *MRC Functional Genetics Unit, University of Oxford, Department of Human Anatomy and Genetics, Oxford, UK*
- MANUELA PRUESS • *European Bioinformatics Institute, Cambridge, UK*
- S. RACKOVSKY • *Department of Biomathematical Sciences, Mt. Sinai School of Medicine, New York, NY*
- THIERRY RABILLOUD • *DBMS/BECP, CEA-Grenoble, France*
- GAJENDRA P. S. RAGHAVA • *Institute of Microbial Technology, Chandigarh, India*
- MELINDA L. RAMSBY • *Department of Medicine, School of Medicine, University of Connecticut Health Center, Farmington, CT*
- DIETHER RECKTENWALD • *Research Department, BD Biosciences, San Jose, CA*

- BURKHARD ROST • *CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia University, Columbia University Center for Computational Biology and Bioinformatics, NorthEast Structural Genomics Consortium, Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY*
- COLETTE J. RUDD • *BD Biosciences, San Jose, CA*
- ROBERT B. RUSSELL • *Structural Bioinformatics EMBL, Heidelberg, Germany*
- STEPHEN RUSSELL • *Functional Genomics and Proteomics Unit, Windber Research Institute, Windber, PA*
- GARY RYMAR • *Michigan Proteome Consortium, University of Michigan, Ann Arbor, MI*
- ANDREJ SALI • *Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, California Institute for Quantitative Biomedical Research, University of California, San Francisco, CA*
- JEAN-CHARLES SANCHEZ • *Biomedical Proteomics Research Group, Central Clinical Chemistry Laboratory, Geneva University Hospital, Geneva, Switzerland*
- BIRTE SCHULENBERG • *Molecular Probes Inc., Eugene, OR*
- FRANÇOISE SEILLIER-MOISEIWITSCH • *Lombardi Comprehensive Cancer Center, Georgetown University, Washington, DC*
- ILYA G. SEREBRIISKII • *Division of Basic Science, Fox Chase Cancer Center, Philadelphia, PA*
- MIN-YI SHEN • *Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, California Institute for Quantitative Biomedical Research, University of California, San Francisco, CA*
- CRAIG D. SHRIVER • *Clinical Breast Care Project, Walter Reed Army Medical Center, Washington, DC*
- PRANAV SINHA • *Institut fur Medizinische und Chemische Labordiagnostik, Landeskrankenhaus Klagenfurt, Klagenfurt, Austria*
- LUCY SKRABANEK • *Department of Physiology and Biophysics and Institute for Computational Biomedicine, Weill Medical College of Cornell University, New York, NY*
- RICHARD I. SOMIARI • *Functional Genomics and Proteomics Unit, ITSI-Biosciences, Johnstown, PA*
- STELLA B. SOMIARI • *Windber Research Institute, Windber, PA*
- DAVID W. SPEICHER • *The Wistar Institute, Philadelphia, PA*
- THOMAS H. STEINBERG • *Molecular Probes, Inc., Eugene, OR*
- ANTHONY G. SULLIVAN • *Functional Genomics and Proteomics Unit, Windber Research Institute, Windber, PA; and Thermoelectron Training Institute, West Palm Beach, FL*
- LAY-PHENG TAN • *Department of Biological Sciences, National University of Singapore, Republic of Singapore*
- CHRISTOPHER M. TAYLOR • *Department of Neuroscience, University of Connecticut Medical School, Farmington, CT*
- JULIE D. THOMPSON • *Laboratoire de Biologie et Genomique Structurales, Institut de Génétique et de Biologie Moléculaire et Cellulaire, France*
- GREGORY C. THORNWALL • *LECB, SAIC • Frederick, Frederick, MD*

- ANNABEL E. TODD • *Department of Biochemistry and Molecular Biology, University College London, London, UK*
- ANDREW E. TORDA • *Zentrum für Bioinformatik, University of Hamburg, Hamburg, Germany*
- JERRY TSAI • *Department of Biochemistry and Biophysics, Texas A&M University, TX*
- PETER UETZ • *Institut für Genetik, Forschungszentrum Karlsruhe, Germany*
- BRIAN BERG VANDAHL • *Department of Medical Microbiology and Immunology, University of Aarhus, Denmark, Loke Diagnostics ApS, Denmark*
- MLADEN VINKOVIĆ • *Astex Technology Ltd., Cambridge, UK*
- SONJA VOORDIJK • *Geneva Bioinformatics S. A., Geneva, Switzerland*
- HANS VOSHOL • *Novartis Institutes for Biomedical Research, Functional Genomics/Proteome Sciences, Basel, Switzerland*
- ANGELA K. WALKER • *Michigan Proteome Consortium, University of Michigan, Ann Arbor, MI*
- JOHN M. WALKER • *School of Life Sciences, University of Hertfordshire, Hatfield, UK*
- DANIEL WALTHER • *Swiss Institute of Bioinformatics, Geneva, Switzerland*
- MARC R. WILKINS • *Proteome Systems Ltd., Sydney, Australia*
- NICHOLE L. WILSON • *Proteome Systems Ltd., Sydney, Australia*
- FRANK A. WITZMANN • *Department of Cellular & Integrative Physiology, Indiana University School of Medicine, Indianapolis, IN*
- YOSHIHIDE YAMAGUCHI • *Department of Molecular Neurobiology, Tokyo University of Pharmacy and Life Science, Tokyo, Japan*
- SHAO Q. YAO • *Department of Biological Sciences, Department of Chemistry, National University of Singapore, Republic of Singapore*
- SU-YIN D. YEO • *Department of Biological Sciences, National University of Singapore, Republic of Singapore*
- EUGENE YI • *Institute for Systems Biology, Seattle, WA*
- JEFF YON • *Astex Technology Ltd., Cambridge, UK*
- MICHAEL F. ZETTEL • *Department of Biochemistry and Biophysics, Texas A&M University, TX*
- HUI ZHANG • *Institute for Systems Biology, Seattle, WA*
- IGOR ZHUKOV • *Institute of Biochemistry and Biophysics, Warsaw, Poland*
- XUN ZUO • *The Wistar Institute, Philadelphia, PA*

Extraction and Solubilization of Proteins for Proteomic Studies

Richard M. Leimgruber

1. Introduction

For any proteomic study involving various control and experimental specimens, several factors need to be in place. A critical one is the extraction and solubilization of all components, regardless of whether a chromatographic (1,2) or two-dimensional (2-D) gel electrophoretic fractionation (3–6) is performed prior to analysis of proteins of interest by mass spectrometry of protein digests. All proteins must not only be extracted, but they must also be completely soluble, free from interacting partners (such as protein–RNA/DNA and protein–protein interactions, metabolites, and so on), and, in the case of 2-D gel electrophoresis, they must remain soluble as they approach their isoelectric points. The solubilization process should extract all classes of proteins reproducibly, such that statistically significant quantitative data can be obtained and correlated with experimental perturbations and the resulting biological responses.

To accomplish this task, various approaches have been presented in the literature (7–11), and many solubilization cocktails are now available commercially. However, it should be noted that currently, despite several attempts by multiple groups, there is no single solubilization cocktail that works perfectly for all conditions and samples, due to sample source-related interfering compounds and a high degree of heterogeneity among samples. This heterogeneity can lead to differing protein solubilities. Also, the presence of highly abundant proteins complicates the extraction, solubilization, and analysis of the less abundant species. Extracts from certain plant tissues also present their own set of unique issues (12). In addition to solubilization of all proteins, the solubilization agents used must also be compatible with the subsequent fractionation/analytical method employed. To date, the most efficient solubilization cocktails consist of a mixture of chaotropic agents, a mixture of detergents containing 13–15 carbon long hydrophobic chains, and a reductant (13–17).

It is important to note that the effectiveness of solubilization is not the only factor that affects the quality of the 2-D protein patterns. Gel strip rehydration, sample application method, sample load, electrophoresis conditions, and so on all have an impact on the quality of the 2-D protein fingerprint or pattern.

Recently, there has been a renewed interest in quantitative protein profiling, a process that is critical for an understanding of biological function (18,19). Because the proteome is a very complex, dynamic process that represents events at the functional

level, automated methods (20) and approaches to correlate quantitative changes in protein levels (including posttranslational modifications) will be required for system biology studies. Such efforts have been reported for breast carcinoma studies (21), the effects of ultraviolet (UV) irradiation of HeLa cells (22), compound-induced liver toxicity (23), and cell-surface protein characterization (24). Studies of reproducibility of cerebrospinal fluid (CSF) analyses (25) indicate that a quantitative 2-D gel electrophoresis approach is a viable one. Sensitive staining methods with greater dynamic ranges (26–28), improvements in peptide detection (29), and image analyses (30–32) also support this approach. Chromatographic approaches, such as those pioneered by the isotope-coded affinity tag labeling (ICAT) approach (1), and commercial efforts in development such as that by Protein Forest, Inc., will play critical roles in the development of these rapid, quantitative approaches.

2. Sample Considerations

2.1. General

When one attempts to extract and solubilize proteins, several factors must be addressed. Among these are time, temperature, pH, protein concentration, salts, metal ions, and cofactors. Because each of these will be fairly specific for a given application, they will not be addressed in this brief review.

2.2. Abundant Proteins

The large dynamic range of proteins in biological samples—up to 10^{10} in bodily fluids such as serum (33)—presents a major problem for whole proteome studies. Because of total protein load limitations by proteomic methods, the most abundant proteins overwhelm the assay and limit or prevent the detection of low-abundance proteins. This issue has been addressed by fractionation procedures (1,34–42) and by depletion strategies (43–47). There are several commercial reagents available for depletion of serum albumin, immunoglobulin (Ig)G, and so on (Agilent Technologies, Amersham Biosciences, Calbiochem, Pierce, Millipore, Sigma). In addition, approaches using ICAT have been employed to detect proteins quantitatively over a broad dynamic range (1,48,49).

It should be noted, however, that removing and discarding of abundant proteins, such as albumin, may not be advisable in the search for clinical biomarkers (50). A recent publication by Mehta et al. (51), which describes an approach to look at the low-molecular-weight peptides bound to serum carrier proteins, clearly demonstrates that these carrier proteins may be rich repositories of biomarkers. This type of affinity-capture approach is also useful for signal transduction studies (52).

2.3. Chaotropes

A significant advance in increasing the solubility of proteins was the use of urea/thiourea mixtures (53,54). The typical mixtures currently in use consist of 5–8 M urea and 2–2.5 M thiourea (13,14,17,41,55). It is interesting to note that although the addition of thiourea to this mixture increases both the number and quality of proteins that are detected, these additional proteins are water-soluble proteins, not membrane proteins (56,57).

2.4. Detergents

Although detergents such as sodium dodecyl sulfate (SDS) are extremely efficient at solubilizing hydrophobic proteins, their anionic nature greatly limits their effectiveness for conventional proteomic analyses. As a result, zwitterionic and nonionic detergents have found widespread use in 2-D electrophoresis (10,13,17,55,58). CHAPS (3-[(3-cholamidopropyl)dimethylamino]-1-propanesulfonate) is one of the most widely used zwitterionic detergents, and it has been shown to promote solubilization and stabilization by shielding hydrophobic zones from nonspecific aggregation and by stabilizing disordered loops to reduce heterogeneity (59). Specific instances have been identified in which various sulfobetaine detergents are the better solubilizing agents and CHAPS is not the best detergent of choice (58). In addition, solubilization of Jurkat membranes with Tween-20, Brij 58, and Lubrol WX yielded significantly more solubilized protein than either Triton or CHAPS (60). It has been reported that Brij 30 and Brij 96 are very efficient under nondenaturing conditions but are not efficient solubilizing agents when utilized in the presence of urea and thiourea (61). However, to date there has not been any single zwitterionic or nonionic detergent or detergent mixture identified that will completely solubilize all proteins. In fact, it has been observed that detergent selection for optimal solubilization for a given set of samples will be an empirical, experimentally determined one (13). It seems logical that the best approach is to combine different types of detergents to obtain a mixture that has the best attributes of each. Approaches to this have been reported (13,14,17). New polymeric surfactants are being developed for protein solubilization (62). These hydrophobically modified pullans (HMCMPs) extracted approx 70% of the total protein, without adversely affecting protein structure or function. Therefore, it may be possible to look at active protein complexes using these newer molecules.

2.5. Membrane Proteins

Membrane proteins play critical roles in cellular communication, transportation of nutrients, metabolites and ions, adhesion, signal transduction, and so on (15). These key proteins are typically not seen or are underrepresented in 2-D protein patterns because the efficient extraction of proteins from membranes is a process that has been difficult to accomplish; but progress is being made (8,14,41,53,63–66). Short-chain phospholipids have been used to isolate functional membrane protein complexes; however, these molecules interact primarily with the membrane lipids, and there is little if any interaction with the integral membrane proteins (67). Detergents act through a series of steps, interacting with and destabilizing lipid components of membranes, yielding detergent-lipid-protein complexes, and then effectively replacing the lipids, such that the detergents now interact with and shield the hydrophobic regions of proteins, resulting in better solubilization of these proteins (68). The ability of detergents to solubilize hydrophobic proteins appears to correlate well with the length of the hydrocarbon chain (16) and hydrophobic lipid balance (HLB) values (61). Recently, progress has been made toward obtaining a better representation of these very hydrophobic molecules in proteomic profiles. Zwitterionic detergents have been synthesized and used to solubilize membrane proteins (13,55,58). Chloroform-methanol extractions of membranes followed by 2-D gel electrophoresis in a detergent/chaotropic mix-

ture resulted in the identification of membrane proteins that had not previously been identified using 2-D gel electrophoresis (42). Differential extractions of purified chloroplast membranes using different chloroform:methanol ratios and detergents also resulted in the identification of previously unidentified membrane proteins using SDS-polyacrylamide gel electrophoresis (PAGE) and amino acid sequencing (64). A sequential fractionation procedure (65) is another possible solution to the study of hydrophilic proteins, but this approach complicates the correlation of quantitative changes within the entire proteome with the biological response. Recently, the enrichment of membrane proteins using carbonate extraction coupled with surfactant-free organic solvent-based solubilization (69) and using a partition phase separation have been reported (70).

2.6. Nucleic Acids

The presence of DNA and RNA in samples for proteomic analyses can present problems with respect to both the quality of 2-D gel patterns and the recovery of DNA- and RNA-binding proteins. The presence of these nucleic acids can result in viscous samples that are difficult to pipet (affecting sample loads) and cause streaking in the first dimension because the nucleic acids tend to act somewhat like ion-exchangers, which can slowly release bound protein. Also, if they are precipitated from solution, any associated proteins may also be lost in this discarded fraction unless efforts are made to extract the nucleic acid fraction with a detergent cocktail such as that described by Giavalisco et al. (57). The DNA and RNA can be digested with DNase and RNase (3) or sheared mechanically with repeated passes through a tuberculin syringe equipped with a 21-gauge needle. A very convenient alternative method of mechanical shear is to place the tissue extract or cell lysate in a QIAshredder (QIAgen) and centrifuge the sample for 1–2 min in a microcentrifuge (17). Any buffer with or without detergents can be used, and in addition to the breaking the nucleic acids, cell debris is eliminated in the pellet, without the apparent loss of protein (Leimgruber, R. M., unpublished results). The QIAshredder has been useful for sample preparation for both one- and 2-D electrophoresis.

2.7. Phenolics

The presence of polyphenolic compounds at varying levels in plant samples (71–74) can adversely affect the 2-D protein patterns, generating streaks. Addition of insoluble polyvinylpyrrolidone (PVP) to the plant extracts effectively removes the phenolics (75,76).

2.8. Reductants

A consideration for obtaining clearly defined, well resolved protein spots on 2-D gels is the complete reduction of each denatured protein, resulting in very homogeneous proteins, which should be detected as well defined, round spots. Maintaining complete reduction has been complicated by the use of reductants such as dithiothreitol (DTT), a weak acid that migrates out of the very basic end of the first-dimension gel. Efforts to minimize this effect by introducing an excess of DTT in the wick at the cathode appear to help extend the pH range for protein resolution (77). Another promising approach to extending the pH range beyond an upper pH of around 8.0 to produce

highly resolved proteins is to reduce samples with tributylphosphine (TBP) and to perform the focusing step in the presence of dithiodiethanol (DTDE) (78,79). Reduction and alkylation of the proteins in this manner also has the advantage of potentially eliminating this step for in-gel tryptic digestions of excised protein spots.

2.9. Plant Tissue

Plant tissues present some of the same challenges for total proteome characterization as those from mammalian sources, including the characterization of membrane proteins (80–83), the identification of low-abundance proteins, and the presence of high-abundance proteins (e.g., ribulose-1,5-bisphosphate carboxylase/oxygenase [rubisco]). However, for the case of plants, there is also the issue of metabolites from secondary metabolism interfering with the separation process (84,85). There also is the issue of high protease activity (86). Because protein levels are low in plant tissues, many of the solubilization procedures involve extraction coupled with trichloroacetic acid (TCA)/acetone precipitation (87,88). As a result, many proteins may not be readily resolubilized, and some may not be captured by the precipitation step. As a result, recent reports have detailed alternative, much improved extraction procedures. In one procedure reported by Wang et al. (40), plant tissue is reground into a very fine powder, washed extensively to remove interfering compounds, and subjected to a phenol/dense SDS extraction. In another approach, a sequential extraction procedure yielding three fractions was utilized (57). The first fraction contains the highly water-soluble proteins in the supernatant resulting from centrifugation of an aqueous extraction of pulverized plant tissue in the presence of protease inhibitors. The resulting pellet sample is extracted with detergents (4% CHAPS and 2% amido sulfobetaine [ASB] 14), followed by the addition of urea and thiourea. The second fraction represents the supernatant after centrifugation of the detergent-solubilized material. Finally, the last pellet is incubated with DNase, followed by the addition of urea and thiourea. Using this approach, the authors detected a threefold increase in the number of total proteins from the stems and leaves of *Arabidopsis*.

2.10. Labeling With Cyanine Dyes (see also Chapter 24)

An important consideration for the extraction/lysis cocktail is whether or not proteins will be labeled with fluorescent dyes prior to the first dimension in 2-D electrophoresis. Cyanine dyes (Cy2, Cy3, and Cy5, Amersham Biosciences) containing an *N*-hydroxysuccinimidyl linker are conjugated to solubilized proteins through a lysine residue under carefully controlled conditions (26). These dyes are designed such that they do not alter the charge on the protein and add only 500 Da. Labeling is performed such that each protein is labeled with only one dye molecule. It is critical that the protein samples do not contain reducing agents, ampholytes, primary amines, or thiols, as they interact with the dye reagent (27). If any of these agents are required for the extraction of protein from tissue, they must be removed prior to the labeling step. The incorporation of an internal pooled standard for differential in-gel electrophoresis (DIGE) analyses has been reported to improve the accuracy of protein quantitation between gels, facilitating the detection of small changes not readily detected with conventional post-electrophoresis staining (89,90). A few recent applications of this technology have been in the areas of oncology (21), inflammation (17), and compound-

induced liver toxicity (91,92). Several previously unidentified proteins were identified, including potential biomarkers of liver toxicity (92). Analyses of the fluorescent 2-D images is performed using either the DeCyder™ software (Amersham Biosciences) or standard image-analysis software (93). Labeling with reactive thiol dyes has also been reported (94).

2.11. *Laser Capture Microscopy (see also Chapter 8)*

Recent advances in sample generation include laser capture microscopy (LCM), which can be utilized to generate large populations of homologous cells from tissue sections, from which the proteins can be solubilized. This approach also has the potential to aid in the characterization of heterogeneous samples, such as some tumor types, and to identify key biomarkers that may be missed when analyzing the entire tumor (96–100). This method can in some cases be utilized as an alternative to histological staining of brain tissue (101). A recent report has utilized LCM for the study of plant cells (102), in which LCM and microarrays were used to analyze global gene expression.

2.12. *Protein Determination*

Because many of the additives employed to solubilize and extract proteins from biological samples interfere with many protein assays, it is often difficult to accurately determine the total protein present in a given sample. These additives at levels typically used tend to interfere with many of the commonly used assays, such as the Bradford and modified Bradford assays. One of the best is that marketed by Cytoskeleton (Advance Protein Assay 01), because it can tolerate relatively high levels of chaotropes, detergents, and reductants. This assay is very rapid, requires little sample, and has a fairly good dynamic range.

3. Current Basic Solubilization Protocol

3.1. *Sample Generation*

3.1.1. *Lysis/Extraction/Rehydration Solution*

3.1.1.1. *GENERAL SOLUBILIZATION COCKTAIL FOR MAMMALIAN TISSUES AND CELLS*

The typical lysis solution (17) consists of the following:

- 5 M Urea.
- 2 M Thiourea.
- 0.25% (v/v) CHAPS (Sigma).
- 0.25% (v/v) Tween-20 (Bio-Rad).
- 0.25% (v/v) sulfobetaine (SB) 3-10 (Sigma).
- 0.25% (w/v) carrier ampholytes (1:1:1:1 mixture of Bio-Lyte 3-10 [Bio-Rad], Servalyte 3-10 [Serva], Ampholine 3.5-9.5 [Amersham Biosciences], and Resolyte 4-8 [BDH]).
- 2 mM Tributylphosphine (TBP).
- 10% Isopropanol.
- 12.5% (v/v) water-saturated isobutanol.
- 5% (v/v) glycerol (Bio-Rad).
- 1 mM Sodium vanadate (phosphatase inhibitor, Sigma).
- 1X complete protease inhibitor cocktail (Boehringer-Mannheim).

This lysis solution can be stored tightly sealed at -80°C for several weeks. Care must be taken with respect to the potential evaporation of the alcohol over prolonged time periods. It may be necessary to change the detergent mixture and level for a given application. In some cases, 0.25–0.5% Triton X-100 yields better results than Tween-20, or increased levels of CHAPS solubilize more proteins. If TBP cannot be used, 100 mM dithiothreitol or 5% 2-mercaptoethanol can be used, but the results are not as good. The ampholyte level may also need to be increased if high amounts of sample are applied to the gel strip. In general, it is best to keep the detergent (0.75–2%) and ampholyte (0.25–1.5%) levels as low as possible for optimal resolution.

3.1.1.2. GENERAL SOLUBILIZATION COCKTAIL FOR PLANT SEEDS

The typical plant extraction/solubilization solution consists of the following:

- 6 M Urea.
- 2 M Thiourea.
- 0.5% (v/v) CHAPS (Sigma).
- 0.25% (v/v) Triton X-100 (Bio-Rad).
- 0.25% (v/v) SB 3-10 (Sigma).
- 0.35% (w/v) carrier ampholytes (1:1:1:1 mixture of Bio-Lyte 3-10 [Bio-Rad], Servalyte 3-10 [Serva], Ampholine 3.5-9.5 [Amersham Biosciences] and Resolyte 4-8 [BDH]).
- 2 mM Tributylphosphine (TBP).
- 16% Isopropanol.
- 5% (v/v) glycerol (Bio-Rad).
- 1X complete protease inhibitor cocktail (Boehringer-Mannheim).

This lysis solution can be stored tightly sealed at -80°C for several weeks. As noted above, care must be taken with respect to the potential evaporation of the alcohol over prolonged time periods. As noted above, this cocktail may need to be modified empirically for specific plant tissue types (Leimgruber, N. L., et al., and Ruebelt, M. C., et al., unpublished data).

3.1.2. Cell Lysates

Cells such as those of the U937 human monocytic cell line are typically solubilized directly in lysis/rehydration solution at a level of approx 20,000 cells/ μL or approx 1.5 mg protein/mL. Much smaller cells (e.g., monocytes and splenocytes) are solubilized at approx 80,000 cells/ μL , which represents approx 1.28 mg protein/mL. Following extraction for 30 min at room temperature on a Nutator mixer, the samples are clarified by centrifugation in a microcentrifuge at 15,300g for 5 min. The presence of alcohol appears to precipitate the nucleic acid out of solution, while the chaotropes-detergents mixture solubilizes and releases the RNA- and DNA-binding proteins.

3.1.3. Serum

Rat serum (typically around 55 mg protein/mL, but ranges from approx 40–100 mg protein/mL) is either analyzed by dilution directly into lysis/rehydration solution or is first depleted with an affinity column to remove albumin, IgGs, and so on. The depleted serum sample is then concentrated back to the starting protein level and diluted into lysis/rehydration solution.

3.1.4. Tissue

Tissue is frozen in liquid nitrogen and pulverized in liquid nitrogen with a BioPulverizer (sizes to accommodate from 10 mg to 10 g of tissue, Biospec Products, Inc.). The pulverized tissue is extracted at a level of approx 2–3 mg of tissue/mL of lysis/rehydration solution for 30 min at room temperature. The extract is clarified by centrifugation, and the supernatant is either analyzed immediately or stored at –80°C. Fresh bone samples can be pulverized and processed in a similar manner using the MicroCryoPulverizer (Biospec Products, Inc.).

3.1.5. Urine

Urine (typically less than 1.5 mg protein/mL in a 24-h collection) is concentrated using membrane filtration devices (Centricon 3, Amicon) or lyophilized prior to dilution into the lysis/rehydration solution at a final concentration of approx 1–2 mg protein/mL.

3.2. Focusing Parameters

Because the solubility of proteins is also dependent upon the amount of salts (and the resulting loss of water as joule heating occurs), the isoelectric focusing steps are slowly ramped up at the start of each electrophoretic run. The rehydrated immobilized pH gradient (IPG) strips are typically focused in Bio-Rad Protean IEF units using the following protocols. In general, the total number of volt-hours should be 50,000–75,000. Narrower range pH gradients require longer focusing times than broad pH ranges.

1. *pH 3.0–10, Linear or Nonlinear, 18-cm-Long IPG Gel Strips, “Normal Samples.”* 150-V rapid ramp for 1 h; 250-V rapid ramp for 1 h; 400-V rapid ramp for 4 hs; 10,000-V linear ramp for 14 h; 10,000-V rapid ramp as a hold step for additional volt-hours. The typical total volt-hours for focusing are 60,000–75,000.
2. *pH 3.0–10, Linear or Nonlinear, 18-cm-Long IPG Gel Strips, Samples Containing Salts up to 150 mM.* 50-V rapid ramp for 4 h; 150-V rapid ramp for 1 h; 250-V rapid ramp for 1 h; 400-V rapid ramp for 4 h; 10,000-V linear ramp for 12 h; 10,000-V rapid ramp as a hold step for additional volt-hours. The typical total volt-hours for focusing are 60,000–75,000.
3. *pH 3.0–10, Linear or Nonlinear, 11-cm-Long IPG Gel Strips “Normal Samples.”* 150-V rapid ramp for 1 h; 250-V rapid ramp for 1 h; 400-V rapid ramp for 4 hs; 8,000-V linear ramp for 14 h; 8,000-V rapid ramp as a hold step for additional volt-hours. The typical total volt-hours for focusing are 55,000–60,000.
4. *pH 3.0–10, Linear or Nonlinear, 11-cm-Long IPG Gel Strips, Samples Containing Salts up to 150 mM.* 50-V rapid ramp for 4 h; 150-V rapid ramp for 1 h; 250-V rapid ramp for 1 h; 400-V rapid ramp for 4 h; 5,000-V linear ramp for 14 h; 5,000-V rapid ramp as a hold step for additional volt-hours. The typical total volt-hours for focusing are 55,000–60,000.

3.3. In Strip Equilibration of Focused Proteins for the Second Dimension

Once the proteins are focused in the first dimension, the gel strips can be frozen and stored sealed at –80°C or equilibrated immediately for electrophoresis in the second dimension (SDS-PAGE). If each gel strip is to be analyzed directly, it is equilibrated directly in 2 mL of a solution containing 62.5 mM Tris-HCl, pH 6.8, 2.3 % SDS, 5% 2-mercaptoethanol (or 100 mM DTT), trace of bromphenol blue for 3 min at room

temperature. If reduction alkylation is to be performed, the strips are incubated for 6 min in 2 mL of a solution containing 5 M urea, 50 mM Tris-HCl (pH 6.8), 2.3% SDS, 20–50 mM DTT, 5% glycerol, trace of bromphenol blue. The first equilibration solution is removed and replaced with the same solution lacking DTT and containing 40–100 mM iodoacetamide. For either case, the IPG strips are incubated in the second solution for 12 min protected from light. For either case, the gel strips are embedded on top of the second-dimension gel with 1% agarose in 5 M urea, 50 mM Tris-HCl (pH 6.8), 2.3% SDS, 20–50 mM DTT, 5% glycerol, trace of bromphenol blue.

4. 2-D Gel Electrophoresis Images

Representative 2-D protein patterns resulting from solubilization of proteins using the chaotropes-detergents cocktail described in **Subheading 3.** are shown in **Figs. 1–5.** Analyses of conditioned media from two sister human stromal cell lines are presented in **Fig. 1.** These proteins are highly water soluble proteins secreted by the cells into the medium. In **Fig. 2**, a profile of a typical cell lysate is shown, using U937 cells following experimental treatment. In panel A, the proteins were labeled with Cy5 prior to separation by 2-D gel electrophoresis, and the Cy5 signal is detected following the 2-D fractionation. The same gel was subsequently fixed and stained with the fluorescent dye SYPRO Orange (**103**), and the resulting image is seen in panel B. Detection of U937 cellular proteins by staining with the fluorescent dye SYPRO Orange after electrophoresis is shown in panel C. There is a good representation of proteins covering the entire pH and molecular-weight ranges, although very high molecular weight proteins are not present. The absence of the high-molecular-weight proteins is likely due to the lack of their diffusion into the gel and/or inefficient extraction. We have found that by using slightly higher than typical rehydration volumes, it is possible to increase the representation of higher-molecular-weight proteins. However, these proteins remain under-represented. Analysis of sera from two different rats resulted in a well-resolved 2-D protein pattern using the solubilization cocktail described in **Subheading 3.1.1.1.** (**Fig. 3A,B**). Similarly, use of this cocktail efficiently solubilized proteins from rat pancreas (**Fig. 3C**) and rat heart (Leimgruber, R. M., unpublished data).

The two-dimensional pattern of proteins extracted from soybean, *Arabidopsis*, and wheat seeds are shown in **Fig. 4**. Well-resolved protein patterns are obtained in high yield. It is usually much more difficult to obtain well-resolved two-dimensional protein patterns from leaf tissue, due to the presence of interfering substances. The results of two different extraction/solubilization cocktails are shown for *Arabidopsis* leaf proteins in **Fig. 5A,B**. Leaf proteins were precipitated with TCA/acetone as described previously (**104**). Much more protein is extracted under the conditions employed for panel **A**. Although it has been previously reported with mammalian samples that the presence of alcohol does not appear to have adverse affects on extraction efficiency (**17**), in this case for *Arabidopsis*, the presence of alcohol (panel **B**) appears to decrease the solubilization efficiency, but has no adverse effect on the quality of the 2-D pattern. However, the overall representation of *Arabidopsis* proteins is very similar for leaves extracted in the presence and absence of isopropanol. Note also the presence of the very abundant protein, rubisco, and some lower-molecular-weight forms of rubisco near the middle of the pH range.

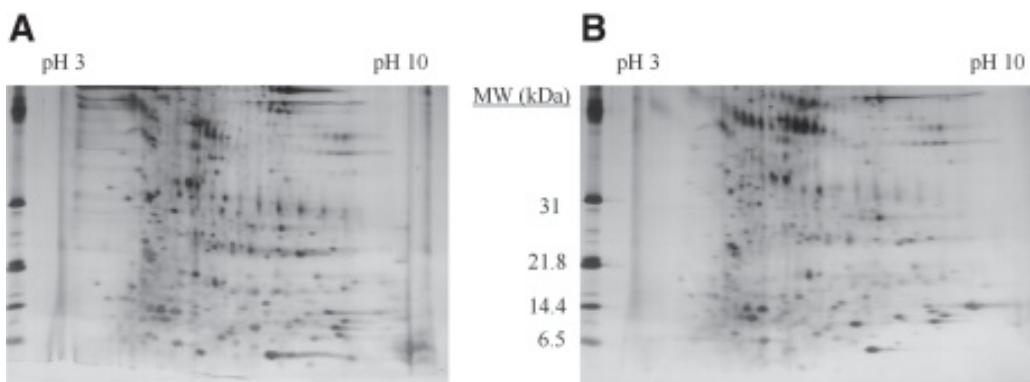


Fig. 1. Solubilization of proteins in conditioned media from human stromal cell lines. Focusing was performed using a nonlinear pH 3.0–10.0 gradient (Amersham Biosciences immobilized pH gradient [IPG] gel strips) for 65,000 V-hours. Second dimension analyses were performed using 10–20% polyacrylamide DALT (25 × 20 cm × 1.5 mm) gels. Protein (approx 800 µg) was detected with SYPRO® orange. (A) Cell line 1. (B) Cell line 2.

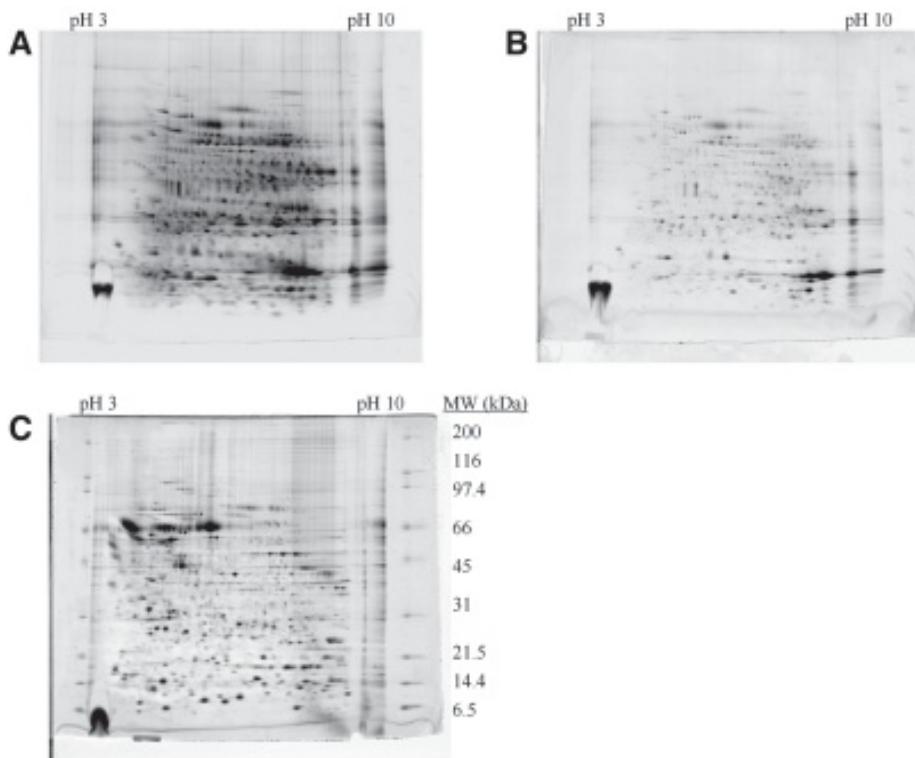


Fig. 2. Lysis and solubilization of U937 cells. Focusing was performed using a nonlinear pH 3.0–10.0 gradient (Amersham Biosciences immobilized pH gradient [IPG] gel strips) for 65,000 V-hours. Second dimension analyses were performed using 10–20% polyacrylamide DALT (25 × 20 cm × 1.5 mm) gels (Leimgruber, R. M., and Malone, J. P.). (A) protein (approx 450 µg) was detected by staining with Cy5. (B) Protein was detected by staining gel in panel A with SYPRO® Orange. (C) Protein (approx 900 µg) was detected by staining with SYPRO Orange.

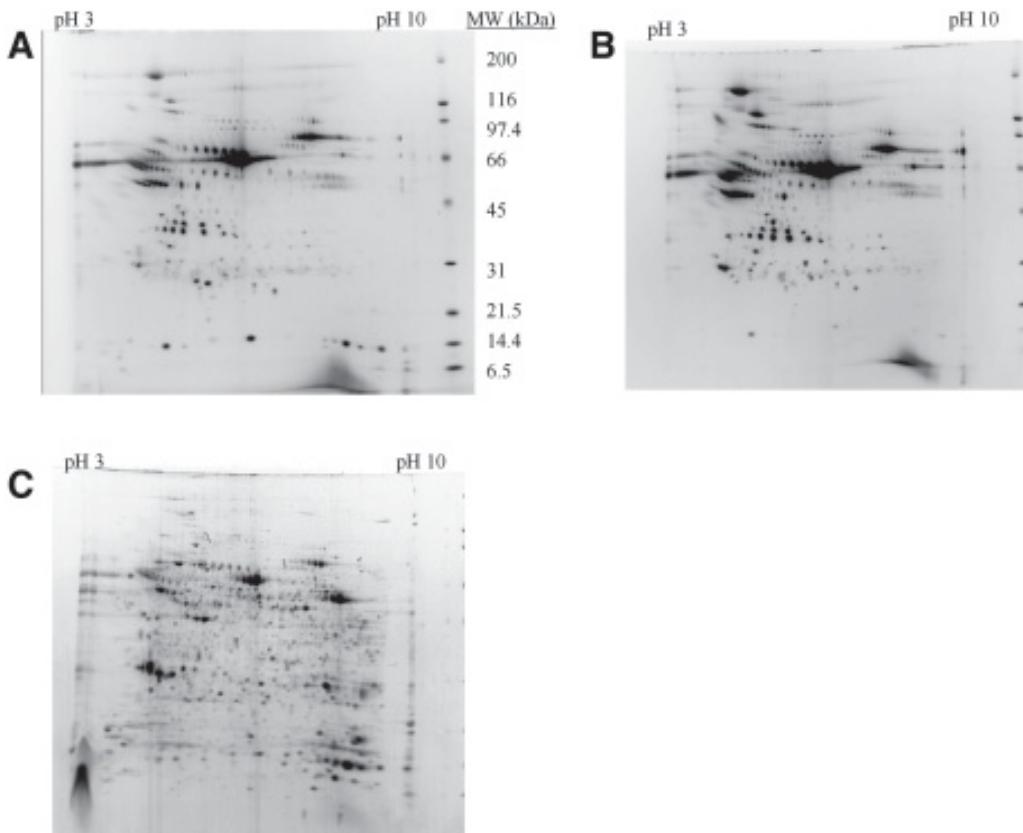


Fig. 3. Solubilization of proteins in rat serum and rat pancreas. Focusing was performed using a nonlinear pH 3.0–10.0 gradient (Amersham Biosciences immobilized pH gradient [IPG] gel strips) for 60,000 V-hours. Second dimension analyses were performed using 10–23% polyacrylamide DALT (25 × 20 cm × 1.5 mm) gels. Protein was detected by staining with SYPRO® Ruby (A,B) and SYPRO Orange (B). (Gels are courtesy of Cabonce, M. C., Pfizer, Inc., Pfizer Global Research and Development.)

5. Conclusion

Unfortunately, no magic method or solution has been identified that solubilizes all proteins completely and reproducibly, free of interfering substances from all sample sources. Progress is being made, however, toward the identification of new detergents that are compatible with downstream analyses, and fractionation procedures are being developed to facilitate protein solubilization, as well as to address the issues associated with the large dynamic range seen for protein levels in a given sample.

Acknowledgments

The author wishes to express gratitude to Marc A. Cabonce, Nancy L. Leimgruber, and Martin C. Ruebelt for generously sharing their 2-D gel images and approaches.

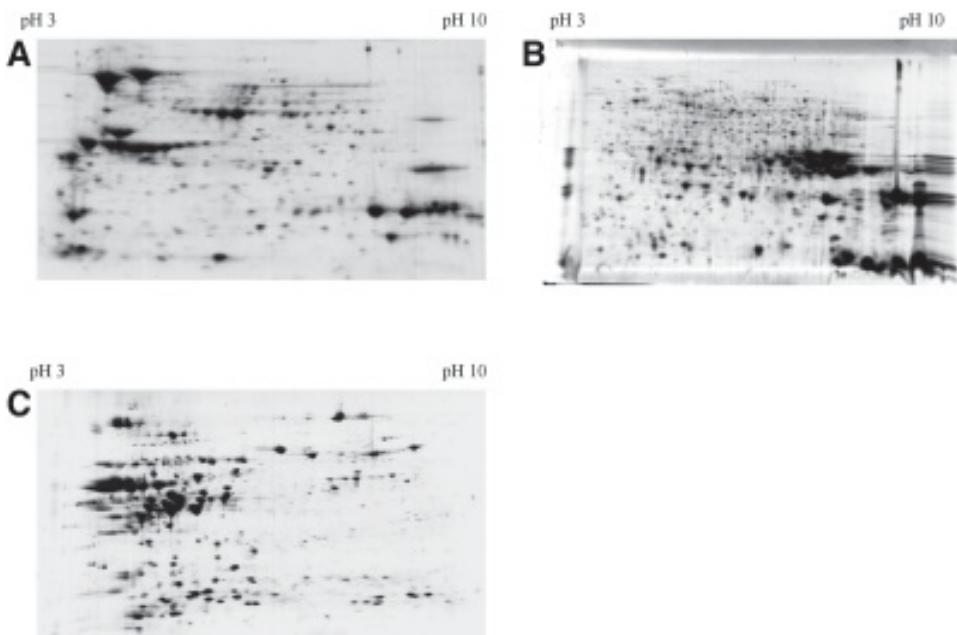


Fig. 4. Solubilization of proteins from plant seeds. Focusing of 150 μ g of protein/gel was performed using a nonlinear pH 3.0–10.0 gradient (Bio-Rad immobilized pH gradient [IPG] gel strips) for 50,000 V-hours. Second dimension analyses were performed using 10–20% (soybean and wheat) and 8–16% (arabidopsis) Criterion Tris-HCl polyacrylamide gels (Bio-Rad). Proteins were detected by staining with colloidal Coomassie brilliant blue. (A) *Glycine max* L. Merr (soybean). (B) *Arabidopsis thaliana* [ecotype Columbia] (gel is courtesy of Ruebelt, M. C., Monsanto Company, Regulatory Sciences). (C) *Triticum aestivum* (bread wheat).

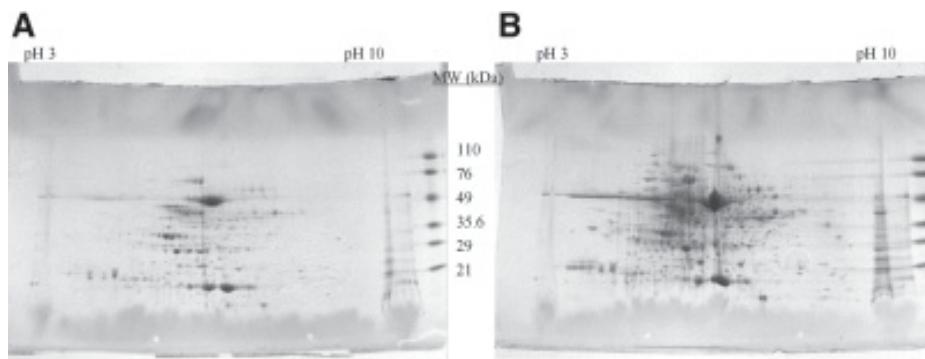


Fig. 5. Solubilization of proteins from *Arabidopsis thaliana* [ecotype Columbia] plant leaves. Sample loading per gel was from the extraction of 2.7 mg of plant leaf tissue. Focusing was performed using a linear pH 3.0–10.0 gradient (Bio-Rad immobilized pH gradient [IPG] gel strips) for 50,000 V-hours. Second dimension analyses were performed using 10–20% Criterion Tris-HCl polyacrylamide gels (Bio-Rad). Proteins were detected by staining with colloidal Coomassie brilliant blue. (Gels are courtesy of Ruebelt, M. C., Monsanto Company). (A) leaves, trichloroacetic acid (TCA)/acetone precipitation, re-solubilized and analyzed in 7 M urea, 2 M thiourea, 2.0% CHAPS, 2.0% Triton X-100, 0.4% carrier ampholyte mixture, 20 mM dithiothreitol (DTT). (B) leaves, TCA/acetone precipitation, resolubilized and analyzed in 7 M urea, 2 M thiourea, 1.0% CHAPS, 1.0% Triton X-100, 1% SB 3–10, 0.5% carrier ampholyte mixture, 2 mM TBP, 20% isopropanol.

References

1. Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., and Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotech.* **17**, 994–999.
2. Patterson, S. D. and Aebersold, R. H. (2003) Proteomics: The first decade and beyond. *Nature Genetics* **33**, 311–323.
3. Garrels, J. (1979) Two-dimensional gel electrophoresis and computer analysis of proteins synthesized by clonal cell lines. *J. Biol. Chem.* **254**, 7961–7977.
4. Gorg, A., Obermaier, C., Boguth, G., et al. (2000) The current state of two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis* **21**, 1037–1053.
5. Rabilloud, T. (2002) Two-dimensional gel electrophoresis in proteomics: Old, old fashioned, but it still climbs up the mountains. *Proteomics* **2**, 3–10.
6. Lefkowits, I., Kettman, J. R., and Frey, J. R. (2000) Global analysis of gene expression in cells of the immune system. I. Analytical limitations in obtaining information on polypeptides in two-dimensional gel spots. *Electrophoresis* **21**, 2688–2693.
7. Herbert, B. (1999) Advances in protein solubilization for two-dimensional electrophoresis. *Electrophoresis* **20**, 660–663.
8. Molloy, M.P. (2000) Two-dimensional electrophoresis on membrane proteins using immobilized pH gradients. *Anal. Biochem.* **280**, 1–10.
9. Rabilloud, T. (1996) Solubilization of proteins for electrophoretic analyses. *Electrophoresis* **17**, 813–829.
10. Rabilloud, T. (1999) Solubilization of proteins in 2-D electrophoresis: An outline. *Methods Mol. Biol.* 112 2-D Proteome Analysis Protocols (Ed. Link, A. J.), 9–19.
11. Rabilloud, T., Adessi, C., Giraudeau, A., and Lunardi, J. (1997) Improvement of the solubilization of proteins in two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis* **18**, 307–316.
12. Kersten, B., Burkle, L., Kuhn, E. J., et al. (2002) Large-scale plant proteomics. *Plant Mol. Biol.* **48**, 133–141.
13. Rabilloud, T., Blismick, T., Heller, M., et al. (1999) Analysis of membrane proteins by two-dimensional electrophoresis: Comparison of the proteins extracted from normal or *Plasmodium falciparum* infected erythrocyte ghosts. *Electrophoresis* **20**, 3603–3610.
14. Chevallet, M., Santoni, V., Poinas, A., et al. (1998) New zwitterionic detergents improve the analysis of membrane proteins by two-dimensional electrophoresis. *Electrophoresis* **19**, 1901–1909.
15. Santoni, V., Molloy, M., and Rabilloud, T. (2000) Membrane proteins and proteomics: un amour impossible? *Electrophoresis* **21**, 1054–1070.
16. Tastet, C., Charmont, S., Chevallet, M., Luche, S., and Rabilloud, T. (2003) Structure-efficiency relationships of zwitterionic detergents as protein solubilizers in two-dimensional electrophoresis. *Proteomics* **3**, 111–121.
17. Leimgruber, R. M., Malone, J. P., Radabaugh, M. R., LaPorte, M. L., Violand, B. N., and Monahan, J. B. (2002) Development of improved cell lysis, solubilization and imaging approaches for proteomic analyses. *Proteomics* **2**, 135–144.
18. Molloy, M. P. and VanBogelen, R. A. (2003) Exploring the proteome: Reviving emphasis on quantitative profiling. *Proteomics* **3**, 1833–1834.
19. Molloy, M. P., Brzezinski, E. E., Hang, J., McDowell, M. T., and VanBogelen, R. A. (2003) Overcoming technical variation and biological variation in quantitative proteomics. *Proteomics* **3**, 1912–1919.
20. Hille, J. M., Freed, A. L., and Watzig, H. (2001) Possibilities to improve automation, speed and precision of proteome analysis: A comparison of two-dimensional electrophoresis and alternatives. *Electrophoresis* **22**, 4035–4052.

21. Somiari, R. I., Sullivan, A., Russell, S., et al. (2003) High-throughput proteomic analysis of human infiltrating ductal carcinoma of breast. *Proteomics* **3**, 1863–1873.
22. Decker, E. D., Zhang, Y., Cocklin, R. R., Witzmann, F. A., and Wang, F. (2003) Proteomic analysis of differential protein expression induced by ultraviolet light radiation in HeLa cells. *Proteomics* **3**, 2019–2027.
23. Thome-Kromer, B., Bonk, I., Klatt, M., et al. (2003) Toward the identification of liver toxicity markers: A proteome study in human cell culture and rats. *Proteomics* **3**, 1835–1862.
24. Jang, J. H. and Hanash, S. (2003) Profiling of the cell surface proteome. *Proteomics* **3**, 1947–1954.
25. Terry, D. E. and Desiderio, D. M. (2003) Between-gel reproducibility of the human cerebrospinal fluid proteome. *Proteomics* **3**, 1962–1979.
26. Swatton, J. E., Prabakaran, S., Karp, N. A., Lilley, K. S., and Bahn, S. (2004) Protein Profiling of human post-mortem brain using 2-dimensional fluorescence difference gel electrophoresis (2-D DIGE). *Mol. Psychiatry* **9**, 128–143.
27. Tonge, R., Shaw, J., Middleton, B., et al. And Davison (2001) Validation and development of fluorescence two-dimensional differential gel electrophoresis proteomics technology. *Proteomics* **1**, 377–396.
28. Patton, W. F. (2002) Detection technologies in proteome analysis. *J. Chromatog. B* **771**, 3–31.
29. Choi, B.-K., Cho, Y.-M., Bae, S.-H., Zouboulis, C. C., and Paik, Y.-K. (2003) single-step perfusion chromatography with a throughput potential for enhanced peptide detection by matrix-assisted laser desorption/ionization-mass spectrometry. *Proteomics* **3**, 1955–1961.
30. Raman, B., Cheung, A., and Marten, M. R. (2002) Quantitative comparison and evaluation of two commercially available, two-dimensional electrophoresis image analysis software packages, Z3 and Melanie. *Electrophoresis* **23**, 2194–2202.
31. Rubinfeld, A., Keren-Lehrer, T., Hadas, G., and Smilansky, Z. (2003) Hierarchical analysis of large-scale two-dimensional gel electrophoresis experiments. *Proteomics* **3**, 1930–1935.
32. Rosengren, A. T., Salmi, J. M., Aittokallio, T., et al. (2003) Comparison of PDQuest and Progenesis software packages in the analysis of two-dimensional electrophoresis images. *Proteomics* **3**, 1936–1946.
33. Anderson, N. L. and Anderson, N. G. (2002) The human plasma proteome: History, character and diagnostic prospects. *Mol. Cell. Proteomics* **1**, 845–867.
34. Lopez, M., Lopez, M. F., Kristal, B. S., et al. (2000) High-throughput profiling of the mitochondrial proteome using affinity fractionation and automation. *Electrophoresis* **21**, 3427–3440.
35. Corthals, G. L., Molloy, M. P., Herbert, B. R., Williams, K. L., and Gooley, A. A. (1997) Prefractionation of protein samples prior to two-dimensional electrophoresis. *Electrophoresis* **18**, 317–324.
36. Hamler, R., Zhu, K., Buchanan, N. S., et al. (2004) A two-dimensional liquid-phase separation method coupled with mass spectrometry for proteomic studies of breast cancer and biomarker identification. *Proteomics* **4**, 562–577.
37. Klose, J. (1999) Fractionated extraction of total tissue proteins from mouse and human for 2-D electrophoresis. *Methods Enzymol.* **112**, 67–85.
38. Klose, J. (1999) Large-gel 2-D electrophoresis. In:: Link, A. (ed), *2-D Proteome Analysis Protocols*. Humana, Totowa, NJ: 147–172.
39. Rothemund, D. L., Locke, V. L., Liew, A., Thomas, T. M., Wasinger, V., and Rylatt, D. B. (2003) Depletion of the highly abundant protein albumin from human plasma using the Gradiflow. *Proteomics* **3**, 279–287.
40. Wang, W., Scali, M., Vignani, R., et al. (2003) Protein extraction for two-dimensional electrophoresis from olive leaf, a plant tissue containing high levels of interfering compounds. *Electrophoresis* **24**, 2369–2375.

41. Molloy, M. P., Herbert, B., Walsh, B. J., et al. (1998) Extraction of membrane proteins by differential solubilization for separation using two-dimensional gel electrophoresis. *Electrophoresis* **19**, 837–844.
42. Molloy, M., Herbert, B. R., Williams, K. L., and Gooley, A. A. (1999) Extraction of *Escherichia coli* proteins with organic solvents prior to two-dimensional electrophoresis. *Electrophoresis* **20**, 701–704.
43. Pieper R., Su, Q., Gatlin, C. L., Huang, S.-T., Anderson, N. L., and Steiner, S. (2003) Multi-component immunoaffinity subtraction chromatography: an innovative step towards a comprehensive survey of the human plasma proteome. *Proteomics* **3**, 422–432.
44. Pieper, R., Gatlin, C. L., Makusky, A. J., et al. (2003) The human serum proteome: display of nearly 3700 chromatographically separated spots on two-dimensional electrophoresis gels and identification of 325 distinct proteins. *Proteomics* **3**, 1345–1364.
45. Wang, Y. Y., Cheng, P., and Chan, D. W. (2003) A simple affinity spin tube filter method for removing high-abundant common proteins or enriching low-abundant biomarkers for serum proteomic analysis. *Proteomics* **3**, 243–248.
46. Ahmed, N., Barker, G., Oliva, K., et al. (2003) An approach to remove albumin for the proteomic analysis of low abundance biomarkers in human serum. *Proteomics* **3**, 1980–1987.
47. Haney, P. J., Draveling, C., Durski W., Romanowich, K., and Qoronfleh, M. W. (2003) SwellGel: a sample preparation affinity chromatography technology for high throughput proteomic applications. *Protein Exp. Purif.* **28**, 270–279.
48. Gygi, S. P., Han, D. K., Gingras, A. C., Sonnenberg, N., and Aebersold, R. (1999) Protein analysis by mass spectrometry and sequence database searching: tools for cancer research in the post-genomic era. *Electrophoresis* **20**, 310–319.
49. Smolka, M., Zhou, H., and Aebersold, R. (2002) Quantitative protein profiling using two-dimensional gel electrophoresis, isotope-coded affinity tag labeling and mass spectrometry. *Mol. Cell. Proteomics* **1**, 19–29.
50. Liotta, L. A., Ferrarri, M., and Petricoin, E. (2003) Written in blood. *Nature* **425**, 905.
51. Mehta, A. I., Ross, S., Lowenthal, M. S., et al. (2003–2004) Biomarker amplification by serum carrier protein binding. *Disease Markers* **19**, 1–10.
52. Berman, D. M., Shih, I.-M., Burke, L.-A., et al. (2004) Profiling the activity of G proteins in patient-derived tissues by rapid affinity-capture of signal transduction proteins (GRASP). *Proteomics* **4**, 812–818.
53. Rabilloud, T., Adessi, C., Giraudel, A., and Lunardi, J. (1997) Improvement of the solubilization of proteins in two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis* **18**, 307–316.
54. Rabilloud, T. (1998) Use of thiourea to increase the solubility of membrane proteins in two-dimensional electrophoresis. *Electrophoresis* **19**, 758–760.
55. Luche, S., Santoni, V., and Rabilloud, T. (2003) Evaluation of nonionic and zwitterionic detergents as membrane protein solubilizers in two-dimensional electrophoresis. *Proteomics* **3**, 249–253.
56. Lanne, B., Potthast, F., Hogland, A., et al. (2001) Thiourea enhances mapping of the proteome from murine white adipose tissue. *Proteomics* **1**, 819–828.
57. Giavalalisco, P., Nordhoff, E., Lehrach, H., Gobom, J., and Klose, J. (2003) Extraction of proteins from plant tissues for two-dimensional electrophoresis analysis. *Electrophoresis* **24**, 207–216.
58. Henningsen, R., Gale, B. L., Straub, K. M., and DeNagel, D. C. (2002) Application of zwitterionic detergents to the solubilization of integral membrane proteins for two-dimensional gel electrophoresis and mass spectrometry. *Proteomics* **2**, 1479–1488.
59. Gall, A.-L., Ruff, M., and Moras, M. (2002) The dual role of CHAPS in the crystallization of stromelysin-3 catalytic domain. *Acta Cryst. D* **59**, 603–606.

60. Schuck, S., Honsho, M., Ekroos, K., Shevchenko, A., and Simons, K. (2003) Resistance of cell membranes to different detergents. *Proc. Natl. Acad. Sci. USA* **100**, 5795–5800.
61. Umbreit, J. N. and Strominger, J. L. (1973) Relation of detergent HLB number to solubilization and stabilization of D-alanine carboxypeptidase from *Bacillus subtilis* membranes. *Proc. Natl. Acad. Sci. USA* **70**, 2997–3001.
62. Duval-Terrie, C., Cosette, P., Molle, G., Muller, G., and De, E. (2003) Amphiphilic biopolymers (amphibiopolis) as new surfactants for membrane protein solubilization. *Protein Science* **12**, 681–689.
63. Stevens, S. M., Jr., Zharikova, A. D., and Prokai, L. (2003) Proteomic analysis of the synaptic plasma membrane fraction isolated from rat forebrain. *Mol. Brain Res.* **117**, 116–128.
64. Seigneurin-Berny, D., Rolland N., Garin, J., and Joyard, J. (1999) Differential extraction of hydrophobic proteins from chloroplast envelope membranes: a subcellular-specific proteomic approach to identify rare intrinsic membrane proteins. *Plant J.* **19**, 217–228.
65. Santoni, V., Kiefer, S., Desclaux, D., Masson, F., and Rabilloud, T. (2000) Membrane proteomics: Use of additive main effects with multiplicative interaction model to classify plasma membrane proteins according to their solubility and electrophoretic properties. *Electrophoresis* **21**, 3329–3344.
66. Molloy, M., Phadke, N. D., Maddock, J. R., and Andrews, P. C. (2001) Two-dimensional electrophoresis and peptide mass fingerprinting of bacterial outer membrane proteins. *Electrophoresis* **22**, 1686–1696.
67. Hauser, H. (2000) Short-chain phospholipids as detergents. *Biochim. Biophys. Acta* **1508**, 164–181.
68. Le Maire, M., Champeil, P., and Moller, J. V. (2000) Interaction of membrane proteins and lipids with solubilizing detergents. *Biochim. Biophys. Acta* **1508**, 86–111.
69. Blonder J., Goshe, M. B., Moore, R. J., et al. (2002) Enrichment of integral membrane proteins for proteomic analyses using liquid chromatography-tandem mass spectrometry. *J. Proteome Res.* **1**, 351–360.
70. Qoronfleh, M. W., Benton, B., Ignacio, R., and Kaboord, B. (2003) Selective enrichment of membrane proteins by partition phase separation for proteomic studies. *J Biomed. Biotechnol.* **4**, 249–255.
71. Dry, I. B. and Robinson, S. P. (1994) Molecular cloning and characterization of grape berry polyphenol oxidase. *Plant Mol. Biol.* **26**, 495–502.
72. Baxter, N. J., Lilley, T. H., Haslam, E., and Williamson, M. P. (1997) Multiple interactions between polyphenols and a salivary proline-rich protein repeat result in complexation and precipitation. *Biochemistry* **36**, 5566–5577.
73. Veljovic-Jovanovic, S., Noctor, G., and Foyer, C. H. (2003) Are leaf hydrogen peroxide concentrations commonly overestimated? The potential influence of artefactual interference by tissue phenolics and ascorbate. *Plant Physiol. Biochem.* **40**, 501–507.
74. Molina, M. C., Crespo, A., Vicente, C., and Elix, J. A. (2003) Differences in the composition of phenolics and fatty acids of cultured mycobiont and thallus of *Physconia distorta*. *Plant Physiol. Biochem.* **41**, 175–180.
75. Davidsen, N. B. (1995) Two-dimensional electrophoresis of acidic proteins isolated from ozone-stressed Norway spruce needles (*Picea abies* L. Karst): Separation method and image processing. *Electrophoresis* **16**, 1305–1311.
76. Koonjul, P. K., Brandt, W. F., Farrant, J. M., and Lindsey, G. G. (1999) Inclusion of polyvinylpyrrolidone in the polymerase chain reaction reverses the inhibitory effects of polyphenolic contamination of RNA. *Nucleic Acids Res.* **27**, 915–916.
77. Hoving, S., Gerrits, B., Voshol, H., Muller, D., Roberts, R. C., and van Oostrum, J. (2002) Preparative two-dimensional gel electrophoresis at alkaline pH using narrow range immobilized pH gradients. *Proteomics* **2**, 127–134.

78. Herbert, B. R., Molloy, M. P., Gooley, A. A., Walsh B. J., Bryson, W. G., and Willaims, K. L. (1998) Improved protein solubility in two-dimensional electrophoresis using tributyl phosphine as reducing agent. *Electrophoresis* **19**, 845–851.
79. Luche, S., Diemer, H., Tastet, C., et al. (2004) About thiol derivatization and resolution of basic proteins in two-dimensional electrophoresis. *Proteomics* **4**, 551–561.
80. Santoni, V., Rabilloud, T., Doumas, P., et al. Towards the recovery of hydrophobic proteins on two-dimensional gels. *Electrophoresis* **20**, 705–711.
81. Santoni, V., Doumas, P., Rouquie, D., Mansion, M., Rabilloud, T., and Rossignol, M. (1999) Large Scale characterization of plant plasma membrane proteins. *Biochimie* **81**, 655–661.
82. Kamo, M., Kawakami, T., Miyatake, N., and Tsugita, A. (1995) Separation and characterization of *Arabidopsis thaliana* proteins by two-dimensional gel electrophoresis. *Electrophoresis* **16**, 423–430.
83. Rouquie, D., Peltier, J. B., Marquis-Mansion, M., Tournaire, C., Doumas, P., and Rossingnol, M. (1997) Construction of a directory of tobacco plasma membrane proteins by combined two-dimensional gel electrophoresis and protein sequencing. *Electrophoresis* **18**, 654–660.
84. Tsugita, A. and Kamo, M. (1999) N-terminal amino acid sequencing of 2-DE spots. *Methods Enzymol.* **112**, 95–97.
85. Porubleva, L., Vander Veldin, K., Kothari, S., Livier, D. J., and Chitnis, P. R. (2001) The proteome of maize leaves: use of gene sequences and expressed sequence tag data for identification of proteins with peptide mass fingerprints. *Electrophoresis* **22**, 1724–1738.
86. Gegenheimer, P. (1990) Preparation of extracts from plants. *Methods Enzymol.* **182**, 174–193.
87. Bak-Jensen, K. S., Laugesen, S., Roepstorff, P., and Svensson, B. (2004) Two-dimensional gel electrophoresis pattern (pH 6–11) and identification of water-soluble barley seed and malt proteins by mass spectrometry. *Proteomics* **4**, 728–742.
88. Damerval, C., de Vienne, D., Zivy, M., and Thiellement, H. (1986) Technical improvements in two-dimensional electrophoresis increase the level of genetic variation detected in wheat-seedling proteins. *Electrophoresis* **7**, 52–54.
89. Alban, A., David, S. O., Bjorkesten, L., et al. (2003) A novel experimental design for comparative two-dimensional gel analysis: Two-dimensional gel electrophoresis incorporating a pooled internal standard. *Proteomics* **3**, 36–44.
90. Yan, J. X., Devenish, A. T., Wait, R., Stone, T., Lewis, S., and Fowler, S. (2002) Fluorescence two-dimensional difference gel electrophoresis and mass spectrometry based proteomic analysis of *Escherichia coli*. *Proteomics* **2**, 1682–1698.
91. Ruepp, S. U., Tonge, R. P., Shaw, J., Wallis, N., and Pognan, F. (2002) Genomics and proteomics analysis of acetoaminophen toxicity in mouse liver. *Toxicological Sciences* **65**, 135–150.
92. Kleno, T. G., Leonardsen, L. R., Kjeldal, H. O., Laursen, S. M., Jensen, O. N., and Baunsgaard, D. (2004) Mechanisms of hydrazine toxicity in rat liver investigated by proteomics and multivariate data analysis. *Proteomics B* **4**, 868–880.
93. Von Eggling, F., Gawriljuk, A., Fiedler, W., et al. (2001) Fluorescent dual colour 2D-protein gel electrophoresis for rapid detection of differences in protein pattern with standard image analysis software. *Int. J. Mol. Med.* **8**, 373–377.
94. Tyagarajan, K., Pretzer, E., and Wiktorowicz, J. E. (2003) Thiol-reactive dyes for fluorescence labeling of proteomic samples. *Electrophoresis* **24**, 2348–2358.
95. Rekhter, M. D. and Chen, J. (2001) Molecular analysis of complex tissues is facilitated by laser capture microdissection: critical role of upstream processing. *Cell. Biochem. Biophys.* **35**, 103–113.
96. Banks, R. E., Dunn, M. J., Forbes, M. A., et al. (1999) The potential use of laser capture microdissection to selectively obtain distinct populations of cells for proteomic analysis—preliminary findings. *Electrophoresis* **20**, 689–700.

97. Craven, R. A. and Banks, R. E. (2001) Laser capture microdissection and proteomics: possibilities and limitation. *Proteomics* **1**, 1200–1204.
98. Jain, K. K. (2002) Recent advances in oncoproteomics. *Curr. Opin. Mol. Ther.* **4**, 203–209.
99. Ornstein, D. K., Gillespie, J. W., Paweletz, C. P., et al. (2000) Proteomic analysis of laser capture microdissected human prostate cancer and in vivo prostate cell lines. *Electrophoresis* **21**, 2235–2242.
100. Wu, S.-L., Hancock, W. S., Goodrich, G. G., and Kunitake, S. T. (2003) An approach to the proteomic analysis of a breast cancer cell line (SKBR-3). *Proteomics* **3**, 1037–1046.
101. Mouledous, L., Hunt, S., Harcourt, R., Harry, J., Williams, K. L., and Gutstein, H. B. (2003) Navigated laser capture microdissection as an alternative to direct histological staining for proteomic analysis of brain samples. *Proteomics* **3**, 610–615.
102. Nakazono, M., Qiu, F., Borsuk, L. A., and Schnable, P. S. (2003) Laser-capture microdissection, a toll for the global analysis of gene expression in specific plant types: identification of genes expressed differentially in epidermal cells or vascular tissues of maize. *Plant Cell* **15**, 583–596.
103. Malone, J. P., Radabaugh, M. R., Leimgruber, R. M., and Gerstenecker, G. S. (2001) Practical aspects of fluorescent staining for proteomic applications. *Electrophoresis* **22**, 919–932.
104. Ruebelt, M. C., Lipp, M., Jany, KI.-D., et al. (2003) Novel Foods—Safety Assessment: Method Development for Proteome Analysis of *Arabidopsis* Seeds Produced by Different Ecotypes (Accessions) and by Transgenic Events, *Proceedings EURO FOOD CHEM XII*, Strategies for Safe Food: Challenges in Organization and Communication, 24–26 September 2003, Brugge, Belgium, 189–192, ISBN number 90-804957-2-7.

Preparation of Bacterial Samples for 2-D PAGE

Brian Berg Vandahl, Gunna Christiansen, and Svend Birkelund

1. Introduction

Sample preparation is a very crucial step in two-dimensional (2-D) gel electrophoresis, in which the proteins of the sample must be brought into a state where they can be separated by isoelectric focusing in the first dimension. That is, they must be denatured, reduced, and solubilized, and they must be kept so during electrophoresis without changing their pI. The sample buffer for this purpose is traditionally called the lysis buffer.

In most bacterial studies it is the aim to solubilize as many proteins as possible to obtain the best possible representation of the total protein content or protein expression under the investigated biological circumstances. However, prefractionation or the successive application of different chemical reagents can be used to investigate bacterial proteins with certain characteristics. Be aware that the gels will reflect the proteome of the bacteria at the time the proteins are solubilized, and that preceding centrifugations or other manipulations may stress the bacteria and thus influence the protein profile.

The solubilization procedure is highly dependent on the nature of the sample. Some bacteria are readily lysed by the constituents of the lysis buffer, whereas others must be disrupted mechanically, and for some it may be necessary to remove the cell wall by enzymatic digestion prior to mechanical disruption. In **Subheading 3.1.**, a general protocol for solubilization will be given. It is important to stress that both the lysis buffer and the protocol always must be optimized for the sample in question. All reagents are described in the notes, together with common alternatives.

Some samples may contain nonprotein substances in amounts that are incompatible with first or second dimensional electrophoresis and thus have to be removed from the sample prior to the addition of lysis buffer. Salts and most other compounds that may disturb the first dimension can be removed by protein precipitation, as described in **Subheading 3.2.1.** If nucleic acids are present in high amounts, these may have to be removed by enzymatic digestion, which is described in **Subheading 3.2.2.**

Also, highly abundant proteins may cause problems by preventing optimal focusing in the first dimension or by masking large areas of the gel. In such cases, it may be necessary to carry out prefractionations or to specifically remove the abundant proteins by immunoprecipitation. Prefractionation can be obtained by isolation of specific organelles, by chemical extractions, or by chromatographic or electrophoretic techniques (**1**), but these methods fall beyond the scope of this chapter.

All procedures should be kept as simple as possible to ensure reproducibility and because proteolytic degradation must be considered a risk. When bacteria are disrupted, proteases that are present in the periplasmic space in a high number will be released and start degrading proteins in the sample if not inhibited. As folded proteins are less susceptible to proteolysis than denatured proteins, and as proteases are often more resistant to denaturation than most other proteins, solubilization in urea will often make the problem worse. However, most proteases will be inactivated by disruption in lysis buffer containing thiourea, in 2% sodium dodecyl sulfate (SDS), or in precipitation solution containing 10% trichloroacetic acid (TCA). Still, all handling of the sample after disruption of the bacteria should be carried out as quickly as possible and on ice to minimize proteolytic degradation, and it may be necessary to add protease inhibitors.

Several reagents used for the sample preparation are toxic and/or carcinogenic. For safety reasons, use protective gloves and glasses, and work in a fume hood when mixing the lysis buffer and handling samples in lysis buffer, during presolubilization with reducing agents, during precipitation, and when working with protease inhibitors.

2. Materials

1. Lysis buffer (*see Note 1*): 7 M urea (2.10 g) (*see Note 2*), 2 M thiourea (0.76 g) (*see Note 3*), 65 mM dithioerythritol (DTE) (650 μ L of a 0.5 M stock) (*see Note 4*), 4% (W/V) CHAPS (0.2 g) (*see Note 5*), 2% (v/v) Pharmalyte pH 3.0–10.0 (*see Note 6*), 40 mM Tris base (*see Note 7*), a trace of bromophenol blue (*see Note 8*).
2. Presolubilization solution: 2% SDS, 65 mM DTE.
3. Precipitation solution: 10% TCA in acetone, 20 mM DTE.
4. DNase/RNase solution: 1 mg/mL DNase I, 0.25 mg/mL RNase A, 50 mM MgCl₂.

3. Methods

3.1. General Solubilization Protocol

It is crucial that all bacteria be disrupted, so that the lysis buffer gains access to all proteins. In studies where multiple extractions with different chemical reagents are employed sequentially, it will mask the result if more and more bacteria are disrupted during the procedure. The best method for disruption is dependent on the type of bacteria. In most cases, disruption by sonication will do, but it may be necessary to add lysosyme to break down cell walls. It is advantageous to perform the disruption in SDS or in lysis buffer containing thiourea, in which most proteases are denatured (2). If prolonged manipulation, such as fractionation by different steps of centrifugation, must be performed, protease inhibitors should be added (*see Note 9*). In the procedure described below, the bacteria are sonicated in 2% SDS, 65 mM DTE, and boiled in order to enhance the protein solubilization in general (3). It has been suggested (4) that SDS used for presolubilization does not interfere with first-dimensional electrophoretic separation because it forms micelles with the nonionic detergent of the lysis buffer and migrates out of the strip. Still, the amount of SDS should be kept low compared to the amount of detergent in the lysis buffer (5) (*see Note 10*). **Figure 1** shows a silver-stained gel loaded with 100 μ g of *Chlamydia pneumoniae* protein that was presolubilized by boiling in 1% SDS, 50 mM Tris-HCl, and diluted in the described lysis buffer to a final concentration of 0.1% SDS (*see Note 11*).

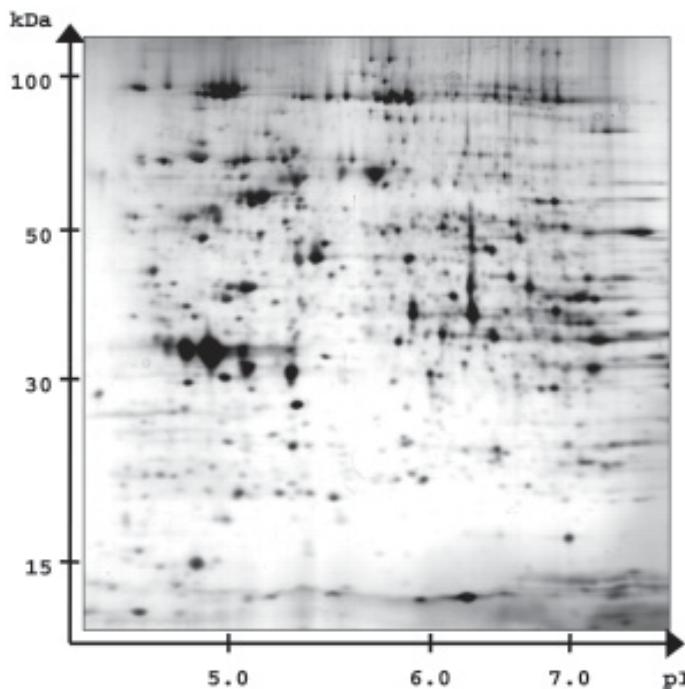


Fig. 1. A pellet of purified *Chlamydia pneumoniae* elementary bodies (12) was resuspended in 1% SDS, 50 mM Tris-HCl, pH 7.0, sonicated and boiled for 5 min. Cooled sample was diluted 1:10 in lysis buffer, sonicated briefly, left at room temperature for 1 h and centrifuged at 20,000g for 15 min. 350 μ L of the supernatant containing 100 μ g protein was loaded onto a pH 3.0–10.0 nonlinear immobilized pH gradient strip (Amersham Biosciences) and focused for 120,000 Volt hours. Second dimension was 9–16% linear gradient SDS-PAGE. The gel was silver stained.

The protocol below describes the solubilization of proteins from pelleted bacteria. If proteins have been precipitated as a purification step, the lysis buffer should be added directly (**step 8**) in the highest possible amount (*see Note 12*).

1. Start out with an appropriate amount of bacteria as pellet (*see Note 13*).
2. Add four times the pellet volume of 2% SDS, 65 mM DTE (*see Note 14*).
3. Sonicate three times for 2–20 s depending on sample size (*see Note 15*).
4. Spin briefly to collect the sample (*see Note 16*).
5. Resuspend any pellet that may have formed.
6. Boil for 5 min (*see Note 17*).
7. Allow the sample to cool.
8. Add 8 vols of lysis buffer to one vol of extract (*see Note 18*).
9. Sonicate three times for 5 s, cool between sonications (*see Note 19*).
10. Leave the sample on a rocking table for 30 min.
11. Spin at 20,000g for 15 min and collect the supernatant (*see Note 20*).
12. Assess the protein concentration (*see Note 21*).
13. Run first dimension immediately or store the sample at -70°C for several months (*see Note 22*).

3.2. Sample Purification

Common contaminants in 2-D PAGE studies are salts, small ionic compounds, polysaccharides, nucleic acids, and lipids. Salt is the most likely reason if bad first-dimensional focusing is observed. Enhanced conductivity and water migration in the strip due to high concentrations of salts will cause horizontal streaks. The concentration of salt should be below 10 mM when samples are loaded by strip rehydration. Small charged substances may likewise disturb the isoelectric focusing. Polysaccharides may clog the gel of the strip and may complex proteins by electrostatic interactions. Lipids may also clog the gel but are mainly a problem due to complexing of hydrophobic proteins and binding of detergent. Nucleic acids may clog the gel, bind proteins through electrostatic interactions, and cause streaking, especially in silver staining.

Dialysis and precipitation (**Subheading 3.2.1.**) are straightforward and effective ways to reduce the concentration of salt and small ionic compounds to an acceptable level. Dialysis causes a minimal loss of sample, but requires relatively large volumes of solute and is rather time consuming. Spin dialysis using, for instance, Amicon Ultra from Millipore is faster and requires no extra volume of solute, but protein may be lost by adsorption onto the dialysis membrane. Precipitation may also be used to remove polysaccharides and to some extent lipids. Large polysaccharides can be removed by ultracentrifugation. If lipids are causing major problems, the amount and nature of detergent must be optimized for the particular sample. High amounts of nucleic acids may require treatment with DNase/RNase (**Subheading 3.2.2.**).

The presence of proteases is likely to be a problem during sample purification, and in that case protease inhibitors must be added (*see Note 9*). It must be stressed that sample purification preceding addition of lysis buffer should be carried out only if necessary and not as a standard part of the sample preparation.

3.2.1. Precipitation

Precipitation is very efficient for removal of most contaminants, including salts, but no precipitant will precipitate all proteins, and some proteins will be difficult to resuspend following precipitation. This is especially a problem when a picture of the total protein content is desired.

A combination of TCA and acetone is the most common precipitant in 2-D PAGE studies, as it is more effective than either of these reagents alone. Besides, very few proteases are active in 10% TCA. Resolubilization is easier after precipitation with acetone alone (75% final concentration), but this gives a less complete precipitation. The TCA/acetone precipitation described here is essentially as in **ref. 6**.

1. Add 10% TCA in ice-cold acetone with 20 mM DTE to the sample (*see Note 23*).
2. Leave at -20°C for 2 h (*see Note 24*)
3. Centrifuge at 10,000g for 10 min.
4. Wash with cold acetone containing 20 mM DTE.
5. Repeat wash.
6. Let the pellet dry to remove residual acetone.
7. Resuspend pellet in lysis buffer (*see Subheading 3.1.*).

3.2.2. DNase/RNase Treatment

If nucleic acids are present in high amounts, the sample will appear viscous and a smear will be seen after silver staining. If ultracentrifugation does not solve the problem, enzymatic digestion will.

1. Add 1/10 of the sample volume of a solution containing 1 mg/mL DNase I, 0.25 mg/mL RNase A, and 50 mM MgCl₂ (see **Note 25**).
2. Incubate on ice for 20 min.

4. Notes

1. Absolute amounts are to make 5 mL. All reagents must be analytical grade. Use doubly distilled water. The solution is best mixed in a 10-mL tube on a rotating device. The solution should be made fresh before use or alternatively frozen in aliquots at -70°C and only thawed once. The solution must not be heated above 37°C.
The composition of the lysis buffer is essential for the final result of 2-D PAGE, and different lysis buffers will be optimal for different samples, and for different proteins in one sample. The function of the lysis buffer described here is to bring as many proteins in the sample as possible into solution and keep them in solution during electrophoresis. As isoelectric focusing is best carried out under denaturing and reducing conditions, the lysis buffer should solubilize, denature, and reduce the proteins of the sample. At the same time, the lysis buffer must not change the pI of the proteins, and it must not be highly conductive; hence, uncharged components are preferred. Most lysis buffers are still based on that introduced by O'Farrell in 1975 (4), containing urea as denaturing agent, a detergent, a reducing agent, and carrier ampholytes. The standard lysis buffer described here is based on (3), and the characteristics of each reagent are described in the following notes.
2. Urea—(NH₂)₂CO—is a noncharged chaotrope that disrupts noncovalent bonds and thereby denatures proteins. It is the main denaturant in all lysis buffers used in 2-D PAGE, and it can be brought into solution in concentrations up to 9.8 M if no thiourea is added. Urea in solution is in equilibrium with ammonium cyanate, which in the form of isocyanic acid will react with amino groups of lysine and arginine residues and the amino terminus of proteins causing carbamylation. The carbamylation of an amine group removes a positive charge from the protein, causing a shift towards the acidic side in the gel. Furthermore, it prevents N-terminal sequencing and some enzymatic digests. To avoid carbamylation, use only freshly prepared urea solutions. A urea solution should not be left at room temperature for long periods and should never be heated above 37°C (7).
3. Thiourea—(NH₂)₂CS—improves the solubilization of especially hydrophobic proteins during first dimension (8), and in combination with urea it can be used in concentrations up to 2.5 M. The addition of thiourea reduces the solubility of urea, and combinations of 7 M urea and 2 M thiourea or 8 M urea and 0.5 M thiourea are most common. The addition of thiourea to the lysis buffer has a pronounced inhibitory effect on proteases, which may still be active in high concentrations of urea alone (2). As thiourea can hinder the binding of SDS to proteins, it should not be included in the buffers used to equilibrate strips prior to second dimension (8).
4. DTE—MW: 154.3. Make a 0.5 M stock solution and store at -20°C. DTE has the same strong reducing power as dithiothreitol (DTT), and both can be used in concentrations from 10–100 mM. At alkaline pH, both DTE and DTT are charged and migrate towards the anode during first dimension, which may leave the basic end of the strip without reducing agent and hence cause streaking due to reoxidation and precipitation. Also, 2-mercaptoethanol can be a problem to use in first dimension due to ionization at alkaline pH. Besides, 2-mercaptoethanol does not have the same reducing power as DTE and DTT.

An alternative and very strong reducing agent is tributyl phosphine (TBP). It can be used in concentrations as low as 2 mM and is noncharged, meaning that it keeps all proteins reduced throughout the first dimension, thereby enhancing the resolution (9). TBP is stable, but spontaneously inflammable in air. Make a 200-mM stock in anhydrous isopropanol and store under nitrogen at 4°C (9).

5. CHAPS—3-[(3-cholamidopropyl)dimethylammonio]-1-propanesulfonate—is a zwitterionic detergent that is used in many 2-D PAGE studies. Zwitterionic or nonionic detergents are preferred to anionic detergents like SDS, which interfere with the isoelectric focusing in first dimension (see Note 1). The efficiency of many zwitterionic detergents has been investigated, and sulfobetaines with a hydrophobic tail of 12–16 alkyl carbons and an empirically determined linker in between have been found to be good alternatives to CHAPS and superior for some samples (10). Which detergent is best for a given sample can still not be predicted. The zwitterionic agent amido sulfobetaine (ASB)-14 ($C_{22}H_{46}N_2O_4S$, Calbiochem), with a 14-carbon alkyl tail; the nonionic Triton X-100 ($C_{14}H_{22}O(C_2H_4O)_n$, with an average number (n) of ethylene oxide of 9 to 10); or the maltoside *n*-dodecyl β -D-maltoside ($C_{24}H_{46}O_{11}$, Sigma-Aldrich) would be good first-choice alternatives if CHAPS does not give satisfactory results (11).
6. Pharmalyte 3–10 can be used for most immobilized pH gradient strips, but if narrow strips or very basic strips are used, carrier ampholytes that match the pH range of the strip should be chosen. Ask the strip supplier if in doubt. Pharmalyte 3–10 is a mixture of carrier ampholytes with pI between 3 and 10. These are small amphoteric compounds with a molecular weight below 1 kDa that have a high buffering capacity at their pI but do not bind proteins due to their high hydrophilicity. When using 2% v/v of Pharmalyte 3–10 (Amersham Pharmacia), it gives a final concentration of carrier ampholytes of 0.72% in the lysis buffer, since Pharmalyte 3–10 is 36% (w/v). Carrier ampholytes help keep proteins in solution during first dimension and especially prevent hydrophobic interactions between proteins and the immobilized pH gradient in the basic end of the strip. Furthermore, the precipitation of nucleic acids is improved by carrier ampholytes.
7. Tris base is added to raise pH of the lysis buffer to 8.5. Without the addition of base, pH of the lysis buffer would be about 5.5. At alkaline pH, more proteins will be anionic and thus not bind to DNA. However, the pH for optimal solubilization will vary between samples, and Tris base is left out in many studies.
8. Bromphenol blue should be added in a small amount to color the solution lightly blue. The color will move towards the anode during first dimension, which can be used to check that the isoelectric focusing is ongoing. However, the color will disappear before the first dimension is finished and cannot be used as an indicator of when to stop.
9. Protease-inhibitor cocktails are available from most commercial laboratory reagent suppliers, but most proteases will be inhibited by adding 2 mM ethylenediaminetetraacetic acid (EDTA), 1 mM phenylmethylsulfonyl fluoride (PMSF), 1 μ M Pepstatin A, and 13 μ M Bestatin. EDTA chelates free metal ions, thereby inhibiting metalloproteases; make a 0.5-*M* stock solution in water, pH 8.0. PMSF inhibits serine proteases and some cysteine proteases; make a 100-mM stock solution in methanol. Pepstatin A inhibits aspartic proteases; make a 1-mM stock solution in methanol. Bestatin inhibits aminopeptidases; make a 13-mM stock solution.
10. When NP-40 is used as detergent in the lysis buffer, it has been reported (5) that the ratio of NP-40 to SDS should be at least 8 to avoid streaking. NP-40 (Nonidet P-40) is a nonionic detergent that is very similar to Triton X-100, and the properties are often reported as being identical. NP-40 (Roche) is $C_{15}H_{24}O(C_2H_4O)_n$, where *n* = 9–10 on average.
11. No reducing agent was added during presolubilization.

12. If the sample is applied by strip rehydration of 18-cm strips, the maximum amount is 350 μ L per strip.
13. For most bacteria, 25–100 μ g of protein is appropriate for silver staining, 100–150 μ g for immunoblotting, and 0.5–2 mg for preparative gels when 18-cm immobilized pH gradient strips in the pH range of 3.0–10.0 or similar broad range intervals are used.
14. Be sure not to add more SDS than can be diluted to 0.25% in lysis buffer. If all the sample is to be used for one gel using 350 μ L lysis buffer to rehydrate an immobilized pH gradient strip, this means that no more than 40 μ L of 2% SDS should be added. However, if streaking is observed, try lowering the amount of SDS used for presolubilization. For some samples, pH must be buffered to optimize solubilization. Use for instance 50–100 mM Tris-HCl, pH 7.0. The optimal pH may vary from sample to sample. Be aware that the final concentration of salt should not exceed 10 mM when samples are loaded by strip rehydration (see **Subheading 3.2.**).
15. Adjust the amplitude of the sonicator so that microbubbles are formed, and keep the tip of the probe deep in the sample to avoid too much foam formation. The sample should be cooled between sonifications.
16. It cannot be avoided that some foam is formed during sonication, and this should be spun down before the sample is boiled, in order to avoid protein coagulation in drying bubbles.
17. DTE develops toxic gas upon heating. Boil in a fume hood.
18. Dilute the sample as much as possible in lysis buffer (see **Notes 13** and **14**).
19. When the sample is in lysis buffer containing urea, it is important to keep the temperature below 37°C in order to avoid carbamylation of proteins. If the SDS/boiling step is left out, the sonication may have to be extended.
20. The centrifugation step is important to remove cell debris and precipitated DNA, and it should not be left out.
21. Several constituents of the lysis buffer may cause problems for assessment of protein concentration. Carrier ampholytes, CHAPS, and other detergents will bind most dyes, and reduction of cupric ion cannot be employed in the presence of thiourea and DTE. Hence, the protein must be selectively precipitated and then measured. This may be done with the 2-D Quant Kit from Amersham Pharmacia. Alternatively, the protein may be estimated in a parallel sample that is not solubilized in lysis buffer. This may not give the actual protein concentration in the lysis buffer but will in most cases provide adequate information to determine the load.
22. Repeated freeze-thaw cycles should be avoided due to the risk of carbamylation and because the solubility of some proteins may be changed by the process.
23. The bacteria must be disrupted beforehand by an appropriate method (see **Subheading 3.1.**). If the bacteria are disrupted directly in the precipitation buffer, most proteases will be inactivated by the TCA.
24. The precipitation time should not be longer than the minimal time required for satisfactory precipitation. For some samples, 15 min will do, while others must be incubated overnight. Be aware that prolonged exposure to the very acidic solution may cause protein degradation.
25. The bacteria must be disrupted beforehand by an appropriate method (see **Subheading 3.1.**). If active proteases are present, it may be necessary to add protease inhibitors (see **Note 9**) even if the DNase/RNase treatment is carried out on ice.

References

1. Righetti, P. G., Castagna A., Herbert B., Reymond F., and Rossier J. S. (2003) Prefractionation techniques in proteome analysis. *Proteomics* **3**, 1397–1407.

2. Castellanos-Serra, L., and Paz-Lago, D. (2002) Inhibition of unwanted proteolysis during sample preparation: evaluation of its efficiency in challenge experiments. *Electrophoresis* **23**, 1745–1753.
3. Harder, A., Wildgruber, R., Nawrocki, A., Fey, S. J., Larsen, P. M., and Gorg, A. (1999) Comparison of yeast cell protein solubilization procedures for two-dimensional electrophoresis. *Electrophoresis* **20**, 826–829.
4. O'Farrell, P. H. (1975) High resolution two-dimensional electrophoresis of proteins. *J. Biol. Chem.* **250**, 4007–4021.
5. Ames, G. F. and Nikaido, K. (1976) Two-dimensional gel electrophoresis of membrane proteins. *Biochemistry* **15**, 616–623.
6. Jacobs, D. I., van Rijssen, M. S., van der Heijden, R., and Verpoorte, R. (2001) Sequential solubilization of proteins precipitated with trichloroacetic acid in acetone from cultured *Catharanthus roseus* cells yields 52% more spots after two-dimensional electrophoresis. *Proteomics* **1**, 1345–1350.
7. McCarthy, J., Hopwood, F., Oxley, D., et al. (2003) Carbamylation of proteins in 2-D electrophoresis—myth or reality? *J. Proteome Res.* **2**, 239–242.
8. Rabilloud, T., Adessi, C., Giraudel, A., and Lunardi, J. (1997) Improvement of the solubilization of proteins in two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis* **18**, 307–316.
9. Herbert, B. R., Molloy, M. P., Gooley, A. A., Walsh, B. J., Bryson, W. G., Williams, and K. L. (1998) Improved protein solubility in two-dimensional electrophoresis using tributyl phosphine as reducing agent. *Electrophoresis* **19**, 845–851.
10. Tastet, C., Charmont, S., Chevallet, M., Luche, S., and Rabilloud, T. (2003) Structure-efficiency relationships of zwitterionic detergents as protein solubilizers in two-dimensional electrophoresis. *Proteomics* **3**, 111–121.
11. Luche, S., Santoni, V., and Rabilloud, T. (2003) Evaluation of nonionic and zwitterionic detergents as membrane protein solubilizers in two-dimensional electrophoresis. *Proteomics* **3**, 249–253.
12. Vandahl, B. B., Birkelund, S., and Christiansen, G. (2002) Proteome analysis of *Chlamydia pneumoniae*. *Methods Enzymol.* **358**, 277–288.

Preparation of Yeast Samples for 2-D PAGE

Joakim Norbeck

1. Introduction

Yeasts are the focus of much research, both in their role as pathogens and as biotechnically important organisms, and not the least in their role as model systems for eukaryotic cells. In particular, *Saccharomyces cerevisiae* has also been the object of several proteomics-related efforts. However, the preparation of protein extract from yeast is complicated by the presence of a cell wall of mainly chitin and glucans, which needs to be disrupted in the extraction process. Several methods presenting solutions to this problem, in connection with sample preparation for two-dimensional polyacrylamide gel electrophoresis (2-D PAGE), have been described (1,2), in which the cell wall is broken either by vortexing in the presence of glass beads or by sonication.

We have found the method described below to be robust and to yield reproducible results in several studies (3–6). It is furthermore easy to perform and does not require any specialized equipment.

2. Materials

1. Sample buffer I: sodium dodecyl sulfate (SDS) (0.3 g), β -mercaptoethanol (5.0 mL), Tris-HCl (0.444 g), Tris base (0.266 g), MilliQ (or equivalent) water (to a final volume of 10 mL).
2. Sample buffer II (see Note 1): 1.5 M Tris base (80 μ L), 1.5 M Tris-HCl (1585 μ L), 1 M $MgCl_2$ (250 μ L), DNase I (Worthington Biochemical Corp., NJ) (5 mg), RNase A (Worthington Biochemical Corp., NJ) (1.25 mg), MilliQ (or equivalent) water (to a final volume of 5 mL).
3. Immobilized pH gradient (IPG)-rehydration buffer (urea/thiourea buffer): urea (4.8 g; gives a final concentration of 8 M; see Note 2), Triton X-100 or Nonidet P40 (100 μ L; see Note 3), 1 M DTT (100 μ L), IPG-buffer/ampholine (50 μ L; see Note 4), bromophenol blue (trace amount approx 0.01% w/v), MilliQ (or equivalent) water (to a final volume of 10 mL).

3. Method

The method described in the following sections can be divided into (1) an initial cell disruption step; (2) a protein solubilization step; (3) a nuclease treatment step; and (4) a final phase in which protein extract is diluted in IPG-rehydration buffer immediately prior to application on the first dimension of 2-D PAGE.

The procedure is carried out in 1.5-mL microcentrifuge tubes. A suitable starting material is a pellet of yeast cells from 10 mL of culture with a density of 5–10 million cells/mL, corresponding to an optical density (at 610 nm) of approx 0.5, which will typically yield a pellet of 5–10 μ L of cells. The method described below is adjusted to this amount of cells.

The protocol can be scaled up or down; however, care should be taken to not use a final extract volume of more than 500 μ L or less than 50 μ L, since this will reduce the efficiency of the cell-disruption step.

3.1. Cell Disruption

1. Add 160 μ L of ice-cold milliQ-quality water containing protease inhibitors (e.g., CompleteTM, Roche, Inc.) to the cell pellet.
2. Add 0.25 g of chilled glass beads (diameter 0.5 mm) to the sample.
3. Vortex 4 \times 30 s on a table shaker at maximum speed (approx 2500 rpm) with intermittent placement of samples on ice for at least 1 min.

3.2. Protein Solubilization

1. Add 20 μ L of sample buffer I and vortex the tube(s) briefly to mix.
2. Place tube(s) at 95°C for 5 min. Make sure to secure the lid of the tube, alternatively to make a small hole in the lid, prior to the heating step.
3. Cool samples on ice for 5 min.

3.3. Nuclease Treatment

1. Add 20 μ L of sample buffer II and vortex the tube(s) briefly to mix.
2. Incubate on ice for 10 min.
3. Centrifuge samples at full speed (approx 15,000g) in a microcentrifuge at 4°C.
4. Aspire the supernatant (constituting the protein extract) to a new microcentrifuge tube and freeze at –20°C, or use immediately.

3.4. Dilution in IPG-Rehydration Buffer

1. Dissolve sample in required volume of IPG-rehydration buffer (see **Note 5**).
2. Incubate the sample at 37°C for 10 min.
3. Spin down sample (15,000g, 10 min, room temperature) to remove any particles that might remain.

4. Notes

1. The DNase and RNase should be dissolved in the buffer as the final step.
2. 8 M Urea can be substituted by a combination of 7 M urea and 2 M thiourea. The chaotropic agent thiourea is the most highly beneficial addition to the IPG-rehydration buffer; this addition can strongly improve the solubility of many proteins which may produce “streaking” or which are completely absent on 2-D PAGE gels run with normal urea-based buffer in the first dimension (7). However, thiourea requires a special permit from inspecting authorities in many countries due to its suspected carcinogenic properties.
3. Triton X-100 can be substituted for 2% (w/v) of CHAPS. CHAPS is considered to be the preferred detergent for the IPG-rehydration buffer (7). Unfortunately, it is also considerably more expensive than Triton X-100 and Nonidet P-40. We normally use Triton X-100, since in our hands the choice of detergent in the rehydration buffer has only a minor influence on the final protein resolution.

4. IPG buffer or ampholine should be chosen for each type of IPG strip, or for the desired separation interval.
5. The final volume to which the sample is diluted is determined by the system used for running this first dimension. We most often use the precast pH-gradient strips (supplied by, for example, Amersham Biosciences and BioRad, Inc.), for which the rehydration volume is commonly in the range of 125–500 μ L. However, the protein extract is also compatible with the original glass tube-based system for running 2-D PAGE (8), or for application on IPG strips using sample cups, in which the sample volume will typically be <100 μ L.

In the simplest case, the desired amount of protein extract to be loaded is less than 10% of the final volume and contains low amounts of salt. The extract can then be mixed directly with IPG-rehydration buffer to the required volume immediately prior to application on the first dimension.

If the desired amount of protein extract is greater than 10% of the final volume, or if the sample contains large amounts of substances that may disturb the isoelectric focusing (e.g., salts), additional steps such as precipitation or dialysis are required:

A simple and straightforward way to concentrate the sample and remove low-molecular-weight compounds is the use of precipitation, either by acetone or trichloroacetic acid, or a combination of the two (9). The precipitate is then dissolved in a proper amount of IPG-rehydration buffer.

Alternatively, the protein extract can be adjusted to a final concentration of 8 M urea (or 7 M urea/2 M thiourea), followed by dialysis against a large volume of IPG-rehydration buffer. In the latter case, a micro-dialysis kit with a cutoff of 1 kDa should be used (e.g., PlusOneTM Micro Dialysis kit, Amersham Biosciences). Following dialysis, the sample is adjusted to the required volume, using IPG-rehydration buffer.

However, it should be cautioned that both precipitation and dialysis may cause the loss of certain proteins from the extract.

The amount of extract to be diluted also depends on whether a constant amount of protein (in μ g) or a constant amount of radioactivity (in dpm of radioactively labeled amino acid) is desired. When measuring protein concentration in the extract (**Subheading 3.3.4.**), it is important to use a method that tolerates the presence of detergents and mercaptoethanol (e.g., a method which incorporates an initial protein precipitation step).

References

1. Blomberg, A. (2002) Use of two-dimensional gels in yeast proteomics. *Methods Enzymol.* **350**, 559–584.
2. Harder, A., Wildgruber, R., Nawrochi, A., et al. (1999) Comparison of yeast cell protein solubilization procedures for two-dimensional electrophoresis. *Electrophoresis* **20**, 826–829.
3. Norbeck, J. and Blomberg, A. (1995) Gene linkage of two-dimensional polyacrylamide gel electrophoresis resolved proteins from isogene families in *Saccharomyces cerevisiae* by microsequencing of in-gel trypsin generated peptides. *Electrophoresis* **16**, 149–156.
4. Norbeck, J. and Blomberg, A. (1997) Two-dimensional electrophoretic separation of yeast proteins using a nonlinear wide range (pH 3–10) immobilized pH gradient in the first dimension; reproducibility and evidence for isoelectric focusing of alkaline (pI >7) proteins. *Yeast* **13**, 1519–1534.
5. Norbeck, J. and Blomberg, A. (1997) Metabolic and regulatory changes associated with growth of *Saccharomyces cerevisiae* in 1.4 M NaCl. Evidence for osmotic induction of glycerol dissimilation via the dihydroxyacetone pathway. *J. Biol. Chem.* **272**, 5544–5554.

6. Blomberg, A., Blomberg, L., Norbeck, J., et al. (1995) Interlaboratory reproducibility of yeast protein patterns analyzed by immobilized pH gradient two-dimensional gel electrophoresis. *Electrophoresis* **16**, 1935–1945.
7. Rabilloud, T., Adessi, C., Giraudel, A., and Lunardi, J. (1997) Improvement of the solubilization of proteins in two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis* **18**, 307–316.
8. O'Farrell, P. H. (1975) High resolution two-dimensional electrophoresis of proteins. *J. Biol. Chem.* **250**, 4007–4021.
9. Ozols, J. (1990) Amino acid analysis. *Methods Enzymol.* **182**, 587–601.

Preparation of Mammalian Tissue Samples for Two-Dimensional Electrophoresis

Frank A. Witzmann

1. Introduction

The true power of two-dimensional gel electrophoresis (2-DE) requires the careful preparation of protein samples to minimize sample variability and maximize solubilization. This is of particular relevance when 2-DE is relied upon to characterize the differential expression of mammalian tissue proteomes in large experiments with large numbers of samples. One of the weaknesses of the 2-DE approach relates to its depth of field—that is, its limited ability to resolve reproducibly those proteins with extreme isoelectric points (pI) (e.g., <3 and >9), hydrophobicity, and low abundance. Second, the initial step in 2-DE of protein mixtures, isoelectric focusing, is susceptible to a number of problems that cause variability in the final protein pattern, interferences that must be avoided. Hence, a simple yet reproducible method of preparing cell and tissue samples for consistent 2-DE results, involving as little manipulation as possible, is of utmost importance.

Realistically, no single preparation method can be applied to all possible samples. Ideally, a complex protein mixture like rodent liver should be treated in a manner that completely disrupts the cells, denatures all protein complexes by breaking noncovalent interactions, denatures complex protein secondary and tertiary structure, prevents differential oxidation or other protein modifications, and solubilizes hydrophobic proteins. Whole tissue samples are disrupted by homogenization (e.g., ground-glass) and/or sonication, best accomplished directly in a “lysis buffer” for 2-DE. This buffer generally consists of a mixture of the following components:

1. Nonionic or zwitterionic detergents such as Nonidet® P40, Triton X-100, Igepal CA-630 ([octylphenoxy]polyethoxyethanol), CHAPS (3-[(cholamidopropyl)dimethylammonio]-propanesulfonate), or the nondetergent sulfobetaines aid in the solubilization of hydrophobic proteins. Having the least hydrophilicity of these detergents, CHAPS is said to be most adept at preparing hydrophobic proteins for separation on immobilized pH-gradient gel strips.
2. The chaotrope urea is often used to denature the secondary, tertiary, and quaternary protein structure, but can be combined with thiourea in samples containing a prominent membrane protein component.
3. Reducing agents such as dithiothreitol (DTT, Cleland’s reagent), its isomer dithioerythritol (DTE), or uncharged reducing agents such as tributyl phosphine (TBP) (1) are essential

for preventing thiol group oxidation and for reducing disulphides to dithiols, thereby avoiding horizontal streaking.

4. Carrier ampholytes are often included, especially in tube gel isoelectric focusing (IEF), where they have been claimed to aid in protein solubilization and in nucleic acid precipitation. Narrow-range alkaline mixtures (i.e., pH 8.0–10.5) also help maintain an alkaline pH in the solubilized sample (subverting acid protease activity) and assisting in the stabilization of the alkaline end of the tube gel during focusing. In preparing lysis buffers for immobilized pH-gradient (IPG) strip IEF, it is a good idea to include low amounts (<0.5%) of carrier ampholytes that correspond exactly to the IPG strip pH gradient, as these ionic species will migrate to their pI (slowing down the IPG run) and can generate visible gaps in the focused pattern if, for example, an excessively high percentage of a pH 8.0–10.5 mixture is used in broad range (pH 3.0–10.0) strips.
5. Finally, some researchers choose to avoid proteolysis with the addition of protease inhibitor cocktails. Typically, unless lengthy tissue-sample fractionation steps are necessary (2), these inhibitors are not essential.

This chapter describes general-purpose sample preparation methods for mammalian tissue and monolayer cell-culture samples destined for tube-gel and IPG-strip first-dimension isoelectric focusing. Preparation of whole tissues for 2-DE is generally rather straightforward. Tissue-culture samples are more problematic. During recovery of cells from monolayer cell-culture, scraping, washing, and centrifugal pelleting followed by solubilization results in the introduction of significant variability between samples. Consistently uniform protein recovery for 2-DE is accomplished by first removing the culture medium and then directly solubilizing the adherent cells (3). This approach exploits the small surface area of flat-bottom, multi-well culture plates (e.g., 6-well BD Falcon™ Cell Culture Plates) with or without collagen, and has been applied successfully to 2-DE analysis in primary hepatocyte cell culture (3), HeLa cell culture (4), and keratinocytes (NEHK) (5).

This chapter does not address special cases such as prefractionated tissue samples (e.g., membranes), in vitro preparations established in various cell lysis buffer formulations, or precipitated protein samples. It should be understood that while the approaches described in this chapter have been used successfully by the author to analyze whole-tissue lysates on 2-D gels in a variety of mammalian tissues (for quasi-global differential expression proteomics), one size never fits all, and optimization for unique samples is always a good idea.

2. Materials

2.1. Equipment

1. 6-well BD Falcon Cell Culture Plates.
2. 50-mL beakers, surgical scissors (12 cm, straight sharp/sharp).
3. 3- or 5-mL DUALL® ground-glass tissue grinders (no. 21 or no. 22, <http://www.kimble-kontes.com/html/pg-885450.html>).
4. Beckman TL-100 ultracentrifuge (or equivalent).
5. Beckman polyallomer centrifuge tubes (1/2 × 2 in).
6. TLA 100.3 rotor, 100K rpm max.
7. 2-mL polypropylene flat top microcentrifuge tubes (Fisherbrand).
8. Fisher Sonic Dismembranator.

2.2. Buffers

1. Lysis buffer A (for tube gel IEF with broad-range carrier ampholytes): 9 M urea, 4% Igepal CA-630 (I-3021, Sigma-Aldrich), 1% DTT, and 2% carrier ampholytes (pH 8.0–10.5) (Pharmalyte, 17-0455-01, Amersham Biosciences). This mixture should be placed in Eppendorf tubes in 1-mL aliquots and stored at –80°C. Repeated refreezing-thawing of the lysis buffer should be avoided.
2. Lysis buffer B (for IPG strips of any pH range): 9 M urea, 4% Igepal CA-630 (I-3021, Sigma-Aldrich), 1% DTT, and 2% carrier ampholytes (pH 3.0–10.0) (Pharmalyte, 17-0456-01, Amersham Biosciences)
3. For tissue homogenates (e.g., liver, kidney, and so on), 10–20 µL of solubilized sample is added to 480–490 µL of a rehydration buffer containing: 8 M urea, 2% CHAPS, 0.23% DTT, 0.04% ampholytes (pH 3.0–10.0) (Pharmalyte, 17-0456-01, Amersham Biosciences), and 0.02 mg/mL Orange G (O-1625, Sigma-Aldrich). The total 500 µL volume is used later to rehydrate (re-swell) a 24-cm IPG strip for IEF.

3. Methods

3.1. Fresh Tissue

1. Fresh tissues should be freed of connective tissue and fat, preferably perfused with ice-cold saline prior to excision or at least rinsed briefly after excision, and 250 mg placed in a 50-mL beaker at room temperature (RT). Rapidly add 8 vols of lysis buffer (RT) onto the tissue and thoroughly mince with surgical scissors. Cooling the urea-based lysis buffer will result in urea crystallization and is not recommended.
2. The minced tissue-sample slurry is then placed in a 3- or 5-mL DUAL ground-glass tissue grinder and manually homogenized. The duration and extent of manual homogenization depends on the connective-tissue content of the sample and should be continued until a uniform homogenate is formed and no tissue pieces are visible. Motor-driven homogenization should be avoided, as rapid pestle-rotation in ground glass will result in significant sample heating and protein carbamylation (*see Note 4*).
3. The resulting 12.5% homogenate is then allowed to remain at RT for 120 min, after which the samples are placed in Beckman polyallomer centrifuge tubes (1/2 × 2 in) and centrifuged at 100,000g for 30 min at 22°C using a Beckman TL-100 ultracentrifuge to remove nucleic acid and insoluble materials.
4. The resulting clarified supernate should then be aliquoted into small volumes that can be stored at –45°C or –80°C. Although tissue protein content (mg/g wet wt) can be quite variable across mammalian tissue types, the resulting protein concentration should be approx 15–20 mg/mL. Because tissue fluid abnormalities resulting from experimental conditions (e.g., edema, dehydration, and so on) can alter this estimate, it is desirable to measure protein concentration in the solubilized sample. This can be accomplished using any of a number of commercially available urea/detergent-compatible protein assays such as the Bradford Method (Pierce), RC DC Protein Assay (BioRad), and the Noninterfering Protein Assay™ (Genotech), keeping in mind their rather narrow linear range of detection.

3.2. Cultured Cells

1. The culture medium is first removed from the multi-well plate by aspiration. 400 µL of lysis buffer A or B (*see Notes 2 and 3*) is added directly to each well.
2. The culture plates are then placed in a 37°C incubator for 1 h with intermittent (every 15 min) manual agitation.

3. Following the 1-h solubilization, the entire volume is removed from each well and placed in 2-mL Eppendorf tubes.
4. Each sample is then sonicated with a Fisher Sonic Dismembranator using 3×2 -s bursts at instrument setting #3. Sonication is conducted every 15 min for 1 h. Avoid contacting the sides of the tube with the sonicator probe as this generates bubbles/suds and potential sample loss.
5. At this point, samples may be ultracentrifuged (as above), after which the solubilized samples are transferred to microcentrifuge tubes for storage at -45°C or -80°C until thawed for analysis.

4. Notes

1. Tissues containing a large proportion of collagenous connective tissue can be frozen in liquid N_2 , ground to a fine powder in a mortar and pestle with liquid N_2 added, and the powdered sample placed in the DUALL grinding tube with lysis buffer and solubilized as described above.
2. For buffers A and B, all reagents should be ultrapure. Given the critical nature of urea as a chaotrope and the importance of avoiding carbamylation, the author recommends the use of BDH Aristar (item no. 452046C).
3. When solubilizing monolayer-cultured cells as described above for IPG-based IEF, do not use lysis buffer A. The presence of high concentrations (2%) of carrier ampholytes in the 8.5–10.0 pH range together with the high sample application volumes needed for proper protein loading (40–50 μL sample in 450–460 μL rehydration buffer) results in a major vertical gap in the 2-D pattern near the basic end. We have determined the width of this gap is proportional to sample volume (unpublished results). The solution is to use lysis buffer B when IPG IEF is planned.
4. The 400- μL lysis buffer volume suggested above is a convenient volume for solubilizing hepatocytes, HeLa cells, and NEHK, all of which were estimated to number approx $1-2 \times 10^6$ cells/well in the various experiments. This volume was determined by trial and error to provide optimum cell lysis and solubilization on each surface (approx 10 cm^2) in the 6-well plate without diluting the cellular proteins to such an extent that loading volumes become impractically large. Thus, a 40- μL aliquot that contains approx 150–200 μg protein (representing approx 100,000–200,000 cells), depending on the tissue type, can be loaded on a tube gel or IPG strip (for large format 2-DE). Because many cells contain far less protein than this (i.e., human peripheral lymphocytes, approx 5- to 10-fold less), optimization is strongly encouraged.
5. It is well known that urea + heat + protein = carbamylation due to the formation of isocyanic acid from urea and its reaction with the N-terminus, arginines, lysines, and cysteines of proteins. Carbamylation, in turn, generates a variety of protein charge modifications, major artifacts in the horizontal dimension of a 2-D gel pattern. Based on our own observations, supported by a recently published evaluation (6), neither the RT sample treatment nor the brief incubation at 37°C in the urea-based lysis buffers described above have any artifactual effect on protein charge modification via carbamylation.

Acknowledgments

The author gratefully acknowledges the technical assistance of Heather Ringham, Sheng Liu, and Kevin Geiss; and careful reading of the manuscript by Dr. Mu Wang. Much of this work was supported by the Air Force Office of Scientific Research through grants F49620-96-1-0156 and F49620-99-1-0153, and ManTech/Geo-Centers Joint Venture Contract F33615-00-C-6060.

References

1. Herbert, B. R., Molloy, M. P., Gooley, A. A., Walsh, B. J., Bryson, W. G., and Williams, K. L. (1998) Improved protein solubility in two-dimensional electrophoresis using tributyl phosphine as reducing agent. *Electrophoresis* **19**, 845–851.
2. Righetti, P. G., Castagna, A., Herbert, B., Reymond, F., and Rossier, J. S. (2003) Prefractionation techniques in proteome analysis. *Proteomics* **3**, 1397–1407.
3. Witzmann, F. A., Clack, J. W., Geiss, K., et al. (2002). Proteomic evaluation of cell preparation methods in primary hepatocyte cell culture. *Electrophoresis* **23**, 2223–2232.
4. Decker, E. D., Zhang, Y., Cocklin, R. R. Witzmann, F. A., and Wang, M. (2003). Proteomic analysis of differential protein expression induced by ultraviolet light radiation in HeLa cells. *Proteomics* **3**, 2019–2027.
5. Witzmann, F. A. and Li, J. (2002). Cutting-edge technology. II. Proteomics: core technologies and applications in physiology. *Am. J. Physiol. Gastrointest. Liver Physiol.* **282**, G735–G741.
6. McCarthy, J., Hopwood, F., Oxley, D., et al. (2003). Carbamylation of proteins in 2-D electrophoresis—myth or reality? *J. Proteome Res.* **2**, 239–242.

Differential Detergent Fractionation of Eukaryotic Cells

Melinda L. Ramsby and Gregory S. Makowski

1. Introduction

Differential detergent fractionation (DDF) represents an alternative method for cell fractionation that employs sequential extraction of cells or tissues with detergent-containing buffers to partition cellular proteins into structurally and functionally intact and distinct compartments (1–5). Relative to cell fractionation by differential pelleting, DDF has the advantage of preserving the integrity of microfilament and intermediate-filament cytoskeletal networks, and is especially applicable to use with limited quantities of biomaterial (4–6). In addition, DDF is simple, highly reproducible, labor sparing, and ultracentrifuge independent. DDF is appropriate for a variety of investigations, including those aiming to: (1) enhance the detectability of low-abundance species or semi-purify components of known subcellular localization; (2) define the subcellular localization of enzymes, regulatory, or structural proteins as well as nonprotein metabolites; (3) monitor physiologic fluxes and compartmental redistribution of biomolecules under basal and stimulated conditions; (4) identify cytoskeletal-associated and interacting proteins; and (5) investigate the role of cytoskeletal networks in the subcellular localization of endogenous and exogenous factors, including mRNA, viral components, and heat-shock proteins—interactions relevant to understanding mechanisms of infection, protein turnover, and the stress response (7–15).

The DDF protocol detailed here reproducibly partitions cellular proteins into distinct cytoskeletal and noncytoskeletal compartments that can be directly analyzed by two-dimensional gel electrophoresis. This straightforward methodology is applicable to a variety of studies and cell types relevant to basic and clinical research. Overall, the protocol entails the sequential extraction of cells (suspension cultures or monolayers) with three detergent-containing buffers: (1) digitonin/ethylenediaminetetraacetic acid (EDTA); (2) Triton X-100/EDTA; and (3) Tween-40/deoxycholate (see **Fig. 1**). All solutions are pH-adjusted at 4°C and contain the organic buffer PIPES, as well as the neutral serine protease inhibitor PMSF.

DDF reproducibly yields four electrophoretically distinct fractions, which contain (1) cytosolic proteins and soluble cytoskeletal elements, (2) membrane and organellar proteins, (3) nuclear membrane and soluble nuclear proteins, and (4) detergent-resistant cytoskeletal filaments with nuclear matrix proteins (see **Fig. 2** and **Table 1**). Biochemical, immunochemical, and electrophoretic characterization of these fractions were recently described (16). We have used this procedure to successfully fractionate a variety of cell types incubated in suspension culture (hepatocytes, neutrophils) or grown

From: *The Proteomics Protocols Handbook*
Edited by: J. M. Walker © Humana Press Inc., Totowa, NJ

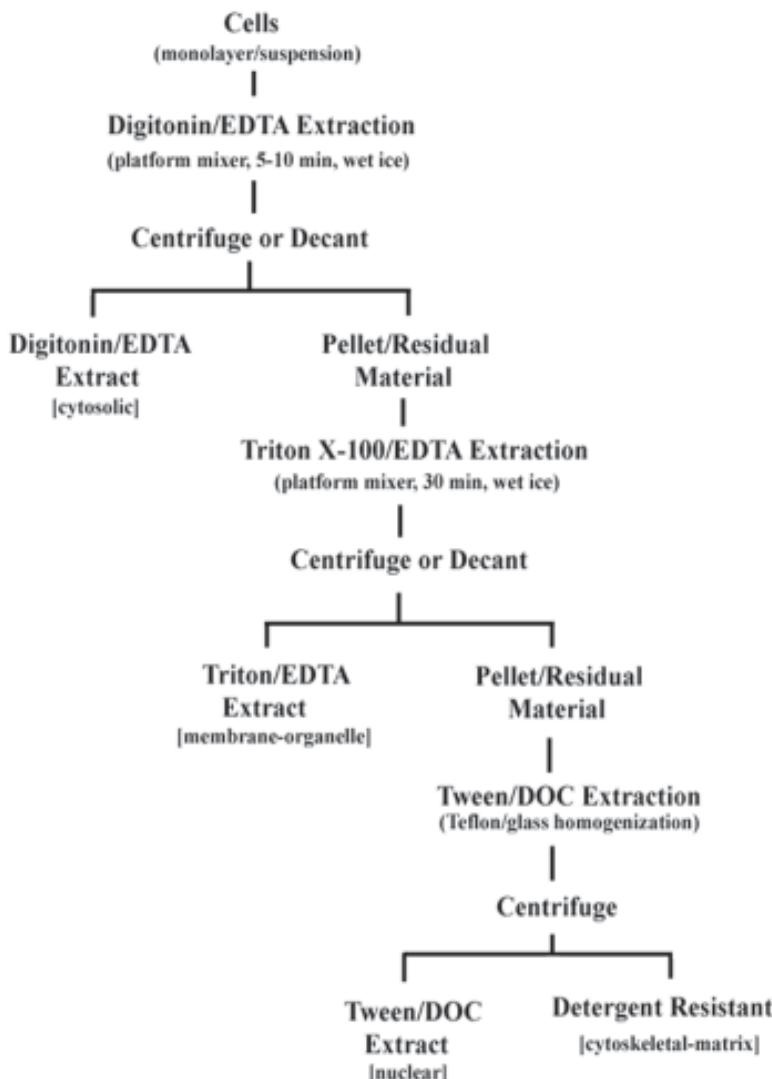


Fig. 1. Simplified schematic for differential detergent fractionation (DDF) of cells cultured in suspension or in monolayers (see Methods section for details).

Table 2
Composition of Stock Buffers

Reagent	FW (g/mol)	4X Stock Buffer		10X Stock Buffer	
		per 250 mL (g)	mM (1X)	per 100 mL (g)	mM (1X)
Sucrose	342	103	300	—	—
NaCl	58.4	5.8	100	0.58	10
PIPES	302	3	10	3	10
MgCl ₂ .6H ₂ O	203	0.64	3	0.20	1

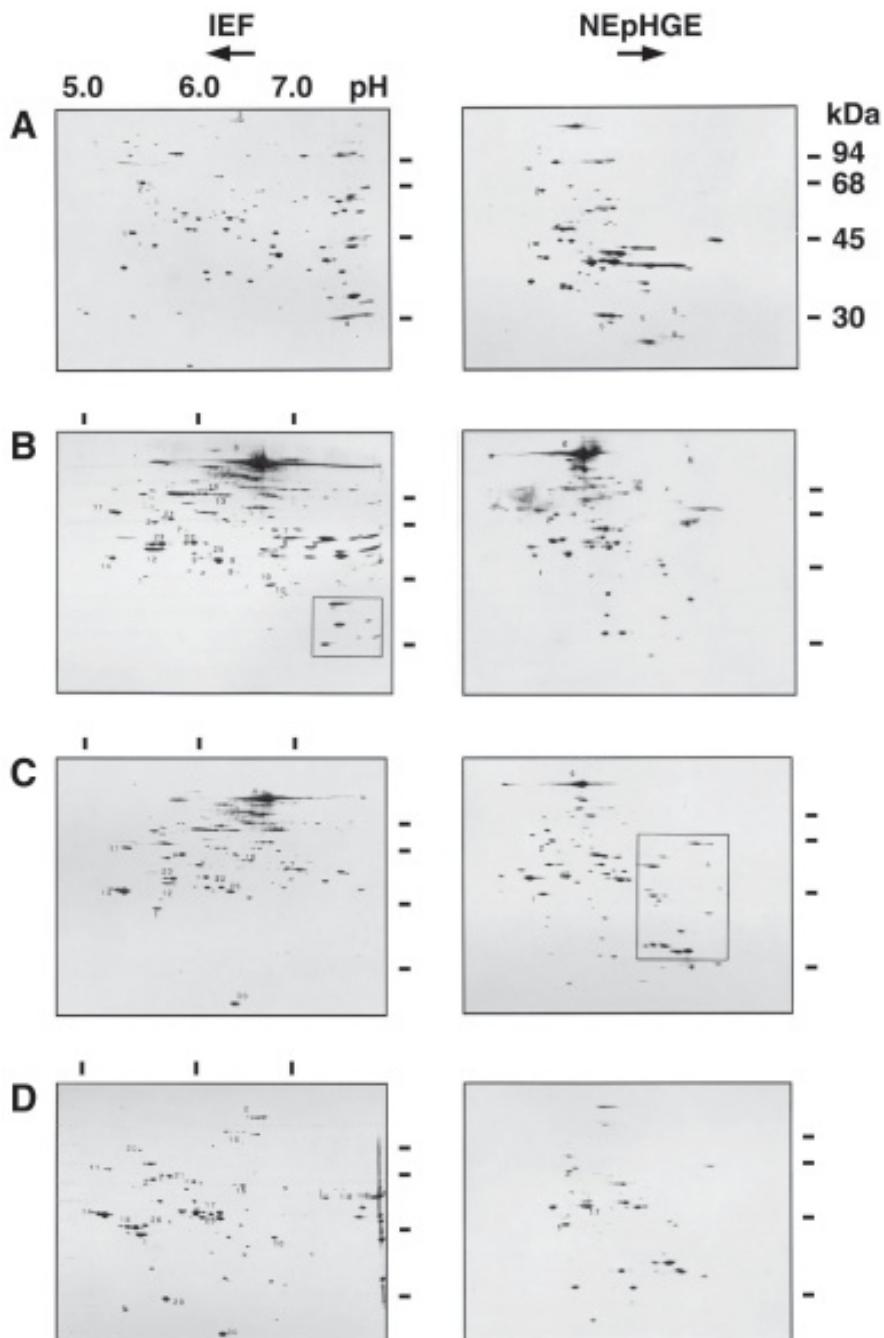


Fig. 2. Silver-stained 2-D isoelectric focusing and nonequilibrium pH gradient electrophoresis gels of differential detergent fractionation (DDF) obtained from suspension-cultured rat hepatocytes. DDF extracts: (A) digitonin/ethylenediaminetetraacetic acid (EDTA) (cytosolic); (B) Triton X-100/EDTA (membrane-organelle); (C) Tween-40/DOC (nuclear); and (D) detergent-resistant (cytoskeletal-matrix) fraction. Isoelectric point (pI) values shown at top and molecular weight (kDa) values shown on left. (Reproduced with kind permission from VCH Publishers; for details, see ref. 16.)

Table 1
Subcellular Protein Distribution in Differential Detergent Fractions^a

Constituent	Digitonin/EDTA (cytosolic)	Detergent fraction			Detergent resistant (cytoskeletal)
		Triton/EDTA (membrane-organelle)	Tween/DOC (nuclear)		
Protein (% of total)	35	50	5		10
Marker Enzymes ^b (% of total)	LDH (90) GST (84)	β-Gal (92) GDH (98) cyo p450 (92)	5'-nucleotidase (18)		
Noncytoskeletal proteins	carbonic anhydrase acetyl CoA carboxyl hsp-70 hsp-47	pyruvate dehydrog carbamoyl phos synth cyto b5 reductase GRP-78	p38 nucl prot	ribophorin docking prot	
	fatty acid bind prot calpain I & II calpastatin	carnit palmit transf prot disulf isom α-connexin			
Cytoskeletal proteins	desmoplakin II myosin vinculin α-tubulin β-tubulin	desmoplakin I desmoplakin II vinculin nuclear lamins cytokeratin A/D	actin	actin	
	actin			actin	

^aNote: distribution profiles were obtained by marker enzyme analysis or immunoblotting; these distributions reflect the fractionation profile for hepatocytes (*16*), a highly metabolic cell type. Fractionation profiles may vary with less metabolically active or specialized cell types.

^bEDTA, ethylenediaminetetraacetic acid; LDH, lactate dehydrogenase; GST, glutathione-S-transferase; hsp-70, heat shock protein-70; β-Gal, β-galactosidase; cyo p450 red, NADPH, cytochrome p450 reductase; GRP-78, glucose-regulated protein-78 (*ref. 16*).

as monolayers (corneal and vascular endothelial cells, normal and malignant osteoblasts, fibroblasts). In addition, we have been able to zymographically analyze the intracellular and extracellular localization of fibrinolytic enzymes (17), as well as purify intact RNA from each fraction for studies of mRNA distribution and turnover (18). Thus, the potential applications of this method are numerous.

2. Materials

2.1. Equipment

1. Ultrapure water (doubly distilled, deionized, $>18\text{ M}\Omega$) is used for all reagent preparation.
2. Reagent grades should be of the highest quality, as appropriate for intended use (cell culture, molecular biology, ultrapure electrophoretic).

2.2. Reagents

2.2.1. Stock Solutions

1. 100 mM EDTA: dissolve 3.36 g EDTA in 100 mL water (final volume). Store at room temperature (RT).
2. 100 mM phenylmethylsulfonyl fluoride (PMSF): dissolve 174 mg PMSF in 10 mL isopropanol. Sonicate if necessary. Store at RT in dark.
3. Piperazine-*N,N*-bis(2-ethanesulfonic acid) (PIPES) stock buffers (4X and 10X):

Filter through 0.45 μm sterile filter. Store at 4°C in dark.

2.2.2. Detergent Extraction Buffers (see Note 2)

1. Digitonin extraction buffer (0.015%, pH 6.8 at 4°C): dissolve by heating 18.75 mg digitonin in 10 mL 4X stock buffer in a small flask with a stir bar, then add 1 mL PMSF. Combine with remaining reagents: 15 mL 4X stock buffer and 5 mL EDTA (see Note 3). Add water to 100 mL (final vol).
2. Triton X-100 extraction buffer (0.5%, pH 7.4 at 4°C): combine 25 mL 4X stock buffer, 1 mL PMSF, 3 mL EDTA, and 5 mL freshly prepared 10% Triton X-100 (see Note 4). Add water to 100 mL (final vol).
3. Tween-40/deoxycholate extraction buffer (1% Tween/0.5% DOC, pH 7.4 at 4°C): separately dissolve 0.5 g DOC in 2.5 mL 10X stock buffer and 1 mL Tween-40 in 2.5 mL 10X stock buffer (warm to dissolve if necessary). Combine and add 5 mL 10X stock buffer and 1 mL PMSF (see Note 5). Add water to 100 mL (final vol).
4. Cytoskeleton solubilization buffer (5% sodium dodecyl sulfate [SDS], 10 mM sodium phosphate, pH 7.4): for nonreducing buffer, solubilize 0.5 g SDS in 5 mL 20 mM sodium phosphate buffer, pH 7.4. Add water to 10 mL (final vol). For denaturing buffer, add 1 mL β -mercaptoethanol. Adjust water appropriately.

2.2.3. Sample Lysis Buffers

1. O'Farrell lysis buffer (1X): dissolve 5.7 g ultrapure electrophoretic-grade urea, 0.2 mL NP-40, 0.2 mL ampholines (0.16 mL pH 5.0–7.0 and 0.04 mL pH 3.0–10), 0.5 mL β -mercaptoethanol in ultrapure water. Bring to 10 mL final vol with water. Solution may be warmed to facilitate solubilization. Aliquot (1 mL) and store at -70°C.
2. O'Farrell lysis buffer (10X): combine 0.2 mL NP-40, 0.2 mL ampholines (0.16 mL pH 5.0–7.0 and 0.04 mL pH 3.0–10.0), 0.5 mL β -mercaptoethanol in water. Bring to 1 mL final vol with water. Solution may be warmed to facilitate solubilization. Aliquot (100 μL) and store at -70°C.

3. Methods

3.1. Cell Preparation

DDF can be employed to fractionate cells cultured in suspension or in monolayers. Following DDF, all extracts are stored frozen at -70°C . Save an aliquot of DDF buffers at -70°C to use for sample normalization and as a control material in enzymatic and protein analyses (see **Subheadings 3.3. and 3.4.2.**). To eliminate culture media effects, cells should be washed twice in ice-cold saline, PBS, or other non-detergent buffer prior to DDF.

3.1.1. Suspension-Cultured Cells

1. The volume of DDF solutions for suspension-cultured cells is based on wet weight or cell number. To determine wet weight, transfer an aliquot of suspension-culture to a preweighed plastic tube and centrifuge briefly.
2. Decant the culture media and determine wet weight of the cell pellet. Digitonin/EDTA extraction buffer (5 vol/g wet weight) can be added directly to cell pellets. For volumes of other DDF buffers, see **Subheading 3.2.**

3.1.2. Monolayer-Cultured Cells

1. The volume of DDF buffers for monolayer cell cultures is determined per surface area or cell number. For a typical T25 culture flask (approx 5×10^6 cells), 1 mL of digitonin/EDTA buffer is used initially.
2. Following removal of culture medium, DDF can be performed in the culture flask. For volumes of other DDF buffers see **Subheading 3.2.**

3.2. Detergent Fractionation (see Note 6)

3.2.1. Digitonin Extraction (Cytosolic Fraction)

1. Add ice-cold digitonin extraction buffer to washed cell pellets (5 volumes/g wet weight, gently resuspend by swirling) or monolayers (1 mL/T25 flask) (see **Notes 7 and 8**).
2. Incubate cells on ice with gentle agitation (platform mixer) until 95–100% of cells are permeabilized (5–10 min) as assessed by trypan blue exclusion.
3. For suspension-cultured cells, centrifuge the extraction mixture (480g) and remove supernatant.
4. For cell monolayers, tilt the culture flask and remove extract (cytosolic proteins) with a pipet. Record extract volume, aliquot, and store at -70°C .

3.2.2. Triton X-100 Extraction (Membrane/Organelle Fraction)

1. Carefully resuspend digitonin-insoluble pellets in ice-cold Triton X-100 extraction buffer in a volume equivalent to that used for digitonin extraction (5 vols relative to starting wet weight) to obtain a homogeneous suspension (see **Note 9**).
2. For monolayer cultures, add 1 mL Triton extraction buffer per T25 flask equivalent (approx 5×10^6 cells).
3. Incubate on ice with gentle agitation (platform mixer) for 30 min.
4. Remove Triton extract (membrane and organellar proteins) by centrifuging suspensions (10 min, 5000g) or tilting and decanting monolayers.
5. Measure volume of the extract, aliquot, and store at -70°C .

3.2.3. Tween/DOC Extraction (Nuclear Fraction)

1. Resuspend the Triton-insoluble pellets from suspension cultures in Tween/DOC extraction buffer at one-half the volume used for Triton extraction; resuspend using a Teflon smooth-walled glass homogenizer (five strokes, medium speed) (see **Note 10**).

Table 3
Sample Preparation for 2-D Electrophoresis

Sample	Sample (μ L)	Digitonin/EDTA buffer (μ L)	Urea (mg)	10X Lysis buffer (μ L)	Load (μ L)
1	100	0	85	15	20
2	90	10	85	15	20
3	80	20	85	15	20

EDTA, ethylenediaminetetraacetic acid.

2. Remove Tween/DOC extract (nuclear proteins) by pelleting detergent-resistant residue (6780g).
3. Extract cell monolayers with 0.5–1 mL Tween/DOC buffer per T25 flask equivalent.
4. Record the volume, aliquot, and store at –70°C.

3.2.4. Detergent-Resistant Residue (Cytoskeletal/Nuclear Matrix Fraction)

1. The detergent-resistant pellet is washed in ice-cold PBS (pH 7.4, 1.2 mM PMSF) by resuspension (Teflon/glass homogenizer) and centrifugation (12,000g) to mechanically shear DNA (see Note 11).
2. Pellets from suspension cultures are washed once with –20°C 90% acetone, lyophilized, and weights determined in tared Eppendorf centrifuge tubes. Samples are stored at –70°C.
3. Monolayers are rinsed *in situ* with PBS, and the detergent-resistant residue is suspended directly into nondenaturing cytoskeleton (CSK) solubilization buffer without β -mercaptoethanol by titration. Store at –70°C.

3.3. Protein Determination

1. Thaw aliquots of detergent buffers and detergent extracts on ice.
2. Dilute an aliquot of digitonin and Triton X-100 extracts with 4 vols of ultrapure water (extracts obtained from monolayer culture may require less dilution). The Tween/DOC extract is used without dilution.
3. Solubilize lyophilized CSK pellets in CSK solubilization buffer minus β -mercaptoethanol (10 mg dry weight/mL) (see Note 11). CSK preparations from monolayers are diluted as necessary.
4. Assay 20–50 μ L of diluted or undiluted sample from each fraction, in duplicate, using the Folin-phenol method of Peterson (19) (see Note 12).
5. Use oven-dried bovine serum albumin prepared in each detergent buffer to generate standard curves.

3.4. Two-Dimensional Gel Electrophoretic Analysis

In our experience, DDF samples obtained from a variety of cell types can be utilized for two-dimensional polyacrylamide gel electrophoresis (PAGE) under both isoelectric focusing (IEF) and nonequilibrium pH gradient electrophoresis (NEPHGE). Samples to be compared by 2-D gel analysis are first normalized to contain equal protein in equal volumes (see Subheading 3.4.2.).

3.4.1. Sample Preparation

1. Fresh or defrosted DDF samples are kept on ice.
2. Samples obtained from digitonin, Triton X-100, and Tween-40/DOC extracts are brought to 9.5 M urea by addition of solid urea.

3. For 100 μ L sample, add 85 mg urea and 15 μ L 10X O'Farrell lysis buffer and warm to RT (see Note 13).
4. For the dried CSK extract, solubilize directly in 1X O'Farrell lysis buffer (20,21) (see Subheading 2.2.3., item 2).
5. For CSK extracts in nonreducing SDS buffer, bring to 9.5 M urea by addition of solid urea, and add 10X lysis buffer.

3.4.2. Sample Normalization

Samples for comparison are normalized with respect to protein concentration prior to addition of lysis buffers or urea by volume normalization using the appropriate detergent extraction buffer. For example, to normalize three digitonin/EDTA samples containing protein at 8, 9, and 10 μ g/ μ L make the following additions:

3.4.3. 2-D Electrophoresis

DDF samples are subjected to 2-D gel electrophoresis by established methods (19,20) (see Chapter 13) (see Note 14). IEF gels contain a total of 3.5% ampholines (2% pH 5.0–8.0, 1% pH 3.0–10.0, 0.5% pH 2.0–5.0), and samples are electrophoresed for a total of 9800 V·h with hyperfocusing at 800 V for the final hour (16,22). NEPHGE gels contain 2% pH 3.0–10.0 ampholines, and samples are electrophoresed for 2400 V·h (16).

4. Notes

1. For 4X stock solutions, solubilize sucrose and NaCl together in water. Solubilize PIPES separately in a small vol of 1 M NaOH before mixing with remaining ingredients. MgCl₂ can be added directly to sucrose solution or final buffer as a solid or from a concentrated stock solution prepared in ultrapure water, as convenient. Stock buffers should be sterile filtered (0.45 μ m) and stored at 4°C in the dark. Stock solutions are stable for up to 2 mo. Maintain aseptic technique when diluting stock for preparation of 1X working solutions. Alternately, stock buffer solutions may be stored in volumes appropriate for single-use aliquots.
2. Stability of detergent extraction buffers varies. Digitonin solutions are reportedly stable in buffer for approx 3 h at 0°C (23). Triton X-100 and Tween-40 are nonionic detergents and decompose in aqueous buffer to form peroxide radicals, which oxidize sulphydryl groups (1,24,25). In our experience, 1X working digitonin and Triton X-100 buffers, if prepared with EDTA, remain stable longer and can be aliquoted, and fresh frozen at -70°C. Such solutions should be thawed on ice and not refrozen.
3. The 1X working digitonin extraction buffer should be prepared by adding solid digitonin to a small amount of buffer and carefully boiling to dissolve. PMSF is solubilized in the warm digitonin solution by slow addition with constant stirring (magnetic stir bar). The digitonin/PMSF solution is then added to the remaining buffer ingredients. The solution is cooled to 4°C and the pH adjusted to 6.8 with dilute HCl. The digitonin extraction solution is brought to volume with water and kept on ice until use, or aliquoted and stored frozen at -70°C as noted above.
4. The 1X working Triton X-100 extraction buffer should be prepared using freshly made 10% Triton X-100 in ultrapure water. The solution is cooled to 4°C, pH adjusted to 7.4, and brought to final volume. The solution is kept on ice or stored frozen as noted above.
5. The 1X working Tween/DOC buffer is prepared by dissolving Tween-40 and DOC separately in a small volume of buffer with heating to dissolve. The solutions are mixed, cooled to 4°C, pH adjusted to 7.4, and brought to volume. The solution is used fresh. Unused buffer is discarded.

6. The extraction protocol described here represents a modification of a method described by Fey et al. for fractionation of MDCK cells (4). Modifications include the addition of a digitonin extraction step, the inclusion of EDTA in digitonin and Triton buffers and the exclusion of a nuclease digestion step (DNA is denatured by shear force in the presence of SDS).
7. The digitonin extraction protocol was formulated in accord with considerations described in the literature (23,26–28). Briefly, digitonin is a steroidal compound believed to complex with plasma membrane cholesterol, resulting in membrane permeabilization and the rapid release of soluble cytosolic components, leaving behind intact cell ghosts and heavy organelles. At concentrations of 0.015%, digitonin preserves the ultrastructure of ER and mitochondrial membranes (27), which at higher concentrations (approx 0.1%) are damaged secondary to solubilization of membrane phospholipid (27,28).
In our experience, the inclusion of EDTA significantly enhances the effectiveness of low concentrations of digitonin, as evidenced by an increased rate of membrane permeabilization (10 min +EDTA vs 40 min –EDTA). In addition, EDTA is beneficial for inhibiting calcium-dependent neutral proteases, typically enriched in cytosolic extracts (16), and thus avoids artifactual proteolysis. Consistent with the reports of others, digitonin releases proteins larger than 200 kDa (28), as evidenced by the presence of myosin (>220 kDa), desmoplakin II (>250 kDa), and the calpain-inhibitor calpastatin (approx 300 kDa) in cytosolic extracts (16).
In brief, digitonin/EDTA extraction as described here yields a cytosolic extract representing approx 35% total cellular protein, which is enriched in cytosolic markers (90% lactate dehydrogenase [LDH] activity, 100% carbonic anhydrase immunoreactivity) and essentially devoid of mitochondrial, lysosomal, and ER markers (16); it therefore represents a significant improvement over previous low-concentration digitonin extraction methods, which released only 60% of LDH concomitant with approx 30% contamination from organelle markers (10,29).
8. The selective fractionation of cytoskeletal tubulins in the digitonin extract likely reflects the effect of both cold and EDTA on inducing microtubule depolymerization, and has been capitalized on as a first-step method for preparing polymerization-competent microtubules (unpublished investigations).
9. Triton is a nonionic detergent which solubilizes membrane lipids and releases organelle contents. It has been used in hyper- or hypotonic buffers to prepare cytoskeletal preparations enriched in intermediate filaments (31). Lower concentrations of Triton concomitant with iso-osmolar, isotonic buffer composition, preserves nuclear and microfilament integrity (4). Thus, as used here, Triton extracts are enriched in markers for membrane and organelle proteins (16) and constitute the bulk of cellular proteins (approx 50% total protein). Triton extraction is possible using either X-100 or X-114 series. Fractionation with X-114 allows subfractionation of peripheral vs integral membrane proteins (32,33), and lends further flexibility to fractionation goals.
10. DOC is a weakly ionic detergent that destroys nuclear integrity (34) and solubilizes actin and other cytoskeletal elements (4,5). Tween/DOC buffer extracts approx 5% of total cell protein and contains exclusively, immunoreactivity for the nuclear protein p38 (16), thereby verifying that nuclear integrity persists through Triton extraction. Consideration of marker enzyme profiles (see **Table 1**), Western blot analysis, and 2-D electrophoreograms (16) suggests that although Tween/DOC buffer extracts proteins common to both the membrane/organelle and detergent-resistant cytoskeletal fractions, specific distinctions are apparent suggesting this fraction represents a metabolically distinct, possibly more labile, protein compartment. However, for studies in which nuclear parameters are

- not of interest, the DDF protocol may be simplified by omitting the Tween/DOC step; this may be especially warranted when fractionating limited amounts of sample.
11. The detergent-resistant fraction accounts for approx 7–10% of cellular protein and is enriched in intermediate filaments, actin, and various cytoskeletal associating/interacting proteins (16). It also contains the nuclear matrix proteins and DNA. DNA causes a viscous, difficult-to-manage extract but can be readily denatured by mechanical shear (35) using either Teflon/glass homogenization (for suspension pellets) or tituration with a pipet (for monolayer residues). This fraction is intact, as evidenced by the absence of staircase patterns (indicative of proteolytic degradation) on 2-D PAGE. Staircase patterns are obtained if PMSF or EDTA is absent from extraction buffers. Solubilization of detergent-resistant samples in SDS-containing phosphate buffer in the absence of mercaptoethanol allows direct assay of protein content by the method of Peterson (19).
 12. The Folin-phenol method of Peterson has been extensively detailed elsewhere (19). Briefly, this method is considered the method of choice for international laboratory standardization of protein values and is not susceptible to interference by detergents, thus enabling direct analysis of samples. In contrast, assay by the standard Lowry method results in detergent-induced flocculation (personal observation). Direct analysis avoids the cumbersome practice of protein precipitation prior to analysis that was used in earlier detergent fractionation protocols (4). We have verified the lack of detergent interference by comparing blanks and standards prepared in detergent buffers at various dilutions in triplicate assays.
 13. Soluble DDF extracts (digitonin, Triton, and Tween/DOC) may contain low concentrations of protein. To minimize the dilutional effects of the lysis buffer, solid urea and 10X O'Farrell lysis buffer are added to these samples. The sample may be warmed slightly to facilitate urea solubilization. Caution: excessive warming (increased temperature and/or prolonged heating) may result in carbamylation artifacts.
 14. Typically, 25–100 µg protein in 15–60 µL provides adequate sample for visualization by Coomassie blue or autoradiography; lesser concentrations may be analyzed by silver staining. In our experience the volume of detergent buffer contained in samples of these volumes did not adversely affect the linear range of the pH gradient in IEF gels. It should be noted, however, that a slight shift to more acidic values may occur with DDF samples containing Tween/DOC.

References

1. Lenstra, J. A. and Bloemendal, H. (1983) Topography of the total protein population from cultured cells upon fractionation by chemical extractions. *Eur. J. Biochem.* **135**, 413–423.
2. Lenk, R., Ransom, L., Kaufman, Y., and Penman, S. (1977) A cytoskeletal structure with associated polyribosomes obtained from HeLa cells. *Cell* **10**, 67–78.
3. Reiter, T. and Penman, S. (1983) “Prompt” heat shock proteins: translationally-regulated synthesis of new proteins associated with nuclear matrix-intermediate filaments as an early response to heat shock. *Proc. Natl. Acad. Sci. USA* **80**, 4737–4741.
4. Fey, E. G., Wan, K. M., and Penman, S. (1984) Epithelial cytoskeletal framework and nuclear matrix-intermediate filament scaffold: three-dimensional organization and protein composition. *J. Cell Biol.* **98**, 1973–1984.
5. Reiter, T., Penman, S., and Capco, D. G. (1985) Shape-dependent regulation of cytoskeletal protein synthesis in anchorage-dependent and anchorage-independent cells. *J. Cell Sci.* **76**, 17–33.
6. Katsuma, Y., Marveau, N., Ohta, M., and French, S. W. (1988) Cytokeratin intermediate filaments of rat hepatocytes: different cytoskeletal domains and their three-dimensional structure. *Hepatology* **8**, 559–568.

7. Cervera, M., Dreyfuss, G., and Penman, S. (1981) Messenger RNA is translated when associated with the cytoskeletal framework in normal and VSV-infected cells. *Cell* **23**, 113–120.
8. Bird, R. C. and Sells, B. H. (1986) Cytoskeleton involvement in the distribution of mRNP complexes and small cytoplasmic RNAs. *Biochim. Biophys. Acta* **868**, 251–225.
9. Bag, J. and Pramamik, S. (1987) Attachment of mRNA to the cytoskeletal framework and translational control of gene expression in rat L6 muscle cells. *Biochem. Cell. Biol.* **65**, 565–575.
10. Doherty, F. J., Wassell, J. A., and Mayer, R. J. (1987) A putative protein sequestration site involving intermediate filaments for protein degradation by autophagy. Studies with microinjected purified glycolytic enzymes in 3T3-L1 cells. *Biochem. J.* **241**, 793–800.
11. Bonneau, A.-M., Darveau, A., and Sonenberg, N. (1985) The effect of viral infection on host protein synthesis and mRNA association with the cytoplasmic cytoskeletal structure. *J. Cell Biol.* **100**, 1209–1218.
12. Belin, M.-T. and Boulanger, P. (1985) Cytoskeletal proteins associated with intracytoplasmic human adenovirus at an early stage of infection. *Exp. Cell Res.* **160**, 356–370.
13. Ciampi, F. (1988) The role of the cytoskeleton and nuclear matrix in viral replication. *Acta Virol.* **170**, 338–350.
14. Tanquay, R. M. (1983) Genetic regulation during heat shock and function of heat-shock proteins: a review. *Can. J. Biochem. Cell. Biol.* **61**, 387–394.
15. Welch, W. J. and Suhan, J. P. (1985) Morphological study of the mammalian stress response: characterization of changes in cytoplasmic organelles, cytoskeleton, and nucleoli, and appearance of intranuclear actin filament in rat fibroblasts after heat shock. *J. Cell Biol.* **101**, 1198–1211.
16. Ramsby, M. L., Makowski, G. S., and Khairallah, E. A. (1994) Differential detergent fractionation of isolated hepatocytes: biochemical, immunochemical and two-dimensional gel electrophoresis characterization of cytoskeletal and noncytoskeletal compartments. *Electrophoresis* **15**, 265–277.
17. Ramsby, M. L. and Kreutzer, D. L. (1993) Fibrin induction of tissue plasminogen activator expression in corneal endothelial cells in vitro. *Invest. Ophthalmol. Vis. Sci.* **34**, 3207–3219.
18. Ramsby, M. L. and Makowski, G. S. (2003) Differential detergent fractionation of eukaryotic cells and additional protocols- precipitation of tubulins and MAPs (microtubule-associated proteins) using magnesium and isolation of RNA from detergent extracts. In Simpson, R. J. (ed), *Proteins and Proteomics: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY: 126–137.
19. Peterson, G. L. (1983) Determination of total protein. *Meth. Enzymol.* **91**, 95–119.
20. O'Farrell, P. H. (1975) High resolution two-dimensional electrophoresis of proteins. *J. Biol. Chem.* **250**, 4007–4021.
21. O'Farrell, P. Z., Goodman, H. M., and O'Farrell, P. H. (1977) High resolution two-dimensional electrophoresis of basic as well as acidic proteins. *Cell* **12**, 1133–1142.
22. Duncan, R. and Hershey, J. W. B. (1984) Evaluation of isoelectric focusing running conditions during two-dimensional isoelectric focusing/sodium dodecyl sulfate-polyacrylamide gel electrophoresis: variation of gel patterns with changing conditions and optimal isoelectric focusing conditions. *Anal. Biochem.* **138**, 144–145.
23. Zuurendonk, P. F. and Tager, J. M. (1974) Rapid separation of particulate components and soluble cytoplasm of isolated rat-liver cells. *Biochim. Biophys. Acta* **333**, 393–399.
24. Lever, M. (1977) Peroxides in detergents as interfering factors in biochemical analysis. *Anal. Biochem.* **83**, 274–284.

25. Chang, H. W. and Bock, E. (1980) Pitfalls in the use of commercial nonionic detergents for the solubilization of integral membrane proteins: sulfhydryl oxidizing contaminants and their elimination. *Anal. Biochem.* **104**, 112–117.
26. Mackall, J., Meredith, M., and Lane, L. M. (1979) A mild procedure for the rapid release of cytoplasmic enzymes from cultured animal cells. *Anal. Biochem.* **95**, 270–274.
27. Fiskum, G., Craig, S. W., Decker, G. L., and Lehninger, A. L. (1980) The cytoskeleton of digitonin-treated rat hepatocytes. *Proc. Natl Acad. Sci. USA* **77**, 3430–3434.
28. Weigel, P. H., Ray, D. A., and Oka, J. A. (1983) Quantitation of intracellular membrane-bound enzymes and receptors in digitonin-permeabilized cells. *Anal. Biochem.* **133**, 437–449.
29. Earl, R. T., Mangiapane, E. H., Billett, E. E., and Mayer, R. J. (1987) A putative protein sequestration site involving intermediate filaments for protein degradation by autophagy. Studies with transplanted Sendai-viral envelope proteins in HTC cells. *Biochem. J.* **241**, 809–815.
30. Morgenstern, R., Meijer, J., Depierre, J. W., and Ernster, L. (1980) Characterization of rat-liver microsomal glutathione-S-transferase activity. *Eur. J. Biochem.* **104**, 167–174.
31. Franke, W. W., Schmid, E., Osborn, M., and Weber, K. (1978) The intermediate-sized filaments in rat kangaroo PtK2 cells. II. Structure and composition of isolated filaments. *Cytobiol. Eur. J. Cell Biol.* **17**, 392–411.
32. Bordier, C. (1981) Phase separation of integral membrane proteins in Triton X-114 solutions. *J. Biol. Chem.* **256**, 1604–1607.
33. Pryde, J. G. and Phillips, J. H. (1986) Fractionation of membrane proteins by temperature-induced phase separation in Triton X-114. *Biochem. J.* **233**, 525–533.
34. Capco, D. G., Wan, K. M., and Penman, S. (1982) The nuclear matrix: three-dimensional architecture and protein composition. *Cell* **29**, 847–858.
35. Franke, W. W., Mayer, D., Schmid, E., Denk, H., and Borenfreund, E. (1981) Differences of expression of cytoskeletal proteins in cultured rat hepatocytes and hepatoma cells. *Exp. Cell Res.* **134**, 345–365.

Serum or Plasma Sample Preparation for Two-Dimensional Gel Electrophoresis

**Anthony G. Sullivan, Stephen Russell, Henry Brzeski,
Richard I. Somiari, and Craig D. Shriver**

1. Introduction

The importance of serum/plasma as a source of clinically relevant biomarkers/surrogate markers of human disease has increased significantly over the last decade (1,2), and modern proteomic methods have evolved and been adapted to meet the demand. The specific challenges facing serum analysis include the wide dynamic range in the concentration of individual components and the tremendous number of potential variants of glycosylated proteins (3). The most dominant plasma proteins, albumin and immunoglobulin (Ig)G, typically comprise up to 70% of the plasma proteome in abundance. To enable the majority of the remaining, far less abundant proteins to be better visualized by two-dimensional gel electrophoresis (2-DE), these two proteins must first be removed, or at least depleted in relative concentration. There are a number of currently available commercial products from a range of suppliers that enable albumin depletion by chemical affinity, exploiting the remarkable albumin-binding ability of structures closely related to the reactive dye molecule Cibacron blue 3GA (4), and the IgG binding properties of protein G (5). The blue dye has been shown to have a special affinity for proteins containing the dinucleotide fold, a structural feature that is common to several classes of proteins (4). Albumin can be separated from other plasma proteins using lectin affinity, as it is not normally glycosylated, while the majority of classical plasma proteins are. This approach allows both enrichment of lower-abundance proteins, and the study of differences in glycoprotein profiles (6). Highly effective depletion of albumin using monoclonal antibody selection has also been demonstrated, and coupled with protein G/IgG depletion (7). Recent research has identified 325 distinct proteins from 1800 2-D gel protein features following multi-component immunoaffinity extraction and further comprehensive chromatographic fractionation (8).

2. Materials

1. Blue dye affinity resin: Affi-gel, (Biorad), Mimetic Blue SA (Prometic).
2. 20 mM Na₂HPO₄, pH adjusted to 7.0 with HCl.
3. Lysis buffer: 7 M urea, 2 M thiourea, 30 mM Tris, 5 mM magnesium acetate, 4% CHAPS, 1% NP-40.

4. Trichloroacetic acid (TCA)/acetone protein precipitation: 2-D Cleanup Kit (Amersham Biosciences, Piscataway, NJ).
5. Biomax 10K NMWL membrane centrifugal filter (Millipore, Bedford, MA).
6. WGA-agglutinin lectin (Sigma, St. Louis, MO).
7. Handee Mini Spin Columns (Pierce Biotechnology, Inc., Rockford, IL) or Macro Spin Columns (NEST Group, Southborough, MA).
8. Phosphate buffer: 50 mM Na₂HPO₄, 0.2 M NaCl, pH adjusted to 7.0 with HCl.
9. Phosphate buffer with SDS: 50 mM Na₂HPO₄, 0.2 M NaCl, 0.05% SDS, pH adjusted to 7.0 with HCl.
10. Sugar solution: 0.3 M N-acetyl glucosamine or neuraminic acid (Sigma, St. Louis, MO), in phosphate buffer.
11. Resin slurry: Albumin and IgG Removal Kit (Amersham Biosciences, Piscataway, NJ).
12. POROS Affinity depletion cartridges, (Applied Biosystems, Framingham, MA).
13. AccuGENE 10X PBS solution (Cambrex, East Rutherford, NJ).

3. Methods

3.1. Affinity Depletion of Serum Albumin

3.1.1. Batch Use of Dye-Agarose Affinity Resins

Here we describe a generic method, based on manufacturer's instructions, adaptable for the majority of available dye resin slurries. The steps involved are equilibration, binding, washing, and stripping.

1. Add 250 µL of resin (see Note 1) to a 1.5-mL microcentrifuge tube (see Note 2), centrifuge for 2 min at 0.2g, and remove liquid using a gel-loading pipet tip.
2. Add 200 µL of 20 mM sodium phosphate, pH 7.0, and shake gently for 10 min. Centrifuge at 1000 rpm for 2 min and remove liquid. Repeat the equilibration twice.
3. Take 25 µL of clarified plasma/serum, dilute it with 175 µL of the 20 mM sodium phosphate, pH 7.0 buffer, and mix thoroughly.
4. Add the diluted plasma to the conditioned resin, vortex briefly, and shake gently for 10 min to allow serum albumin to bind to the dye resin. Centrifuge for 2 min at 1000 rpm and remove supernatant using a gel-loading pipet tip. Transfer solution, containing unbound albumin-depleted proteins, to a clean 1.5-mL microcentrifuge tube.
5. Wash the resin three times with 150 µL 20 mM sodium phosphate, pH 7.0 buffer, shake for 10 min, centrifuge for 2 min at 1000 rpm, remove the supernatant, and combine with the previous solution.
6. Elute the proteins bound to the resin by washing three times with lysis buffer (see Note 3) as described in Subheading 3.1.1.2.. Combine stripped fractions into a separate 1.5-mL tube.
7. Concentrate the albumin-depleted fraction (approx vol 650 µL) to a final vol of approx 100 µL using a 10,000 MWCO (molecular weight cut-off) membrane filter. Desalt and prepare for 2-DE by TCA/acetone precipitation (see Note 4). For the albumin-rich fraction, remove a 150-µL aliquot from the stock solution and perform TCA/acetone precipitation directly.

An example of a 2-D gel showing serum prepared by this method is shown in [Fig. 1](#).

3.2. Lectin Affinity Serum Albumin Removal Using WGA/Agarose

1. Prepare the Handee Mini Spin column for use by ensuring the frit at the bottom of the main chamber is firmly in place by pushing down with a paper clip. Resuspend the lectin in the buffer supplied by the manufacturer to give a homogeneous lectin slurry, then pipet 200 µL into the chamber, centrifuge briefly, and remove liquid.

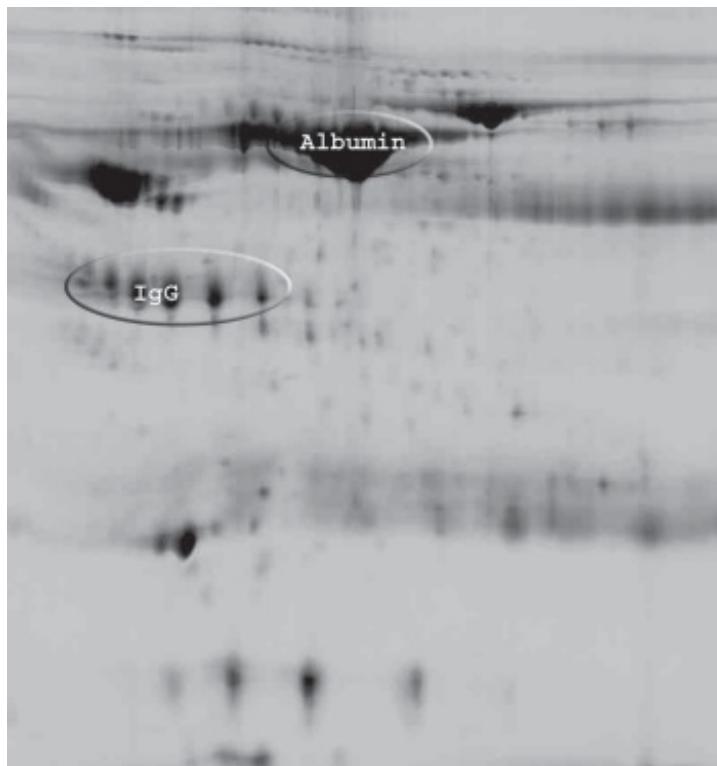


Fig. 1. An example of a two-dimensional gel showing serum prepared using the method outlined in **Subheading 3.1.1.**

2. Add 500 μ L of phosphate buffer and shake gently for 10 min. Centrifuge at 1000 rpm for 2 min, remove liquid, and repeat twice.
3. Prepare plasma by diluting 50 μ L of plasma with 750 μ L of phosphate buffer containing 0.05% SDS. This amount will typically yield 50 μ g of albumin-free protein.
4. Transfer diluted plasma to the column containing the beads. Shake gently using a rotary mixer, ensuring good mixing, for 5 min. Centrifuge briefly and collect eluent. Wash beads three times with 200 μ L of phosphate buffer. Combine wash eluents (albumin-rich fraction).
5. Elute the glycosylated proteins with 200 μ L sugar solution (see **Note 6**), shaking for 5 min on a rotary shaker. Wash three times with 150 μ L of sugar solution and combine, total volume approx 650 μ L.
6. Reduce the albumin-depleted fraction to a final vol of approx 100 μ L using a 10,000 MWCO membrane filter. Desalt and prepare for 2-DE by TCA/acetone precipitation (see **Note 4**).

An example of a 2-D gel showing serum prepared by the method in **Subheading 3.2.1.** is shown in [Fig. 2](#).

3.3. Antibody Serum Albumin/Immunoglobulin G Removal

3.3.1. Use of Antibody/Protein G Resin

1. Pipet 20 μ L of plasma/serum into a 1.5-mL microcentrifuge tube and add 750 μ L of resin slurry. Close cap and mix on a rotary shaker, ensuring strong mixing is taking place, for 30 min at room temperature.

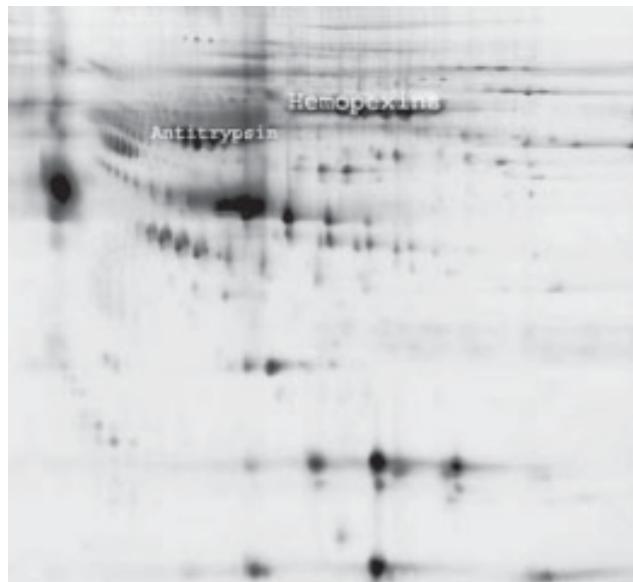


Fig. 2. An example of a two-dimensional gel showing serum prepared using the method outlined in **Subheading 3.2.1.**

2. Transfer entire slurry to the upper chamber of a microspin column and insert column into a suitable microcentrifuge tube. Centrifuge at 1000 rpm for 5 min. Remove liquid (approx 500 μ L) and desalt/concentrate as described in **Subheading 3.1.1., step 7.**
An example of a 2-D gel showing serum prepared by the method in **Subheading 3.3.1.** is shown in **Fig. 3.**

3.3.2. Use of Antibody/Protein G Cartridges (see **Note 7**)

1. Connect a syringe to the anti-SA cartridge using a needle-port adapter and equilibrate packing material with at least five cartridge volumes of PBS solution, pH 7.4.
2. Dilute the serum/plasma 10-fold with PBS solution and inject an amount less than the maximum binding capacity into the cartridge at a steady flow rate (*see Note 8*).
3. Collect the flowthrough in a clean 0.5-mL microcentrifuge tube. Wash cartridge with three column vols of PBS solution and combine wash fractions in a separate tube.
4. Connect the needle-port adapter to the protein G cartridge and condition the packing material with at least five cartridge vols of PBS solution, pH 7.4.
5. Inject the albumin-depleted fraction from **step 3** at a steady flow rate and collect the flowthrough in a fresh 1.5-mL tube.
6. Inject the combined wash fractions from **step 3** at a steady flow rate and combine with liquid from **step 6**.
7. Concentrate the albumin-IgG-depleted fraction to a final volume of approx 100 μ L using a 10,000 MWCO membrane filter. Desalt and prepare for 2-DE by TCA/acetone precipitation (*see Note 4*).

An example of a 2-D gel showing serum prepared by the method in **Subheading 3.3.2.** is shown in **Fig. 4.**

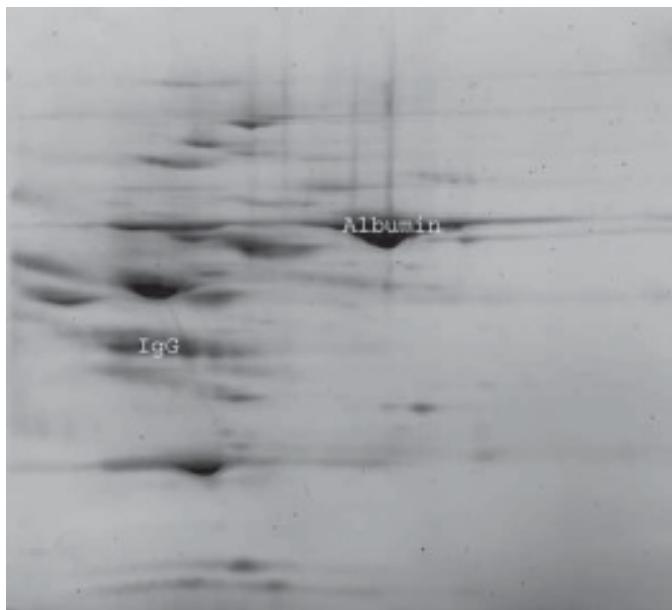


Fig. 3. An example of a two-dimensional gel showing serum prepared using the method outlined in **Subheading 3.3.1**.



Fig. 4. An example of a two-dimensional gel showing serum prepared using the method outlined in **Subheading 3.3.2**.

4. Notes

1. The amount of resin used here ensures that the total albumin loaded to the resin is less than its albumin-binding capacity (typically approx 11 mg/mL).
2. As an alternative to the procedure listed here, the resin may be applied to the upper chamber of a spin column, and the removal of liquid facilitated by centrifugation. In our hands, this variation led to less efficient albumin removal due to inconsistent residence times.
3. Alternatives to lysis buffer include solutions of sodium chloride or guanidinium chloride (5–6 M).
4. Desalting/concentrating is absolutely required for the isoelectric focusing step. The use of MWCO membrane filters followed by protein precipitation was found in our hands to be the most effective method combination. The initial preconcentration to volumes of 100 µL depletes salt levels and reduces the amount of protein precipitation solutions required.
5. Columns are available in several sizes and therefore total protein loading capacities.
6. *N*-acetyl glucosamine can be replaced by *N*-acetyl neuraminic acid depending on the type of glycosylation under investigation.
7. These cartridges can be used separately as outlined here, or in series, either manually with the aid of a syringe, or automatically using a solvent delivery system on a liquid chromatograph.
8. The albumin-binding capacity depends on the column size and will be supplied by the manufacturer. The albumin content of the plasma can only be estimated, unless specific albumin assay methods are available.

References

1. Steel, L. F., Trotter, M. G., Nakajima, P. B., Mattu, T. S., Gonye, G., and Block, T. (2003) Efficient and specific removal of albumin from human serum samples. *Mol. Cell Proteomics* **2**, 262–270.
2. Tabar, L., Dean, P. B., Kaufman C. S., Duffy, S. W., and Chen, H. H. (2000) A new era in the diagnosis of breast cancer. *Surg. Oncol. Clin. N. Am.* **9**, 233–277.
3. Anderson, N. L. and Anderson, N. G. (2002) The human plasma proteome, history character and diagnostic prospects. *Mol. Cell Proteomics* **1**, 845–867.
4. Thompson, S. T., Cass, K. H., and Stellwagen, E. (1975) Blue dextran-sepharose: an affinity column for the dinucleotide fold. *Proc. Nat. Acad. Sci. USA* **72**, 669–672.
5. Reis, K. J., Ayoub, E. M., and Boyle, M. D. P. (1984) Streptococcal Fc receptors. II. Comparison of the reactivity of a receptor from a group C streptococcus with staphylococcal protein A. *J. Immunol.* **132**, 3098–3102.
6. Brzeski, H., Katenhusen, R. A., Sullivan, A. G., et al. (2003) Albumin depletion method for improved plasma glycoprotein analysis by 2-dimensional difference gel electrophoresis. *Biotechniques* **35**, 1128–1132.
7. Steel, L. F., Trotter, M. G., Nakajima, P. B., Mattu, T. S., Gonye, G., and Block, T. (2003) Efficient and specific removal of albumin from human serum samples. *Mol. Cell Proteomics* **2**, 262–270.
8. Pieper, R., Gatlin, C. L., Makusky A. J., et al. (2003) The human serum proteome: display of nearly 3700 chromatographically separated protein spots on two-dimensional electrophoresis gels and identification of 325 distinct proteins. *Proteomics* **3**, 1345–1364.

Preparation of Plant Protein Samples for 2-D PAGE

David W. M. Leung

1. Introduction

A critical step in the application of O'Farrell's (1) two-dimensional polyacrylamide gel electrophoresis (2-D PAGE) in plant biology research is the preparation of plant protein samples free of artefactual protein modifications and without adversely affecting gel resolving power and reproducibility. However, it is important to note that many problems could be encountered in this step. A large part of the plant cell volume is often occupied by the vacuole. Numerous compounds found therein, upon release during disruption of plant cells, might have detrimental or interfering effects on protein extraction. These include organic acids, phenolic compounds, proteases, pigments, polysaccharides, and so on.

Another major problem that could impact on the quality of 2-D PAGE (resulting in protein-spot streaking and smearing) might arise from the incomplete disruption of protein complexes, and by protein aggregation during sample solubilization. In particular, complete solubilization of membrane proteins is a frequent concern, and the possible interaction between nucleic acids and cellular proteins during tissue homogenization could adversely affect protein resolution during electrophoresis.

Here, a protocol largely based on the original paper of Mayer et al. (2) to extract "total proteins" from whole plant tissue that should result in satisfactory 2-D PAGE analysis is described. In addition, since extracellular proteins present in intercellular fluids of leaves might be of interest for 2-D PAGE analysis, a method derived from the original papers of De Wit and Spikman (3), and Jeknic and Chen (4) for preparing such proteins is also given.

2. Materials

1. Solution A: 2% (v/v) 2D Pharmalyte pH 3.0–10.0, 0.3 M NaCl, 1 mM ethylenediamine-tetraacetic acid (EDTA)-Na, 1 mM EGTA, 1% (v/v) Nonidet P-40, 1% (w/v) CHAPS, 5 mM ascorbic acid, 0.1 M dithiothreitol, 10 µg/mL leupeptin, 10 µg/mL α_2 -macroglobulin. These reagents have been chosen because they do not interfere with the first dimensional isoelectric focusing (IEF) gel electrophoresis. The first six reagents help solubilization of proteins, the next two minimize oxidation of proteins, and the last reagents inhibit proteolysis.
Prepare 10 mL of this solution and store 1-mL aliquots at -80°C until needed.
2. Solution B: 0.7% (v/v) 2-mercaptoethanol in acetone (of analytical grade). This solution should be kept at -20°C at least several hours before use.

3. Solution C: 9.5 M urea, 2% (w/v) CHAPS, 0.4% (v/v) Pharmalyte pH 3.0–10.0, 1.6% (v/v) Resolyte pH 4.0–8.0, 5% (v/v) 2-mercaptoethanol. Prepare 10 mL of this solution and store 1-mL aliquots at –80°C until needed. All solutions should be prepared with deionized or distilled water.
4. Protamine sulfate.
5. Urea (ultrapure).

3. Methods (see Note 1)

3.1. Extraction From Whole Plant Tissue

1. Plant tissue samples should be ground to a fine powder with a pestle and mortar prechilled in liquid nitrogen.
2. Determine the weight of a 1.5-mL Eppendorf tube before cooling it in liquid nitrogen.
3. Place the powder (from step 1) into the Eppendorf tube and weigh again.
4. Add solution A to the powder for protein extraction. The solution to powder ratio of 1 to 2 (v/w) should be adequate. Vortex immediately to mix the powder with the solution.
5. Add solid protamine sulfate (1 mg/mL).
6. Incubate the mixture on ice for 15 min with occasional vortexing.
7. Centrifuge (13,000g at 2°C) for 15 min (see Note 2). Transfer the supernatant to a new Eppendorf tube and then add solid urea to a final concentration of 9 M. For 5 min, vortex the mixture intermittently at room temperature.
8. The protein-urea extract can be used immediately or divided into 70-μL aliquots, which should be stored at –80°C until required for the first-dimensional IEF gel electrophoresis (see Note 3).

3.2. Extraction From Intercellular Fluids

3.2.1. Collection of Intercellular Fluids

1. Immerse entire leaves (100 mg to 1 g, fresh weight) in distilled water in a beaker placed inside a desiccator connected to a vacuum pump. Vacuum filtration of leaf materials is judged finished when the leaf blades have shown a glossy dark green appearance.
2. Gently blot the leaves dry using paper towels before placing the leaves in a 2.5-mL plastic syringe barrel, which should be sitting on top of an Eppendorf tube (with its lid cut off). This whole syringe–Eppendorf tube assembly is then placed inside a 30-mL centrifugation tube (we use Oak Ridge centrifuge tubes, Nalgene, Rochester, NY).
3. Centrifuge (3000g at 4°C) for 10 min. The intercellular fluid is collected in the Eppendorf tube. It can be used immediately or stored at –80°C until required for the first-dimensional IEF gel electrophoresis.

3.2.2. Protein Extraction

1. Add 5 vols of solution B to 1 vol of intercellular fluid (see Notes 4 and 5). Incubate the mixture at –20°C overnight.
2. Centrifuge (15,000g at 4°C) for 15 min. Discard the supernatant and dry the pellet under vacuum at room temperature.
3. Resuspend the pellet with at least 50 μL of solution C. Vortex vigorously and incubate on ice for 30 min.
4. Centrifuge (14,000g at 4°C) for 15 min. Discard the insoluble material; aliquots of the supernatant should be used immediately and loaded directly onto first-dimensional IEF gels.

4. Notes

1. These techniques were used routinely in preparing protein extracts from petunia leaf discs, callus cultures, intensely pigmented begonia leaf and petiole materials, root-forming hypocotyl segments of *Pinus radiata* D. Don., and intercellular fluids of rose leaves. Satisfactory 2-D PAGE analysis, particularly of polypeptides with apparent molecular masses of 45 KDa or lower (6), was achieved despite the presence of high levels of phenolic compounds, carbohydrates, or pigments in some of these plant materials.
2. A general principle is that protein extracts must be free of any particulate or insoluble materials before they are applied to top of first-dimensional IEF gels. Another general principle is that protein extracts should be kept at a low temperature, particularly after urea has been added. Otherwise, undesirable chemical modifications of proteins might result.
3. Loading of 20 µL of extracts containing 20 µg of proteins using the methods described here should be appropriate for gels to be silver stained following 2-D PAGE.
4. We found that 10–100 mg (fresh weight) of whole plant tissue yielded a sufficient amount of proteins for 2-D PAGE analysis. There is generally no need for a protein concentration step. In contrast, a major problem encountered with intercellular fluids is the low abundance of proteins. The acetone precipitation step is effective to concentrate extracellular proteins for subsequent 2-D PAGE.
5. Like many whole plant tissue samples, intercellular fluids may contain a high level of substances, particularly phenolic compounds, that could be problematic for generating well-separated 2-D PAGE protein profiles. The acetone precipitation step incorporating 2-mercaptoethanol is important for both protein concentration and minimizing chemical modifications of proteins. The low level of 2-mercaptoethanol does not seem to have any detrimental effect on subsequent gel analysis. Incubation overnight at –20°C was found to be satisfactory in the case of intercellular fluids of rose leaves (5). This might need to vary with different source materials.

Acknowledgment

The techniques described here were adapted from the appropriate original papers and used in the work of D. Burritt, L. Boul, M. Li, and Y. Suo, all past research students in my laboratory supported by research grants from the University of Canterbury.

References

1. O'Farrell, P. H. (1975) High resolution two-dimensional electrophoresis of proteins. *J. Biological Chem.* **250**, 4007–4021.
2. Mayer, J. E., Hahne, G., Palme, K., and Schell, J. (1987) A simple and general plant tissue extraction procedure for two-dimensional gel electrophoresis. *Plant Cell Reports* **6**, 77–81.
3. De Wit, P. J. G. M. and Spikman, G. (1982) Evidence for the occurrence of race- and cultivar-specific elicitors of necrosis in intercellular fluids of compatible interactions of *Cladosporium fulvum* and tomato. *Physiol. Plant Path.* **21**, 1–11.
4. Jeknic, Z. and Chen, T. H. H. (1999) Changes in protein profiles of poplar tissues during the induction of bud dormancy by short-day photoperiods. *Plant Cell Physiol.* **40**, 25–35.
5. Suo Y. and Leung, D. W. M. (2002) Accumulation of extracellular pathogenesis-related proteins in rose leaves following inoculation of in vitro shoots with *Diplocarpon rosae*. *Scientia Hort.* **93**, 167–178.
6. Li, M. and Leung, D. W. M. (2001) Protein changes associated with adventitious root formation in hypocotyls of *Pinus radiata*. *Biol. Plant.* **44**, 33–39.

Laser-Assisted Microdissection in Proteomic Analyses

**Darrell L. Ellsworth, Stephen Russell, Brenda Deyarmin,
Anthony G. Sullivan, Henry Brzeski, Richard I. Somiari, and Craig D. Shriver**

1. Introduction

Current research on the molecular basis of human reproductive cancers involves mapping cellular pathways and identifying molecular alterations associated with disease onset and progression. The ability to characterize changes in protein expression that occur during the transition from normal to benign disease to invasive cancer requires homogeneous populations of cells free from contamination by other cell types. With advances in proteomic technologies, proteins can now be quantified in a relatively small number of target cells obtained from a limited amount of tissue. The increased demand for selective isolation of pure cell populations from small quantities of tissue calls for refined protocols for tissue preparation and efficient methods of tissue microdissection that preserve protein integrity.

Laser-assisted microdissection is an emerging technology with increasing applications in proteomic research. Laser microdissection permits selective transfer and recovery of cells from histological tissue sections with far greater speed and precision than manual dissection methods. Homogeneous populations of specific cell types can be isolated from complex tissues and the captured material verified microscopically. Laser microdissection has become an important component of research on human reproductive cancers because neoplastic cells can be segregated from normal cells in heterogeneous carcinomas to better characterize protein expression differences between diseased and normal cells. However, detailed protein studies are feasible only if the microdissection process maintains the molecular integrity of the native proteins (1).

2. Materials

2.1. Histology

1. 70% Ethanol (AAPER Alcohol; Shelbyville, KY).
2. 100% Ethanol (AAPER Alcohol; Shelbyville, KY).
3. Sub X (xylene substitute) (Surgipath; Richmond, IL).
4. Toluidine blue stain, 1% aqueous (LabChem; Pittsburgh, PA).
5. Mayers hematoxylin (Sigma Diagnostics; St. Louis, MO).
6. Scotts tap water (bluing reagent) (Fisher Scientific; Pittsburgh, PA).
7. Uncharged glass slides (Fisher Scientific; Pittsburgh, PA).
8. Super Frost™ plus glass slides (Fisher Scientific; Pittsburgh, PA).

9. Membrane-based laser microdissection slides (Leica Microsystems; Wetzlar, Germany).
10. Disposable microtome blades, HP35n, noncoated (ThermoShandon; Pittsburgh, PA).
11. Protease inhibitor (Calbiochem[®]; San Diego, CA).

2.2. Saturation Labeling

1. Lysis buffer: 7 M urea, 2 M thiourea, 5 mM magnesium acetate, 4% CHAPS, 1% NP-40, 30 mM Tris-HCl (pH 8.0)—store buffer aliquots at -20°C (see Note 1).
2. 2-D Clean-up kit (Amersham Biosciences; Piscataway, NJ).
3. Tris (2-carboxyethyl) phosphine hydrochloride (TCEP), 2 mM solution (Pierce Chemical; Rockford, IL).
4. Cy3 and/or Cy5 maleimide dye (Amersham Biosciences; Piscataway, NJ).
5. Immobilized pH gradient (IPG) buffer (Amersham Biosciences; Piscataway, NJ).
6. 1X rehydration buffer: 7 M urea, 2 M thiourea, 4% CHAPS, 1% NP-40, 10% isopropanol, 5% glycerol, 1% IPG buffer, 130 mM DTT.
7. 2X rehydration buffer: 7 M urea, 2 M thiourea, 4% CHAPS, 1% NP-40, 10% isopropanol, 5% glycerol, 2% IPG buffer, 130 mM DTT.
8. Dithiothreitol (DTT) (Amersham Biosciences; Piscataway, NJ).
9. Chelex 100 (Bio-Rad Laboratories; Hercules, CA).

2.3. Unlabeled Samples

1. Lysis buffer: 7 M urea, 2 M thiourea, 5 mM magnesium acetate, 4% CHAPS, 1% NP-40, 30 mM Tris-HCl (pH 8.0).
2. 2-D Clean-up kit (Amersham Biosciences; Piscataway, NJ).
3. 1X rehydration buffer: 7 M urea, 2 M thiourea, 4% CHAPS, 1% NP-40, 10% isopropanol, 5% glycerol, 0.5% IPG buffer, 3 mg/mL DTT.
4. 2X rehydration buffer: 7 M urea, 2 M thiourea, 4% CHAPS, 1% NP-40, 10% isopropanol, 5% glycerol, 0.5% IPG buffer, 6 mg/mL DTT.
5. SYPRO Ruby Protein Gel Stain (Molecular Probes; Eugene, OR).

2.4. Isoelectric Focusing (IEF)

1. IPGphor IEF electrophoresis unit (Amersham Biosciences; Piscataway, NJ).
2. Ceramic strip holders (Amersham Biosciences; Piscataway, NJ).
3. Immobiline IEF strips (Amersham Biosciences; Piscataway, NJ).
4. Cover fluid (Amersham Biosciences; Piscataway, NJ).

2.5. Sodium Dodecyl Sulfate-Polyacrylamide Gel Electrophoresis (SDS-PAGE)

1. Equilibration buffer A: 6 M urea, 30% glycerol, 2% SDS, 0.5% DTT, 50 mM Tris-HCl (pH 8.8).
2. Equilibration buffer B: 6 M urea, 30% glycerol, 2% SDS, 4.5% iodoacetamide, 50 mM Tris-HCl (pH 8.8).
3. SDS-PAGE running buffer: 25 mM Tris, 192 mM glycine, 0.2% SDS.
4. Mini SDS-PAGE electrophoresis equipment (Invitrogen; Grand Island, NY).
5. Precast Tris-glycine minigels, 2-D well (Invitrogen; Grand Island, NY).
6. Agarose sealing solution: 0.5% low melting point agarose in SDS-PAGE running buffer.
7. DeCyderTM 2-D image analysis software (Amersham Biosciences; Piscataway, NJ).
8. Iodoacetamide (IAA) (Acrōs Organics; Geel, Belgium).

3. Methods

3.1. Mounting and Staining OCT-Embedded Sections on Glass or Membrane Slides

The procedure for mounting and staining OCT-embedded sections is outlined in **Subheadings 3.1.1.–3.1.3.**. It is important to add protease inhibitor (50 µL per 50 mL of solution) to all reagents except the Sub X (xylene substitute). Add additional protease inhibitor every 4 h. Steps denoted by asterisk should be used only in the preparation of tissue sections on glass slides.

3.1.1. Mounting (see **Notes 2 and 3**)

1. Place the OCT-embedded sample in the specimen holder.
2. Section at 8 µm and mount on an uncharged glass slide or a membrane slide.
3. Keep slides frozen at –80°C until ready to stain.
4. Stain only enough slides to complete microdissection within 1 h.

3.1.2. Hematoxylin Staining (Keep All Staining Reagents on Ice)

1. Immerse the slide sequentially in 70% ethanol for 1 min, ddH₂O for 1 min, and Mayer's Hematoxylin for 1 min.
2. Rinse with ddH₂O until clear, then immerse in bluing reagent for 30 s.
3. Rinse in ddH₂O, then immerse sequentially in 100% ethanol, two times, for 1 min each, then Sub-X, two times, for 5 min each.
4. Air-dry at room temperature before beginning microdissection.

3.1.3. Toluidine Blue Staining

1. Immerse the slide sequentially in 70% ethanol for 1 min, ddH₂O for 1 min, toluidine blue (1% aqueous) for 1 min, 100% ethanol rinse, two times, for 30 s each, and Sub-X, two times, for 5 min each.
2. Air-dry at room temperature before beginning microdissection.

3.2. Laser-Assisted Microdissection

3.2.1. Instrumentation

The instrument currently used in our laboratory for the majority of laser-assisted microdissection is the ASLMD Laser Microdissection system (Leica Microsystems). The ASLMD microdissection platform uses laser ablation technology in which histological sections are mounted on a membrane, cells of interest are visualized and outlined on a computer screen using the mouse, a pulsed ultraviolet laser cuts the membrane along the user-drawn outline, and the material excised by the laser falls by the force of gravity into a capture tube located directly beneath the slide. A PixCell® II laser-capture microdissection apparatus (Arcturus Engineering, Mountain View, CA) is also available for specific microdissection applications. Laser microdissection using the PixCell II entails placing a CapSure™ transparent plastic cap containing a thermoplastic membrane over a section of tissue, visualizing the tissue microscopically, and selectively adhering cells of interest to the membrane with a short, focused pulse from an infrared laser (2). Tissue sections are normally mounted on glass microscope slides to provide support for the CapSure cap during microdissection, but we have found that the efficiency of laser microdissection when using the PixCell II apparatus can be increased by mounting tissue sections on membrane-based slides rather than on glass microscope slides recommended for the PixCell II (3).

3.2.2. Microdissection Methods

A dense carcinoid tumor of the ovary was obtained from the Tissue Repository at the Windber Research Institute and subjected to laser-assisted microdissection using the methods outlined above (*see Notes 4–6*). Approximately 5000 cells were collected in separate tubes and stored at –80°C until proteomic analysis could be performed.

3.3. Proteomic Analysis

3.3.1. Sample Preparation

1. To solubilize the microdissected tissue, add 100 µL of lysis buffer supplemented with 1X protease inhibitor. Place the tube on ice for approx 30 min.
2. After the 30-min incubation, sonicate the microdissected sample in a bath sonicator for approx 5 s. A bath sonicator is used to minimize sample heating. Be sure to return the sample to ice immediately following sonication for at least 1 min. Repeat the entire sonication process two times.
3. After sonication, centrifuge the sample at 15,000g for 10 min at 4°C.
4. Carefully remove the supernatant and transfer to a clean tube.
5. The sample lysate will be dilute and unpurified; therefore, the sample will need to be concentrated and purified prior to labeling. Use the 2-D cleanup kit following the manufacturer's instructions.

3.3.2. Saturation Labeling

1. To the entire volume of lysate, add 1 µL of 2 mM TCEP (2 nmoles) and mix thoroughly with a pipet.
2. Briefly centrifuge the tube, then incubate at 37°C for 1 h, avoiding prolonged exposure to light.
3. Add the Cy3 (or Cy5) dye to the tube. The volume of Cy3 (or Cy5) dye added should be twice the volume of the TCEP (2 µL of Cy dye [4 nmoles] when 1 µL of TCEP is used). Mix well using a pipet.
4. Briefly centrifuge the tube, then incubate at 37°C for 1 h, protecting the lysate from light.
5. Calculate the liquid volume in the tube and stop the labeling reaction by adding an equal volume of 2X rehydration buffer. Mix well with a pipet.
6. The labeled sample can be used immediately or stored at –70°C for 30 d.

3.3.3. First-Dimension Gel Electrophoresis (Isoelectric Focusing)

To assay differential protein expression between two samples, the lysates from each sample can be labeled with different Cy dyes, then mixed in equal amounts to enable direct comparisons of the samples on one gel with no gel-to-gel variations. *See Chapter 24 for a discussion of differential in-gel electrophoresis (DIGE).*

FOR LABELED SAMPLES

1. Increase the labeled sample to an appropriate volume with 1X rehydration buffer. The final volume depends on the length of the immobiline strip to be used and the manufacturer's recommendation. For 7-cm strips, the final volume should be 125 µL.
2. Add the labeled sample to the bottom of a 7-cm strip holder and place a 7-cm (pH 3.0–10.0) immobiline IEF strip on top of the sample, gel side down.
3. Rehydrate the IEF gel for 12 h (10 h minimum) while applying 30 V.
4. Focus the 7-cm (pH 3.0–10.0) IEF gel for 12,600 Vh as follows: 500 V for 30 min, 1000 V for 30 min, a 30-min linear gradient up to 5000 V, and a final 10,000 Vh step at 5000 V.
5. The focused strip can then be processed for the second dimension immediately or stored at –70°C in a sealed container for up to 2 wk.

FOR UNLABLED SAMPLES

1. Dilute the entire volume of lysate 1:1 with 2X rehydration buffer. Increase the unlabeled sample to an appropriate volume with 1X rehydration buffer. The final volume depends on the length of the immobiline strip to be used and the manufacturer's recommendation. For 7-cm strips, the final volume should be 125 μ L.
2. Add the unlabeled sample to the bottom of a 7-cm strip holder and place a 7-cm (pH 3.0–10.0) immobiline IEF strip on top of the sample, gel side down.
3. Rehydrate the IEF gel for 12 h (10 h minimum) while applying 30 V.
4. Focus the 7-cm (pH 3.0–10.0) IEF gel for 12,600 Vh as follows: 500 V for 30 min, 1000 V for 30 min, a 30-min linear gradient up to 5000 V, and a final 10,000 Vh step at 5000 V.
5. The focused strip can then be processed for the second dimension immediately or stored at –70°C in a sealed container for up to 2 wk.

3.3.4. Second-Dimension Gel Electrophoresis**FOR LABELED SAMPLES**

1. Place the focused strip in a container so the gel comes in direct contact with equilibration buffer A. For a 7-cm strip we recommend a screw-top 15-mL centrifuge tube.
2. Add equilibration buffer A to the tube, making sure the plastic backing of the strip is in contact with the vessel wall and the gel side is facing out. Equilibrate on a rocker for 15 min at room temperature. For a 7-cm strip, use 3 mL of equilibration buffer.
3. Place the IEF strip in a mini SDS-PAGE gel with either one well or a 2-D well (*see comment below*). If necessary, the strip can be trimmed up to the point where the electrodes from the strip holder contact the gel. Overlay the strip with melted agarose sealing solution, making certain there are no bubbles between the IEF strip and the second-dimension gel. (Comment: If there are no wells for molecular-weight markers, the markers can still be run on the gel by absorbing them onto a strip of filter paper and inserting the filter paper in the well at either end of the gel. The paper strip has to be small enough to be completely covered in agarose sealing solution.)
4. Assemble the electrophoresis unit and run the gel at constant voltage (40 V) for 15 min or until the blue dye enters the gel.
5. After 15 min, raise the voltage to 125 V, and run the gel until the blue dye migrates to the bottom of the gel.

FOR UNLABLED SAMPLES

1. Place the focused strip in a container so the gel comes in direct contact with equilibration buffer A. For a 7-cm strip we recommend a screw-top 15-mL centrifuge tube.
2. Add equilibration buffer A to the tube, making sure the plastic backing of the strip is in contact with the vessel wall and the gel side is facing out. Equilibrate on a rocker for 15 min at room temperature. For a 7-cm strip, use 3 mL of equilibration buffer.
3. Pour off equilibration buffer A and add equilibration buffer B. Equilibrate as in **step 2** above.
4. Place the IEF strip in a mini SDS-PAGE gel with either one well or a 2-D well (*see comment below*). If necessary, the strip can be trimmed up to the point where the electrodes from the strip holder contact the gel. Overlay the strip with melted agarose sealing solution, making certain there are no bubbles between the IEF strip and the second-dimension gel. (Comment: If there are no wells for molecular-weight markers, the markers can still be run on the gel by absorbing them onto a strip of filter paper and inserting the filter paper in the well at either end of the gel. The paper strip has to be small enough to be completely covered in agarose sealing solution.)
5. Assemble the electrophoresis unit and run the gel at constant voltage (40 V) for 15 min or until the blue dye enters the gel.

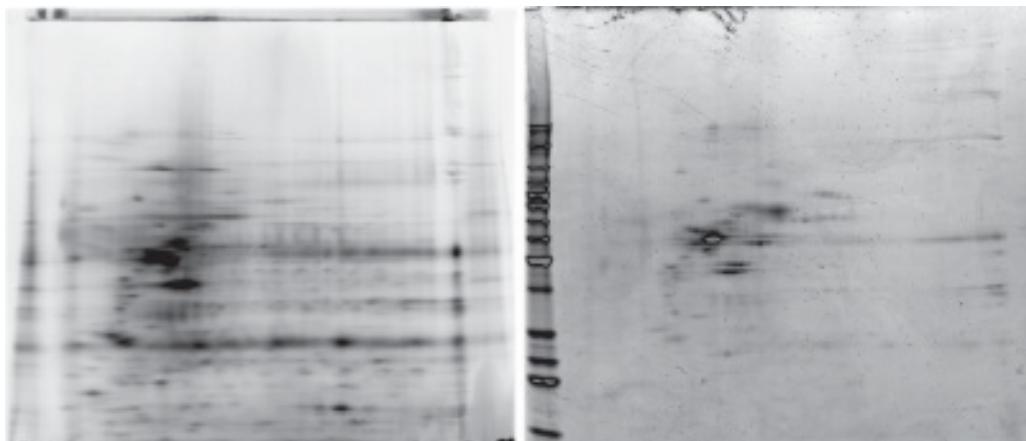


Fig. 1. Two-dimensional gels showing resolution of proteins from approx 5000 cells using Cy5 maleimide saturation dye (left panel) and SYPRO Ruby Protein Gel Stain (right panel). Note the difference in the number of detectable spots.

6. After 15 min, raise the voltage to 125 V, and run the gel until the blue dye migrates to the bottom of the gel.

3.3.5. Image Analysis

FOR LABELED SAMPLES

1. Image the resulting gel on a fluorescence scanner capable of scanning at the wavelengths necessary to image the Cy dye (for Cy 3, 532 λ excitation, 580 λ emission; for Cy 5, 633 λ excitation, 670 λ emission).
2. Analyze the scanned image using DeCyder analysis software.

FOR UNLABELED SAMPLES

1. After second-dimension electrophoresis, the gel can be stained with a variety of available protein stains. A fluorescent stain, such as SYPRO Ruby (4), has a high dynamic range of detection and is as sensitive as silver stain.
2. Scan and view the gel per the manufacturer's suggestions and analyze the image using DeCyder analysis software.

3.3.6. Results

Two-dimensional gels resulting from analysis of approx 5000 cells derived from a carcinoid tumor of the ovary following the procedures outlined above are shown in **Fig. 1**. Note that the Cy5 maleimide saturation dye is more sensitive and allows automated detection of a greater number of spots than the SYPRO Ruby protein stain (**Fig. 2**). Visualization of a randomly chosen spot indicates that the Cy5 saturation dye provides better spot morphology and less background (**Fig. 3**).

4. Notes

1. We routinely pass our buffers containing thiourea through a weak cation chelating resin (Chelex 100) before the addition of buffers or salts to greatly diminish electroendosmosis. Electroendosmosis is the process by which water is actively transported to the ends of the IPG strip, and is greatly exaggerated by the presence of charged molecules in the rehydra-

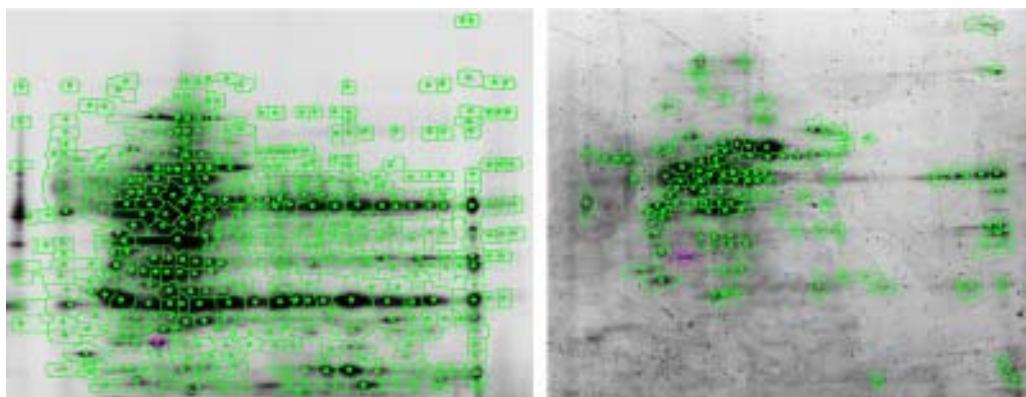


Fig. 2. Automated spot detection on two-dimensional gels visualized with Cy5 maleimide saturation dye (left panel) and SYPRO Ruby Protein Gel Stain (right panel) using DeCyder Differential Analysis Software DIA (differential in-gel analysis). The number of spots detected was 440 for the Cy5 maleimide saturation dye and 123 for the SYPRO Ruby stain.

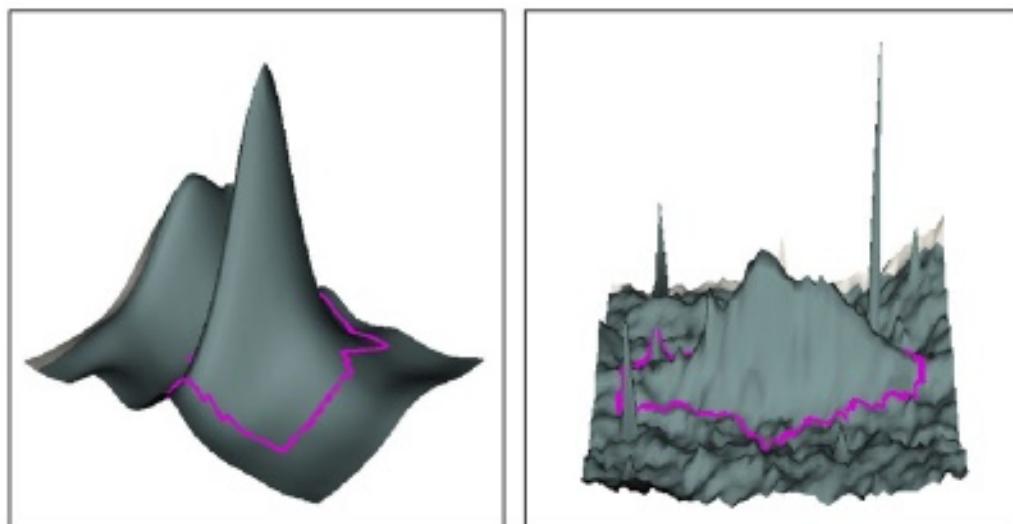


Fig. 3. Three-dimensional images of a randomly chosen spot from the two-dimensional gels shown in **Figs. 1** and **2**. The image in the left panel was stained with Cy5 maleimide saturation dye and the image in the right panel was stained with SYPRO Ruby protein stain.

tion buffer. Removing charged molecules will minimize electroendosmosis and greatly reduce distortions, streaking, and scorching of the gel.

2. When sectioning OCT samples on the cryostat, the Leica foil-on-glass slides (PEN slides) work better than the membrane-based slides with a stainless steel frame (PET slides). If you must use the PET slides, we have found that supporting the back of the slide with a glass slide improves tissue adherence.
3. We recommend mounting the first and last sections from the series of tissue sections on Super Frost™ plus slides, staining with hematoxylin and eosin, and adding a cover slip.

We examine the first section under the microscope to determine the number of additional sections needed based on our own estimation of the number of desired cells present.

4. We generally target and cut areas for microdissection at $\times 10$ magnification, but we may cut at $\times 4$ magnification (after carefully inspecting at $\times 10$) if we are working with a large homogeneous tumor or tissue type.
5. Be sure to calibrate the laser each morning—additional calibrations may be needed if the laser tends to drift from the user-defined cut lines.
6. With Laser Microdissection LMD software version 4.0, the laser control settings we typically use are: aperture, 12–16; intensity, 40–46; speed, 1–3 for clean cuts; offset, 22; bridge, small; aperture differential, 6.

Acknowledgments

John Yerger and Harold Ashcraft of the Windber Medical Center Pathology Department provided access to archival tissue samples and contributed invaluable guidance. Laurel Hoffman provided laser microdissection assistance.

References

1. Fend, F. and Raffeld, M. (2000) Laser capture microdissection in pathology. *J. Clin. Pathol.* **53**, 666–672.
2. Emmert-Buck, M. R., Bonner, R. F., Smith, P. D., et al. (1996) Laser capture microdissection. *Science* **274**, 998–1001.
3. Ellsworth, D. L., Shriver, C. D., Ellsworth, R. E., Deyarmin, B., and Somiari, R. I. (2003) Laser capture microdissection of paraffin-embedded tissues. *BioTechniques* **34**, 42–46.
4. Berggren, K., Chernokalskaya, E., Steinberg, T. H., et al. (2000) Background-free, high sensitivity staining of proteins in one- and two-dimensional sodium dodecyl sulfate-polyacrylamide gels using a luminescent ruthenium complex. *Electrophoresis* **21**, 2509–2521.

Purification of Cellular and Organelle Populations by Fluorescence-Activated Cell Sorting for Proteome Analysis

William L. Godfrey, Colette J. Rudd, Sujata Iyer, and Diether Recktenwald

1. Introduction

A goal of proteomic research is to detect all proteins in a particular biological subsystem, such as a cell type or cellular subfraction (1). This effort requires a combination of technologies—first, for subfractionation and purification of the cells and cellular components of interest, and second, a sensitive method for detection and identification of all possible proteins. The challenge of proteomics is the wide range of abundance and sizes of cellular proteins, as well as the number of different proteins and posttranslational modifications. Clearly, maximizing both the specificity of the sample purification method and the sensitivity of the protein detection method is essential. In addition, targeting the analysis to specific sub-components of the cell can both enhance the sensitivity of the analysis and contribute to the functional analysis of the proteins (2).

Antibody-based purification methods for proteins are well established, and have proven particularly useful for the isolation of blood cell populations. Two powerful cell purification methods that incorporate antibodies for specificity are fluorescence-activated cell sorting (FACS) and immunomagnetic separation (3,4). Using FACS, fluorescent-labeled antibodies bound to cell-specific antigens serve as guides for directing targeted cells to the collection vessel (Fig. 1). Other markers, both internal and external, can also be used to characterize and separate cell types with current FACS technology. With magnetic separation, the cells or subcomponents of interest must be labeled with a specific antibody. Both methods are capable of yielding millions of purified cells of a specific subtype (4).

In this chapter, we describe our approach using the specificity and practicality of FACS-based cell purification methods to purify human leukocyte populations, murine liver-cell populations, and mitochondria for proteomic analysis. CD4 and CD8 monoclonal antibodies were used for selection of the T-cells. CD45 and propidium iodide were used to select viable cell subsets from liver-cell preparations. Anti-metaxin was used to sort mitochondria from liver-cell homogenates. Protein identification was accomplished using liquid chromatography coupled to an “ion trap” mass spectrometer (LC-MS/MS), a method that allows identification of many proteins in a single analysis without a requirement for prior protein purification (5). For the analysis, proteins must be cleaved into multiple peptides, ideally ranging in size from 1000 to 3000 Da. As the

From: *The Proteomics Protocols Handbook*
Edited by: J. M. Walker © Humana Press Inc., Totowa, NJ

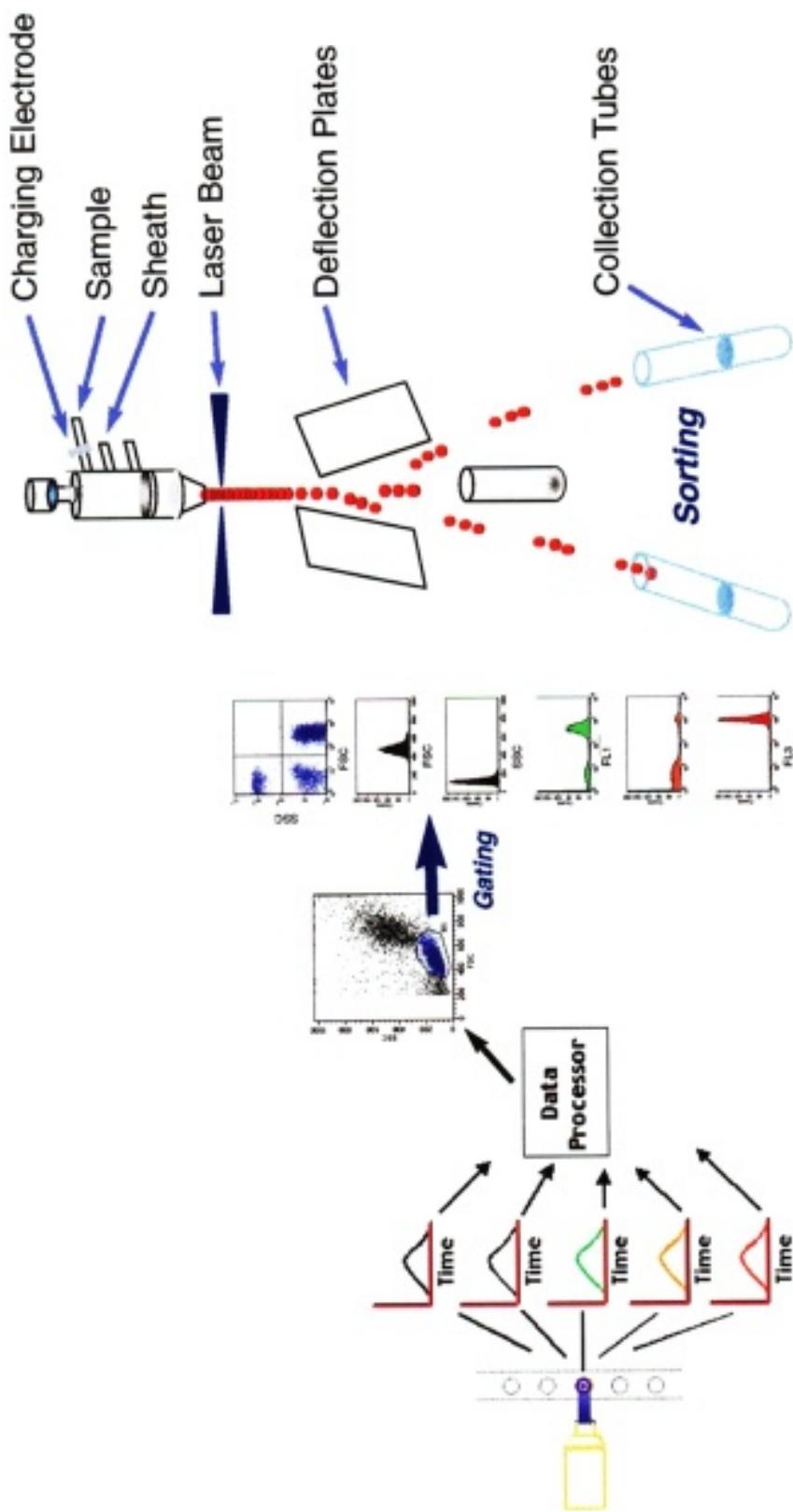


Fig. 1. Fluorescence-activated cell-sorter operation.

peptides are eluted into the mass spectrometer, ions within narrow mass ranges are sequentially collected and fragmented via collision-induced dissociation (CID). The fragmentation ions are used to identify each peptide, which is then matched to its parent protein using protein or genomic databases. LC-MS/MS analysis using an ion trap is relatively sensitive, allowing detection of less than one femtomole of some peptides (6).

Since proteins are localized to different parts of the cell, isolation and concentration of subcellular components can potentially enable the detection of otherwise undetectable protein constituents. The cell surface is an area of interest for study, and its accessibility to extracellular proteases facilitated the development of a peptide-based method of analysis. Because peptides can be identified using LC-MS/MS, it was practical to produce and analyze tryptic peptides from proteins associated with the cell surface of purified leukocytes, while minimizing the presence of abundant cytoplasmic proteins. This method, described in detail in this chapter, combines the power of FACS with the capability of LC-MS/MS for peptide identification, enabling the identification of cell-surface proteins on specific subsets of human T-cells.

In summary, FACS provides a powerful tool to separate subpopulations of cells in sufficient quantity for global proteome analysis by LC-MS/MS. Proteins expressed at under 100,000 copies per cell, including lineage-specific proteins, may be detected with the LC-MS/MS approach in this chapter. The ability to identify an individual protein depends on multiple factors, including the protein's concentration on the cell; its post-translational modifications; the number, concentration, and sizes of peptides generated from the protein; the number of co-eluting peptides during the chromatography; and the sophistication of the software for peptide matching. Although the complexity of the problem is challenging, development of new methods promises to increase the sensitivity of the proteomic analysis and lead to even better identification of proteins in the future.

2. Materials

2.1. Cell Preparation Reagents

2.1.1. Blood Cells

1. Normal human blood, drawn in an ethylenediaminetetraacetic acid (EDTA) BD Vacutainer® tube (BD Vacutainer, BD Preanalytical Systems, Franklin Lakes, NJ).
2. Ficoll-paque Plus (Pharmacia, Uppsala, Sweden).
3. 50-mL Polyethylene conical tubes (BD Biosciences Discovery Labware, Bedford, MA).
4. PBS-P: 10 mM sodium phosphate, 150 mM sodium chloride, 0.5% (w/v) Pluronic® F-68 (BASF Corporation, Mount Olive, NJ), pH 7.4. Filter through 0.4 μ m and store at 4°C. Stable for 1 wk.
5. Anti-CD4, FITC-conjugated, anti-CD8, RPE-conjugated (BD Biosciences Immuno-cytometry Systems, San Jose, CA). Store at 4°C, protected from light.

2.1.2. Murine Liver Cells

1. Liberase (Promega, Madison, WI).
2. 1 M CaCl_2 , in RO/DI water.
3. 2 M MgCl_2 , in RO/DI water.
4. Hanks balanced salt solution (HBSS; GIBCO, Grand Island, NY).
5. Digestion medium: 5 mL HBSS, 100 μ L (1.4 units) liberase, 25 μ L 1 M CaCl_2 , 25 μ L 2 M MgCl_2 .
6. 37°C Waterbath.

7. Trypan blue (GIBCO).
8. Microscope and hemocytometer.
9. Nylon mesh, 100 μ m openings (Small Parts, Miami, FL).
10. Fetal bovine serum (FBS; GIBCO).
11. HBSS supplemented with 5% FBS.
12. Anti-CD45, allophycocyanin (APC) conjugated (BD Biosciences Immunocytometry Systems). Store at 4°C, protected from light.
13. Propidium iodide (Molecular Probes, Eugene, OR). Store desiccated at 4°C, protected from light.

2.1.3. Mitochondria

1. Homogenization medium (HM): 200 mM mannitol, 50 mM sucrose, 10 mM KCl, 10 mM HEPES, 1 mM EDTA, pH 7.4. Store for up to 2 d at 4°C.
2. Nylon mesh, 10 μ m openings (Small Parts).
3. Glycerol.
4. Anti-metaxin (BD Biosciences Pharmingen, San Diego, CA). Store at 4°C.
5. Anti-mouse kappa chain, RPE-conjugated (BD Biosciences Immunocytometry Systems). Store at 4°C, protected from light.

2.2. Fluorescence-Activated Cell Sorting

1. Instruments: BD FACS Vantage™-SE or BD FACSAria™ cell sorter (BD Biosciences Immunocytometry Systems).
2. FACS Flow sheath fluid (BD Biosciences Immunocytometry Systems).
3. Alignment particles appropriate to cytometer, per manufacturer's recommendations.
4. AccuDrop Beads (BD Biosciences Immunocytometry Systems).

2.3. LC-MS/MS Reagents

1. Tris (2-carboxyethyl) phosphine (TCEP; Pierce, Rockford, IL).
2. Sequencing-grade modified trypsin (Promega).
3. HBSS (GIBCO).
4. Sterile phosphate-buffered saline (PBS; GIBCO).
5. Trypan blue (GIBCO).

3. Methods

3.1. Cell Preparation

3.1.1. Preparation and Labeling of Blood Cells

1. Collect normal human blood in EDTA Vacutainer tubes.
2. Dilute 10 mL whole blood with 20 mL PBS-P and underlay with 10 mL Ficoll-Paque Plus in a 50-mL centrifuge tube. Centrifuge at 400g for 20 min at room temperature.
3. Aspirate the plasma layer (top) and collect the mononuclear cells from the plasma/ficoll interface, expecting a recovery of approx 1×10^7 peripheral blood mononuclear cells (PBMCs) per mL of blood.
4. Wash cells twice with 20 mL PBS-P in a centrifuge at 800g for 30 min.
5. Resuspend the cells in 10 mL PBS-P and check cell concentration with a hemocytometer.
6. Adjust cell concentration to approx 10^7 cells per mL PBS-P. Reserve an aliquot of unlabeled cells on ice for instrument set up.
7. For 5 mL of cell suspension, add 500 μ L of anti-CD4-FITC and anti-CD8-RPE and react for 15 min at room temperature in the dark (see Note 1).
8. Wash the cells by centrifuging at 800g for 15 min and resuspend in the original volume of PBS-P. Hold on ice until ready to put onto a FACS instrument.

3.1.2. Preparation and Labeling of Liver-Cell Suspensions

1. Dissect the liver from a young (8–10 wk old) mouse; expect 1 to 2 g of tissue weight. Place the liver in a chilled Petri plate with 10–15 mL HBSS (*see Note 2*).
2. Decant the HBSS and mince the tissue with a razor blade to pieces approx 1 mm across. Suspend the pieces in 20 mL cold HBSS and centrifuge at 500g for 5 min at 4°C (*see Note 3*).
3. Decant the supernatant and wash minced tissue by resuspending the pellet in 20 mL cold HBSS, then centrifuging at 500g for 5 min at 4°C.
4. Decant the supernatant and resuspend the minced tissue in 5 mL of warm digestion medium. Allow digestion to proceed in a 37°C waterbath for 10 to 20 min with occasional agitation, checking progress by observing aliquots in trypan blue on a microscope (*see Note 3*).
5. When digestion is largely complete, stop the process with 500 µL of FBS.
6. Filter the digest through 100-µm nylon mesh. If desired, the retained material in the mesh can be further treated with digestion medium.
7. Centrifuge the filtrate at 800g for 5 min at 4°C.
8. Resuspend the pellet in 2 mL cold HBSS with 5% FBS. If large pieces of tissue are present, they may be allowed to settle out, or the suspension may be centrifuged at 50g for 1–2 min to remove them.
9. Determine the cell concentration using a hemocytometer and adjust the concentration to about 2×10^7 cells/mL with HBSS + 5% FBS. Reserve an aliquot of unlabeled cells on ice for instrument setup.
10. To stain 1 mL of cell suspension, add 20 µg anti-CD45-APC or other appropriate fluorescent antibody and incubate for 15–30 min on ice (*see Note 4*).
11. After staining, add 1 mL HBSS + 5% FBS and centrifuge at 800g for 5 min at 4°C.
12. Resuspend in 5 mL of HBSS + 5% FBS (about 2×10^6 cells/mL) and hold on ice until ready to put onto a FACS instrument. Just before sorting, add propidium iodide to 12 µM to stain dead cells.

3.1.3. Preparation and Labeling of Mitochondrial Suspensions

1. Dissect the liver from a young (8–10 wk old) mouse. Expect 1 to 2 g of tissue weight. Place the liver in a chilled Petri plate with 10–15 mL HM (*see Note 5*).
2. Decant the HM and mince the tissue with a razor blade to pieces 2–3 mm across. Suspend the pieces in 20 mL cold HM and allow to settle.
3. Decant the supernatant. Suspend the tissue in 20 mL cold HM and transfer to a chilled dounce homogenizer vessel.
4. Set pestle speed at approx 500 rpm and homogenize the tissue with five to six up-and-down strokes.
5. Pass the homogenate through nylon mesh with 10-µm openings and collect the filtrate. Store at 4°C for use within 1 d, or add glycerol to 40% and freeze for longer storage.
6. For staining, dilute the fresh homogenate 1:400 with PBS-P; dilute thawed homogenate 1:200. This dilution should be sufficient to yield at least 1000 mitochondrial events per microliter, but yield will depend on source tissue and efficiency of homogenization (*see Note 6*).
7. Add 20 µg of anti-metaxin or other primary antibody to 500 µL diluted homogenate and incubate for 30 min on ice. Add 2 mL PBS-P and centrifuge at 3000g for 5 min to wash. If the primary antibody is fluorescently conjugated, the pellet may be resuspended for sorting at this point.
8. Resuspend the pellet with 500 µL of PBS-P and add 20 µg of phycoerythrin-conjugated anti-mouse kappa chain. Incubate for 30 min on ice. Add 2 mL PBS-P and centrifuge at 3000g for 5 min.
9. Resuspend the pellet in 2 mL PBS-P and hold on ice until put on a FACS instrument.

3.2. Cell Sorting

3.2.1. Fluorescence-Activated Cell Sorter Set-Up

1. Cells were sorted using either a BD FACSVantage™ SE or a BD FACS Aria™ system. Refer to instrument user manual for detailed instructions on instrument setup and use (see Note 7).
2. Select sample tip, sheath pressure, drive frequency, and event rate as discussed below under individual applications.
3. Once the sorting stream has been set up, check the analysis performance and watch the drop break-off point for any changes.
4. Adjust the drop delay using the AccuDrop system or by performing a drop delay profile. A drop delay profile involves sorting fluorescent beads onto separate spots on a microscope slide with incremental changes to the drop delay setting, then determining the setting that gives the best bead recovery.

3.2.2. Blood-Cell Sorting

1. For the human leukocytes, use a 70-micron nozzle tip with sheath pressure at 40 psi sheath pressure and a drop drive frequency of approx 64 kHz. Use a sample rate of 10,000 events per second.
2. Set up cytometer with unlabeled cells by setting forward and side scatter to linear amplification and by adjusting the detectors to place the lymphoid population in the lower center of a forward scatter by side scatter plot (Fig. 2). Use a forward scatter threshold. Set fluorescein and phycoerythrin detectors to log amplification and adjust to put the unlabeled cells low in the first decade.
3. Run stained cells and adjust compensation so that stained CD4 and CD8 populations appear as in Fig. 2. Draw sort gates around the desired populations (see Note 8).
4. Use Normal-R sort mode, or a sort mode appropriate for normal recovery.
5. An aliquot of sorted cells may be rerun to determine sort purity.

3.2.3. Hepatocyte Sorting

1. For murine liver cells, use a 100-micron tip with sheath pressure set at 15 psi and drop drive frequency at approx 30 kHz. Since the drop frequency is lower than above, the event rate needs to be lowered accordingly. A rate of 5000 events per s is a good starting point.
2. Set up cytometer with unlabeled cells so that the largest-size cells are in the third decade on log-amplified forward scatter and side scatter (Fig. 3). Use a forward scatter threshold. Set propidium iodide and allophycocyanin detectors in log amplification so that unstained cells are low in the first decade.
3. Run the stained cells. Leucocytes should appear low in forward scatter and bright in allophycocyanin. Dead cells will appear bright for propidium iodide. For lymphocytes, set sort gates to include allophycocyanin-positive events, but to exclude propidium iodide-positive events. For hepatocytes, set sort gate on the population high in forward and side scatter, but exclude propidium iodide-positive events (Fig. 3) (see Note 8).
4. Use Normal-R sort mode, or a sort mode appropriate for normal recovery.
5. An aliquot of sorted cells may be rerun to determine sort purity.

3.2.4. Mitochondria Sorting

1. For mitochondria, use a 70- μ nozzle tip at 40 psi sheath pressure and drop drive frequency of approx 64 kHz. Use a sample event rate of 10,000 events per s. For mitochondria, anticipate performing a two-pass sort to separate mitochondria from small cytosolic material.
2. Set a forward or side scatter threshold. Observing a scatter histogram, adjust scatter so that events just appear above the threshold. Continue to increase the detector voltage set-

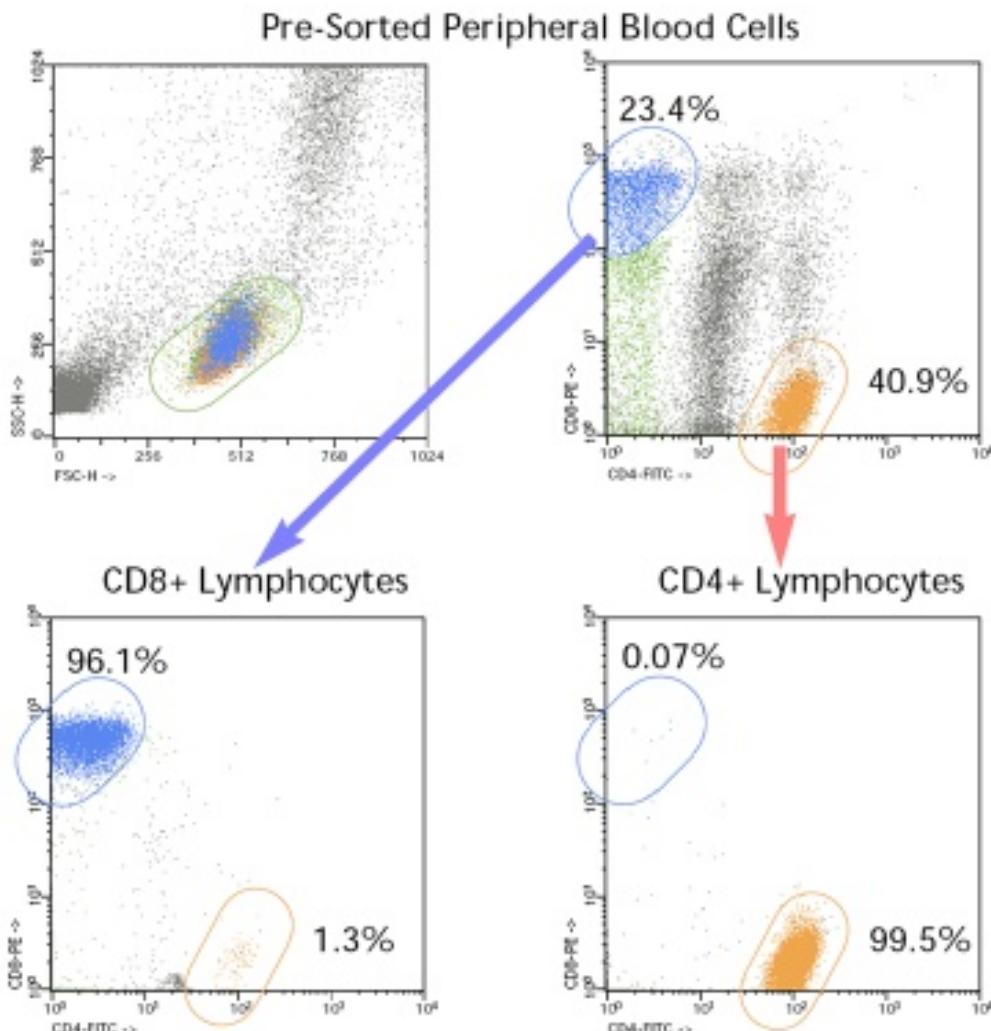


Fig. 2. Sort strategy for CD4 and CD8 cells. The dot plots show the sorting strategy used for stained peripheral blood cells and population purity after sorting for CD4⁺ and CD8⁺ cells. CD4 cells were gated on scatter and FITC fluorescence; CD8 bright cells were gated on scatter and RPE fluorescence. Sorted populations showed >95% purity.

ting until the noise population is seen to rise sharply. Decrease the detector voltage so that a peak or shoulder can be observed on the noise.

- With the stained sample, gate on the stained mitochondria by fluorescence and scatter as in **Fig. 4** (see Note 9).
- Use the Enrich sort mode for the first sort pass and the Normal-R mode for the second sort.
- An aliquot of sorted events may be rerun to determine sort purity.

3.3. LC-MS/MS

3.3.1. Preparation of Tryptic Peptides

- Wash 5–20 × 10⁶ sorted cells twice with 5 mL HBSS and resuspend at approx 10⁷ cells per mL HBSS.

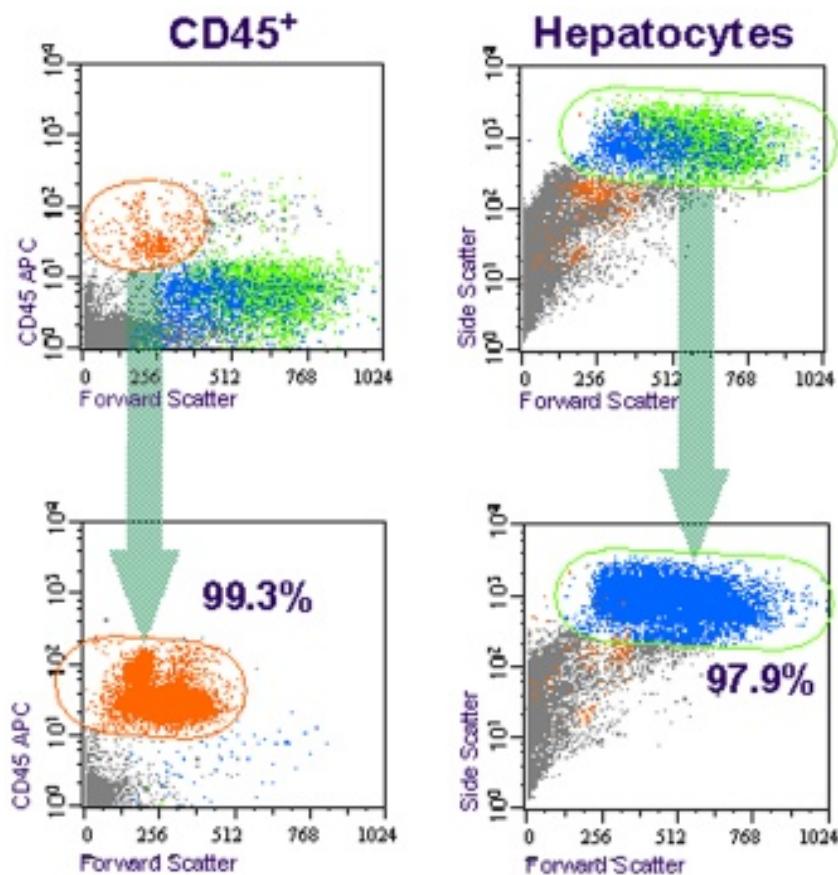


Fig. 3. Sort strategy for cells from liver. Cells were stained with anti-CD45-APC and propidium iodide, then sorted based on scatter and fluorescence. Hepatocytes were gated based on forward scatter and side scatter, and viability (PI^-). Viable leukocytes were gated using scatter, CD45⁺ staining, and PI^- staining. Recovered populations are expressed as a percentage of total gated (hepatocyte and leucocyte).

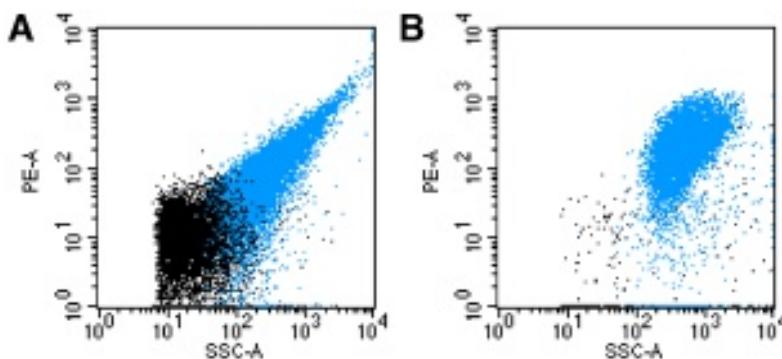


Fig. 4. Sort strategy for mitochondria. Liver homogenate was stained with anti-metaxin and phycoerythrin-conjugated anti-mouse kappa antibodies. Mitochondria were gated on side scatter and phycoerythrin fluorescence (A). The initial sort was sorted a second time to remove residual cytosolic components (B).

2. Add cells to a 1.5-mL microcentrifuge tube and centrifuge at 2000g for 1 min (room temperature).
3. Decant the supernatant and resuspend the cells in 1 mL of 10 mM TCEP in HBSS, pH 7.0 (up to 20×10^6 cells per mL) at room temperature for 5 min to reduce extracellular disulfide bonds.
4. Wash cells once with 1 mL HBSS and resuspend the cell pellet in 0.1 mL of sterile PBS containing 20 μ g of sequencing-grade trypsin. Incubate for up to 30 min at room temperature. Check cell viability with trypan blue and a hemocytometer, limiting the incubation time to 15 min if necessary to maintain high viability.
5. Centrifuge the cells at 2000g for 1 min and transfer the supernatant to a fresh microfuge tube.
6. Incubate the supernatant at 37°C overnight. Store fully digested sample at -80°C until LC-MS/MS analysis.

3.3.2. Peptide Analysis by LC-MS/MS

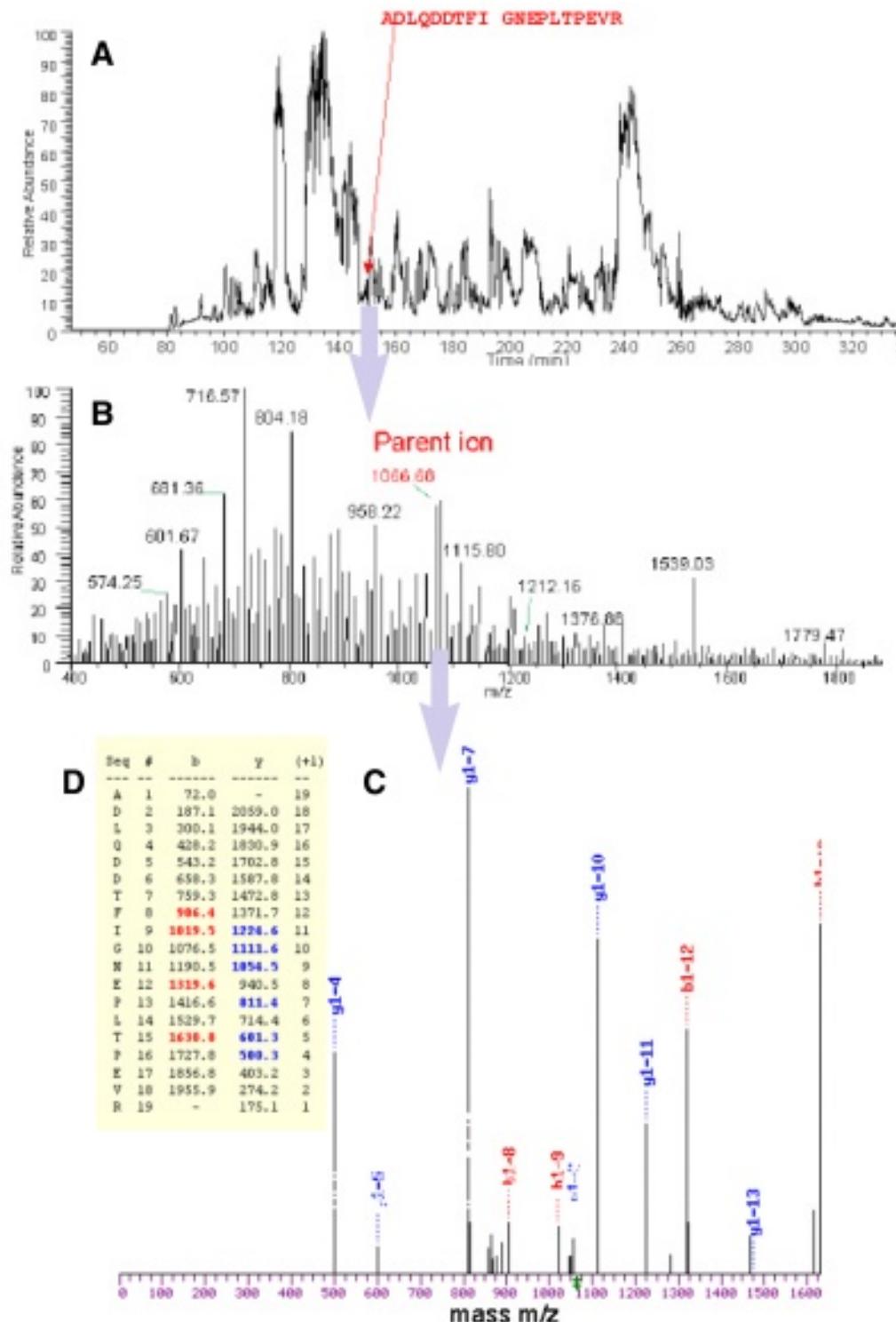
1. An LC-MS/MS system, such as a Surveyor® HPLC (Thermo Electron, San Jose, CA) equipped for nanoflow operation, coupled to an ion-trap mass spectrometer (LCQ™ Deca, Thermo Finnigan) as described in (6) is necessary for identification of peptides released from the cells.
2. Reduce the supernatant by adding 1 mM TCEP, then heat for 5 min at 90°C. Centrifuge sample in microfuge at high speed for 1 min to remove any precipitates. Concentrate the peptide sample (approx 100 μ L) on a polystyrene/divinylbenzene trap and then elute the peptides onto a reversed-phase C18-BioBasic packed-tip column (75 μ m i.d. \times 10 cm) at a flow rate of 100 nL/min and with a gradient of 0–40% B (B = 0.1% formic acid in acetonitrile, A = 0.1% formic acid in water) over 6 h.
3. Acquire data using a mass spectrometer with a feature such as Dynamic Exclusion™, with three MS/MS scans for every full MS scan, to maximize the detection of lower-abundance ions. For the data analysis, compare the observed fragmentation spectra with theoretical spectra based on peptides derived from the human-protein database, allowing for up to two missed cleavages per peptide (see Notes 10 and 11).

4. Notes

1. For optimal separation of blood-cell populations, antibody reagents should be titrated before use. It is important to keep the reagent concentration constant when increasing the sample volume. Keep in mind that for high-avidity antibodies and high-density staining, the binding of the reagent to cells will decrease the effective reagent concentration.
2. For liver-cell preparation, starve the mouse overnight to reduce glycogen content of the liver. The mouse can be exsanguinated by cardiac puncture to reduce the number of erythrocytes in the liver.
3. Tissue can also be chopped with a mechanical chopper. Digestion can be accelerated by using a wrist-action shaker in a 37°C waterbath. Alternate digestion protocols that provide viable single cells can be used. Viability of the cell preparation during digestion can be monitored by microscope with trypan blue.
4. Antibodies should be titrated for optimal staining; titration can be done with smaller reaction volumes. As a rule of thumb, start with about 1 μ g antibody for 10^6 cells. Avoid long and vigorous centrifugation to minimize damage to hepatocytes.
5. The cell homogenization protocol was adapted from (7,8). Any procedure that produces intact mitochondria can be used. For preparation of mitochondria, starve the mouse overnight to reduce the glycogen content of the liver. Tissue can also be chopped with a mechanical chopper. Protease inhibitors may be added to the mitochondrial suspension to improve stability during storage.

6. Anti-metaxin has been found to label the outer surface of murine and human mitochondria (unpublished data). Antibodies should be titrated for optimal staining; titration can be done with smaller reaction volumes. Aggregation during staining can be reduced by greater dilution of the homogenate before staining. The amount of homogenate to be stained will depend on the number of events to be sorted.
7. FACS instrument set-up parameters and optimal sorting rates vary across instrument manufacturers and models. It is best to start with recommended settings for the instrument used. For more information on expected sort performance, see the sort calculator on the Purdue University Cytometry Laboratory website (<http://www.cyto.purdue.edu/flowcyt/software/data/jscript/recovery.htm>). It is possible that the drop drive will add noise to the scatter parameters, which can affect threshold and gate efficiencies. A slight adjustment to the obscuration bar can correct these effects. Use a single-drop sorting mode for better side-streams.
8. For blood- and liver-cell sorting, setting compensation to correct for spectral overlap aids in identifying populations of interest, but is not necessary as long as target populations can be resolved and gated.
9. Mitochondria demonstrate a fairly broad range of scatter and will not be well resolved from the other cytosolic components in the homogenate. Event rate may need to be decreased if the mitochondrial shoulder cannot be seen.
10. The results of LC-MS/MS analyses of both CD4⁺ (12–15 × 10⁶ cells) and CD8⁺ (5 × 10⁶ cells) sorted-cell-sample digests revealed large numbers of peptides (Fig. 5). Trypsin digestion of TCEP-treated cells decreased the staining of CD4, CD8, and CD3 by 98%, 82%, and 33%, respectively, indicating that peptide release is not uniform across protein targets (data not shown). A total of 387 proteins were identified in the CD8⁺ cells, and out of these, 44 were known to be membrane bound and 14 had CD designations. The CD4⁺ cells yielded a total of 347 protein-identifying MS/MS spectra, 36 of which were membrane proteins and 9 of which had CD designations. The presence of intracellular proteins in the samples is probably the result of highly abundant proteins coming from small numbers of dead cells in the samples, because the trypsin treatment was shown by FACS analysis to have minimal effect on viability and membrane integrity.
11. Two to five million mitochondrial events should be sorted at a minimum for proteome analysis. Because of the small size of cytosolic contaminants, it is likely that a significant amount of cytosolic components may be trapped in sorted drops. If FACS analysis of sorted material shows significant unstained debris, the sample should be sorted an additional time.

Fig. 5. (opposite page) Typical LC-MS/MS run. Peptide mixtures were separated by reverse-phase high-performance liquid chromatography (A) as described in Methods. Eluted peptides were subjected to electrospray injection into the mass spectrometer and analyzed for their mass/charge ratio (m/z value) (B). Selected ions were collected in the ion trap. These parent ions were cracked by collision ion dissociation to produce a range of fragment sizes (C) that were compared to predicted peptide sequences in the human database using TurboSEQUEST™ (D).



Acknowledgments

The authors would like to acknowledge Pierce Norton and David Houck for sharing their extensive expertise in cell sorting. They also acknowledge the technical work of Rana Alsharif, Jitakshi De, and Martin Tapia.

References

1. Service, R. F. (2003) Public projects gear up to chart the protein landscape. *Science* **302(5649)**, 1316–1318
2. Link, A. J., Eng, J., Schieltz, D. M., et al. (1999) Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* **17(7)**, 676–682.
3. Shapiro H. (1994) Practical Flow Cytometry, Third Edition. Alan R. Liss, New York, NY.
4. Recktenwald, D. J., and Radbruch, A. (eds). (1998) Cell Separation Methods and Applications. Marcel Dekker, Inc., New York, NY.
5. Chelius, D., Huhmer, A. F., Shieh, C. H., et al. (2002) Analysis of the adenovirus type 5 proteome by liquid chromatography and tandem mass spectrometry methods. *J. Proteome Res.* **1(6)**, 501–13.
6. Chelius, D. and Bondarenko, P. V. (2002) Quantitative profiling of proteins in complex mixtures using liquid chromatography and mass spectrometry. *J. Proteome Res.* **1(4)**, 317–23.
7. Hovius, R., Lambrechts, H., Nicolay, K., and Kruijff, B. (1990) Improved methods to isolate and subfractionate rat liver mitochondria. Lipid composition of the inner and outer membrane. *Biochim. Biophys. Acta* **1021**, 217–226.
8. Reinhart, P. H., Taylor, W. M., and Bygrave, F. L. (1982) A procedure for the rapid preparation of mitochondria from rat liver. *Biochem. J.* **204**, 731–735.

Purification of Nucleoli From Lymphoma Cells and Solubilization of Nucleolar Proteins for 2-DE Separation

Régis Dieckmann, Yohann Couté, Denis Hochstrasser, Jean-Jacques Diaz, and Jean-Charles Sanchez

1. Introduction

Differential screening of entire cell proteomes by two-dimensional gel electrophoresis (2-DE) often leads to the identification of several differentially expressed but functionally unrelated target proteins. Hence, this approach often fails to give a comprehensive picture of the underlying biological mechanism. Thus, the level of complexity of the entire cell proteome needs to be reduced (1). Cell fractionation is both a way of reducing complexity and a way to target a specific cellular compartment. The purification of nucleoli from three Hodgkin nonadherent lymphoma cell lines is presented here. It is based on an original method from Muramatsu and Onishi (2) and has been adapted from a protocol developed for HeLa cells (3).

Because apoptosis or necrosis has a dramatic effect on nucleolar protein composition, nucleoli should be purified from exponentially growing cells (**Subheading 3.1.1.**) and cell viability should be checked before purification (**Subheading 3.1.2.**). The first step of the purification procedure is the preparation of intact nuclei free of cytoplasmic contamination. For this, cells are incubated in a hypotonic buffer before the addition of a detergent that destabilizes plasma membrane. Nuclei are then released by homogenization in a Dounce before being purified through a sucrose cushion. Controlled sonication of purified nuclei releases nucleoli, which are purified by centrifugation through a second sucrose cushion (**Subheading 3.1.3.**). A critical aspect of this procedure is the concentration of divalent cations in the hypotonic buffer, because it influences the efficiency of the nuclei purification. Thus, an optimal concentration has to be determined for each cell type.

The purity of nucleoli is high, as demonstrated by Western blot and transmission electron microscopy analyses, and therefore well suited for proteome analysis. However, in order to determine an efficient way to prepare nucleolar proteins for 2-DE separation, different techniques have been tested. The best results are obtained when purified nucleoli are resuspended in the recently described TFE buffer (4). TFE buffer contains 2,2,2-trifluoroethanol, a co-solvent used to stabilize secondary structures of proteins analyzed by nuclear magnetic resonance. The use of this co-solvent in 2-DE solubilization buffers improves resolution for all samples analyzed, including membrane fractions.

From: *The Proteomics Protocols Handbook*
Edited by: J. M. Walker © Humana Press Inc., Totowa, NJ

2. Materials

1. Hodgkin cell lines L-428, L-540, or KM-H2 (DSMZ).
2. Cell-culture medium: RPMI 1640 with 2 mM glutamine, 100 U/mL penicillin, 100 µg/mL streptomycin, fetal bovine serum 10% (L-428 and KM-H2) or 20% (L-540).
3. PBS, pH 7.4, without Ca²⁺ or Mg²⁺.
4. Trypan blue stain, 0.4%, possibly carcinogenic.
5. NP-40.
6. Saccharose, 1 M, sterile filtered.
7. Reticulocyte standard buffer (RSB): 0.01 M Tris-HCl (pH 7.4), 0.01 M NaCl, x mM MgCl₂ (concentration varying according to cell type) (see Note 1).
8. Saccharose, 0.25 M (store at 4°C): 0.25 M saccharose, 10 mM MgCl₂.
9. Saccharose, 0.34 M (store at 4°C): 0.34 M saccharose, 0.05 mM MgCl₂.
10. Saccharose, 0.88 M (store at 4°C): 0.88 M saccharose, 0.05 mM MgCl₂.
11. Dounce homogenizer with a small-clearance pestle (0.4 mm).
12. Light microscope with a minimum magnification of ×40.
13. Tip sonicator.
14. TFE buffer: 5 M urea (Merck), 2 M thiourea (Sigma, toxic), 4% CHAPS (Apollo), 65 mM 1,4-dithioerythritol (DTE) (Merck), 2% ampholytes e.g. Resolyte Electran 4-8 (BDH), 50% (v/v) 2,2,2-trifluoroethanol (TFE) (Fluka).
15. NL 3.0–10.0 immobilized pH gradient (IPG) strips (Pharmacia).
16. Immobiline DryStrip Reswelling tray (Pharmacia).
17. Paraffin.

3. Methods

The methods described below outline (1) the purification of nucleoli from Hodgkin lymphoma cells; (2) the validation of sample purity; and (3) the extraction of the proteins from purified nucleoli for 2-DE separation.

3.1. Purification of Nucleoli From Hodgkin Lymphoma Cells

3.1.1. Cell Culture and Harvesting

1. The cells are resuspended in fresh medium at 0.5–0.7 × 10⁶ cells/mL between 4 and 12 h before harvesting.
2. Just before harvesting, the cells are counted.
3. The cell viability is estimated with a trypan blue staining as described in Subheading 3.1.2.
4. Pellet up to 300 × 10⁶ cells (see Note 2) by centrifugation at 272g for 5 min.
5. The cells are then washed three times with cold PBS and gathered in a conical tube. A pellet of intact cells can be saved and frozen for further analysis.

3.1.2. Trypan Blue Staining

1. Take a cell aliquot of 1 mL.
2. Pellet the cell at 272g for 5 min.
3. Resuspend the cells in 750 µL PBS.
4. Add 250 µL trypan blue 0.4% in PBS.
5. Incubate for 5 min at room temperature.
6. Count the number of dead blue cells versus the total number of cells to evaluate the viability of the cell culture.

3.1.3. Purification of Nucleoli

1. All buffers used in this chapter as well as samples and centrifuge are refrigerated at 4°C or kept on ice.

2. Pellet the cells.
3. Add 15 cell volumes (cvol) of RSB and incubate 45 min on ice (see **Note 1** and **Fig. 1A**).
4. Add NP-40 to a final concentration of 0.3%.
5. Homogenize the cell suspension and transfer it in a Dounce homogenizer on ice.
6. Begin with three strokes of Dounce pestle. Check an aliquot under light microscope. The intact nuclei must be free of surrounding cytoplasm. If not, homogenize further. Normally, five to six strokes are sufficient. Avoid doing too many strokes, otherwise nuclei will break and aggregate. Nucleoli appear as bright white balls within nuclei (**Fig. 1B**).
7. Centrifuge the suspension 5 min at 1200g at 4°C. Collect the supernatant (cytoplasmic fraction) and freeze it for further analysis.
8. Resuspend the nuclear pellet in 10 cvol of the 0.25 M sucrose solution. Homogenize to break any nuclear aggregate.
9. Underlay a 0.88 M sucrose cushion (10 cvol) under the resuspended nuclear pellet and centrifuge 10 min at 1200g at 4°C.
10. Resuspend the purified nuclear pellet in 10 cvol of the 0.34 M sucrose solution. Check an aliquot under light microscope. Nuclei must be free of any cytoplasmic debris. They should be round and individualized. Nucleoli have generally lost their brightness (**Fig. 1C**). A pellet of purified nuclei can be saved and frozen for further analysis.
11. Sonicate the homogenized solution on ice until no intact nucleus is visible under microscope. Time and power of sonication depends on the cell type and on the number of cells. Typically, the sonication time will be 6×15 s for 20×10^6 cells, and is increased to 10×30 s when the power is set at 20 W. The power and time are increased to 10×30 s at 40 W for 300×10^6 cells.
12. Underlay 10 cvol of the 0.88 M sucrose solution under the sonicate and centrifuge 20 min at 2000g at 4°C.
13. Collect the supernatant (post-nuclear fraction) and freeze it for further analysis. Resuspend the nucleolar pellet in a small volume of 0.34 M sucrose buffer. Check an aliquot under microscope. No intact nucleus should be seen (**Fig. 1D**).

3.2. Validation of the Nucleolar Enrichment

Observations under light microscope are immediate and essential checkpoints throughout purification procedure. A quick validation is then made by analyzing the different collected fractions on a one-dimensional sodium dodecyl sulfate (SDS)-polyacrylamide gel electrophoresis (PAGE) (**Fig. 2**). The protein load is normalized according to the number of nucleoli, nuclei, and cells. The protein pattern of the various cellular fractions is clearly different. The nuclear fraction is enriched in histones and nucleolar fraction in nucleophosmin (**Fig. 2**). The enrichment in nucleolar proteins should also be monitored by Western blotting (**Fig. 3** and **Note 3**). Electron microscopy can be also used to demonstrate that the ultrastructure of nucleoli is conserved after isolation.

3.3. Preparation of Purified Nucleoli for 2-DE Separation

1. Resuspend directly a nucleolar pellet corresponding to 25×10^6 starting cells in 450 μ L TFE buffer (see **Note 4**). This amount of nucleoli corresponds to 50–100 μ g of proteins according to Bradford protein assay.
2. NL 3-10 IPG strip (18 cm) are rehydrated by in-gel sample rehydration in a rehydration strip tray for 6 h or overnight at room temperature and covered with paraffin oil. First dimension is run at 50 kVh. The proteins are then transferred on a 12.5% T polyacrylamide gel using an agarose layer. Second dimension is run at 40 mA per gel. The gels are silver stained with a MS non-compatible sensitive method described on the web site

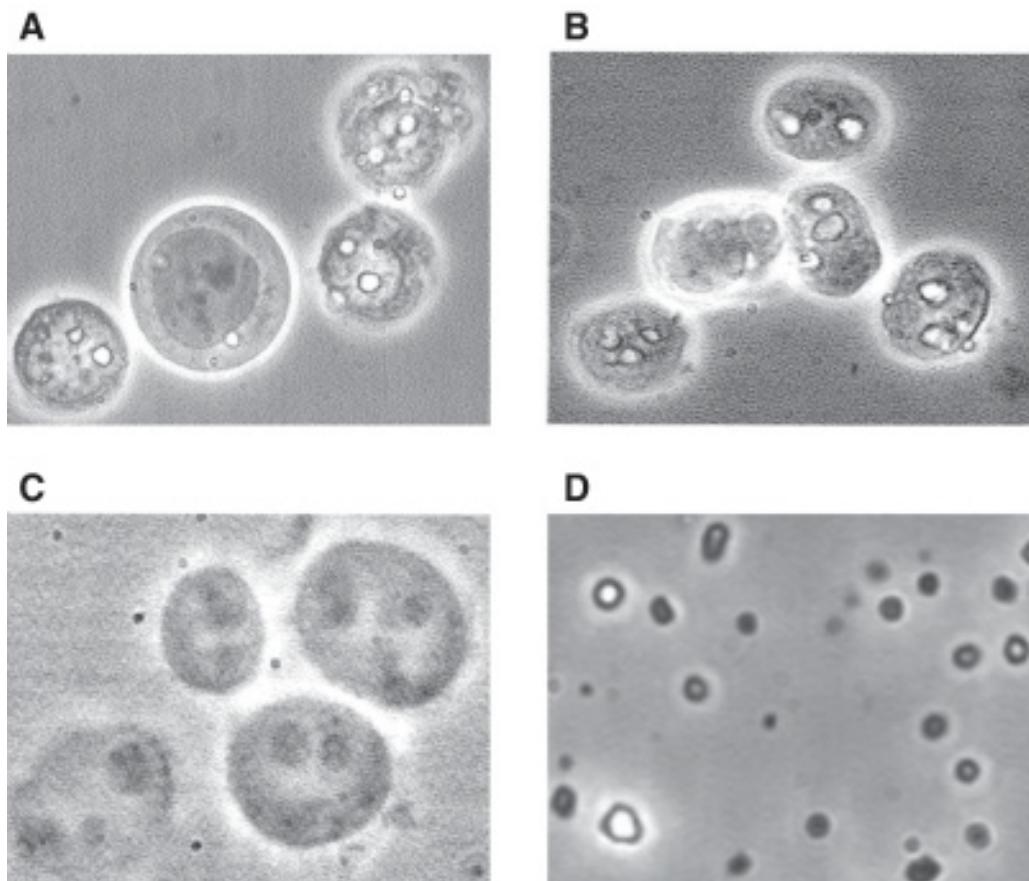


Fig. 1. Light microscopy images of the different purification steps of nucleoli of KM-H2 cells. (A) Cells after 45 min incubation in the hypotonic buffer RSB. (B) Broken cells releasing distorted nuclei after Dounce homogenization. (C) Purified nuclei. (D) Purified nucleoli.

<http://us.expasy.org/ch2d/protocols/>. Other silver staining, such as the one described in Chapter 18 can be used. The obtained 2DE gel is shown in **Fig. 4**.

4. Notes

1. The divalent cations concentration (Mg^{2+}) reinforce the nuclear membrane and the nucleolar structure. Not enough Mg^{2+} leads to the breakage of the structures, i.e., nuclear aggregation and lysis or nucleolar disintegration, and hence poor yield. With a too high Mg^{2+} concentration, nuclei cannot be fully broken by sonication and hence contaminate nucleolar preparation. The Mg^{2+} concentration is critical and should be adjusted for each type of cells. KM-H2 and L-540 cell lines are purified with 3 mM $MgCl_2$ RSB buffer, and L-428 cell line with 4 mM $MgCl_2$, respectively.
2. Below 10×10^6 cells, the technique is not really efficient. Above 300×10^6 cells, the quality of the preparation begins to be randomized. Especially, the nuclear breakage by sonication is not efficient in terms of yield over quality of the preparation.
3. Western blot allows screening for a specific location marker throughout the different sub-cellular fractions. The sensitivity of this technique gives an in-depth view of the cyto-

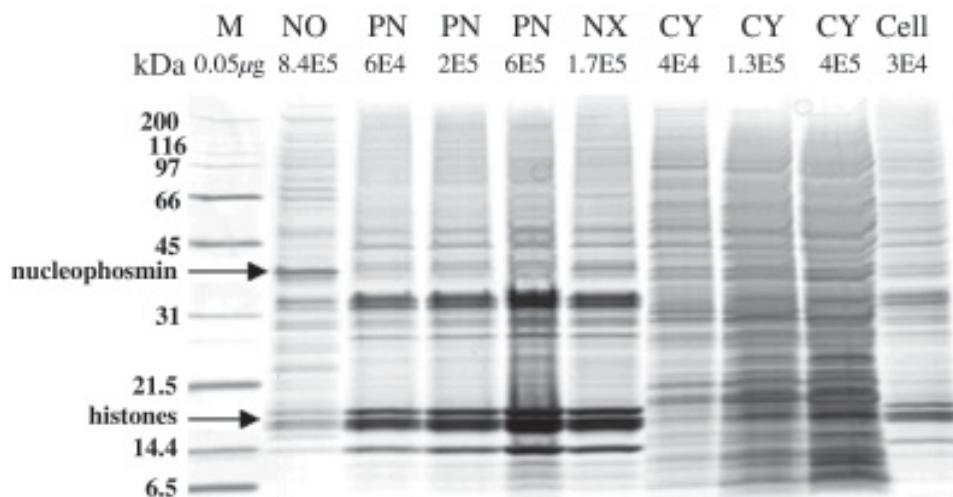


Fig. 2. Silver-stained sodium dodecyl sulfate-polyacrylamide gel electrophoresis gel of the fractions collected during the purification. The different fractions are the entire cell lysate (Cell), the cytoplasmic fraction (CY), purified nuclei (NX), post-nuclear fraction (PN), and purified nucleoli (NO). The markers (M) are the BioRad Broad Range markers. Quantities loaded on the gel are indicated as an equivalent of the initial cell count.

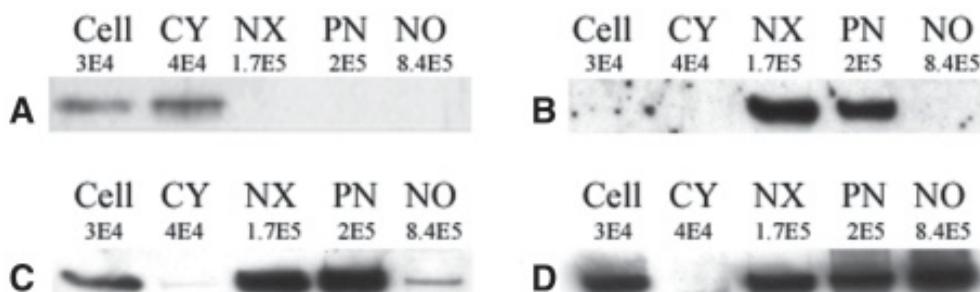


Fig. 3. Western blot analyses of the different fractions collected during the purification. The different fractions are the entire cell lysate (Cell), the cytoplasmic fraction (CY), purified nuclei (NX), post-nuclear fraction (PN), and purified nucleoli (NO). Quantities loaded on the gel are indicated as an equivalent of the initial cell count. Cellular fractions were separated by sodium dodecyl sulfate-polyacrylamide gel electrophoresis and transferred onto a polyvinylidene difluoride membrane essentially as described in (9). Proteins were revealed using an anticalreticulin antibody (A), an antinucleoporin p62 goat polyclonal antibody (B), an antilamin B goat polyclonal antibody (C), and an antinucleolin monoclonal antibody (D). Anti-calreticulin antibody was a kind gift from K. H. Krause, University of Geneva, Geneva. Other antibodies were purchased from Santa Cruz Biotechnology, Santa Cruz, CA. All primary antibodies were used with 1/1000 dilutions.

plasm-nucleus and nucleus-nucleoli separation steps as long as the location markers are well chosen. The cytoplasm-nucleus separation is visualized by detecting, for example, calreticulin, a specific marker of the endoplasmic reticulum and phagosomes. According

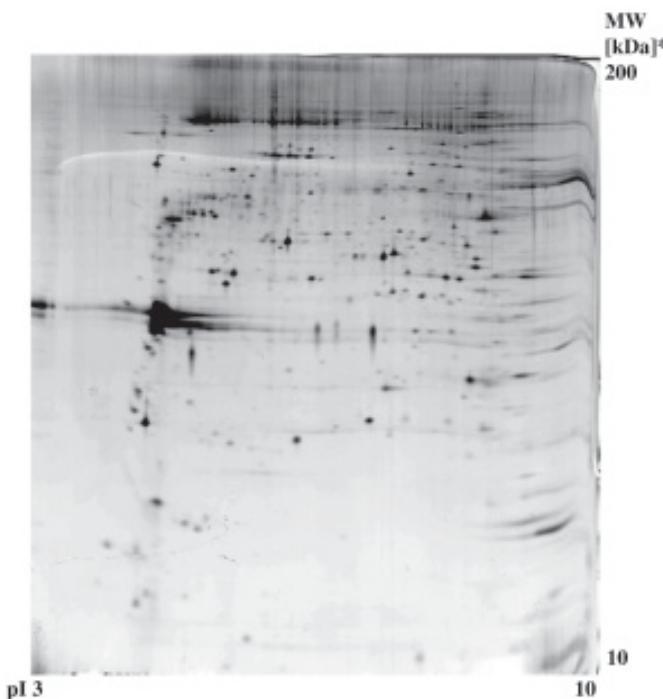


Fig. 4. Silver-stained two-dimensional electrophoresis gel of nucleoli purified from 25×10^6 KM-H2 Hodgkin lymphoma cells.

to the actual hypothesis that proteins are moving freely in the nucleus and are preferentially found in only one nuclear domain, nucleoli can contain minimal amounts of proteins from other nuclear compartments. In order to follow the separation of nuclei from nucleoli, two well-suited proteins are the nucleoporin p62 and lamin B. Nucleoporin p62 is located at the nuclear pore. Lamin B is a major component of the nuclear matrix. Therefore, these proteins are used as nuclear markers. The enrichment in specific nucleolar proteins such as nucleophosmin or nucleolin can be monitored. Nucleophosmin is seen on SDS-PAGE directly. Therefore, nucleolin is preferentially used in Western blot. As nuclear proteins are diffusible, nucleolin should be enriched in the nucleolus but can give a signal in the nucleoplasm. Agyrophilic staining could also be used on blots to monitor nucleophosmin and nucleolin.

4. Different protocols have been tested on NL3-10 IPG strips but none gave a better result than the one described above. In-gel and cup loading gave comparable resolution. TFE buffer showed better resolution than the urea buffer (<http://us.expasy.org/ch2d/protocols/>) or the urea-thiourea buffer described by Rabilloud (5). As described by Muramatsu and Busch, nucleoli purified by this method contain approx 10% RNA, 7% DNA, and 83% protein (6). Thus, several 2DE-compatible DNA/RNA removal protocols were tested: (1) endonuclease digestion; (2) acetic acid extraction (3); (3) spermine treatment (7); and (4) spermine + polyethyleneimine (PEI) treatment (8). Resolution was not improved by endonuclease digestion. Acetic acid extraction resulted in a general protein loss but no change in the protein profile. Spermine and PEI treatments resulted in a dramatic shift of all proteins to the basic side due to the absence of precipitation of spermine or PEI with DNA and resulting removal of nucleic acids.

References

1. Dreger, M. (2003) Subcellular proteomics. *Mass Spectrom. Rev.* **22**, 27–56.
2. Muramatsu, M. and Onishi, T. (1978) Isolation and purification of nucleoli and nucleolar chromatin from mammalian cells. *Methods Cell. Biol.* **17**, 141–161.
3. Scherl, A., Coute, Y., Deon, C., et al. (2002) Functional proteomic analysis of human nucleolus. *Mol. Biol. Cell* **13**, 4100–4109.
4. Deshusses, J. M., Burgess, J. A., Scherl, A., Wenger, Y., Walter, N., Converset, V., Paesano, S., Corthals, G. L., Hochstrasser, D. H., and Sanchez, J. C. (2003) Exploitation of specific properties of trifluoroethanol for extraction and separation of membrane proteins. *Proteomics* **3**, 1418–1424.
5. Rabilloud, T., Adessi, C., Giraudel, A., and Lunardi, J. (1997) Improvement of the solubilization of proteins in two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis* **18**, 307–316.
6. Muramatsu, M. and Busch, H. (1982) Isolation, composition, and function of nucleoli of tumors and other tissues. *Methods Cancer Res.* **2**, 303–359.
7. Jung, E., Hoogland, C., Chiappe, D., Sanchez, J. C., and Hochstrasser, D. F. (2000) The establishment of a human liver nuclei two-dimensional electrophoresis reference map. *Electrophoresis* **16**, 3483–3487.
8. Burgess, R. R. (1991) Use of polyethyleneimine in purification of DNA-binding proteins. *Methods Enzymol.* **208**, 3–10.
9. Burkhard, P. R., Sanchez, J. C., Landis, T., and Hochstrasser, D. F. (2001) CSF detection of the 14-3-3 protein in unselected patients with dementia. *Neurology* **56**, 1528–1533.

Prefractionation of Complex Protein Mixture for 2-D PAGE Using Reversed-Phase Liquid Chromatography

Volker Badock and Albrecht Otto

1. Introduction

In the last years, great progress has been achieved by development of multidimensional liquid chromatographic separation methods for separation of protein complexes after enzymatic digestion and subsequent identification of the proteins with mass spectrometric techniques (1–4). However, high-resolution two-dimensional gel electrophoresis (2-DE) is until now the only technique that allows separation of thousands of proteins in a gel (5–8). It has been estimated that the proteome of a given cell contains at least 10,000–30,000 different proteins, but only 2000–10,000 proteins can be visualized on a silver-stained 2-DE gel, depending on the 2-DE method applied (8), and only a proportion of them are present at levels sufficient for mass-spectrometric identification. The introduction of two-dimensional differential gel electrophoresis (2-D DIGE) by Ünlü et al. (9) allows running two or more differently fluorescent-labeled protein samples on a single 2-DE gel. Moreover, the dynamic range of the covalent fluorescent protein staining is higher than silver staining. This increases the reproducibility and saves material and time of the experiment.

However, the most decisive drawback lies in the deficiency for visualizing low-copy-number gene products in the presence of highly abundant proteins. Many disease-associated proteins or drug targets are low-abundance proteins that are present in femtomole amounts or less and are therefore difficult to identify. Increasing the quantity of protein loaded onto the gel might be a possible way to increase sensitivity. The introduction of immobilized pH gradient (IPG) isoelectric focusing allows a loading capacity of several milligrams of total protein on a single gel (10,11). However, a higher protein amount loaded on the gel results in less resolved 2-DE patterns, as spots of very abundant proteins overlay the spots of less abundant proteins. Therefore, enrichment of low-abundance proteins should be combined with a suitable prefractionation method.

There are in principle two alternative approaches to enrich and prefractionate low-abundance proteins when using cell lines or tissue: subcellular fractionation and protein prefractionation. The preparation of subcellular organelles is based commonly on size or density differences. Another procedure uses free-flow isoelectric focusing (FF-IEF), a technique that exploits the charge differences of organelles (12). FF-IEF can also be used as a protein prefractionation technique. By this means, up to 9 mg protein/h can be separated continuously, resulting in 80 fractions. However, the hydroxypropyl-

methylcellulose (HPMC) in the FF-IEF separation solution has to be removed, because it is detrimental to the first dimension of 2-DE (13). Several other approaches are based on the principle of affinity chromatography, such as heparin affinity (14,15), hydroxyapatite affinity (16), dye ligand (17), or immobilized metal affinity chromatography (IMAC) (18). These methods have in common that they enrich only proteins with an affinity to the chromatographic material applied. Therefore, these approaches are not appropriate for a global protein analysis. Molloy et al. (19) exploit the different solubility of cell lysate proteins in buffer solutions with increasing solubilizing ability. Using a three-step extraction protocol, they obtained three protein mixtures, which were applied to 2-DE.

Görg et al. (20) have developed a prefractionation procedure based on flat-bed IEF in Sephadex gels containing urea, thiourea, detergents, DTT, and carrier ampholytes. Gel fractions alongside the pH gradient are removed with a spatula and directly applied onto the surface of the corresponding narrow-range IPG strips as the first dimension of 2-DE. Zuo and Speicher (21) have done prefractionation using microscale solution isoelectrofocusing prior to narrow pH range two-dimensional electrophoresis. Butt et al. (22) have described the application of anion-exchange chromatography prior to 2-DE to simplify the proteome and enrich proteins up to the 13-fold. However, this method has the restriction that it fractionates only soluble proteins, since urea is not used in the extraction buffer. Hydrophobic interaction chromatography was successfully used to enrich low-copy-number gene products of the soluble protein fraction of *Haemophilus influenzae*, and eluted protein fractions were analyzed by 2-DE and matrix-assisted laser desorption/ionization (MALDI) mass spectrometry (23).

We have introduced a simple approach for prefractionating of complex protein samples such as whole lysates of cells and tissue prior to high-resolution 2-DE (24). Reversed phase high-performance liquid chromatography (RP-HPLC) has been successfully applied for this purpose using a multistep elution gradient of increasing acetonitrile concentrations. Van den Bergh et al. (25) have successfully used this approach in combination with 2-D DIGE for enrichment of lowly expressed proteins.

In this chapter we give a detailed protocol of the prefractionation of a crude lysate of human breast epithelial cell line HBL-100 containing 7 M urea, 2 M thiourea, and other reagents, that was loaded directly onto a reversed phase column.

The great benefit of this simple approach is that in principle every protein of a complex mixture is accessible for enrichment, in contrast to other methods, which isolate certain proteins due to their affinity to a matrix. This method is, therefore, ideal for a global protein analysis, and, due to its protein recovery and reproducibility, is appropriate for studying differential protein expression by means of subtractive analysis. Furthermore, the reproducibility allows the pooling of any number of consecutive HPLC runs with the aim of enriching low-abundance proteins. By relatively simple modification of the elution gradient, the proteome can be divided up to any number of 2-D gels. Even single fractions of the chromatographic run can be selected in order to isolate a protein of interest or to perform 2-DE with gels of small size or narrow pH gradient. It is conceivable to combine this technique with other chromatographic systems such as ion-exchange chromatography, a combination which is effectively applied in liquid chromatography (LC)-mass spectrometry (MS) (two-dimensional LC-MS).

2. Materials

1. Ultracentrifuge Optima TL 100, Rotor TLA 100.3 (Beckman).
2. Vacuum centrifuge: SpeedVac® (Savant).
3. Stirrer (such as Vortex).
4. A standard system for reversed phase HPLC is needed. We used a Shimadzu LC-6 system using a T-junction instead of the static mixing chamber.
5. Fraction collector (an LKB 7000 was used).
6. Vydac C4 reversed phase column (150 × 2.1 mm, 5 µm, 300 Å) with Vydac C4 guard column (Vydac, Hesperia, CA).
7. HPLC solvent A: 0.1% trifluoroacetic acid (TFA) in water.
8. HPLC solvent B: 0.1% TFA in acetonitrile (Merck, gradient grade).
9. CHAPS (Fluka).
10. Dithiotreitol (DTT, Merck, 1.4 M stock solution in water).
11. Pepstatin A (Boehringer, 100 µg/mL stock solution).
12. PMSF (Sigma, 200 mM stock solution in ethanol).
13. Thiourea (ultra pure) (Merck).
14. Trifluoroacetic acid (TFA, Fluka).
15. Tributylphosphine (TBP, Sigma).
16. Servalyt pH 2.0–4.0 (Serva).
17. Urea (ultra pure) (BioRad).
18. Sample preparation buffer: 80 mM Tris-HCl (pH 7.1), 2 mM ethylenediaminetetraacetic acid (EDTA), 2 mM KCl, 2 mM benzamidine, and 4.2 mM leupeptin.
19. 2-DE buffer: 7 M urea, 2 M thiourea, 4% CHAPS, 40 mM Tris-HCl (pH 7.1), and 5 mM TBP.

2.1. Cell Culture

The human breast epithelial cell line HBL-100 was cultured in DMEM HAMs E12 (Biochrom, Berlin, Germany) supplemented with 5% fetal calf serum (Life Technologies, Karlsruhe, Germany), 10 µg/mL insulin (Biochrom), 10 µg/mL transferrin (Life Technologies), 1.8 µg/mL hydrocortisol, 100 U/mL penicillin, and 100 µg streptomycin.

3. Methods

3.1. Sample Preparation

1. Harvest the HBL-100 cells by centrifugation at 2000g at 4°C for 10 min, wash twice in PBS and once in sample preparation buffer, spin down again to a “wet pellet,” and determine the weight (W) of the pellet (mg). Freeze pellet at –80°C until use.
2. Lyse the cell pellets by addition of $0.841 \times W$ mg urea and $0.304 \times W$ thiourea, corresponding to 7 M urea and 2 M thiourea, respectively. Resulting volume (V) was assumed to be $2 \times W$.
3. Add 4% CHAPS and reduce the disulfide bonds by addition of $0.05 \times V$ DTT (1.4 M stock solution, 70 mM final concentration)
4. Add 2.5% carrier ampholytes (Servalyte, pH 2.0–4.0) and additional protease inhibitors PMSF (1 mM final concentration) and pepstatin A (1.4 µM final concentration) and stir gently for 30 min at room temperature.
5. Clear the lysate by centrifugation at 200,000g for 20 min and freeze supernatant at –80°C until further use (see Note 1).

3.2. Reversed Phase High-Performance Liquid Chromatography

Liquid chromatography was performed with a Shimadzu LC-6 system using a Vydac C4 reversed phase column (150 × 2.1 mm, 5 µm, 300 Å) connected to a Vydac C4

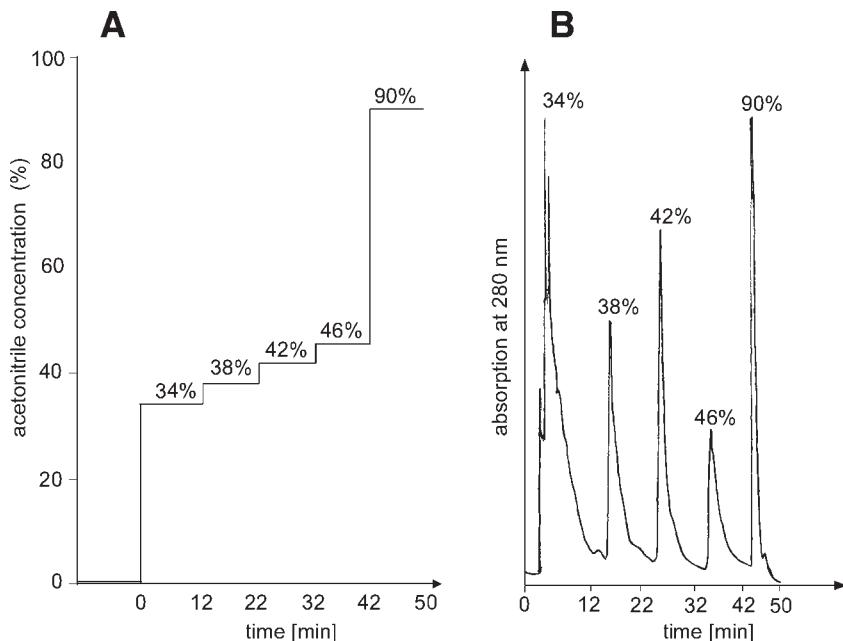


Fig. 1. (A) Scheme of the five-step gradient with increasing concentrations of solvent B (0.1% TFA in acetonitrile). The duration of each step is indicated on the x axis. (B) Corresponding high-performance liquid chromatography chromatogram monitored at a wavelength of 280 nm.

guard column at a total flow rate of 0.5 mL/min (**Note 2**). Solvent A: 0.1% TFA in water; solvent B: 0.1% TFA in acetonitrile. The chromatograms were recorded at 280 nm.

1. Equilibrate the column with solvent A at a flow rate of 0.5 mL/min for 10 min.
2. Inject up to 1 mg of the sample (diluted 1:1 with solvent A) and wash the column carefully with solvent A for 20 min (stable baseline).
3. Start the step gradient that consisted of five steps with increasing concentrations of solvent B in solvent A. Step 1: 34%, 12 min; step 2: 38%, 10 min; step 3: 42%, 10 min; **step 4: 46%, 10 min; step 5: 90%, 5 min** (*see Fig. 1*).
4. Collect fractions automatically every 2 min.
5. Dry all fractions of each step in a SpeedVac concentrator (*see Note 2*).
6. Dissolve dried fractions in 20 μ L SDS gel loading buffer for one-dimensional gel electrophoresis (*see Fig. 2* and **Note 4**)
7. Dissolve dried fractions in 2-DE buffer and pool fractions of each step for 2-D PAGE (*see Note 3*). The total volume of the pooled fractions should not exceed the loading capacity of a rod gel of the first dimension: 10–12 μ L for an analytical gel and 40–50 μ L for a preparative gel.

4. Notes

RP-HPLC

1. For protein quantification, the amino acid composition was determined by vapor-phase hydrolysis with 6 N HCl and 7% thioglycolic acid as an additive for 24 h at 110°C. Protein concentrations of lysates of human breast cells ranged between 5 and 8 μ g/ μ L.

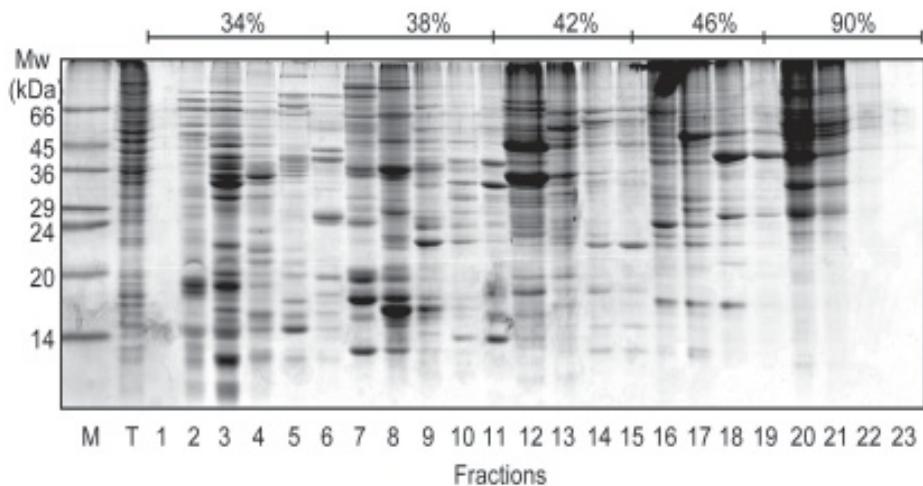


Fig. 2. Sodium dodecyl sulfate (SDS)-polyacrylamide gel electrophoresis (PAGE) analysis of HBL-100 cell lysate eluted from a reversed phase C4 (150 × 2.1 mm, 5 µm, 300 Å) column. The soluble protein fraction of HBL-100 cells was applied onto the column and eluted in a five-step gradient (see Fig. 1). The freeze-dried samples were dissolved in 20 µL SDS sample buffer and loaded onto a 15% SDS gel. The gel was stained with Coomassie blue. The numbers correspond to the numbers of the high-performance liquid chromatography fractions. Fractions that belong to the same step are indicated with the concentration of buffer B. T, total cell lysate (15% of amount loaded onto the column). M, protein size marker (see also Note 4).

2. We used Vydac C4 reversed phase material in our experiment with good success. But, in principle any type of RP material that is suitable for protein separation can be used.
3. Problems in redissolving dried protein samples were avoided by freezing the fractions at -80°C before freeze-drying in a SpeedVac. During this process the vial should be locked and the cap has to be perforated to avoid a loss of freeze-dried protein. The fractions have to be kept frozen during the whole process. For this reason it is essential to apply a good vacuum to the SpeedVac (oil-filled rotary pump).

SDS-PAGE

4. The separation performance of the column was examined by analyzing the collected fractions by SDS-PAGE (Fig. 2). Even very intense bands of highly abundant proteins were eluted mainly in one fraction (see for example Fig. 2, fraction 12); in principle, it was possible to select an individual fraction of interest and to separate the proteins subsequently by 2-DE.

Two-Dimensional Gel Electrophoresis (2-DE)

5. 2DE was performed by the combination of isoelectric focusing (first dimension) and sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) (second dimension). Because it is not the subject of this chapter, we have omitted an extensive description of the preparation of two-dimensional gels. Instead, we refer to two excellent and detailed publications by Klose (7,8) (see also Chapter 13). All gel solutions (ready made) and 2-DE equipment were purchased from WITA GmbH (Teltow, Germany). IEF was performed in rod gels (inner diameter, analytical gels, 0.09 cm, and micropreparative gels, 0.15 cm). Stained gels were scanned in a Hell/Linotype scanner. Gel images were pro-

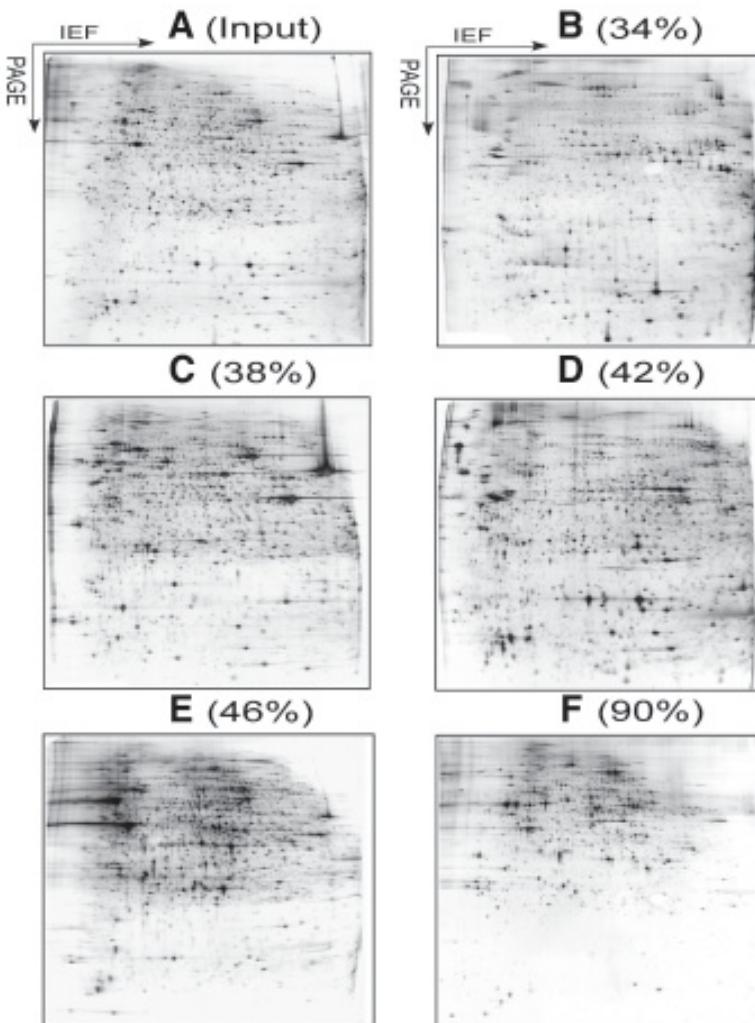


Fig. 3. Silver-stained two-dimensional gel electrophoresis patterns of whole HBL-100 cell lysate (A) (input), and of the pools 34% (B), 38% (C), 42% (D), 46% (E), and 90% (F) of fractions collected after separation by reverse phase high-performance liquid chromatography. 35 µg of total protein were loaded onto gel A while 245 µg of total protein were loaded onto the column (*see also Notes 5 and 6*).

cessed using Adobe Photoshop software. Spot number calculations and gel image calibration were performed using Phoretix v5.0 software.

6. A comparison of the 2-DE images of completely non-concentrated cell lysate and the proteins eluted from the column is shown in **Fig. 3**. The protein amount loaded onto the 2-D gel A in **Fig. 3** was 15% of the amount applied to the RP column resulting in 2-D gels B–F. The comparison of 2-D gels B–F confirms the general tendency of high-molecular-weight proteins to be more hydrophobic than low-molecular-weight proteins or peptides. Furthermore, it was observed that only very intense spots (deriving from so-called house-keeping proteins) overlap between adjacent steps of the gradient. The reproducibility was verified by comparison of the 2-DE gels of different HPLC runs from the same sample

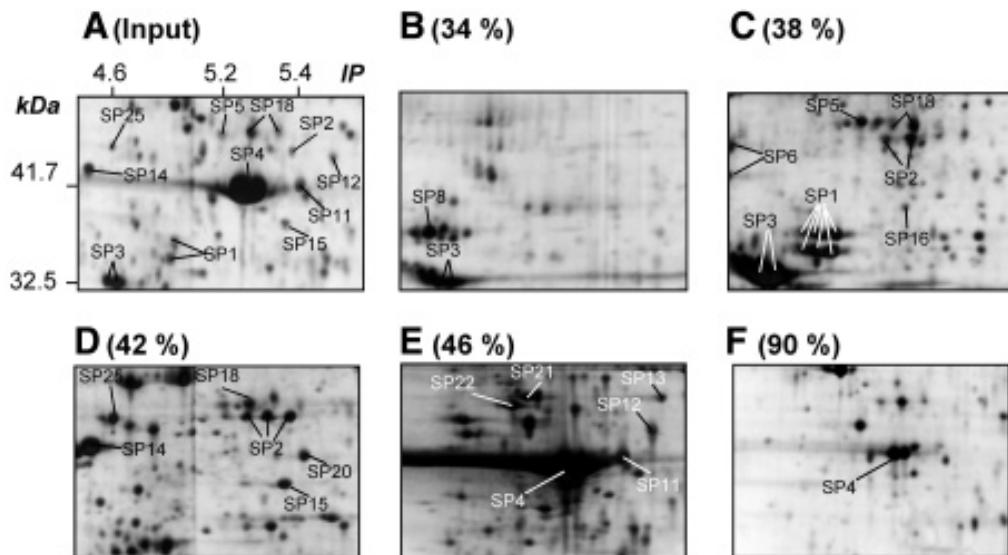


Fig. 4. Silver-stained sections of the two-dimensional electrophoresis gel images of the HBL-100 cell lysate (input) (A), and of the pools 34% (B), 38% (C), 42% (D), 46% (E), and 90% (F) of fractions collected from the RP column (see Note 7). The proteins were identified either by matrix-assisted laser desorption/ionization mass spectrometry (MALDI MS) or by electrospray ionization-MS/MS.

Table 1
Detected Number of Spots
in 2-D Electrophoresis
Gels A–F of Fig. 3

Gel	Number of spots
A (Input)	2700
B (34%)	1200
C (38%)	1750
D (42%)	1800
E (46%)	1340
F (90%)	900

Number of spots was computed by Phoretix 2DE analysis software.

(data not shown). We obtained the same 2-DE pattern from run to run using the same reversed phase column repeatedly. This good reproducibility allows pooling of several consecutive runs of the same sample, resulting in a higher concentration factor. Due to the enrichment effect, the protein patterns of 2-D gels B–F are nearly as complex as the protein pattern of the input (Fig. 4A). This was confirmed by the computer-assisted calculation of the spot numbers of 2-D gels A–F (see Table 1). If the same protein amounts were loaded onto a single 2-D gel as were loaded onto the column, this would result in an extremely complex protein pattern with strong horizontal spot streaking. Since very com-

plex protein patterns of 3000 and more proteins are difficult to analyze by software programs such as Melanie, Phoretix, or PDQuest, prefractionation by RP chromatography simplifies the proteome, thus making the evaluation easier.

7. **Figure 4** shows a few partial 2-DE images of the five pools (B–F) in more detail. A comparison with 2-DE gels of the corresponding part of the 2-DE gel of the unseparated extract (input, A) shows that several protein spots were strongly enriched. For example, nucleophosmin (SP3, **Fig. 4C**), hnRNP C1/C2 (SP1, **Fig. 4C**), and nuclear distribution gene C (SP2, **Fig. 4D**) are represented as strong protein spots, which are seen in the 2-DE gel of the input gel only as relatively weak spots. The comparison also demonstrates the minor overlap of protein spots between adjacent gradient steps. The very intensive spot for actin (SP4), for example, is particularly concentrated in the 46% pool, and only small amounts of it can be detected in the 90% pool. Furthermore, this fractionation procedure should open the possibility of identifying spots that are normally covered by intense spots of abundant proteins.

References

1. Link, A. J., Eng, J., Schieltz, O. M., et al. (1999) Direct analysis of protein complexes using mass spectrometry. *Nature Biotechnol.* **17**, 676–682.
2. Wolters, D. A., Washburn, M. P., and Yates, J. R. III. (2001) An automated multidimensional protein identification technology for shotgun proteomics. *Anal. Chem.* **73**, 5683–5690.
3. Han, D. K., Eng, J., Zhou, H., and Aebersold, R. (2001) Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. *Nature Biotechnol.* **19**, 946–951.
4. Aebersold, R. and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature* **422**, 198–207.
5. Klose, J. (1975) Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues. A novel approach to testing for induced point mutations in mammals. *Humangenetik* **26**, 231–243.
6. O'Farrell, P. H. (1975) High-resolution two-dimensional electrophoresis of proteins. *J. Biol. Chem.* **250**, 4007–4021.
7. Klose, J. and Kobalz, U. (1995) Two-dimensional electrophoresis of proteins: An updated protocol and implications for a functional analysis of the genome. *Electrophoresis* **16**, 1034–1059.
8. Klose, J. (1999) Large gel 2-D electrophoresis in 2-D proteome analysis protocols. In: Link, A. J. (ed), *Methods in Molecular Biology*, Humana Press, Inc., Totowa, NJ, Vol.112, pp 147–172.
9. Ünlü, M., Morgan, M. E., and Minden, J. S. (1997) Difference gel electrophoresis: a single gel method for detecting changes in protein extracts. *Electrophoresis* **18**, 2071–2077.
10. Corthals, G. L., Molloy, M. P., Herbert, B. R., Williams, K. L., and Gooley, A. A. (1997) Prefractionation of protein samples prior to two-dimensional electrophoresis. *Electrophoresis* **18**, 317–323.
11. Sanchez, J. C., Rouge, V., Piste, M., et al. (1997) Improved and simplified in-gel sample application using reswelling of dry immobilized pH gradients. *Electrophoresis* **18**, 324–327.
12. Weber, G. and Bocek, P. (1998) Recent developments in preparative free flow isoelectric focusing. *Electrophoresis* **19**, 1649–1653.
13. Burggraf, D., Weber, G., and Lottspeich, F. (1995) Free flow-isoelectric focusing of human cellular lysates as sample preparation for protein analysis. *Electrophoresis* **16**, 1010–1015.

14. Jungblut, P. and Klose, J. (1986) Composition and genetic variability of heparin-sepharose CL-6B protein fractions obtained from the solubilized proteins of mouse organs. *Biochem. Genet.* **24**, 925–939.
15. Karlsson, K., Cairns, N., Lubec, G., and Fountoulakis, M. (1999) Enrichment of human brain proteins by heparin chromatography. *Electrophoresis* **20**, 2970–2976.
16. Fountoulakis, M., Takacs, M. F., Berndt, P., Langen, H., and Takacs, B. (1999) Enrichment of low abundance proteins of *Escherichia coli* by hydroxyapatite chromatography. *Electrophoresis* **20**, 2181–2195.
17. Jungblut, P. and Klose, J. (1989) Dye ligand chromatography and two-dimensional electrophoresis of complex protein extracts from mouse tissue. *J. Chromatogr.* **482**, 125–132.
18. Jungblut, P., Baumeister, H., and Klose, J. (1993) Classification of mouse liver proteins by immobilized metal affinity chromatography and two-dimensional electrophoresis. *Electrophoresis* **14**, 638–643.
19. Molloy, M. P., Herbert, B. R., Walsh, B. J., et al. (1998) Extraction of membrane proteins by differential solubilization for separation using two-dimensional gel electrophoresis. *Electrophoresis* **19**, 837–844.
20. Görg, A., Boguth, G., Köpf, A., Reil, G., Parlar, H., and Weiss, W. (2002) Sample prefractionation with Sephadex isoelectric focusing prior to narrow pH range two-dimensional gels. *Proteomics* **2**, 1652–1657.
21. Zuo, X. and Speicher, D. W. (2002) Comprehensive analysis of complex proteomes using microscale solution isoelectrofocusing prior to narrow pH range two-dimensional electrophoresis. *Proteomics* **2**, 58–68.
22. Butt, A., Davison, M. D., Smith, G. J., et al. (2001) Chromatographic separations as a prelude to two-dimensional electrophoresis in proteomics analysis. *Proteomics* **1**, 42–53.
23. Fountoulakis, M., Takacs, M. F., and Takacs, B. (1999) Enrichment of low-copy-number gene products by hydrophobic interaction chromatography. *J. Chromatogr. A* **833**, 157–168.
24. Badock, V., Steinhusen, U., Bommert, K., and Otto, A. (2001) Prefractionation of protein samples for proteome analysis using reversed phase high-performance liquid chromatography. *Electrophoresis* **22**, 2856–2864.
25. Van den Bergh, G., Clerens, S., Vandesande, F., and Arckens, L. (2003) Reversed phase high-performance liquid chromatography prefractionation prior to two-dimensional difference gel electrophoresis and mass spectrometry identifies new differentially expressed proteins between striate cortex of kitten and adult cat. *Electrophoresis* **24**, 1471–1481.

Fractionation of Complex Proteomes by Microscale Solution Isoelectrofocusing Using ZOOM™ IEF Fractionators to Improve Protein Profiling

Xun Zuo, Ki-Boom Lee, and David W. Speicher

1. Introduction

All current methods for quantitatively comparing protein profiles, including two-dimensional polyacrylamide gel electrophoresis (2-D PAGE) as well as non-2-D gel methods such as isotope-coded affinity tag methods (ICAT) and other liquid chromatography (LC)/ tandem mass spectrometry (MS/MS) methods, have limited resolution and modest dynamic ranges for detection (1–5). As a result, these proteomic methods can analyze only small portions of very complex proteomes, such as cell or tissue extracts from higher eukaryotes. One strategy to increase the proportion of complex proteomes that can be quantitatively compared by existing protein-profiling methods is to fractionate these samples into a modest number of well resolved pools prior to downstream multi-dimensional protein profiling using 2-D gels or LC/LC-MS/MS methods. Unfortunately, most protein separation methods have relatively low resolving power and often do not separate noncovalent complexes. As a result, low- to moderate-resolution fractionation methods typically result in extensive and variable cross-contamination of many proteins among multiple fractions, which makes quantitative comparisons impractical. In contrast, several recently developed, high-resolution, liquid-phase isoelectric focusing (IEF) methods can fractionate proteomes into a modest number of well-separated fractions under denaturing conditions where specific and nonspecific protein-protein interactions are minimized (6–10).

A solution IEF method that is convenient, relatively simple, economical, and directly compatible with sample sizes commonly used for proteome experiments is microscale solution IEF (MicroSol-IEF), which is capable of efficiently separating microgram to milligram quantities of complex protein mixtures (6,11,12). Complex proteomes are fractionated based upon protein isoelectric points (pI) by utilizing protein-permeable acrylamide partitions that contain covalently bound immobilines at different pHs between tandem separation chambers. After isoelectric focusing, proteins in a given separation chamber will have pIs between the pHs defined by the two boundary membranes. Separation of complex proteomes such as mammalian cell extracts (11) and serum (13,14) results in well-resolved fractions that can usually be directly applied to immobilized pH gradient (IPG) gels or 1-D sodium dodecyl sulfate (SDS) gels without concentration. A commercial device for performing MicroSol IEF separations, the

From: *The Proteomics Protocols Handbook*
Edited by: J. M. Walker © Humana Press Inc., Totowa, NJ

ZOOM™ IEF Fractionator, is now available (Invitrogen Corp.) and is more user-friendly than prior homemade devices. So far, the most extensively utilized downstream analysis method for ZOOM IEF fractions is to employ a series of slightly overlapping narrow pH range 2-D gels (11,14). Compared with direct analysis of unfractionated samples on a series of narrow pH range 2-D gels, fractionation of complex proteomes using ZOOM IEF (1) conserves limited proteome samples; (2) results in higher resolution on narrow pH range gels with fewer artifacts; (3) allows for higher sample loads, which increases dynamic range and detection of lower-abundance proteins; and (4) enables detection of far more total proteins. In addition, ZOOM IEF prefractionation is compatible with most downstream protein profiling methods, including LC-MS/MS and LC/LC-MS/MS methods (15). This chapter describes the use of a ZOOM IEF Fractionator to subdivide complex proteomes using either commercial or custom-made partition membranes followed by comprehensive protein profiling using slightly overlapping narrow pH range 2-D gels as well as several non-2-D gel methods (12,15).

2. Materials

2.1. Casting Custom Acrylamide/Immobiline Partition Membranes at Desired pHs

1. Immobilines (Amersham Biosciences, Uppsala, Sweden).
2. Gel support film (Bio-Rad Laboratories, Hercules, CA).
3. Repel-Silane (Amersham Biosciences).
4. Lid from 1-mL Rainin pipet tip box.
5. 1.5-mm slab gel spacers (Bio-Rad Laboratories).
6. Hydrophilic porous polyethylene—0.67-mm thickness, medium pores, 45–90 μ m Dry Blend Surfactant (DBS). Available as large sheets (POREX, Leicester, MA), or as precut oval-shaped polyethylene disks that fit ZOOM IEF Fractionators (Invitrogen Corp.).
7. 30% T/10% C acrylamide/bisacrylamide: 27 g acrylamide, 3 g bisacrylamide (Bio-Rad Laboratories) in 100 mL final volume, store at 4°C for up to 1 mo.
8. 87% Glycerol (Amersham Biosciences).
9. 4% Ammonium persulfate: 4 g ammonium persulfate (Bio-Rad Laboratories) in 100 mL final volume; prepare immediately before use.
10. TEMED (BioRad Laboratories).
11. 24-well tissue-culture plates (Becton Dickinson Labware).
12. 12% Glycerol.
13. Parafilm or adhesive microtiter plate film.
14. Partition membrane storage solution: 10 mM phosphate buffer (pH 6.8), 0.4% acrylamide, 12% glycerol, and 2 mM sodium azide. Prepare on day of use.

2.2. Assembling the ZOOM IEF Fractionator

1. ZOOM IEF Fractionator (Invitrogen Corporation, Carlsbad, CA).
2. A series of ZOOM Disk partition membranes (Invitrogen Corp.) or custom-made membranes (**Subheading 3.2.**) at desired pHs.
3. 10X IEF anode buffer and 10X cathode buffer (Bio-Rad Laboratories) diluted to 1X and containing 8 M urea, 2 M thiourea (final concentrations).
4. Sample buffer: 8 M urea, 2 M thiourea, 4% CHAPS, 1% dithiothreitol (DTT), and 0.2–0.5% IPG buffer pH 3.0–10.0 L (carrier ampholytes).

5. 0.22 μ m Ultrafree-MC micro-filter (Millipore Corporation, Bedford, MA).
6. Power supply capable of operating at low currents (0.1–1 mA) and up to at least 1000 V.

3. Methods

3.1. Experimental Design and Alternative Setups of ZOOM IEF Fractionator

A major advantage of the ZOOM IEF Fractionator is its flexibility. A photograph of an assembled unit and the major components for the device are shown in **Fig. 1**. Separation conditions, pH ranges, protein loading amounts and volumes, number of fractions, and fraction volumes can be readily altered to meet requirements of various types of proteome samples and research aims (**Table 1**). For example, mammalian plasma, serum, and most other biological fluids contain albumin, which constitutes more than 50% of the total sample protein and typically severely restricts the amount of biological fluid that can be applied to a 2-D gel or analyzed by alternative downstream methods. Hence, a useful experimental design is to isolate albumin in a single chamber with a fairly narrow pH range to enhance detection of less abundant proteins in other fractions; e.g., mouse serum albumin focuses as a series of spots with pIs from approx pH 5.4 to 5.8 if the sample is not alkylated prior to ZOOM IEF prefractionation. As a result, essentially all of the murine albumin can be sequestered in a single narrow-range pool, which dramatically enhances loading capacity of other fractions on narrow-range 2-D gels (**13,14**).

The current version of the ZOOM IEF Fractionator contains seven potential separation chambers, but the actual number of fractions and the pH ranges can be easily varied. A schematic of a unit utilizing six commercially available partition membranes to produce five different pH-range fractions is shown in **Fig. 2**. Either an acrylamide/immobiline partition disk (crosshatched rectangles) or a spacer with an open hole (spacer) can be placed between any two separation chambers or between the terminal separation chambers and electrode chambers. The volumes of individual fractions can be equal to the optimal load volume of individual chambers (650–700 μ L), or two or more chambers can be connected with spacers rather than with pH partition membranes to increase volumes of one or more selected fractions (**Table 1**). Where feasible, experiments should be designed so that fractionated sample volumes and protein concentrations are directly compatible with downstream analysis steps, without the need to concentrate the samples. For example, if narrow pH range IPG gels will be used to analyze fractions, a 700- μ L sample volume plus a 100- μ L rinse of the chamber yields a total fraction volume of 800 μ L, which will allow up to the entire fraction to be analyzed on duplicate 24-cm IPG strips if rehydration loading of samples is used.

3.2. Casting Custom Acrylamide/Immobiline Partition Membranes at Desired pHs

Separation methods using commercially available membranes or custom membranes in the ZOOM IEF Fractionator are identical except for the added steps of casting and storing custom pH membranes, which are described in this section. The acrylamide/immobiline partition membranes can be made with different acrylamide gel concentrations, but in general, gel concentrations should remain as low as possible, typically 3–4% so that very large proteins can be efficiently focused.

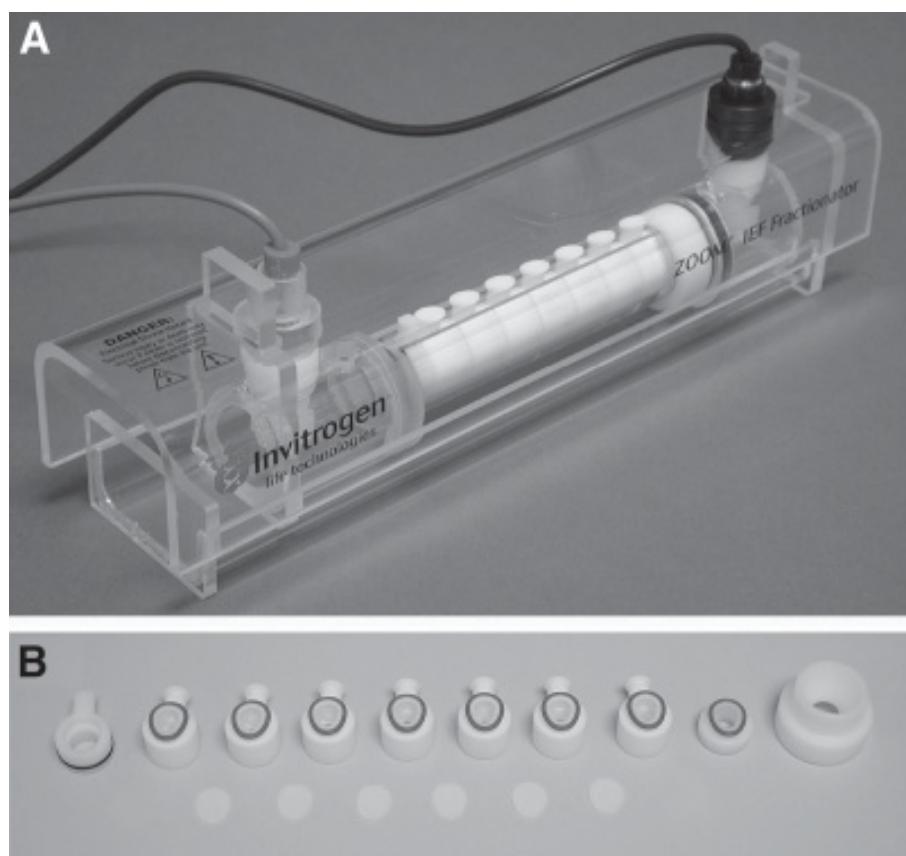


Fig. 1. The ZOOM IEF Fractionator. (A) A completely assembled device. (B) The major internal components of the device needed for prefractionation using five pH ranges: top row (left to right)—anode end sealer, seven oval-shaped chambers (with plugs and O-rings), cathode end sealer, and cathode end screw cap; bottom row—six membrane partitions that are acrylamide/immobiline gels cast on egg-shaped polyethylene disks.

Table 1
Representative Configurations for Zoom IEF Fractionator

Separation mode	Spacer/membrane placement ^a							
	1	2	3	4	5	6	7	8
Commercially available membranes								
5 fractions, uniform volumes	S	3.0	4.6	5.4	6.2	7.0	10.0	S
5 fractions, 2 double volumes	3.0	4.6	S	5.4	S	6.2	7.0	10.0
2 fractions, uniform minimal volumes	S	S	3.0	5.4	10.0	S	S	S
Custom-made membranes								
7 fractions, uniform volumes	3.0	4.4	4.9	5.4	5.9	6.4	8.1	10.0
4 fractions, variable volumes	3.0	5.0	S	S	6.0	S	7.1	10.0

^aEither a spacer (S) or a partition membrane with a specific pH is placed between the chambers at the indicated positions in the unit.

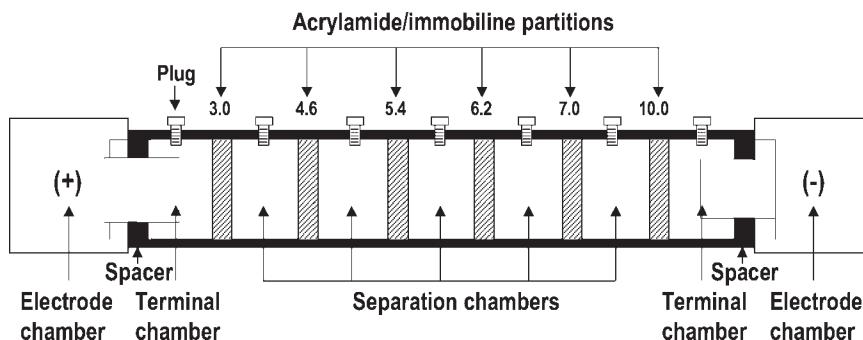


Fig. 2. Schematic illustration of the Zoom IEF Fractionator with five pH ranges. The device consists of seven small Teflon chambers (total volume approx 750 μ L; preferred sample loading volume 650–700 μ L) separated by six acrylamide/immobiline partition membranes (pH values shown above the partitions). In this configuration (five fractions), the two terminal chambers are coupled to the electrode chambers with spacers equal in thickness to partition membranes but with a hole equal to the bore of the chambers. Hence, these terminal chambers and the adjacent electrode chambers are contiguous and are loaded with anode and cathode isoelectric focusing electrode buffers, respectively.

Currently, six different immobilines (pKs 3.6, 4.6, 6.2, 7.0, 8.5, and 9.3) are available as 0.2 M stock solutions from Amersham Biosciences. Various immobilines are mixed to produce buffering capacity at any desired pH between 3.0 and 10.0. Typically, the amount and types of immobilines used for desired pHs are calculated using a computer program, “Doctor pH,” originally developed by Giaffreda et al. (16), which is commercially available from Amersham Biosciences. Alternatively, one can use pH recipes from formula tables that have been precalculated using this software (17). The general method for preparing immobiline/acrylamide partition membranes at desired custom pHs is described below. Water used for all reagents should be a high-quality, research-grade water such as Milli-Q or equivalent.

3.2.1. Preparing Immobiline Mixtures at Specific pHs

1. Prepare immobiline mixtures at desired pHs using “Doctor pH” or precalculated tables. Also, **Table 2** lists recipes for commonly used immobiline mixtures that match commercially available large IPG strips using the strategy described in **Subheading 3.7.1.** (see **Note 1**).
2. Dilute to the indicated final volume with water.
3. Measure pH immediately, using a pH meter with an accuracy of at least 0.01 pH units.
4. If observed pH deviates from desired pH, carefully adjust by adding a small amount of the most acidic or most basic immobiline (either pK 3.6 or 9.3).
5. Titrate the immobiline mixture to a pH of approx 6.5 using either 1 M Tris base or 1 M acetic acid. Tris or acidic acid are not covalently incorporated into the gel matrix and will be washed out after polymerization. Titration of all immobiline mixtures to a pH of approx 6.5 ensures efficient polymerization of the acrylamide/immobiline gel.

3.2.2. Casting Custom Acrylamide/Immobiline Partition Membranes at Desired pHs

1. Coat two sheets of gel support film with Repel-Silane and trim both sheets to fit into the lid of a 1-mL Rainin pipet tip box.
2. Insert one silanized sheet in the lid and place 1.5-mm gel spacers around the perimeter.

Table 2
Preparation of Immobiline^a Mixtures at Specific pHs

	pH 4.4	pH 4.9	pH 5.4	pH 5.9	pH 6.4	pH 8.1
pK 3.6	615 µL	479 µL	331 µL	280 µL	228 µL	—
pK 4.6	236 µL	321 µL	348 µL	438 µL	529 µL	206 µL
pK 6.2	472 µL	363 µL	247 µL	204 µL	161 µL	760 µL
pK 7.0	48 µL	89 µL	107 µL	144 µL	181 µL	299 µL
pK 8.5	—	—	—	—	—	—
pK 9.3	138 µL	258 µL	314 µL	424 µL	532 µL	218 µL
Milli-Q Water	5.88 mL	5.91 mL	6.11 mL	5.98 mL	5.87 mL	17 µL
Total volume	approx 7.5 mL	5.92 mL				
						approx 7.5 mL

^aImmobilines (0.2 M stocks) are used to make approx 12 mM final concentration in each gel membrane. When seven fraction separations are conducted, custom partition membranes are used, which are cast using the above immobiline mixtures at the indicated pHs. In addition, the commercially available 3.0 and 10.0 partition membranes are used to complete the pH series as shown in Table 1.

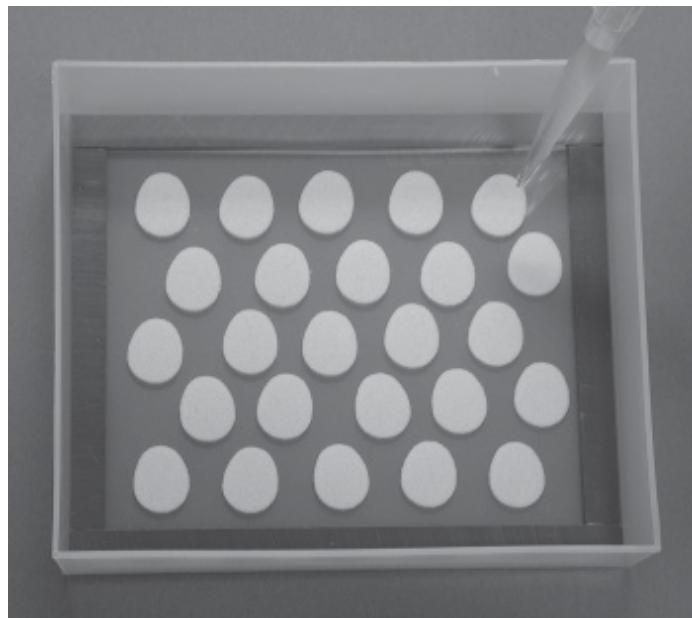


Fig. 3. Casting custom acrylamide/immobiline partition membranes using porous polyethylene discs. The plastic cover from a 1.0-mL Rainin pipet tip box is used as a convenient container. Typical placement of 24 egg-shaped discs on top of a silanized gel support film with four 1.5-mm gel spacers around the perimeter is shown. As illustrated, it is important to pipet the gel solution directly onto the discs to remove air trapped in the disc pores.

Table 3
Casting Acrylamide/Immobiline Gels

	4%T/10%C gel
Immobiline mixture (with specific pH)	7.50 mL
Acrylamide/Bis (30%T/10%C)	3.33 mL
Glycerol (87%)	3.45 mL
Milli-Q water	10.70 mL
Ammonium persulfate (4%)	20 μ L
TEMED	12 μ L
Total Volume	approx 25 mL

3. Arrange precut egg-shaped polyethylene disks (Invitrogen) so that disks do not touch (see **Fig. 3**). Blank disks are precoated with a surfactant; do not rinse or wash prior to use.
4. Prepare acrylamide/immobiline gel working solution using the recipe in **Table 3**, which is sufficient for casting 24 disk membranes.
5. Immediately pipet the acrylamide/immobiline mixture directly onto the disks until the solution is completely absorbed inside the polyethylene pores and all air is displaced.
6. Continue adding gel solution to cover the remaining areas of the film until the solution is uniformly as high as the 1.5-mm spacers on the lid perimeter.
7. Immediately cover the solution with the second silanized support film from **step 1**.
8. Polymerize at room temperature (approx 2 h).

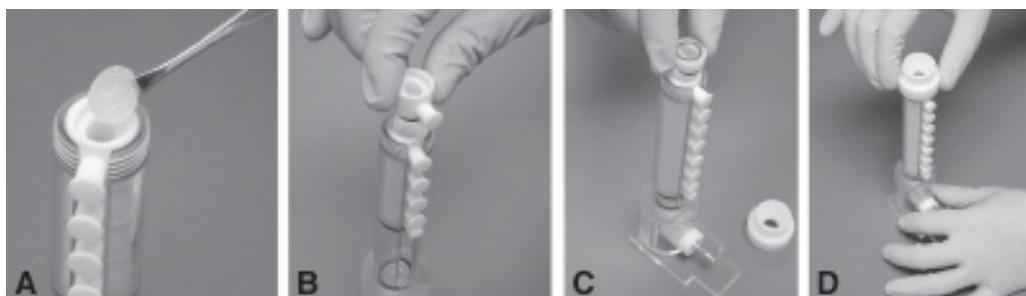


Fig. 4. Assembly of chambers and partition membranes for Zoom IEF Fractionator. **(A)** Insertion of a partition membrane into the groove of a separation chamber. **(B)** After inserting a partition membrane, the next separation chamber is carefully inserted with the O-ring side down. **(C)** After all membranes and separation chambers are inserted, the cathode end sealer is added. **(D)** After inserting all components, the cathode end screw cap is screwed onto the top of the illustrated assembly to compress the O-rings, forming seals between the tandem Teflon chambers and end units.

9. After polymerization is complete, remove the disks and trim excess gel from all surfaces using a scalpel or razor blade.
10. Immediately transfer individual cleaned disks into wells of a 24-well tissue-culture plate.
11. Wash disks three times in 2 mL of 12% glycerol solution for 30 min each with shaking at room temperature to remove the acetic acid or Tris and polymerization byproducts.
12. Add 1 mL of partition membrane storage solution to each well and tightly seal with parafilm or a sheet of adhesive microtiter plate film.
13. Partition membranes can be used immediately or stored at 4°C for up to at least 6 wk without losing effectiveness.

3.3. Assembling the ZOOM IEF Fractionator

The typical arrangement of chambers, partition membranes, and spacers for a Zoom IEF Fractionator with five separation chambers is shown in [Fig. 1B](#) and [Table 1](#). The unit is assembled prior to loading samples following the manufacturer's instructions. As an example, the method for the configuration shown in [Fig. 1B](#) is briefly summarized below.

1. Insert the anode end sealer into the chamber assembly tube followed by a spacer.
2. Insert a separation chamber, which contains a port plug and an O-ring. The O-ring should be facing down. The use of a spacer between the electrode chamber and this first potential separation chamber means that in this configuration the first "separation chamber" will actually contain electrode buffer (see [Fig. 2](#)).
3. Add a pH 3.0 membrane partition disk. The disks are labeled with the pH to minimize the potential that disks will be inserted out of order.
4. Add the next separation chamber, making sure that the port plug is in place and the O-ring is facing down. Push the chamber into the chamber assembly tube until it is flush with the top of the assembly tube.
5. Repeat **steps 3 and 4** for the pH 4.6, 5.4, 6.2, 7.0, and 10.0 membranes and separation chambers ([Fig. 4A,B](#)).
6. Add a spacer followed by the cathode end sealer ([Fig. 4C](#)).

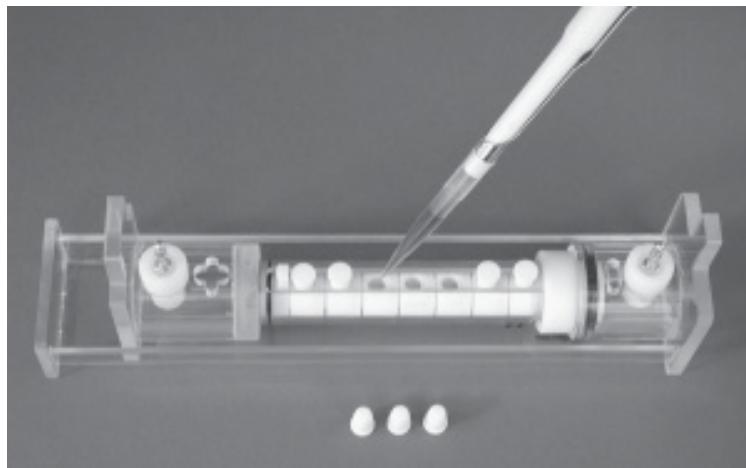


Fig. 5. Loading a protein sample into a Zoom IEF Fractionator. Loading a sample into the center three separation chambers is shown.

7. The cathode end screw cap is then fastened onto the assembly tube (**Fig. 4D**) and the cathode electrode chamber is inserted.
8. Electrode chambers are filled with 1X IEF anode (7 mM phosphoric acid) and cathode buffer (20 mM lysine/20 mM arginine). Typically, 8 M urea and 2 M thiourea (final concentrations) are included in the electrode buffers to minimize electro-osmosis during IEF.

3.4. Loading Protein Samples

In general, protein samples are prepared in a sample buffer that contains 8 M urea, 2 M thiourea, 4% CHAPS, 1% DTT, and 0.2–0.5% IPG buffer (carrier ampholytes, pH 3.0–10.0 L). The sample buffer is also used to fill any separation chambers that do not receive sample. Depending upon the number of separation chambers and the amount and volume of the sample to be used (*see Notes 2 and 3*), the sample prepared in sample buffer can be further diluted with additional sample buffer to a final volume equal to the total volumes of the chambers that will be loaded with sample. For example, if a seven-chamber separation strategy will be used and the sample will be loaded into the five center chambers, a final sample volume of 3.5 mL (5 × 700 µL) can be used. Immediately prior to loading, the sample solution should be filtered through a 0.22 µm Ultrafree-MC microfilter unit in a microcentrifuge at room temperature to remove any particulate material. **Fig. 5** illustrates loading a sample into the three center chambers. After all chambers are filled with either sample in sample buffer or sample buffer only, it is very important that all port plugs be tightly seated to form an air-tight seal (*see Note 4*).

3.5. Conducting Solution IEF Fractionation

Isoelectrofocusing conditions and total time required are very dependent upon sample ionic strength, protein concentration, and number of chambers loaded with the sample (*see Notes 2, 3, and 5*). Typically, prefractionation of a complex proteome such as mammalian cell extracts requires a power supply capable of operating up to at least

1000 V and at currents as low as 0.1 mA. If the power supply has current and power limit capabilities, set maximum current at 1 mA and maximum power at 1 W. Depending upon the programming capacity of the power supply used, voltage can be increased in steps manually, use programmed linear ramps, or use programmed current and voltage limits. The key factors are to limit total power to 1W or less and to ultimately achieve and hold 1000 V until a stable low current (<1mA) is reached. A stable minimum current is defined as <0.1 mA change over 30 min.

A sample program that was used for a 3.0-mg human cell extract loaded in the central five chambers of a seven pH range separation was: 50 V for 1 h, 100 V for 1.5 h, 200 V for 30 min, 350 V for 30 min, 600 V for 20 min, 1000 V until the lowest possible current of 0.2–0.3 mA was obtained and maintained for 30 min (approx 1–1.5 h total at 1000V). The total run time for this sample fractionation was approx 5 h. A simple alternative focusing program limits the current to 1 mA and the voltage to 1000 V. Initially, the sample is focused at the maximum 1-mA current and voltage rises as resistance decreases. When the 1000-V limit is reached, the current begins to fall. As above, focusing is continued until a minimum current is reached and remains unchanged for 30 min. Some simple samples such as prokaryote extracts may be adequately focused, as previously reported (6), and the manufacturer's recommendations are to use a maximum of 600 V. However, many complex samples such as mammalian cell extracts and biological fluids may be incompletely focused unless higher voltages, e.g., 1000 V, are used (see Note 6).

3.6. Collecting Solution Fractions and Recovery of Proteins Trapped in Partition Membranes

After the IEF run is complete, fractionated samples are removed through the fill ports using a pipet and transferred to 1.5-mL microcentrifuge tubes. A small volume of sample buffer (e.g., approx 100 μ L) is used to rinse the inside walls of the chambers and the membrane surfaces, and this rinse is combined with the fraction to minimize protein losses. A few proteins, primarily those with pIs equal to pHs of partition membrane disks are retained in these disks, but can be extracted using two extractions with 400 μ L sample buffer for 30 min with shaking at room temperature (see Note 7). These extracts (total approx 800 μ L) are then pooled. All samples, including the solution fractions and membrane disk extracts, can be used immediately for further analysis or stored in aliquots at –80°C until required.

3.7. Evaluating Effectiveness of ZOOM-IEF Sample Fractionation

3.7.1. Initial Evaluation of Fractionation Efficiency on 1D SDS Mini-Gels

A simple initial check of fractionation results is to run proportional amounts of all fractions and membrane disk extracts on 1-D 10% SDS minigels. After the gels are stained, overall protein recovery and distribution among fractions and membrane partitions can be observed. If desired, estimates of total protein amounts in each lane can be obtained using 1-D gel image analysis software such as Discovery Series Quantity One (version 4.2.0) (Bio-Rad Laboratories). These 1-D gels are particularly useful for preliminary comparison of reproducibility of replicate runs and for comparison of different experimental samples. Representative results from fractionation of a human breast cancer cell extract into five pH ranges using the Zoom IEF Fractionator with the six

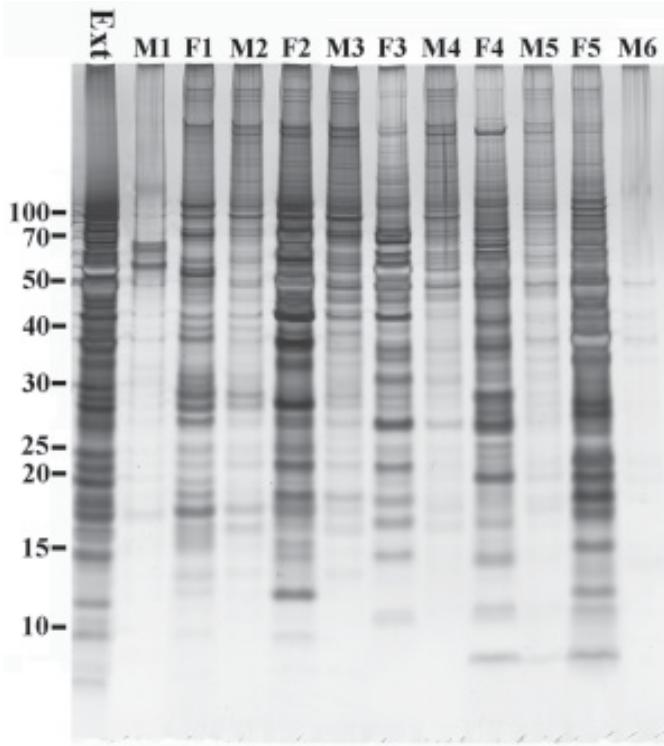


Fig. 6. Evaluation of ZOOM-IEF fractions using a 1-D sodium dodecyl sulfate (SDS) gel. After 1.5 mg of a MCF-7/6 human breast cancer cell extract was fractionated into five fractions (see Figs. 1 and 2), fractions and extracts from partition membranes were analyzed on a 10% SDS Bis-Tris NuPAGE mini gel (Invitrogen) followed by silver staining. Ext: cell extracts without fractionation (2 µg); F1–F5: fractions 1–5; M1–M6: extractions from the partition membranes with pH 3.0, 4.6, 6.2, 7.0, and 10, respectively. Protein loads were amounts derived from approx 8 µg of original cell extract.

commercially available partition membranes (see Figs. 1 and 2) are shown in Fig. 6. Silver stain was used in this case to maximize sensitivity, and as a result of the narrow dynamic range for silver stains, many bands in solution fractions are outside the linear range of the stain. As a result, minor bands in the membrane fractions are somewhat overemphasized. However, it is clear that only M3 has a substantial amount of protein relative to adjacent solution fractions.

3.7.2. Evaluation of Fractionation Efficiency Using Mid-pH Range 2-D Gels

The most reliable and efficient way to determine whether proteins in a fraction are within the intended pH range, and if sharp sample boundaries are observed near the expected pH, is to run small aliquots of each fraction (and membrane disk extracts, if desired) on appropriate 3 pH unit wide 2-D mini gels. Either 7-cm or 11-cm IPG gels could be used for this purpose. However, 11-cm IPG gels and subsequent 13-cm-wide SDS gels (Criterion gel format, Bio-Rad Laboratories) are strongly preferred because these gels have about 1.6-times greater separation distance than 7-cm IPG strips, and these midrange gels are not significantly more difficult or time-consuming to run com-

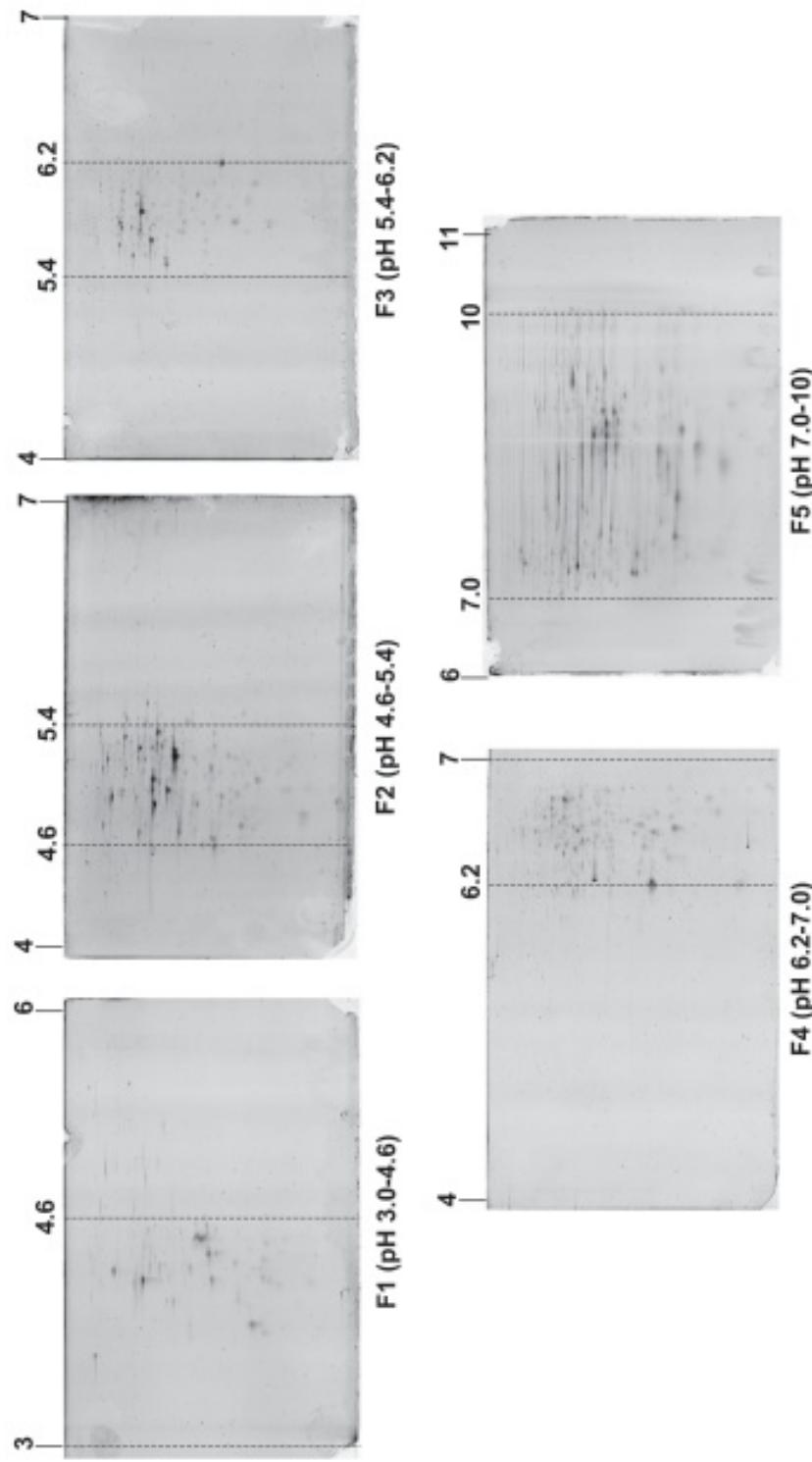


Fig. 7. Evaluation of ZOOM-IEF fractionation using mid-pH range 2-D gels. Samples and conditions were as shown in **Fig. 6**. Fractionated samples were focused on 11-cm immobilized pH gradients (IPGs) followed by Tris-Tricine sodium dodecyl sulfate gel electrophoresis using Criterion gel cassettes (11 × 9 cm). The amount of samples applied to IPG strips (as equivalent to the original cell extract): F1, approx 80 µg; F2, F3, and F4, approx 40 µg; and F5, approx 80 µg. The pH values of the IPG strips used are indicated above the images, and the expected fraction boundaries are indicated by vertical dashed lines. Gels were stained using Sypro Ruby.

pared with mini-gels (see Note 8). Representative 2-D-gel results for analysis of the same experiment as the 1-D gels in **Fig. 6** are shown in **Fig. 7**. These 2-D maps confirm that the extract was separated into five well-resolved fractions, since clear boundaries near the expected pHs were obtained and only minimal amounts of proteins were observed outside the expected pH ranges of each fraction (see Notes 5 and 6).

3.8. Alternative Downstream Protein-Profiling Methods for Fractionated Samples

3.8.1. Narrow pH Range 2-D PAGE

ZOOM IEF fractions interface seamlessly with subsequent narrow pH range IPG gels because similar sample buffers are used and fractions usually do not need to be concentrated prior to 2-D gel analysis (see Note 9). Ideally, a series of overlapping IPG strips, each with slightly wider pH ranges (± 0.1 pH unit) than the ZOOM IEF fractions, should be used to maximize separation distances. Matching 1.0 pH range IPG strips for the three center fractions (each 0.8 pH units) obtained using the six commercially available partition membranes (**Fig. 2** and **Table 1**) are available for mini-gels, i.e., 7-cm IPG strips (Invitrogen Corp.). But unfortunately, the separation distance obtained on these small gels is similar to that obtained by running a single pH 4.0–7.0 2-D gel using 18- or 24-cm gels, thereby, largely negating the advantages obtained by prefractionating the sample. Although longer custom IPG strips could be prepared in individual laboratories, this approach is quite labor intensive, and more importantly, it would likely lead to increased variability between laboratories when an important goal is to develop proteomic platforms that are highly consistent between laboratories.

The seven-fraction fractionation strategy shown in **Table 1** was designed to interface with large, commercially available IPG strips in a manner that would maximize both total separation distance and throughput (see Note 8). The pH ranges of these seven fractions have been selected in part so that the ZOOM IEF fractions could be separated on readily available 24-cm narrow pH range gels. After IEF, regions of the IPG strip that are more than 0.1 pH outside the pH range of the ZOOM IEF fraction are cut off to reduce the IPG strip size to approx 17 to 18 cm. The second-dimension gels can then be run on 18-cm wide gels, which are less labor intensive and less expensive than 24-cm gels. This strategy (**Table 4**) produces a total effective IEF separation distance of about 86 cm over the range of pH 3.0 to 10.0 with no appreciable loss of resolution compared with running 24-cm second-dimension gels (**12,15**).

3.8.2. Large-Pore 1-D PAGE

Analysis of ZOOM IEF fractions on large-pore 1-D gels is highly complementary to the 2-D gel analysis described in the subheading above because 2-D gels generally do not reliably separate and recover large proteins (>100 kDa). In contrast, 1-D gradient acrylamide gels, such as NUPAGE 3–8% Tris-acetate mini-gels (Invitrogen Corp.), provide high resolution to >500 kDa and can be used to rapidly compare and detect differences in fractionated proteins in the >100-kDa range from different samples (**12,15**).

3.8.3. Two-Dimensional Differential In-Gel Electrophoresis (2-D DIGE)

2-D DIGE is a relatively new approach to quantitative proteome analysis using 2-D gels, which involves separately labeling two or three different protein extracts with distinct fluorophores (cyanine dyes) by covalently modifying lysine residues in the

Table 4
Strategy for Interfacing ZOOM IEF Fractions With Large Commercially Available Immobilized pH Gradient (IPG) Strips

Fraction	F1	F2	F3	F4	F5	F6	F7
ZOOM pH	3.0–4.4	4.4–4.9	4.9–5.4	5.4–5.9	5.9–6.4	6.4–8.1	8.1–10
Commercial 24-cm IPGs	3.0–6.0	4.0–5.0	4.5–5.5	5.0–6.0	5.5–6.7	6.0–9.0	7.0–10
pH (trimmed to 18 cm)	3.0–5.25	4.25–5.0	4.75–5.5	5.25–6.0	5.7–6.6	6.2–8.45	7.75–10

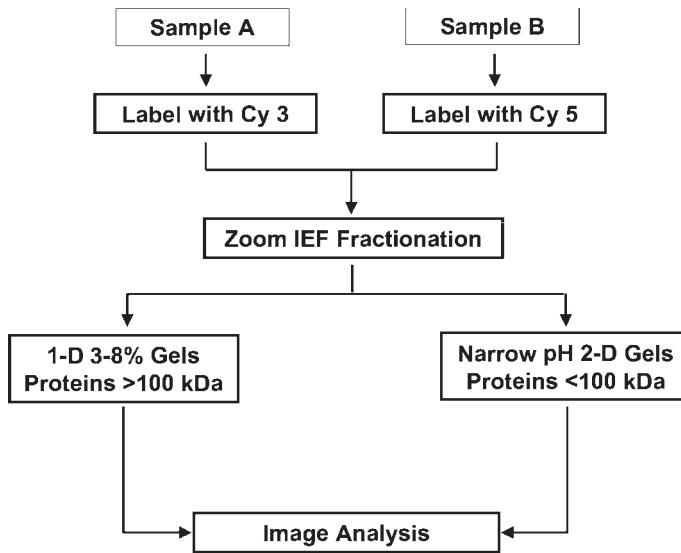


Fig. 8. A strategy combining 2-D differential in-gel electrophoresis (DIGE) fluorescent labeling with ZOOM-IEF fractionation for comprehensive gel-based proteome analysis. The fractionated samples are subsequently analyzed on conventional 1-D and 2-D gels, and images are compared using 2-D DIGE software for quantitative comparisons.

proteins (18). The ZOOM IEF and subsequent slightly overlapping narrow-range 2-D gels plus large-pore 1-D gel methods described in the subheadings above are readily compatible with 2-D DIGE. When these methods are combined (Fig. 8), any subtle variations in ZOOM IEF fractionation and on the 2-D gels are no longer a problem, because the two or three labeled samples are mixed prior to these separation steps. Another advantage is that the number of large, high-resolution, narrow-range gels that are required will be greatly reduced. The major disadvantages are those that are inherent to 2-D DIGE—i.e., the method is often less sensitive than silver staining, and it can be difficult to match observed fluorescent changes with the corresponding unlabeled protein for identification by mass spectrometry.

3.8.4. LC-MS/MS and LC/LC-MS/MS Protein-Profiling Methods

Tandem mass spectrometry (MS/MS) analysis of tryptic peptides after one (LC) or two (LC/LC) dimensions of liquid chromatography is the major alternative to 2-D gel-based methods for protein profiling. Although these methods alone can detect lower-abundance proteins than observed on 2-D gels, they are not currently capable of detecting more than about 1000 to 2000 proteins, and throughput is similar to 2-D gels. Hence, these methods can also benefit from prior prefractionation of complex proteomes using ZOOM IEF. Two alternative schemes for interfacing these methods are shown in Fig. 9. As with other LC-MS/MS methods, stable isotope labeling of protein samples is utilized for quantitative comparisons. The method shown in the bottom left side of Fig. 9 uses a short 1-D SDS gel as a convenient method of removing detergent, ampholytes, and other byproducts that would otherwise interfere with trypsin digestion and/or subsequent MS analysis. The preferred strategy for this combined ZOOM IEF with classical LC/LC-MS/MS analysis is to use commercial ZOOM IEF partition discs

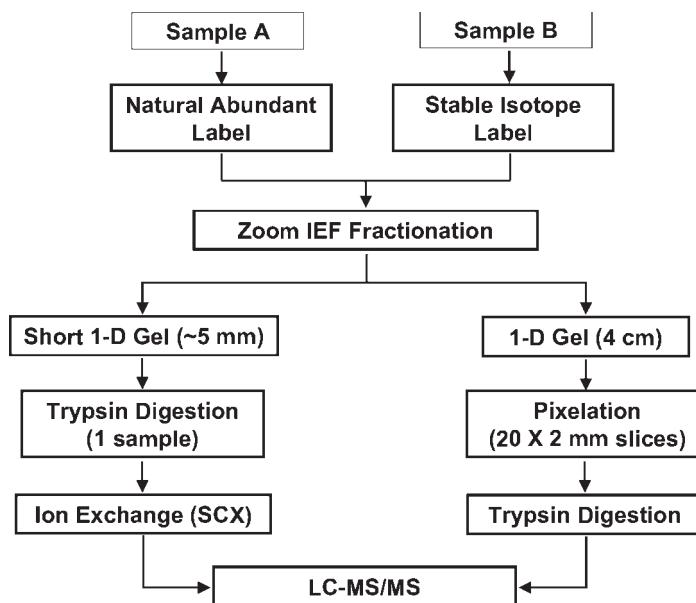


Fig. 9. Protein profiling strategies combining ZOOM IEF prefractionation with liquid chromatography (LC) tandem mass spectrometry (MS/MS) methods. Two protein samples are initially labeled with different stable isotopic tags by metabolic or chemical derivatization. Samples are mixed and fractionated using the ZOOM IEF Fractionator into five to seven or more pH ranges. To interface with conventional LC/LC methods (left side), each fraction is separately electrophoresed into a sodium dodecyl sulfate (SDS) mini-gel for a short distance (approx 5 mm), fixed, and stained with colloidal Coomassie to remove interfering reagents from the ZOOM IEF fractionation. After in-gel digestion of the entire gel sample from each fraction, the digest is analyzed by conventional LC/LC-MS/MS methods. An alternative method is “protein array/pixelation,” where ZOOM IEF fractions are separated for about 4 cm in a 1D SDS mini-gel followed by dividing each lane into uniform 2-mm slices (pixelation) for trypsin digestion and LC-MS/MS analysis.

to separate complex proteomes into five fractions. Although one now has five samples rather than one sample to analyze by LC/LC-MS/MS, this division of the proteome into five well-resolved fractions means that less SCX fractions and possibly shorter reverse-phase run times can be used while still identifying more proteins.

A novel alternative 3-D protein profiling method, batch protein array pixelation, has recently been developed as an alternative method that utilizes ZOOM IEF prefractionation (15). As illustrated in the lower right side of Fig. 9, the five to seven fractions from ZOOM IEF are run on short 1-D gels, and each lane is cut into uniform slices. The combination of IEF in the first dimension and SDS gel in the second dimension produces a two-dimensional grid (or “array”) where every spot (or “pixel”) in the array contains a group of proteins that are subsequently identified by LC-MS/MS after in-gel trypsin digestion. This method has several advantages compared with conventional bottom-up approaches such as conventional LC/LC-MS/MS alone. First, it is more readily scaleable because it exploits the greater diversity of physical properties inherent to proteins compared with tryptic peptides. Second, information about protein prop-

erties is not lost as with bottom-up approaches. That is, all proteins in a pixel are known to have pIs within a specific range and molecular weights within a specific range.

4. Notes

1. One critical challenge with preparing custom acrylamide/immobiline membranes is maintaining reproducible pHs and buffering capacities to ensure reproducible separations. Although theoretically only two immobilines are required to define any pH, in practice more than two immobilines are needed to ensure more uniform buffering capacity (16). Generally, the total immobiline concentration in the acrylamide gel should be between 10 and 20 mM. Higher concentrations would provide great buffering capacity and should decrease the potential for high-abundance proteins with pIs near the membrane pH to overwhelm immobiline buffering capacities and disrupt the separation. However, greater than 20 mM immobilines usually causes gels to swell during washing and use. In addition, as immobiline concentrations increase (e.g., >40 mM), their efficiency of incorporation begins to decrease, and inconsistent incorporation of different immobilines into a covalent acrylamide matrix will skew buffering pH relative to the expected pH (16).
2. There is substantial flexibility in loading protein samples into a Zoom IEF Fractionator. For example, for the five-fraction chamber device illustrated in **Fig. 2**, the initial sample can be loaded into any single chamber, several chambers, or all five separation chambers. The advantages of different loading methods appear to be somewhat sample dependent. For example, it may be advantageous to load an entire serum sample into the single chamber that will contain albumin. Because albumin constitutes >50% of the protein content in serum, less than half the protein needs to migrate through partition membranes with reduced potential for protein precipitation or protein overloading of membrane buffering capacity. However, except for biological fluids with the high albumin problem, loading an entire sample into a single chamber increases the local initial protein concentration compared with more distributed loading strategies, and this higher protein concentration could lead to protein precipitation and aggregation early in the focusing step. If an available sample is too dilute to fit into one or several separation chambers but would fit into all separation chambers, then it is usually preferable to load all chambers rather than risk sample losses during a concentration step. However, some proteins show poor focusing in terminal chambers when the sample is loaded into all separation chambers. This may be because ampholytes migrate more rapidly than proteins, and the pHs in the terminal chambers may reach pH extremes before proteins that are unstable at extreme pHs have sufficient time to migrate out of these terminal chambers, which may result in precipitation and aggregation. In addition, proteins that have pIs at the opposite pH extreme compared with their loading position have the greatest distance to travel. Hence, it is reasonable to expect that in general, focusing times will be greater if sample is loaded in all chambers compared with loading in one or several central chambers. Hence, although sample loading strategies can be highly dependent upon sample type and experimental design, a general initial recommendation is to load samples into several central chambers and, where possible, avoid loading samples into the most acidic and most basic chamber. Hence if five separation chambers are used, start by loading the sample in the central three chambers; and if seven separation chambers are used, start by loading the sample in the central three or five chambers.
3. The maximum sample loading capacity seems to depend primarily upon the type of sample, and secondarily on the loading position (see **Note 2**) and total number of separation chambers. In general, higher loads can be used for simple, clean samples that contain primarily or only proteins, compared with very complex extracts with large amounts of

lipids and nucleic acids. Hence, for complex samples such as mammalian cell or tissue extracts, up to 2- to 3-mg protein loads are suitable for a device with seven separation chambers. Higher loads are possible for less complex proteomes such as prokaryotic cell extracts, cytoplasmic fractions, and biological fluids. Overloaded samples tend to result in protein deposition on partition membrane surfaces, and are usually incompletely separated. The pIs of some proteins will exactly match the pH of any partition membrane, which may be a critical factor in limiting sample load, either by blocking the pores with subsequent deposition of proteins on the membrane surfaces or by overwhelming the buffering capacity of the immobilines in the membrane. However, the present load capacity matches well with the maximum feasible loads for subsequent narrow pH range 2-D gels (11,13,14).

4. It is critical that all chambers have a small head space filled with air and that all ports have air-tight seals. Tightly sealed chambers minimize electro-osmotic flow. If a chamber is not completely sealed, it is very likely that electro-osmotic flow will deplete the liquid in the affected chamber. When the liquid level in a single chamber is nearly depleted, conductivity falls and the observed very low current incorrectly indicates that equilibrium has been reached. Therefore, the first indicator of this problem is that the current falls to a low stable level more rapidly than expected compared with prior analyses with similar samples. Another indication is low sample volumes in one or more chambers after focusing is completed. Small changes in volume relative to the amount loaded usually do not indicate a problem. But if the volume decreases more than 100 μ L relative to load volume, this usually indicates an air leak in that chamber, and if the final volume remaining in one or more chambers is less than half the original volume, a severe leak has probably occurred and samples should be discarded because they will not be properly focused.
5. Although focusing protein samples in a ZOOM IEF Fractionator has a desalting function, protein samples that have high ionic strength should be avoided where possible. For example, extract samples using low ionic strength buffer, or in the case of concentrated biological fluids such as serum, dilute with a low ionic strength buffer or with the ZOOM IEF sample buffer. High salt concentrations in samples will produce high conductivity, and since current is typically limited to 1 mA, voltages will remain low for extended times, and the overall run time increases. Some samples with high conductivity have been successfully focused, and in extreme cases, initial focusing was conducted overnight at a low voltage to clear the salts from the separation chambers, followed by completion of the higher voltage steps the next morning. Similarly, ionic detergents such as SDS should be avoided or kept to a minimum final concentration (preferably <0.1%). In addition, the concentration of carrier ampholytes will affect the initial conductivity, total focusing time, and the minimum current that is reached at the final 1000-V focusing step. This effect must be counterbalanced with the possibility that higher ampholyte concentrations may aid protein solubility at high protein loads. Most separations have been conducted with 0.2% to 0.5% ampholytes. The higher concentration may be advisable for very complex or difficult samples.
6. Incomplete focusing of multiple types of samples has sometimes been observed when only 500–600 V was used as the final voltage. Even when 1000 V was used, longer focusing times at the final voltage usually resulted in more complete transfer of proteins into the correct pH-range chamber. The most common sign of incomplete focusing was appearance of some proteins in a fraction that actually belonged to the next fraction. If the sample was loaded in central chambers, incomplete focusing was most noticeable in the terminal chambers.
7. One problem that can occur is precipitation of proteins on the surfaces of partition membranes. This usually indicates that the protein load should be decreased, or the speed of

- focusing should be decreased, or the sample may contain nonprotein components that may bind to the proteins and interfere with the separation.
8. Minor deviations are sometimes observed for the actual pH ranges of ZOOM IEF fractions compared with the expected pH ranges when fractions are separated on narrow pH (approx 1.0–1.2 unit) range IPG gels. When using custom membranes prepared as described above, most fractions showed fairly consistent shifts of approx 0.1–0.2 pH units below the expected pHs. There are several possible reasons for this minor pH deviation. First, partition membrane pHs were determined in the absence of denaturing agents, while the IPG strip separation is conducted in the presence of high urea and thiourea concentrations. Second, different immobilines may be incorporated into gel matrix with varying efficiencies, which would skew the actual pH. Also, the actual pHs throughout the IPG gels may deviate slightly from their theoretical values, and it is not straightforward to verify pHs on IPG gels since the pH values on polymerized acrylamide/immobiline gels cannot be measured directly even with an advanced surface electrode (16). In practice, the minor pH differences observed are not a serious problem because the ZOOM IEF partition pHs appear to be reproducible between batches when the same procedure is followed, and reproducibility is more critical than precise pHs of fraction boundaries.
 9. Very similar sample solubilization buffers are used for both ZOOM-IEF and IPG-IEF, which has at least two advantages. First, ZOOM-IEF fractions can be applied directly to IPG gels after adding additional ampholytes and fresh reducing reagent, and after compensating for the small pH shift discussed in **Note 8**, protein migration in solution should closely match subsequent focusing in IPG strips. Second, reagents that promote sample solubility and effective focusing in one IEF system should produce similar beneficial effects in the other system. The combination of urea and thiourea in the sample buffer is superior to urea alone for solubilizing and focusing proteins (19). Similarly, DTT is the most commonly used reducing reagent for IEF, despite its mild negative charge, which might cause it to migrate out of high-pH separation chambers during ZOOM-IEF. Addition of a small amount of DTT (e.g., approx 0.2%) to the ZOOM IEF cathode buffer could minimize this depletion, although it does not appear to be a major factor, with the possible exception that it may be needed for very long focusing runs such as those with high initial conductivity. Although TBP was thought to be superior to DTT because it has no electrical charge and it is a more powerful reducing agent (20), we found that TBP is unstable and rapidly degraded during either solution IEF or IPG-IEF, and should be avoided. The presence of soluble ampholytes (IPG buffer) is necessary for optimal separation of proteins in both IEF systems. The actual minimum amount of ampholytes needed for optimal ZOOM IEF appears to vary somewhat with type of sample, protein load, and pH ranges of the fractions. As noted above, 0.2% is a good starting concentration, and up to 0.5% may be needed for more difficult samples.

Acknowledgments

The authors gratefully acknowledge the assistance and scientific contributions of Peter Hembach, Lynn Echan, Nadeem Ali-Khan, Kaye Speicher, Hsin-Yao Tang, and Sandra Harper throughout development of MicroSol/ZOOM-IEF prefractionation and downstream analysis methods. The administrative assistance of Emilie Gross is greatly appreciated. We also thank Invitrogen Corp. for providing precut blank hydrophilic polyethylene discs for casting custom acrylamide/immobiline partition membranes. This work was primarily supported by NIH grants CA77048, CA92725, and CA94360 to D. W. S. The support of institutional grants including an NCI cancer center grant (CA10815) and the Commonwealth Universal Research Enhancement Program, Pennsylvania Department of Health, is also acknowledged.

References

1. Quadroni, M. and James, P. (1999) Proteomics and automation. *Electrophoresis* **20**, 664–677.
2. Corthals, G. L., Wasinger, V. C., Hochstrasser, D. F., and Sanchez, J. (2000) The dynamic range of protein expression: A challenge for proteome research. *Electrophoresis* **21**, 1104–1115.
3. Ali-Khan, N., Zuo, X., and Speicher, D. W. (2002) Overview of proteome analysis. In: Coligan, J. E., Dunn, B. M., Ploegh, H. L., Speicher, D. W., and Wingfield, P. T. (eds), *Current Protocols in Protein Science*. John Wiley & Sons, New York: 22.1.1–22.1.19.
4. Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., and Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* **17**, 994–999.
5. Ranish, J. A., Yi, E. C., Leslie, D. M., et al. (2003) The study of macromolecular complexes by quantitative proteomics. *Nat. Genet.* **33**, 349–355.
6. Zuo, X. and Speicher, D. W. (2000) A method for global analysis of complex proteomes using sample prefractionation by solution isofocusing prior to two-dimensional electrophoresis. *Anal. Biochem.* **284**, 266–278.
7. Righetti, P. G., Castagna, A., and Herbert, B. (2001) Prefractionation techniques in proteome analysis: A new approach identifies more low-abundance proteins. *Anal. Chem.* **73**, 320–326.
8. Righetti, P. G., Castagna, A. C., Herbert, B. H., Reymond, F. R., and Rossier, J. S. (2003) Prefractionation techniques in proteome analysis. *Proteomics* **3**, 1397–1407.
9. Weber, G., Grimm, D., and Bauer, J. (2000) Application of binary buffer systems to free flow cell electrophoresis. *Electrophoresis* **21**, 325–328.
10. Hoffmann, P., Hong, J., Moritz, R. L., et al. (2001) Continuous free-flow electrophoresis separation of cytosolic proteins from the human colon carcinoma cell line LIM 1215: a non two-dimensional gel electrophoresis-based proteome analysis strategy. *Proteomics* **1**, 807–818.
11. Zuo, X., Hembach, P., Echan, L., and Speicher, D. W. (2002) Enhanced analysis of human breast cancer proteomes using micro-scale solution isoelectrofocusing combined with high resolution 1-D and 2-D gels. *J. Chromatog. B* **782**, 253–265.
12. Zuo, X., Lee, K., and Speicher, D. W. (2004) Electrophoretic prefractionation of proteomes for comprehensive analysis of proteomes. In: Speicher, D. W. (ed), *Proteome Analysis: Interpreting the Genome*, Elsevier Science, New York: 93–118.
13. Zuo, X., Echan, L., Hembach, P., et al. (2001). Towards global analysis of mammalian proteomes using sample prefractionation prior to narrow pH range two-dimensional gels and using one-dimensional gels for insoluble and large proteins. *Electrophoresis* **22**, 1603–1615.
14. Zuo, X. and Speicher, D. W. (2002). Comprehensive analysis of complex proteomes using microscale solution isoelectrofocusing prior to narrow pH range two-dimensional electrophoresis. *Proteomics* **2**, 58–68.
15. Speicher, D. W., Lee, K., Tang, H. Y., et al. (2004) Current challenges in proteomics: Mining low abundance proteins and expanding protein profiling capacities. In: Ashcroft, A. E., Brenton, A. G., and Monaghan, J. J. (eds), *Advances in Mass Spectrometry*, Vol. 16, McMillan, New York: 37–57.
16. Giaffreda, E., Tonani, C., and Righetti, P. G. (1993) A pH gradient simulator for electrophoretic techniques in a windows environment. *J. Chromatogr.* **630**, 313–327.
17. Amersham Biosciences (1997) Isoelectric membrane formulas for IsoPrime Purification of proteins. IsoPrime Protocol Guide #1.

18. Unlu, M., Morgan, M. E., and Minden, J. S. (1997) Difference gel electrophoresis: A single gel method for detecting changes in protein extracts. *Electrophoresis* **18**, 2071–2077.
19. Rabilloud, T., Adessi, C., Giraudel, A., and Lunardi, J. (1997) Improvement of the solubilization of proteins in two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis* **18**, 307–316.
20. Herbert, B. R., Molloy, M. P., Walsh, B. J., Gooley, A. A., Bryson, W. G., and Williams, K. L. (1998) Improved protein solubility in 2-D electrophoresis using tributyl phosphine. *Electrophoresis* **19**, 845–851.

Large-Format 2-D Polyacrylamide Gel Electrophoresis

Henry Brzeski, Stephen Russell, Anthony G. Sullivan,
Richard I. Somiari, and Craig D. Shriver

1. Introduction

Proteins are composed of different numbers of neutral, positively and negatively charged amino acids. Therefore, proteins vary widely in size and have either positive, negative, or zero net charge, depending on the pH of their surroundings. The original two-dimensional gel electrophoresis format was developed almost 30 years ago (1) to exploit this variation in protein charge and size for separation purposes. The isoelectric point of a protein (pI) is the pH at which it has a net zero charge, which, for the majority of proteins, lies between pH 4.0 and 8.0. The sizes of proteins vary widely (10–500 kDa), with an average molecular weight of approx 50 kDa. These two mutually independent properties are exploited by firstly denaturing proteins in urea and then subjecting them to an electric field in a pH gradient established in a low-concentration polyacrylamide gel (originally this pH gradient was formed, using ampholytes, *in situ*, but now precast immobilized pH gradient [IPG] strips [2–4] are found to be more reliable). In this case, all but the very largest proteins can migrate freely until they reach a pH at which they have no net charge (isoelectric focusing [IEF]). After completion of the focusing, the proteins are denatured *in situ*, their native charge is saturated with the anionic detergent sodium dodecyl sulphate (SDS), and then the gel is layered, perpendicular to the direction of focusing, on a higher-concentration polyacrylamide gel, and the focused proteins are separated on the basis of size. This gives rise to discrete spots representing one (or perhaps a very small number) of different proteins.

Despite the many advances and improvements in alternative separation technologies over the intervening years, two-dimensional polyacrylamide gel electrophoresis is still the method of choice to investigate a cell's/organism's proteome, as it is capable of resolving perhaps 3000 (5) individual protein spots on a single gel. This capability has led to its almost universal use to resolve the multitude of different proteins in a cell or tissue. Since the spots (features) identified are now essentially pure, these spots can be picked (purified), proteolytically or chemically digested, and the resulting peptides used to identify the protein(s) present in the original spot using a mass spectrometric technique (see Chapter 35). In all cases, the downstream processes rely absolutely on the ability to obtain maximum resolution of as many proteins as possible over the two-dimensional gel to reduce the possibility of cross-contamination with closely migrating protein spots.

The technique can be modified to suit a particular experimental system by (a) using different pH gradients in the first dimension (a single wide range or multiple narrow pH ranges) to resolve proteins with different pIs and (b) using high- and/or low-concentration second-dimension polyacrylamide gels or even gradient gels to resolve proteins of different size ranges.

In order to reduce the inherent variability of the system, various technological improvements have been made over the years, and this chapter will concern itself with describing these improvements to indicate the procedures that now make it possible to easily generate reproducible two-dimensional polyacrylamide gels for comparative proteomics.

2. Materials

1. Trichloroacetic acid (TCA)/acetone protein precipitation: 2-D Cleanup Kit, (Amersham Biosciences, Piscataway, NJ).
2. Lysis buffer (for protein solubilization prior to isoelectric focusing): 30 mM Tris-HCl (pH 8.5) at 4°C, 2 M thiourea (see **Note 1**), 7 M urea, 4% (w/v) CHAPS, 1% NP-40. (See Chapters 1–13 for alternative methods of preparing and solubilizing samples from other biological sources prior to use in two-dimensional polyacrylamide gel electrophoresis.)
3. IEF rehydration buffer (for rehydration of dried, precast IPG strips): 7 M urea, 2 M thiourea, 4% (w/v) CHAPS, 10% (v/v) isopropanol, 5% (v/v) glycerol, 1.2% (v/v) DeStreak (Amersham Biosciences), 1% (v/v) ampholytes with a pH range that is comparable to the pH range of the IPG strip (see **Note 2**), 2 mg/mL dithiothreitol (DTT), 0.002% bromophenol blue solution. The bromophenol blue acts as a visual indicator of the progress of focusing.
4. 2X sample buffer (for correction of cell lysis buffer composition prior to IEF): 7 M urea, 2 M thiourea, 4% (w/v) CHAPS, 1% NP40, 2.4% (v/v) DeStreak (Amersham Biosciences), 2% (v/v) ampholytes with a pH range that is comparable to the pH range of the IPG strip (see **Note 2**). If DeStreak is not used, then 3 mg/mL of DTT should be added to this buffer.
5. SDS equilibration buffer (for equilibration of focused IPG strips prior to second-dimension polyacrylamide gel electrophoresis): 50 mM Tris-HCl (pH 8.8), 6 M urea, 30% (v/v) glycerol, 2% (w/v) SDS, 0.002% (w/v) bromophenol blue.
6. SDS equilibration buffer A (for reduction of proteins prior to IEF): Immediately before use, add 50 mg of DTT to 10 mL of SDS equilibration buffer.
7. SDS equilibration buffer B (for alkylation of proteins after reduction prior to IEF): Immediately before use, add 450 mg of iodoacetamide to 10 mL of SDS equilibration buffer.
8. Agarose sealing solution (for sealing the IPG strip to the second dimension gel): 1X SDS electrophoresis running buffer, 0.5% w/v agarose, 0.002% bromophenol blue. Dispense the individual components into a conical flask and heat gently until the agarose has dissolved completely.
9. Gel storage solution (for storage of second-dimension gels after electrophoresis): 0.375 M Tris-HCl, pH 8.8.
10. 4X resolving gel buffer (for preparation of second-dimension gel): 1.5 M Tris-HCl, pH 8.8.
11. Stock acrylamide solution (for preparation of second-dimension gels): 40% (w/v) acrylamide, 3% (w/v) *N,N'*-methylenebisacrylamide.
12. 10% (w/v) sodium dodecyl sulfate (SDS) (for preparation of second-dimension gel).
13. 10% (w/v) ammonium persulfate (APS) (for preparation of second-dimension gel).
14. 1X SDS electrophoresis running buffer (for electrophoresis of second-dimension gel): 25 mM Tris, 192 mM glycine, 0.2% SDS. This solution does not require pH adjustment.

15. Water-saturated butanol (for second-dimension gel overlay): Add 5 mL of water to 50 mL of *n*- or *t*-butanol and shake vigorously. Allow the two phases to separate and use the upper (butanol) layer.
16. Gel fixing solution: 7% (v/v) acetic acid, 10% (v/v) methanol.
17. Molecular-weight markers (for second-dimension polyacrylamide gel electrophoresis): The molecular-weight markers purchased for use in this laboratory are supplied in the correct buffer for use in polyacrylamide gel electrophoresis. When required, the molecular-weight marker is thawed and then 10 μ L is added directly to a 1 cm \times 2 cm piece of filter paper just prior to loading onto the gel.
18. Glass plate cleaning solution: 5% (v/v) Contrad 70 (manufactured by Decon Laboratories and distributed by Fisher Scientific).

3. Methods

This laboratory carries out a high-throughput approach to proteomic analysis and routinely resolves proteins from tissue-culture cells, plasma, serum, and human tissue samples. Each of these has its own preferred method of sample preparation, depending on the protein population, ease of protein solubilization, and so on. Procedures for preparing protein lysates from a number of different biological sources prior to electrophoresis on two-dimensional gels can be found in other chapters in this volume (*see* Chapters 1–13).

Although it was common in the past to radioactively label proteins, this is not often used today because of issues with (a) safety regulations, which restrict exposure and disposal, and (b) impracticability of metabolically labeling proteins from tissue samples using radioactive amino acids. However, there are a number of fluorescent labels that can be used to tag proteins either prior to electrophoresis, using Cy3 and Cy5 (*see* Chapter 28), or afterwards, using the SYPRO dyes (*see* Chapters 20–23).

The procedures for large-format 2-D polyacrylamide gel electrophoresis are broken down below into multiple steps that involve sample solubilization, sample application to the IPG strip, isoelectric focusing, equilibration of the IPG strip with second-dimension buffer, loading of the IPG strip onto the second-dimension gel, and second-dimension gel electrophoresis. Each one of these steps is explained specifically below. We routinely run multiple two-dimensional polyacrylamide gels at one time. The samples are usually differentially labeled using Cy3 or Cy5, then mixed prior to isoelectric focusing. After electrophoresis in the second dimension the gels are scanned for the presence of fluorescent spots, and then areas of interest are picked, treated with different proteases, and the resulting peptides subjected to matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) or MS/MS (Q-ToF) analysis. The procedures reported below have been written to be essentially generic, although some steps will be equipment specific. Manufacturers of other equipment will provide their own details that may be pertinent to their own technologies. This laboratory uses precast first-dimension IPG strips that are focused in an Ettan IPGphor isoelectric system, which provides an integrated power supply and cooled electrophoresis plate. This laboratory also prepares its own second-dimension polyacrylamide gels and carries out electrophoresis in an apparatus manufactured by Hoefer and sold by Amersham Biosciences.

The following descriptions refer to the preparation and electrophoresis of 24 cm \times 20 cm two-dimensional polyacrylamide gels.

First, an idea of the time scale involved in obtaining a result from a large format 2-D gel analysis—all times are approximate. Sample preparation—1 h; equilibration of sample with first-dimension gel—12 h; focusing of first-dimension gel—12 h; casting a second-dimension gel (homogeneous or a gradient)—1 h; polymerization of a second-dimension gel—12 h (overnight); fixing of gel—30 min to overnight (see Note 15); staining a second-dimension gel—2 to 5 h, depending on the staining procedure used (silver, Coomassie blue, deep purple, and so on). Visualization of protein spots depends very much on the staining procedure—i.e., if the dyes are fluorescent, then scanning dyes at high resolution takes 20 min per dye for a 24 cm × 20 cm gel (Cy2, Cy3, Cy5, or deep purple). If the stain is visible (silver, Coomassie blue, deep purple, and so on) it will require 10 min to take a photo with a Polaroid or digital camera.

3.1. Sample Preparation

1. Instructions for obtaining protein samples from different biological sources have already been presented in other chapters (1 to 13) in this volume. IPG strips have a high protein-loading capacity (6), and we have loaded up to 500 µg of protein per strip so as to obtain mass spectrometric IDs from proteins that are present at very low concentrations (some manufacturers suggest that this loading limit can be increased to 1 mg of protein per 24-cm IPG strip). This laboratory routinely carries out protein loading during rehydration of the IPG strip (7,8).
2. Although it is possible to obtain resolution of the large majority of proteins present in a cell using a single 3–10 or 3–10NL (see Note 3), we routinely use both pH 3.0–7.0 and pH 6.0–11.0 IPG strips for optimal separation of proteins (see Notes 4 and 5).
3. Calculate the volume of sample solution that will contain the required amount of protein.
4. Add an equal volume of 2X sample buffer.
5. Make the total volume up to 450 µL with rehydration buffer.
6. If the protein solution is too dilute or contains excess salt (see Note 7), then the protein should be acetone precipitated and the pellet dissolved in 450 µL of rehydration buffer.
7. Generally, we load 100 µg of protein per IPG strip, but if the required proteins are present at low concentration in the sample, it is possible to load up to 1 mg of total protein per strip.

3.2. Isoelectric Focusing

Sample protein can be loaded in one of three ways. This laboratory routinely loads protein onto IPG strips during rehydration of the dried strip, but it is also possible to rehydrate the IPG strip without applying a voltage and subsequently load the sample using loading cups or paper wicks.

Rehydration and sample absorption is performed in a specially designed ceramic tray, which allows effective thermal contact with the cooled base of the electrophoresis apparatus so that the sample can be focused at very high voltage after absorption. This allows the user to perform concurrent rehydration, sample absorption, and focusing in individual ceramic strip holders (see Note 7).

3.2.1. Preparation of Focusing Apparatus

1. Identify an IPG strip of the required length (24 cm) and pH range.
2. The IPG strip package is opened starting at the acidic (pointed) end, as this end of the gel is usually harder and less easily damaged.
3. Position the IPG strip alongside the ceramic strip holder with the gel side down and the acidic (pointed) end aligned so as to fit into the pointed end of the ceramic strip holder.

3.2.2. Rehydration and Equilibration of Sample With IPG Strips

1. Add 450 μ L of rehydration buffer (containing sample; **Subheading 3.1., step 4**) in the middle of each 24-cm IPG ceramic strip holder and allow the solution to distribute itself over the entire strip holder.
2. Lower the IPG strip onto the rehydration solution containing the sample, starting at the pointed end. The strip should be gently manipulated backwards and forwards to ensure the whole strip is coated and evenly wetted, taking great care not to trap any air bubbles beneath the gel.
3. Ensure that the gel makes contact with the electrodes in the strip holder at each end. Add cover fluid to each strip holder to prevent evaporation and subsequent crystallization of urea.
4. Place the plastic cover over the strip holder and position it in the isoelectric focusing apparatus.
5. Ensure the strip holders are parallel to, and contact, the electrodes.
6. Ensure the pointed end of the IPG strip is at the anode (positive electrode).

3.2.3. Focusing Voltages and Times

We minimize gel swelling and maximize protein uptake (*see Note 2*) into the IPG strip using the following program (*see Note 7*):

1. Allow the gel to rehydrate and sample to absorb for 12 h at 30 V.
2. Increase voltage to 200 V and maintain for 1 h.
3. Increase voltage to 500 V and maintain for 1 h.
4. Increase voltage to 1000 V and maintain for 1 h.
5. Ramp voltage using a 1000–8000 V linear gradient for 3 h
6. Continue focusing at 8000 V for a total (for steps 1–6 in **Subheading 3.2.3.**) of 65,500 V·h. This normally takes about 12 h (for **Subheading 3.2.3., steps 2–6**) but may take longer if there is excess salt in the focusing buffer (*see Note 8*).

Once focusing has been completed, there will be very little movement of ions, and this results in a very low current (*see Note 7*) after about 12 h. However, if the sample contained a large amount of salt, then this will not occur, and complete focusing will not take place unless the focusing time is extended.

3.2.4. Removing IPG Strips From Ceramic Strip Holders

1. Use forceps to remove the IPG strip from the ceramic strip holder.
2. Drain the excess buffer and cover fluid by touching the plastic backing against absorbent, lint-free cloth.
3. If the IPG strip will not be used immediately, then it may be stored at -80°C for up to 1 mo. The strips should be stored in a rigid container, as the frozen strips are very brittle and can be easily broken.

3.3. Preparing IPG Gels for Second-Dimension Electrophoresis

3.3.1. First Equilibration (Reduction)

The purpose of this step is to produce discrete subunits from disulfide linked, multimeric proteins by reducing the cystine disulphide bridges that covalently link them.

1. After draining of cover fluid, or thawing, place each IPG strip in a separate tube, ensuring the plastic backing support is against the wall so that the IPG gel maintains contact with the equilibration buffer.

2. Add 15 mL of SDS equilibration buffer A to each tube, cap it, and place it on a shaking platform for 15 min. Alternatively, each IPG strip can be bent and placed in a 15-cm plastic Petri dish so that the plastic backing film is against the wall of the Petri dish. Add 10 mL of SDS equilibration buffer A to each dish and place it on a shaking platform for 15 min.

3.3.2. Second Equilibration (Alkylation)

The DTT reduction is reversible, so it is necessary to alkylate the reduced cysteines to prevent re-formation of random high-molecular-weight protein aggregates by re-polymerization of reduced cysteines.

1. Remove the SDS equilibration buffer A from the tube or Petri dish, add 15 mL/10 mL of SDS equilibration buffer B, and place it on a shaking platform for a further 15 min.

3.4. Preparing Gel Molds for Electrophoresis of Proteins in the Second Dimension

Note: In this section, the procedure uses acrylamide solutions for preparation of polyacrylamide gels. Acrylamide is a known neurotoxin, and suitable precautions should be taken at all times.

The original procedure that was reported by O'Farrell (1) was based on a discontinuous system originally developed by Laemmli (9). This protocol used a discontinuous buffer system with stacking and resolving gels. This is no longer required for two-dimensional polyacrylamide gel electrophoresis.

This laboratory routinely uses 1-mm-thick 24 cm × 20 cm second-dimension polyacrylamide gels for both staining and for picking spots for digestion and protein identification by mass spectrometry.

1. The glass plates are first wiped with ethanol or isopropanol using a lint-free cloth to remove dust and fibers.
2. The second-dimension gel molds are assembled using a back plate (which has 1-mm spacers permanently fixed to the plate).
3. Smear the spacer with a thin film of silicone grease (see **Note 6**)
4. The slightly smaller front plate (which has no spacers) is then aligned with the back plate.
5. The correctly aligned plates are then inserted into the multiple gel caster.

3.5. Casting Second-Dimension 24 × 20 cm Slab Gels

It is possible to purchase precast second-dimension polyacrylamide gels; however, this laboratory routinely prepares them as and when required. This laboratory routinely detects proteins using fluorescent dyes, and, at the time of writing, it is not possible to obtain precast gels with a nonfluorescent backing. Empty gel molds are assembled as described in **Subheading 3.4**.

3.5.1. Multiple Gel Caster

1. The preassembled gel molds (**Subheading 3.4**) are inserted into the gel caster. Any unused spaces are filled with blanks to avoid excessive use of acrylamide solution.
2. The instructions below can be used to prepare homogeneous (**Subheading 3.5.2**) or gradient polyacrylamide gels (**Subheading 3.5.3**).

3.5.2. Homogeneous Gels

3.5.2.1. PREPARE SOLUTIONS FOR A 12% ACRYLAMIDE GEL

1. Calculate the total volume of acrylamide solution required for the number of gels and the spacer thickness (the manufacturer of your gel-casting system will provide these volumes).

2. Prepare the multiple gel caster and gel molds (**Subheadings 3.5.1. and 3.4.**).
3. In this laboratory the following would suffice for the preparation of fourteen homogeneous 12% polyacrylamide gels:

<i>Solution</i>	<i>Volume</i>
Stock acrylamide	270 mL
4X resolving gel buffer	225 mL
10% ammonium persulfate	9 mL
De-ionized water to 900 mL	

4. Mix and add the entire solution to a vacuum flask.
5. Seal the opening of the flask and apply a vacuum while stirring on a magnetic stirrer. This will remove all dissolved oxygen, which would inhibit acrylamide polymerization.
6. Continue to de-gas until no more bubbles are seen, then carefully release the vacuum.
7. Add 276 μ L of TEMED to initiate acrylamide polymerization.
8. Swirl the flask gently to mix the acrylamide and catalysts, taking care not to introduce any more oxygen (*see Note 9*).
9. Proceed to **Subheading 3.5.2.2.** immediately.

3.5.2.2. POURING HOMOGENEOUS 12% POLYACRYLAMIDE GELS

1. Pour (*see Note 9*) the acrylamide solution prepared in **Subheading 3.5.2.1.** into the reservoir in the gel caster and allow the solution to fill the empty molds, from the bottom, until the solution is about 5 mm below the top of the glass plate.
2. Immediately after pouring the gel, overlay the solution with water-saturated butanol to prevent exposure of the acrylamide to oxygen and so create a flat gel surface (oxygen will inhibit polymerization; *see Note 9*).
3. Allow the gel to polymerize (overnight), dismantle the multiple gel caster, discard excess polyacrylamide gel, remove the butanol by inverting the gel mold, and rinse the gel surface with gel storage solution.
4. If the gel is not required immediately, it can be stored for up to 2 wk by submerging the entire glass mold and gel in gel storage solution at 4°C.
5. Load the IPG strip onto the top of the second-dimension gel (**Subheading 3.6.**)

3.5.3. Gradient Gels

3.5.3.1. PREPARING THE GRADIENT MAKER

1. Assemble the gradient maker.
2. Connect the tubing from the outlet of the gradient maker to the pump.
3. Connect the tubing from the pump to the inlet at the base of the gel caster.

3.5.3.2. PREPARE SOLUTIONS FOR A 4% ACRYLAMIDE GEL

Ensure that both high and low concentration acrylamide solutions for gradient gels have been prepared prior to initiating polymerization.

1. Calculate the total volume of acrylamide solution required for the number of gels and the spacer thickness (the manufacturer of your gel casting system will provide these volumes) and divide this volume by 2 (for linear gradient gels).
2. Prepare the multiple gel caster (**Subheading 3.5.1.**), gel molds (**Subheading 3.4.**) and gradient maker (**Subheading 3.5.3.1.**).
3. In this laboratory the following would suffice for the preparation of fourteen 4%–20% gradient polyacrylamide gels.

<i>Solution</i>	<i>Volume</i>
Stock acrylamide	45.0 mL
4X resolving gel buffer	112.5 mL
10% ammonium persulfate	4.5 mL
De-ionized water to 450 mL	

4. Mix and add the entire solution to a vacuum flask.
5. Seal the opening of the flask and apply a vacuum while stirring on a magnetic stirrer. This will remove all dissolved oxygen, which would inhibit acrylamide polymerization.
6. Continue to de-gas until no more bubbles are seen, then carefully release the vacuum.
7. Prepare and de-gas both high- and low-concentration acrylamide solutions to be used for gradient gels before proceeding with **step 8**.
8. Add 96 μ L of TEMED to initiate acrylamide polymerization.
9. Swirl the flask gently to mix the acrylamide and catalysts, taking care not to introduce any more oxygen (*see Note 9*).
10. Proceed to **Subheading 3.5.3.4.** immediately.

3.5.3.3. PREPARE SOLUTIONS FOR A 20% ACRYLAMIDE GEL

Ensure that both high- and low-concentration acrylamide solutions for gradient gels have been prepared prior to initiating polymerization.

1. Calculate the total volume of acrylamide solution required for the number of gels and the spacer thickness (the manufacturer of your gel casting system will provide these volumes) and divide this volume by 2 (for linear gradient gels).
2. Prepare the multiple gel caster (**Subheading 3.5.1.**), gel molds (**Subheading 3.4**) and gradient maker (**Subheading 3.5.3.1.**).
3. In this laboratory, the following would suffice for the preparation of fourteen 4%–20% gradient polyacrylamide gels:

<i>Solution</i>	<i>Volume</i>
Stock acrylamide	225.0 mL
4X resolving gel buffer	112.5 mL
Glycerol	31.0 mL
10% ammonium persulfate	2.3 mL
De-ionized water to 450 mL	

4. Mix and add the entire solution to a vacuum flask.
5. Seal the opening of the flask and apply a vacuum while stirring on a magnetic stirrer. This will remove all dissolved oxygen, which would inhibit acrylamide polymerization.
6. Continue to de-gas until no more bubbles are seen, then carefully release the vacuum.
7. Prepare and de-gas both high- and low-concentration acrylamide solutions to be used for gradient gels before proceeding with **step 8**.
8. Add 13 μ L of TEMED to initiate acrylamide polymerization.
9. Swirl the flask gently to mix the acrylamide and catalysts, taking care not to introduce any more oxygen (*see Note 9*).
10. Proceed to **Subheading 3.5.3.4.** immediately.

3.5.3.4. POURING LINEAR 4%–20% GRADIENT GELS

1. Pour (*see Note 9*) the 4% acrylamide solution prepared in **Subheading 3.5.3.2.** into that part of the gradient gel maker allocated for the low-concentration solution.
2. Pour (*see Note 9*) the 20% acrylamide solution prepared in **Subheading 3.5.3.3.** into that part of the gradient gel maker allocated for the high-concentration solution.
3. Start the pump and allow the solution to fill the empty molds, from the bottom, until the solution is about 5 mm below the top of the glass plate (*see Note 12*).
4. Immediately after pouring the gel, overlay the solution with water-saturated butanol to prevent exposure of the acrylamide to oxygen and so create a flat gel surface (oxygen will inhibit polymerization—*see Note 9*).
5. Allow the gel to polymerize (overnight), dismantle the multiple gel caster, discard excess polyacrylamide gel, remove the butanol by inverting the gel mold, and rinse the gel surface with gel storage solution.

6. If the gel is not required immediately it can be stored for up to 2 wk by submerging the entire glass mold and gel in gel storage solution at 4°C.

3.6. Loading the IPG Strip onto the Second-Dimension Gel

1. Prior to loading the IPG strips onto the second-dimension gel, prepare pieces of filter paper containing the molecular-weight markers.
2. Remove the gel storage solution from the top of the second-dimension gel by inversion.
3. Add 2 mL of agarose sealing solution.
4. Push one end of the IPG strip down into the molten agarose until it touches the surface of the second-dimension gel.
5. Push the other end of the IPG strip down slowly, ensuring the air bubble beneath the IPG strip is extruded slowly from below the strip.
6. Ensure the IPG strip is in intimate contact with the surface of the second-dimension polyacrylamide gel (see Note 14).
7. Push the filter paper containing the molecular-weight markers prepared in materials step 18 into the molten agarose at one end of the IPG strip.

3.7. Electrophoresis Conditions

1. Assemble the second-dimension electrophoresis tank according to the manufacturer's instructions.
2. Electrophorese the gel overnight at 2 W (constant watts) per gel.
3. Discontinue electrophoresis after the bromophenol blue has migrated from the end of the gel.

3.8. Dismantling the Polyacrylamide Gel Mold

1. After electrophoresis, the electrophoresis apparatus is dismantled.
2. Each gel mold is removed from the apparatus.
3. The gel plates of each mold are separated by insertion of a plastic wedge-shaped device between the two glass plates and prying them apart.
4. At this point, the gel can be processed in many different ways, depending on the samples loaded on the gel and their future use. The gel could be simply stained with Coomassie blue, silver (these will require that the gel be fixed for at least 30 min, or it can be left overnight [see Note 15], then transferred to water and stored in a refrigerator), or a fluorescent dye—e.g., one of the SYPRO range of dyes (see Chapters 20–23). The protein samples may have been prestained with Cy3 and Cy5, and could be examined for differential expression between samples (see Chapter 24). The stained proteins could be visualized after electrophoresis, spots of interest identified, picked, digested chemically or proteolytically, and the resulting peptides identified using a MALDI-TOF or MS/MS (Q-TOF). Many of these techniques are discussed in other chapters, and the reader is directed to these chapters for further information.

3.9. Cleaning Glass Plates

In this laboratory the gels are scanned for fluorescently labeled proteins, and so it is imperative that the glass plates not be scratched, since this interferes with fluorescent detection. For this reason, the cleaning of glass plates is carried out in a manner which is nonabrasive. All residual acrylamide is removed from the glass plates, and then they are submerged in glass plate cleaning solution and allowed to soak overnight. The plates are then gently wiped with a lint-free cloth, rinsed in de-ionized water, and allowed to air dry.

3.10. Comparison of Results From Linear and Gradient Gels

A T-cell leukemia cell line (CEM/C2) was harvested, washed three times with PBS. The pellet was dissolved in lysis buffer, and duplicate samples were subjected to isoelectric focusing as described in **Subheading 3.2**. After equilibration (**Subheading 3.3.**) one IPG strip was layered onto a homogeneous 12% polyacrylamide gel (**Subheading 3.5.2.**) and the second was layered onto a 4% to 20% linear gradient gel (**Subheading 3.5.3.**). The results from these two gels can be seen in **Fig. 1A** (homogeneous 12% gel) and **Fig. 1B** (4%–20% linear gradient gel). A linear 4%–20% gradient was used to clearly illustrate the advantages of using a low-concentration gel to visualize high-molecular-weight proteins.

Examination of these gels highlights a number of interesting points that illustrate the advantages and disadvantages of using a gradient gel.

1. The molecular-weight markers show that the gradient gel (**Fig. 1B**) has a different molecular-weight distribution of proteins, and the high-molecular-weight region occupies a larger fraction of the gel.
2. **Fig. 1B** shows the presence of many faint but visible spots derived from high-molecular-weight proteins (>220 kDa) that are not visible in the homogeneous 12% gel.
3. **Fig. 1B** shows the presence of more proteins with low molecular weight (approx 15 kDa) that were not retained on a homogeneous 12% gel.
4. The spot pattern is tighter on a gradient gel, so that more spots are visible.

However, the gradient gel has two disadvantages. Firstly, the exact shape of the gradient will alter the spot distribution over the gel, making reproducibility of gradient formation an issue. **Figure 1B** was prepared as a linear gradient so that the high-molecular-weight proteins were more easily visualized. However, this will result in the compression of the lower-molecular-weight proteins into a smaller fraction of the gel. For this reason, it will be necessary to use a number of test experiments before the optimal initial and final acrylamide concentrations are found. In addition, it will require multiple trial experiments that vary the shape of the gradient (convex exponential/concave exponential) to determine the optimal gradient shape for the experiment concerned. Having once determined what these conditions are, it will be essential to define them so that they are reproducible from week to week. In addition, the same caveat holds for gradient gels as for wide-range IPG strips (*see Note 4*). If more protein spots are visible on a gel, then there will be many more spots which co-electrophorese. Therefore, the increase in resolution may be illusory, and this co-electrophoresis will become more apparent after high-resolution MS/MS attempts to identify the protein spot, since such techniques readily determine the presence of peptides derived from two, three, or even four proteins.

4. Notes

1. Thiourea provides a more effective method for solubilization of proteins in the tissue samples (**10–12**) regularly processed by this laboratory.
2. Although ampholytes are not required for the generation of a pH gradient during isoelectric focusing, they help to maintain proteins in solution by preventing interactions between proteins that lead to precipitation.
3. 3–10 NL contains a nonlinear pH gradient. This ensures that the majority of proteins in a cell (which have pIs that are generally in the 4–8 range) will be resolved optimally; nevertheless, it is possible to view proteins with pIs up to 10.

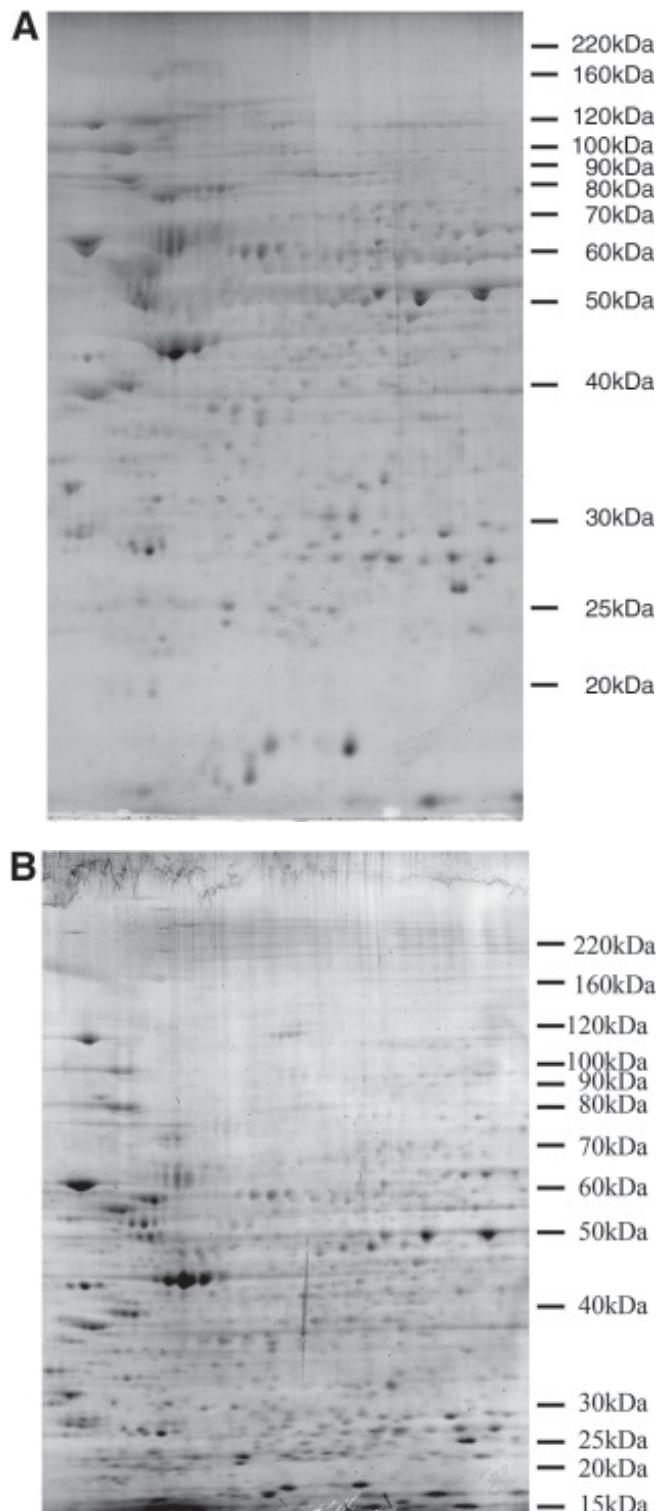


Fig. 1. Comparison of linear and gradient second dimension gels. Duplicate samples from CEM/C2 cells were focused in parallel on wide range pH 3.0–10.0 IPG strips and then run on either a 12% polyacrylamide gel (A) or a 4–20% linear gradient gel (B).

4. The choice of the pH range of the IPG strip used for isoelectric focusing is a compromise between resolving the maximum number of protein features on a single wide-range first-dimension IPG strip (which gives rise to a low-resolution gel in which many features overlap in the second-dimension gel) and running multiple narrow-range IPG strips (which provide a better resolution of the spots, which will not overlap to the same extent as seen in a single gel), but this will require splitting one sample onto two to five first-dimension strips and second-dimension gels.
5. Even under these conditions, it is not unusual to find two or even three protein IDs in a single spot after MS/MS analysis.
6. Sometimes there is a slight gap between the spacer and the front plate, and this allows a small current to flow through this gap. This small current usually does not distort the 2-D gel, but often results in distortion of the molecular-weight standards, which run closer to the edge of the mold.
7. Limit the current to 50 μ A per strip. This prevents burning of the electrodes, which is caused by too high a current flow during isoelectric focusing. The reason for the excessive current is that the sample has too high an ionic strength, derived from high-salt buffers. By limiting the current draw to 50 μ A/IPG strip, the electrodes will not overheat; however, if the ionic strength is too high, then the 50- μ A limit will reduce the voltage, and the time taken to reach the 65,500 V/h limit will be longer (under normal conditions this takes 12 h after the equilibration step). In this case, focusing should be continued longer, until the current drops to 35 μ A per strip.
8. The salt concentration in the sample should be maintained at as low a value as possible so as not to interfere with isoelectric focusing. However, if the salt concentration in the sample buffer used for protein preparation is excessive, it may be necessary to exchange buffers by precipitating the protein or using a molecular-weight cut-off filter.
9. After de-gassing the acrylamide solution used for preparation of second-dimension gels, it is essential to remember not to shake or pour the solution too violently, as this will re-introduce air bubbles (oxygen), which will reduce polymerization and affect spot mobilities.
10. Always wear gloves when handling IPG strips, SDS polyacrylamide gels, or glass plates—firstly, because the stains used for protein detection will pick up fingerprints, and secondly, because if these gels will be used to isolate proteins for mass spectrometric analysis, this will reduce protein contamination that can cause keratin IDs.
11. Perform sample loading under a low-voltage program. It has been reported that low voltages allow high-molecular-weight proteins to enter the IPG strip more effectively (13).
12. Carry out a trial run to ensure the high- and low-concentration acrylamide solutions are in the correct sections of the gradient maker, so that the gel is 4% acrylamide at the top and 20% acrylamide at the bottom.
13. Care must be exercised if the gel will be silver stained afterwards, as ampholytes interfere with this method of staining. It will be necessary to ensure that all residual ampholytes have been removed from the gel by extensive washing.
14. Ensure that no small air bubbles are trapped between the IPG strip and the top of the second-dimension gel, as these will distort the electric field and give rise to uneven spot distribution.
15. If a second-dimension gel requires fixing before staining, then it may be advantageous to use a longer fixing time, as it has been reported that an overnight fixing step removes more SDS and reduces the background.

References

1. O'Farrell, P. H. (1975) High resolution two-dimensional electrophoresis of proteins. *J. Biol. Chem.* **250**, 4007–4021.
2. Westermeier, R., Ek, K., Righetti, P. G., Gianazza, E., Görg, A., and Postel, W. (1982) Isoelectric focusing in immobilized pH gradients: principle, methodology and some applications. *J. Biochem. Biophys.* **6(317)**, 339.
3. Görg, A., Postel, W., Günther, S., and Weser, J. (1985) Improved horizontal two-dimensional electrophoresis with hybrid isoelectric focusing in immobilized pH gradients in the first dimension and laying-on transfer to the second dimension. *Electrophoresis* **6**, 599–604.
4. Görg, A., Postel, W., and Günther, S. (1988) The current state of two-dimensional immobilized pH gradients. *Electrophoresis* **9**, 531–546.
5. Lilley, K. S., Razzaq, A., and Dupree, P. (2002) Two-dimensional gel electrophoresis: recent advances in sample preparation, detection and quantitation. *Curr. Opin. Chem. Biol.* **6(1)**, 46–50.
6. Bjellqvist, B., Sanchez, J.-C., Pasquali, C., et al. (1993) Micropreparative two-dimensional electrophoresis allowing the separation of samples containing milligram amounts of proteins. *Electrophoresis* **14(1375)**, 1375–1378.
7. Sanchez, J.-C., Rouge, V., Pisteir, M., et al. (1997) Improved and simplified in-gel sample application using reswelling of dry immobilized pH gradients. *Electrophoresis* **18**, 324–327.
8. Rabilloud, T., Valette, C., and Lawrence, J. J. (1994) Sample application by in-gel rehydration improves the resolution of two-dimensional electrophoresis with immobilized pH gradients in the first dimension. *Electrophoresis* **15**, 1552–1558.
9. Laemmli, U. K. (1970) Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **227**, 680–685.
10. Molloy, M. P., Herbert, B. R., Walsh, B. J., et al. (1998) Extraction of membrane proteins by differential solubilization for separation using two-dimensional gel electrophoresis. *Electrophoresis* **19**, 837–844.
11. Rabilloud, T., Adessi, C., Giraudel, A., and Lunardi, J. (1997) Improvement of the solubilization of proteins in two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis* **18**, 307–316.
12. Rabilloud, T. (1998) Use of thiourea to increase the solubility of membrane proteins in two-dimensional electrophoresis. *Electrophoresis* **19**, 759–760.
13. Görg, A., Boguth, G., Obermaier, C., Harder, A., and Weiss, W. (1998) 2-D electrophoresis with immobilized pH gradients using IPGphor isoelectric focusing system. *Life Science News* **1**, 4–6.

Analysis of Membrane Proteins by Two-Dimensional Gels

Michael Fountoulakis

1. Introduction

Separation of a protein mixture by two-dimensional (2-D) gel electrophoresis is usually the first step of a proteomic analysis. The second step is the protein identification by mass spectrometry techniques, mainly by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS). Two-dimensional electrophoresis comprises two steps (dimensions): (1) separation of the proteins on the basis of differences in their net charge, called isoelectric focusing (IEF), which is usually performed on immobilized pH gradient (IPG) strips; and (2) separation of the focused proteins on the basis of differences in their molecular masses, which is performed in sodium dodecyl sulfate (SDS) polyacrylamide gels. The major advantage of 2-D electrophoresis is that it enables the simultaneous separation and visualization of thousands of unknown protein forms. No other method can do that at the present time. On the other hand, protein detection in 2-D gels is limited because (1) the major components of a protein mixture are usually visualized and (2) the detection of the low- and high-molecular-mass proteins, as well as of the basic and hydrophobic proteins, is inefficient.

Detection of membrane proteins is of particular interest as they exert important functions in signal transduction pathways, ion transport, cell–cell interaction, and other processes. Many of them are targets for protein interactions or drugs, and consequently of scientific and medical interest. However, detection of hydrophobic and membrane proteins in 2-D gels is associated with certain limitations, and a relatively low number of real hydrophobic and membrane proteins have been detected in 2-D gels (1–4). Using current 2-D electrophoresis products—i.e., nonionic or zwitterionic detergents like CHAPS and nonionic, mild chaotropes like urea or thiourea—mainly the abundant, hydrophilic components of a protein mixture are resolved and visualized (4–6) (see Note 1). In order to overcome this major limitation of proteomics, a large variety of solubilizing agents (4,6–11) and variations in the 2-D electrophoresis technology (12) have been suggested.

In general, for a successful proteomic analysis of membrane proteins (and of all proteins in a broader sense), several prerequisites must be fulfilled: (1) a protein should be brought and kept in solution during the whole two-dimensional separation process; (2) it should belong to the category of proteins that can be separated by 2-D gels—i.e., it should have average pI (usually between pH 4.0 and 10.0) and molecular mass

(between 10 and 120 kDa) values and should not include strong hydrophobic stretches; and (3) it should be present in a minimal concentration, so that the corresponding spot(s) are visible in Coomassie blue-stained gels and identification by mass spectrometry is possible (13–15). Here we will mainly deal with case (1), i.e., the preparation of protein samples for separation by 2-D electrophoresis and the limitations of the current technologies.

2. Materials

2.1. Preparation of Membrane Proteins

1. Lysis buffer A: 20 mM HEPES-NaOH (pH 8.0), containing 10 % glycerol, 150 mM NaCl, 1 mM MgSO₄, 100 units/mL Trasylol®, 5 mM ϵ -aminocaproic acid, and 250 units/mL benzonase.
2. 0.5 M ethylenediaminetetraacetic acid (EDTA) stock solution: suspend EDTA powder in water, neutralize with 10 M NaOH, and adjust to the required volume with water.
3. Dialysis buffer B: 5 mM HEPES-NaOH (pH 8.0), containing 10 % glycerol and 1 mM EDTA-Na₂.
4. Suspension buffer C: 10 mM HEPES-NaOH (pH 8.0), 1 M LiCl, 10% glycerol (v/v), and 1 mM EDTA-Na₂.
5. Hypertonic suspension buffer D: 5 mM HEPES-NaOH (pH 8.0), 20% sucrose (w/v), 10% glycerol (v/v), and 1 mM EDTA-Na₂.
6. Hypotonic suspension buffer E: 5 mM HEPES-NaOH (pH 8.0), 10% glycerol (v/v), and 1 mM EDTA-Na₂.
7. IEF-compatible solubilization buffer F: 20 mM Tris, 7 M urea, 2 M thiourea, 4% CHAPS, and 10 mM dithioerythritol.
8. Strong solubilization buffer G: 10 mM HEPES-NaOH (pH 8.0), containing 1% SDS, 10% glycerol (v/v), 1 mM EDTA, 100 units/mL Trasylol, and 5 mM ϵ -aminocaproic acid. Instead of SDS, the buffer can contain either 1% lithium dodecyl sulfate (LDS), 0.5% sodium cholate, or 6 M guanidine hydrochloride.
9. Ultrafree-15 centrifugal filter devices, equipped with Biomax 10 K NMWL membranes (Millipore, Bedford, MA) for ultrafiltration.
10. Trifluoroacetic acid (TFA). To prepare a 0.1% solution, add 100 μ L TFA to 100 mL water.
11. Reversed-phase column material Poros 20R1 (for desalting of protein samples).
12. 30% acetonitrile, containing 0.1% TFA and 70% acetonitrile, containing 0.1% TFA.

2.2. First-Dimensional Separation (IEF)

1. IPG strips: IPG strips for the first-dimensional separation are commercially available and can be purchased from Amersham Biosciences (Uppsala, Sweden), Bio-Rad Laboratories (Hercules, CA) or other suppliers. They are available in various pH gradients and dimensions. IPG strips are delivered and kept frozen at -20°C, and are thawed just before use.
2. Rehydration solution for IPG strips H: 8 M urea, 2% CHAPS (w/v), 0.4% dithioerythritol (DTE) (w/v), 0.5% IPG buffer (of pH range corresponding to the IPG strips, v/v, Amersham Biosciences), and traces of bromophenol blue. To prepare 100 mL of rehydration solution, mix 48 g urea, 2 g CHAPS, 0.4 g DTE, 0.5 mL IPG buffer, and traces of bromophenol blue, and bring to 100 mL with Milli-Q water. Agitate on an end-over-end rotator until all components are dissolved, filter using a 0.22- μ m filter (Millipore), divide into 20-mL aliquots, and freeze at -20°C. The mixture may not be heated.
3. IEF sample buffer I: 20 mM Tris, containing 7 M urea, 2 M thiourea, 2% CHAPS (w/v), 0.4% DTE (w/v), 1% IPG buffer (v/v), 2 mM tributylphosphine (TBP), and traces of bro-

mophenol blue. To prepare 100 mL of IEF sample buffer, mix 4 mL 0.5 *M* Tris, 42 g urea, 19.4 g thiourea, 2 g CHAPS, 0.4 g DTE, 1 mL IPG buffer (of pH range corresponding to the IPG strips used), and traces of bromophenol blue, and bring to 100 mL with Milli-Q water (no pH adjustment is necessary). Agitate on an end-over-end rotator until all components are dissolved, filter using a 0.22- μ m filter, and divide into 10-mL aliquots. The mixture may not be heated. The aliquots are kept frozen at -20°C until use. Thaw only once and discard unused solution. TBP is added to the sample buffer to a final concentration of 2 mM (100 μ L to 10 mL) just before use (see Note 2).

4. Strip equilibration solution (after IEF) J: 50 mM Tris-HCl (pH 8.8), 6 *M* urea, 30% glycerol (v/v), and 2% SDS (w/v). To the stock equilibration solution, 2% DTE (w/v) or 2.5% iodoacetamide (IAA, w/v) are added for protein reduction and alkylation, respectively, as described below (Subheading 3.3.2.2.). To prepare 1000 mL of strip equilibration solution, mix 33.3 mL 1.5 *M* Tris-HCl (pH 8.8), 360 g urea, 300 mL glycerol (Fluka), and 100 mL 20% SDS solution, bring to 1000 mL with Milli-Q water, and stir at room temperature until all ingredients have been dissolved. Divide into aliquots of 250 mL and keep frozen at -20°C until use. Thaw only once and discard unused solution. To prepare reduction solution, add 5 g DTE to solution J to a final volume of 250 mL and stir until completely dissolved. To prepare alkylation solution, add 6.25 g IAA to solution J to a final volume of 250 mL and stir until completely dissolved.
5. Paraffin oil.
6. Kerosene.

2.3. Second-Dimensional Separation (SDS-PAGE)

1. Acrylamide solution (for SDS-gels): The volume of the acrylamide solution depends on the gel dimensions and number of gels to be prepared. Usually 10–14 gels can be prepared simultaneously in gel casting chambers from Amersham Biosciences or Bio-Rad. The final concentrations of the ingredients are:

Acrylamide/PDA solution (37.5:1, w/v)	depends on the gel strength
Tris-HCl, pH 8.8	375 mM
Sodium thiosulfate	5 mM
TEMED (<i>N,N,N',N'</i> -tetramethylethylenediamine)	0.65 mL/L
10% Ammonium persulfate (APS, w/v)	3.5 mL/L
Water (Milli-Q)	to the required volume

To prepare 2000 mL of 10% acrylamide solution, mix 666 mL stock acrylamide/PDA solution (acrylamide/piperazine di-acrylamide [PDA]) solution [37.5:1, w/v], Biosolve Ltd., Valkenswaard, The Netherlands), 500 mL 1.5 *M* Tris-HCl (pH 8.8), 2.4 g sodium thiosulfate, and bring to 2000 mL with Milli-Q water. De-gas by filtering through a 0.22- μ m filtration device, add 7.0 mL 10% APS (w/v) and 1.3 mL TEMED, mix, and quickly proceed with gel pouring. The solution will stay liquid for approx 45 min (see Note 3).

2. Water-saturated isobutanol.
3. 1% agarose solution: To prepare the agarose solution, mix 1 g of low-molecular-weight agarose with 100 mL of SDS gel running buffer K, previously cooled to 4°C. Heat for about 1 min in a microwave apparatus to dissolve agarose. Unused agarose solution is kept at 4°C. The solidified agarose is heated in the microwave apparatus every time to prepare a solution.
4. SDS gel running buffer K: 25 mM Tris, 192 mM glycine, and 0.1% SDS (w/v). Tenfold concentrated buffer is commercially available. To prepare 10 L of running buffer, dilute 1 L of concentrated stock solution with 9 L of Milli-Q water.
5. Protein fixing solution L: 50% methanol and 5% phosphoric acid in water.

6. Colloidal Coomassie blue.
7. 20% sodium azide.
8. Tricine-SDS running buffer M: 100 mM Tris-base, 100 mM Tricine, 0.1% SDS, final pH 8.3.
9. Benzylidimethyl-*n*-hexadecyl ammonium chloride (16-BAC).

3. Methods

3.1. Preparation of Membrane Proteins

Figure 1 shows the general scheme for the preparation of bacterial membrane proteins. A modified procedure can be followed for preparation of membrane fractions from eukaryotic systems (see **Note 4**).

1. Cell pellet from bacteria culture is kept frozen at -70°C until use.
2. Suspend 20 g (wet weight) of cell paste by gentle stirring in 25 mL of lysis buffer A. The buffer contains the protease inhibitors Trasylol and ϵ -aminocaproic acid to avoid protein degradation, and benzonase to hydrolyze DNA and RNA.
3. Disrupt cells in a precooled French pressure cell (SLM Instruments, Urbana, IL) at 20,000 lb/in².
4. Add EDTA-Na₂ from the 0.5 M stock solution to a 5 mM final concentration.
5. Centrifuge the lysed cell suspension at 3000g for 20 min to sediment intact cells and debris.
6. Centrifuge the supernatant at 150,000g for 90 min to sediment cell membranes.
7. Dialyze the supernatant from this centrifugation step two times against 1 L of buffer B at 4°C for 18 h to reduce the ionic strength and filter the dialyzate through a 0.45- μ m pore-size membrane. The filtrate contains the soluble cytosolic proteins and can be directly used for proteomic analysis or for further fractionation by chromatography. About 1.3 g of total cytosolic proteins are obtained from 20 g of cell paste.
8. Resuspend the membrane pellet in 80 mL of buffer C and centrifuge at 30,000g for 60 min.
9. Resuspend the pellet in 30 mL of hypertonic buffer D.
10. Dilute 10-fold with hypotonic buffer E to remove soluble components trapped in vesicular structures by osmotic shock.
11. Centrifuge at 30,000g for 60 min. The pellet contains the cell membrane proteins and can be either used for 1-D or 2-D gel analysis or kept frozen at -70°C until use. About 175 mg of membrane proteins are usually obtained from 20 g of bacterial cells.

3.2. Solubilization of Membrane Proteins

Membrane proteins need to be solubilized with the use of detergents and/or chaotropes (the cytosolic proteins are already in solution) prior to the proteomic analysis. It is often useful to first analyze the membrane preparation by 1-D electrophoresis in SDS-polyacrylamide gels, as hydrophobic proteins can be detected in this kind of gels (see **Note 5**). For the 2-D gel analysis, the membranes can be solubilized in an IEF-compatible solution or/and in an ionic detergent/chaotropic, IEF-incompatible system (see **Note 6**).

3.2.1. Solubilization in IEF-Compatible Solution

1. Suspend 100 mg of membranes in 0.5 mL of IEF-compatible buffer F by sonication for 30 s.
2. Centrifuge the mixture at 150,000g for 30 min and use the supernatant for 2-D electrophoresis (the solution contains the proteins that are soluble in the IEF-compatible solution).
3. Use strong ionic or chaotropic agents to solubilize the proteins of the pellet and analyze them by 1-D and 2-D gels (see **Subheadings 3.2.2.** and **3.3.**).

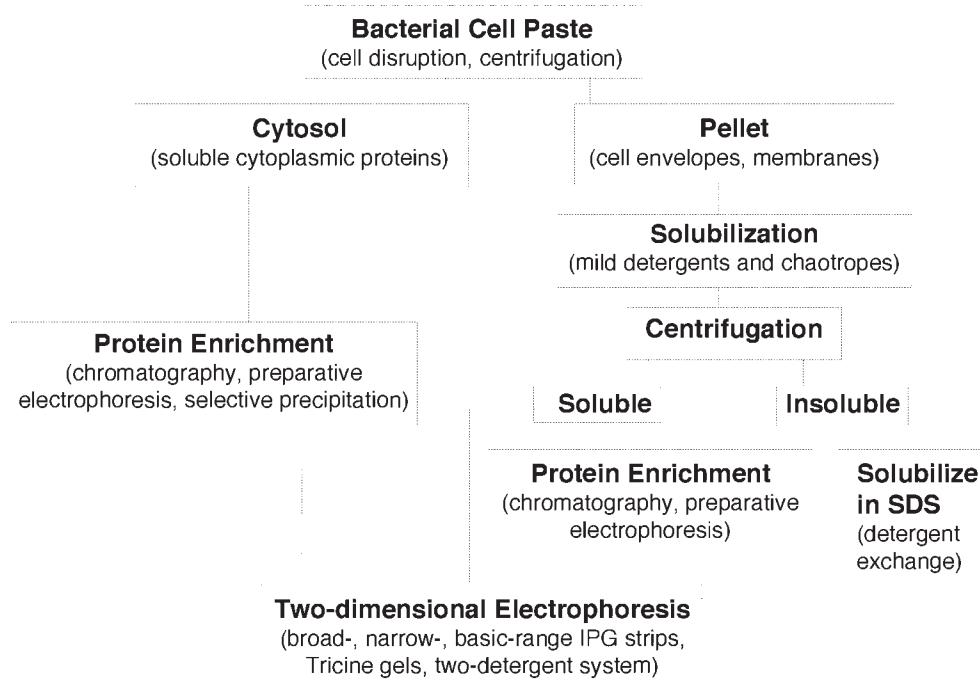


Fig. 1. General scheme of proteomic analysis of bacterial membrane proteins. Similar approaches can be followed for the analysis of membrane proteins from eukaryotic systems.

3.2.2. Solubilization in Strong Agents (Non-IEF Compatible)

1. Suspend 100 mg of membranes in 0.5 mL of buffer G by sonication for 30 s.
2. Centrifuge at 150,000g for 30 min.
3. Dilute the supernatant(s), which contain proteins solubilized with ionic detergents and guanidine hydrochloride and are not compatible with IEF, 10-fold with buffer F, and concentrate to approx 100 μ L by ultrafiltration at 2000g.
4. Repeat the dilution-concentration process for a second time. The concentrate is used for 2-D gel electrophoresis (add IPG buffer and TBP, which are not included in buffer F, to 1% and 2 mM final concentration, respectively).
5. Alternatively, apply the diluted supernatant(s) (**step 3**) onto a column containing 0.4 mL Poros 20R1 material. To prepare the column, suspend 1 g of beads in 3 mL of ethanol and add water to 10 mL (store mixture at 4°C). Apply 1 mL of the suspension into the column and equilibrate with 0.1% trifluoroacetic acid.
6. Apply the protein solution and let the fluid run out of the column. Wash the column with five vols of 30% acetonitrile, containing 0.1% TFA.
7. Elute the proteins with two column vols of 70% acetonitrile, containing 0.1% TFA.
8. Concentrate the solution in a speedvac apparatus to approx 50 μ L. Dilute the concentrate with 1 mL of IEF sample buffer I and concentrate to approx 50 μ L by ultrafiltration at 2000g. The concentrate is used for 2-D gel analysis.

3.3. Two-Dimensional Gel Electrophoresis

Following the solubilization of the membrane proteins by any of the methods described above or similar approaches, the visualization of a protein is usually per-

formed by 2-D gel electrophoresis. We perform 2-D gel electrophoresis essentially as reported earlier (16). Here we describe the general method of performing the 2-D electrophoretic analysis (*see Note 7*). Operations may slightly vary, depending on the equipment used in each laboratory, on developments of the technology, and on new reagents (*see Note 8*).

3.3.1. First-Dimensional Separation

3.3.1.1. REHYDRATION OF IPG STRIPS

1. Take the frozen IPG strips out of the -20°C freezer. Remove the necessary number of strips, which will be used for IEF, from their container (each contains 12 strips) and quickly refreeze the strips that will not be used this time (*see Note 9*). Take two more strips than necessary, because the gel of some IPG strips may be damaged during the operations (*see Note 10*).
2. Let the strips stay at room temperature for about 30 min, then remove the protective plastic foil with a forceps and place the strips gel-down in the groove of the specific rehydration tray, where 1 mL of freshly prepared rehydration solution H has been added (*see Note 11*).
3. Remove air bubbles from the gel. Add 1 mL of paraffin oil to each groove to avoid urea crystallization (*see Notes 12 and 13*).
4. Let the strips rehydrate at room temperature for 16 h (*see Notes 14 and 15*).

3.3.1.2. ISOELECTRIC FOCUSING

1. Transfer the strips from the rehydration tray to the grooves of the plastic aligner in the IEF sample tray. The gel side should face up. The strips should be similarly oriented, with the positive end towards the positive electrode (the red electrode).
2. Place the IEF sample tray onto the cooling plate of the IEF apparatus. Apply a few mL of kerosene onto the cooling surface to establish an efficient heat transfer between the sample tray and the cooling plate.
3. IEF paper electrode wicks, approx 1 cm long, are wetted with deionized water. Excess water is removed with a filter paper, but the wicks should not become completely dry. Place one wick at each end of the IPG strip so that it touches the gel (two electrode wicks are placed on each IPG strip, one at each end) (*see Note 16*).
4. Position the two electrodes so that the wire touches the electrode wicks (the wicks should touch the ends of the gel of the IPG strips and should not touch each other).
5. Place the cup bridge and the cups on the bridge and press them gently down so that they touch the gel surface of all strips. Use two bridges per IEF run to apply the samples at the basic and acidic ends of the IPG strips (*see Notes 17 and 18*).
6. Overlay the IPG strips and fill the cups with low-viscosity paraffin oil. Approx 25 mL of paraffin oil is required to cover the whole surface of the sample tray (*see Note 19*).
7. Pipet the sample into the cups. A total volume of approx 150 μL can be loaded per cup (*see Note 20*).
8. Connect the cooling plate to the cooling apparatus and adjust the temperature to 15° . Place the cover on the IEF apparatus and establish contact with the power source. Start IEF at 200 V and program a 3-V/min voltage increase for 2 h and then a 4-V/min increase until 5000 V is reached. Continue at 5000 V for 4 h. Total run time is about 28 h (*see Notes 21 and 22*).
9. Following IEF, the paraffin oil is removed with a pipet, and the strips can be used for the second-dimensional separation or be frozen at -20°C . When thawed, they are processed in the same way as described below (**Subheading 3.3.2.1.**).

3.3.2. Second-Dimensional Separation (SDS-polyacrylamide gel electrophoresis)

For the second-dimensional separation, the proteins have to be charged by equilibration with SDS. Simultaneously, they should be reduced and alkylated to avoid formation of oligomers and to prepare the sample for a mass-spectrometry analysis.

3.3.2.1. SDS-GEL PREPARATION

1. Assemble the gel casting device, using clean glass plates of the appropriate size and spacers of the proper thickness (see Note 23).
2. Apply a few drops of water to a piece of Whatmann 3MM filter paper with a printed gel number on it, and place it at the lower, right side (basic) of the gel cassette (see Note 24).
3. Insert a plastic sheet between gel sets. Close chamber tightly and make sure that the seal parts are in place.
4. Prepare a fresh acrylamide gel solution (see Subheading 2.3.) and fill the cassette from the bottom by gravity (avoid air bubbles) or by using a pump at 480 mL/h, until the solution level reaches approx 5 mm from the top of the short glass plate.
5. Overlay each gel with 1 mL of water-saturated isobutanol. Apply isobutanol gently, so that it does not enter the acrylamide phase.
6. Let the gels polymerize at room temperature for at least 4 h, or better overnight.
7. Remove the gels from the casting chamber, wash the plates from the outside, and wash out the isobutanol layer completely with de-ionized water (see Note 25).

3.3.2.2. IPG STRIP EQUILIBRATION

1. The strips are removed from the sample tray and placed into labeled plastic 10-cm-diameter Petri dishes, the gel side facing in (see Note 26).
2. 10 mL of equilibration buffer J containing 2% DTE (w/v) are added per Petri dish. The Petri dishes are gently shaken in a rotating shaker at 35 rpm for 15 min.
3. The solution is drawn off carefully so as not to damage the gel, and replaced with 10 mL of equilibration solution J containing 2.5% iodoacetamide (w/v). The Petri dishes are gently shaken at 35 rpm for 15 min.

3.3.2.3. GEL RUNNING

1. After equilibration with DTE and IAA, take the IPG strip with a forceps and place it on top of the SDS gel, with the acidic side of the strip at the left.
2. Establish contact between strip and gel with a few milliliters of a molten 1% agarose solution. Remove air bubbles between strip and gel.
3. Place the gels carrying strips in the gel running apparatus; fill with the appropriate volume of Tris-glycine running buffer K (see Note 27), according to the recommendations of the device supplier, and start run (see Note 28).
4. Stop electrophoresis when the bromophenol blue has reached the front of the gels. Gels can be stained or used for blotting.
5. For stain, disassemble the gel cassette and place the gel in 200 mL of protein fixing solution L for 2 h.
6. Wash the gel twice with water for 20 min each time.
7. Stain the gel with colloidal Commassie blue overnight, following the guidelines of the supplier (see Note 29).
8. Remove the dye and wash the gel five times with de-ionized water for 15 min.
9. In the last wash bath, add 1 mL of 20% sodium azide and incubate the gel for 15 min. Seal the gel in a plastic foil and keep it at 4°C for up to 1 yr (see Note 30).

3.4. Alternative Approaches

3.4.1. Detection of Low-Molecular-Mass Proteins

Low-molecular-mass proteins are underrepresented in 2-D gels. For the detection of such proteins, use of Tricine gels is more efficient than of Tris-glycine gels. The first-dimensional separation is performed as described in **Subheading 3.3.1**. We prepare the Tricine gels according to a modified method of Schägger and von Jagow (18). The gels consist of a lower, separating gel of 10.4% acrylamide and a 20-mm-high stacking gel of 5.4% acrylamide concentration. The separating gel contains 6.2 M urea. The stacking gel does not include urea. The solutions are degassed before gel preparation. The anode buffer is running buffer K, and the cathode buffer is running buffer M. The electrophoresis is performed in the Protean II system (Bio-Rad) at 40 mA per gel for approx 4 h. Proteins are fixed with solution L for 16 h. The gels are stained with colloidal Coomassie blue for 24 h. The Tricine gels have smaller dimensions (approx 160 × 160 × 15 mm) than the Tris-glycine gels (approx 200 × 180 × 15 mm).

3.4.2. Detection of Hydrophobic Proteins

The membrane fraction can also be separated using a discontinuous, two-detergent gel system (12). The first-dimensional separation is performed in 7.5% polyacrylamide gels in the presence of 250 mM benzylidimethyl-*n*-hexadecylammonium chloride (16-BAC), and the second-dimensional separation in 11% polyacrylamide gels in the presence of 0.1% SDS. This system has the advantage that hydrophobic proteins can be detected (which usually do not enter the IPG strips). It has the disadvantage that a relatively low number of spots can be resolved and spot overlapping may occur.

4. Notes

1. Membrane proteins are characterized by hydrophobicity scores, the Grand Average Hydrophobicity Values (GRAVY) (19). GRAVY values usually vary in a range of ± 2 . Positive scores indicate hydrophobic, and negative scores hydrophilic proteins. The GRAVY values characterize the hydrophobicity of the whole molecule and do not represent a reliable criterion whether the protein will enter the IPG strip or not (4).
2. Reagents for the first-dimensional separation are kept at 4°C. Care should be taken when working with thiourea and tributyl phosphine, which are suspected to be carcinogenic. Always wear gloves and open the containers in a hood.
3. Reagents for the second-dimensional separation are kept at 4°C. Acrylamide/PDA solutions can be purchased ready made. Acrylamide is a potent neurotoxin, and purchase of solution should be preferred, to avoid exposure to dust. Wear gloves and protective clothes whenever working with acrylamide.
4. The separation of the protein mixture into cytosolic and membrane fractions prior to the 2-D electrophoretic analysis is usually the first step to increase the chance of detecting low-copy-number gene products of either fraction. Cell-envelope and -membrane proteins usually represent approx 15% of the total cellular proteins. The probability of detecting a low-abundance membrane protein increases by about 10-fold if prefractionation is employed. In eukaryotic systems, the separation of additional organelles, such as mitochondria, peroxisomes, and so on, is very helpful in detecting proteins specific to the corresponding organelle. Preparation of membrane proteins has been extensively described in the literature. Therefore, we will not enter in more details about preparation of pure membranes or isolation of organelles.

5. An efficient approach for a proteomic analysis of membrane proteins is the detailed separation of the original protein sample into subfractions and organelles, and the analysis of the precipitates by one-dimensional SDS gels and MALDI-TOF MS. The method has the advantage of the distribution of the proteins into smaller fractions and the solubilization of the membrane proteins with a strong detergent. It has the disadvantage of the poor resolution in only one dimension, but it is often the method of choice if the protein mixture mainly contains hydrophobic proteins. For the SDS-PAGE, the membranes can be solubilized in buffer G.
6. Nonionic or zwitterionic detergents and nonionic chaotropes are compatible with IEF, but they are not usually very efficient in dissolving difficult-to-solubilize proteins. The strong detergents SDS and LDS are useful in solubilizing most membrane proteins; however, they are ionic and need to be exchanged against zwitterionic detergents, such as CHAPS, prior to IEF. Removal of SDS is difficult, and most likely the free SDS and not the protein-bound detergent is removed.
7. Further information about electrophoresis techniques can be found in Westermeier (17).
8. Here we describe how to perform IEF with the Multiphor II apparatus of Amersham Biosciences. The operations with the apparatuses from other manufacturers are similar. Technical variations may exist, and the researcher should follow the directions of the suppliers. The additional equipment mentioned here (rehydration cassette, bridges, sample cups, and so on) has been purchased from Amersham Biosciences as well.
9. Wear gloves during all operations. Rinse gloves with de-ionized water to remove powder and dry them with paper towels.
10. There are three main strip rehydration variations: rehydration in a tray (**Subheading 3.3.1.1.**), in a cassette and in a solution containing the protein sample (**Note 14**). Thus, instead of the rehydration tray, a rehydration cassette with spacers at the three sides (Amersham Biosciences) may be used. Place the strip with the gel side up. Use a few drops of water to stick the strips to the plate. Assemble the rehydration cassette with the second glass plate and hold the plates together with clamps. Add 15 mL of freshly prepared rehydration solution through the silicon tube of the second glass plate, using a syringe. Remove air bubbles between gel and glass plate and from the strip sides. Let the strips rehydrate at room temperature for 16 h. To avoid drying of the rehydration solution during the overnight rehydration, place the cassette in an upright position in a closed plastic box whose bottom is covered with water. We usually perform strip rehydration in a cassette.
11. Care should be taken that the gel not be touched (the gel may not touch a dry surface). The gel is sticky, and if touched, a gel piece may be damaged, which will have consequences for the IEF performance. Discard damaged strips.
12. Addition of paraffin oil is optional. When we follow this rehydration approach, we perform rehydration without adding paraffin oil.
13. Instead of preparing the necessary amount of rehydration solution each time, a large volume of it can be prepared (for example, 500 mL), divided into 20-mL aliquots, and kept frozen at -20°C in 50-mL screw-cap tubes. The necessary number of tubes will be thawed directly before use. Excess solution should not be frozen again.
14. Alternatively, the IPG strips can be rehydrated at 23°C for 16 h in rehydration solution containing the protein sample to be analyzed (total approx 0.3 mL). Thus, the proteins enter the strips during rehydration. In this case, the step of sample application in cups is omitted (**Subheading 3.3.1.2.**). Users of the IPGphor apparatus (Amersham Biosciences) can apply a low voltage during the rehydration step as recommended by the instrument supplier. Although protein loading during the strip rehydration should theoretically result in the detection of a larger number of spots in 2-D gels, for reasons not well understood,

- this sample application method usually results in detection of a lower number of spots in comparison with cap loading. In our hands, cap loading is the most efficient sample application method.
15. Many laboratories use a 6-h rehydration procedure. There does not seem to be a difference between the 6- and the 16-h rehydration method. Prolonged rehydration times beyond 16 h should be avoided because of the possible chemical modifications of urea during the long incubation time at room temperature.
 16. In the Multiphor apparatus, long electrode strips may be used and placed so that they touch the gel ends of the strips. However, a cross-contamination from proteins moving from IPG strips to other strips through the solid electrode strips may occur.
 17. Correct placement of the sample cups so that they touch but do not damage the gel on the IPG strip is technically the most critical step of the 2-D electrophoresis. Cups may not be pressed too strongly on the strips because the sample cannot move along the strip or the gel may be damaged and sample leakage may occur. On the other hand, cups may not be placed loosely with insufficient contact to the gel surface, as sample leakage may again occur. This step requires a certain amount of practice before starting with precious samples.
 18. Sample application at both sides of the strip usually results in the detection of more spots. We apply two-thirds of the sample volume at the basic end and one-third at the acidic end. Application at one strip side only (usually basic) can also be performed. Use of narrow pH range (pH 4.0–7.0 or pH 6.0–11.0) or very narrow pH range (for example pH 5.0–6.0) IPG strips and application of a relatively large amount of protein (about 2 mg) may in certain cases lead to the detection of new spots. These spots usually represent modifications of already identified gene products.
 19. At this stage, one can detect whether the cups are tight or they are leaking. A small volume of IEF sample buffer (about 25 μ L) can be applied into the cups (before the sample application) to confirm that the blue-colored solution is not leaking. The sample buffer does not need to be removed before sample application.
 20. The protein amount applied on a strip depends on the goal of the experiment and the experience of the laboratory. As a general rule, 0.5–1.0 mg of total protein can be applied if the gel will be used for mass spectrometry analysis afterwards. Lower amounts can be applied; however, the low-abundance components may not be visible if the gel is stained with Coomassie blue. If larger amounts than 1.0 mg are applied, a significant percentage of the proteins does not enter the IPG strips.
 21. IEF conditions may vary considerably, depending on the equipment, experiment, protein amount, IPG strip type, and other factors. The given conditions have been tested and work well. It is important that at the beginning, the voltage be kept low and increase gradually. The total IEF time may vary between 18 and 48 h. Longer IEF times may result in protein loss. Some power supplies can reach only 3500 V. If the protein amount is relatively low (less than 0.5 mg), the lower voltage will not affect IEF.
 22. Samples should be salt free in 20 mM Tris; otherwise, ultrafiltration or protein precipitation followed by solubilization in IEF sample buffer should be applied to remove salt and other disturbing agents from the samples prior to IEF (20).
 23. There are several gel systems available for the simultaneous preparation of up to 12 gels. Researchers should also follow the recommendations of the equipment suppliers. We usually use the Bio-Rad system (Protean MultiCell) to prepare gradient gels and the Amersham Biosciences system (Ettan Dalt II) to prepare uniform gels.
 24. Because in high-throughput proteomics approaches, a large number of gels are produced, each gel should be identified with a number so that it can be tracked afterwards.

25. Gel preparation should be planned the day before IEF is to be finished or early in the morning, to allow sufficient time for polymerization. Gels can be used immediately after acrylamide polymerization or can be overlaid with water and kept at 4°C for approx 2 wk.
26. Care should be taken that the gel does not touch the Petri dish surface; otherwise, it can be damaged. At this stage, before equilibration, Petri dishes containing strips that will not be used for the second-dimensional separation this time can be labeled with the sample name and experiment number, sealed with Parafilm, and stored at -20°C for up to 3 mo.
27. The molecular mass can be determined by running standard protein markers at the right side of selected gels. The size markers cover the range of 10–220 kDa. Place a Teflon 2-D comb about 10 mm from the strip end, before pouring the hot agarose, to prepare a well for application of the size markers. The proposed orientations (acidic end of the strips at the left, the markers at the right side of the gel) are arbitrary. It is helpful if the orientation is consistent, in particular when handling with new samples.
28. Recommended running conditions: Ettan Dalt II apparatus (for 12 gels), 20 W/gel, 20°C, run time about 6 h or 10 mA/gel for overnight run. Bio-Rad Protean MultiCell apparatus, 45 mA/gel, 8°C, run time 5 h or 8 mA/gel for overnight run (18 h).
29. Stain with Coomassie blue is rather insensitive; however, it is usually the method of choice if identification of the spots by MS will follow.
30. Caution: concentrated sodium azide solutions may not come in contact with strong acids (danger of explosion).

References

1. Herbert, B. (1999) Advances in protein solubilisation for two-dimensional electrophoresis. *Electrophoresis* **20**, 660–663.
2. Molloy, M. P. (2000) Two-dimensional electrophoresis of membrane proteins using immobilized pH gradients. *Anal. Biochem.* **280**, 1–10.
3. Santoni, V., Molloy, M., and Rabilloud, T. (2000) Membrane proteins and proteomics: un amour impossible? *Electrophoresis* **21**, 1054–1070.
4. Fountoulakis, M., and Gasser, R. (2003) Proteomic analysis of the cell envelope fraction of *Escherichia coli*. *Amino Acids* **24**, 19–41.
5. Fountoulakis, M. (2000) Two-dimensional electrophoresis. In *Encyclopedia of Separation Science, II/Electrophoresis*, Academic Press, London, pp. 1356–1363.
6. Fountoulakis, M. and Takács, B. (2001) Effect of strong detergents and chaotropes on the protein detection in two-dimensional gels. *Electrophoresis* **22**, 1593–1602.
7. Chevallet, M., Santoni, V., Poinas, A., et al. (1998) New zwitterionic detergents improve the analysis of membrane proteins by two-dimensional electrophoresis. *Electrophoresis* **19**, 1901–1909.
8. Herbert, B. R., Molloy, M. P., Gooley, A. A., Walsh, B. J., Bryson, W. G., and Williams, K. L. (1998) Improved protein solubility in two-dimensional electrophoresis using tributyl phosphine as reducing agent. *Electrophoresis* **19**, 845–851.
9. Rabilloud, T., Blisnick, T., Heller, M., et al. (1999) Analysis of membrane proteins by two-dimensional electrophoresis: comparison of the proteins extracted from normal or *Plasmodium falciparum*-infected erythrocyte ghosts. *Electrophoresis* **20**, 3603–3610.
10. Ferro, M., Seigneurin-Berny, D., Rolland, N., et al. (2000) Organic solvent extraction as a versatile procedure to identify hydrophobic chloroplast membrane proteins. *Electrophoresis* **21**, 3517–3526.
11. Carboni, L., Piubelli, C., Righetti, P. G., Jansson, B., and Domenici, E. (2002) Proteomic analysis of rat brain tissue: comparison of protocols for two-dimensional gel electrophoresis analysis based on different solubilizing agents. *Electrophoresis* **23**, 4132–4141.

12. Hartinger, J., Stenius, K., Högemann, D., and Jahn, R. (1996) 16-BAC/SDS-PAGE: a two-dimensional gel electrophoresis system suitable for the separation of integral membrane proteins. *Anal. Biochem.* **240**, 126–133.
13. Fountoulakis, M. (2001) Proteomics: Current technologies and applications in neurological disorders and toxicology. *Amino Acids* **21**, 363–381.
14. Fountoulakis, M. and Takács, B. (2002) Enrichment and proteomic analysis of low-abundance bacterial proteins. *Methods Enzymol.* **358**, 288–306.
15. Fountoulakis M. (2004) Application of proteomics technologies in the investigation of the brain. *Mass Spectrom. Rev.* **23**(4), 231–258.
16. Langen, H., Roeder, D., Juranville, J.-F., and Fountoulakis, M. (1997). Effect of the protein application mode and the acrylamide concentration on the resolution of protein spots separated by two-dimensional gel electrophoresis. *Electrophoresis* **18**, 2085–2090.
17. Westermeier R. (ed). (1993) *Electrophoresis in Practice* VCH Verlagsgesellschaft, Weinheim, pp. 215–263.
18. Schaegger, H. and von Jagow, G. (1987) Tricine-sodium dodecyl sulfate-polyacrylamide gel electrophoresis for the separation of proteins in the range from 1 to 100 kDa. *Anal. Biochem.* **166**, 368–379.
19. Kyte, J. and Doolittle, R. F. (1982) A simple method for displaying the hydrophobic character of a protein. *J. Mol. Biol.* **157**, 105–132.
20. Jiang, L., He, L., and Fountoulakis, M. (2004) Comparison of protein precipitation methods for sample preparation prior to proteomic analysis. *J. Chromatogr. A* **1023**, 317–320.

2-D PAGE of High-Molecular-Mass Proteins

Masamichi Oh-Ishi and Tadakazu Maeda

1. Introduction

Many high-molecular-mass proteins (MW>100 kDa) are known to be involved in cytoskeleton, defense and immunity, transcription, and translation in higher eukaryotic organisms. Though a variety of protein separation techniques have been described, at the moment purification of a high-molecular-mass protein remains a difficult task. O'Farrell (1) was the first to devise a two-dimensional gel electrophoresis (2-DE) technique which could detect more than 1000 spots in a gel. Though it is evident that this method was quite powerful in his day, the method is not particularly suitable in analyzing high-molecular-mass proteins larger than 200 kDa. When skeletal muscle, for example, is simultaneously analyzed by sodium dodecyl sulfate (SDS)-polyacrylamide gel electrophoresis (PAGE) and O'Farrell's 2-DE method, the myosin heavy chain (approx 200 kDa) is clearly seen in the former gel, but not in the latter. Hirabayashi (2) was the first to develop a 2-DE method that could analyze high-molecular-mass proteins, including myosin heavy chain and dystrophin, as large as 500 kDa (3). His trick was the use of agarose gel instead of polyacrylamide gel for the first-dimensional isoelectric focusing (IEF). Agarose gel, when used for IEF, can analyze much larger proteins than the polyacrylamide gel can. Oh-Ishi and Hirabayashi (4) further improved the method by adding 1 M thiourea and 5 M urea in an agarose IEF medium. Thiourea is a potent protein solubilizing reagent especially effective for high-molecular-mass proteins that could enter the first-dimensional agarose IEF gel. Here, we describe the 2-DE method with agarose gels in the first dimension (agarose 2-DE) that is compatible with analyzing high-molecular-mass proteins.

2. Materials

2.1. Equipment

1. Apparatus for first-dimensional agarose IEF (AE-6300 electrophoresis unit) (ATTO, Tokyo, Japan).
2. Glass tubes (260 mm × 3.4 mm ID) for agarose IEF (ATTO).
3. Dialysis membranes.
4. Rubber bands.
5. Syringe to fit thin 30-cm-long polyethylene tubing.
6. Electrophoresis power supply (Crosspower 3500) (ATTO).
7. Horizontal electrophoresis apparatus (Multiphor II, Amersham Biosciences).
8. Immobiline DryStrips pH 3–10.0 L (Amersham Biosciences).

From: *The Proteomics Protocols Handbook*
Edited by: J. M. Walker © Humana Press Inc., Totowa, NJ

9. Perista pump (SJ-1211) (ATTO).
10. Glass tubes (300 mm × 5 mm ID) for protein fixation on agarose IEF gels.
11. Polyvinyl chloride (PVC) tubing (400 mm × 2 mm ID).
12. Vertical type apparatus for second-dimensional SDS-PAGE (NA-1200 electrophoresis unit) (Nihon Eido, Tokyo, Japan).
13. Glass plates with a 1.5-mm-thick frame (size 165 mm × 240 mm) for second-dimensional SDS-PAGE (Nihon Eido).
14. Plain glass plates (size 165 mm × 240 mm) for second-dimensional SDS-PAGE (Nihon Eido).

2.2. Reagents

1. Extraction medium: 7 M urea, 2 M thiourea, 2% CHAPS, 0.1 M dithiothreitol (DTT), 2.5% Pharmalyte, and protease inhibitors (Complete Mini ethylenediaminetetraacetic acid [EDTA]-free; Roche).
2. Overlaying solution: 4 M urea, 1 M thiourea.
3. Cathode buffer: 0.2 M NaOH.
4. Anode buffer: 0.04 M DL-aspartic acid.
5. Protein fixing solution: 10% trichloroacetic acid, 5% sulfosalicylic acid.
6. Incubation medium: 2% SDS, 10% glycerol, 5% 2-mercaptoethanol, 0.02% bromophenol blue, and 0.05 M Tris-HCl (pH 6.5).
7. Staining solution: 0.02% PhastGel Blue R (Coomassie Brilliant Blue R-350; Amersham Biosciences), 30% methanol, 10% acetic acid.
8. Destaining solution: 30% methanol, 10% acetic acid.

3. Methods

3.1. Agarose Gels for IEF

3.1.1. Gel Preparation

1. We prepared the first-dimensional agarose IEF gels following the same procedure as described in our previous report (5), to which we added the modifications described here. Three agarose IEF gels (180 mm in length and 3.4 mm in diameter) containing 1 M thiourea and 6 M urea were prepared by the present protocol.
2. Agarose IEF (0.10 g) and D-sorbitol (1.2 g) were put into a 50-mL beaker, and dissolved in 5.8 mL distilled water at room temperature (solution A) (see Note 1).
3. Solution A was boiled 10 times for 15 s in a microwave oven until the solution became clear, and was kept in a 70°C water bath for 5 min.
4. A mixture of urea (3.6 g) and thiourea (0.76 g) powder was put into solution A at 70°C, which we shall call solution B hereafter (see Note 2 and Table 1). The volume of solution B was then adjusted to 9.0 mL with distilled water, and was kept at room temperature.
5. Solution B was divided into three test tubes: 1.0 mL for acidic, 4.5 mL for neutral, and 2.5 mL for basic solutions.
6. Four kinds of Pharmalyte (pH 2.5–5.0 for acidic, pH 3.0–10.0 and pH 4.0–6.5 for neutral, and pH 8.0–10.5 for basic solutions) were added to the respective tubes according to the protocol shown in Table 1.
7. A glass tube (standard size: 260 mm × 3.4 mm ID) was prepared beforehand, the bottom of which was covered with a piece of dialysis membrane and tied with a rubber band.
8. The glass tube was attached to AE-6300 electrophoresis unit (apparatus for agarose IEF produced by ATTO).
9. The acidic, neutral, and basic solutions were sucked in with one syringe each through thin 30-cm-long polyethylene tubing. First, the acidic solution was gently injected into the

Table 1
Protocols for Making Agarose Isoelectric Focusing (IEF) Gels

Agarose IEF	0.10 g
D-sorbitol	1.20 g
Distilled water	5.80 mL
Dissolved at 100°C→Solution A	
Set at 70°C	
Urea	3.60 g
Thiourea	0.76 g
Dissolved at 50°C→Solution B	
Add distilled water until solution B comes to 9.0 mL	
Set at room temperature	
Solution B	1.0 mL
Pharmalyte	↓
pH 2.5–5.0	100 µL
pH 3.0–10	—
pH 4.0–6.5	—
pH 8.0–10.5	— ↓
Acidic solution	Neutral solution
	2.5 mL ↓
	— —
	300 µL 150 µL
	— ↓
	250 µL ↓
	Basic solution

glass tube until the meniscus reached a height of 20 mm from the bottom of the tube. Next, the neutral solution was slowly injected until the meniscus reached a height of 135 mm from the bottom, and then the basic solution was carefully injected until the meniscus was 180 mm from the bottom. A 10 µL overlaying solution was gently layered on top of the agarose solution, which made it easier for proteins to enter the agarose IEF gel. The glass tube was filled with the chamber and kept there at least 6 h, until the agarose solution gelled (*see Note 3*).

3.1.2. Sample Preparation

1. Freshly obtained mammalian tissues were cut into small blocks, quickly frozen in liquid nitrogen, and stored at –80°C until use. Each frozen tissue piece (about 10 mg) was homogenized with a Teflon glass homogenizer in extraction medium (20 times the volume of the tissue pieces) (e.g., brain, muscle, kidney, and liver). Homogenates of the tissues were centrifuged at 112,000g for 20 min, and the clear supernatant was subjected to the agarose IEF gel.

3.1.3. Sample Application

1. We added 10–200 µL of protein sample solution at the cathodic end of the gel, and gently filled the overlaying solution above the sample solution to the top of the glass tube. We added anode buffer to the lower reservoir, and cathode buffer to the upper reservoir. First-dimensional IEF was conducted at 600 V for 18 h at 4°C.
2. Then the agarose gel was transferred onto the top of the second-dimensional SDS gel, either directly or after proteins in the gel were fixed (*see Note 4*).

3.2. Second-Dimensional SDS-PAGE

1. Slab gels for second-dimensional electrophoresis were 12% polyacrylamide gels (200 mm × 120 mm × 1.5 mm). Second-dimensional SDS-PAGE was carried out according to the stacking system of Laemmli (6) with the slight modification of adding 1% SDS both in the stacking and separation gels.
2. The first-dimensional agarose IEF gel was loaded on top of the stacking gel without SDS equilibration (see Note 5) and covered with 1% agar solution to keep the agarose IEF gel in place. An incubation medium was overlaid on the IEF gels.
3. The second-dimensional gel electrophoresis was started with a constant current at 40 mA for 1 h and continued at 70 mA until the end of the run.
4. The slab gels were first soaked and shaken overnight in a destaining solution, which removed Pharmalytes from the gel (see Note 6). The slab gels were then stained with a staining solution containing PhastGel Blue R and destained with the destaining solution.

3.3 Comparison of IPG-Dalt With Agarose 2-DE

1. In this study, we compared the agarose 2-DE system with IPG-Dalt (immobilized pH gradients for IEF in the first dimension, SDS-PAGE in the second dimension) to assess their capability of analyzing high-molecular-mass proteins.
2. The first-dimensional IEF with IPGs was performed essentially as described by Gorg et al. (7). IPG dry strips (Immobiline DryStrips pH 3.0–10.0 L) and Multiphor II (Amersham Biosciences) were used in this study.
3. Coomassie-stained 2-DE patterns of rat duodenum (Fig. 1) revealed that high-molecular-mass proteins larger than 150 kDa could be analyzed with the agarose 2-DE but not with IPG-Dalt.

4. Notes

1. Shake the beaker gently until a mixture of agarose IEF and D-sorbitol powder is completely suspended.
2. Immediately after the mixture of urea and thiourea powder was put into the beaker, solution B was stirred with a magnetic stirrer until the urea and thiourea were completely dissolved.
3. As a side effect of thiourea, a thiourea-urea agarose IEF solution does not gel at room temperature but at 4°C, and the gel formed at 4°C does not melt even when the gel is returned to room temperature. From the practical point of view of an experimental scientist, the 1 M thiourea/5 M urea agarose IEF gel was a tremendous improvement over the 7 M agarose IEF gel originally used (2), because the agarose solution temperature no longer needed to be kept above 40°C when preparing agarose IEF gels.
4. When proteins in the agarose gel were fixed prior to the second-dimensional electrophoresis, the gel was extruded into a 300-mm-long, 5-mm-diameter glass tube filled with a protein fixing solution containing 10% trichloroacetic acid and 5% sulfosalicylic acid. Both ends of the 5-mm-diameter tube were connected to a Perista pump with PVC tubing to form a closed circuit filled with the fixing solution. When more than two gels were to be fixed, each of the gels was respectively extruded in a 5-mm-diameter glass tube, which were serially connected to one another to form a bigger closed circuit filled with the fixing solution. The proteins in the gel were fixed by 1 h circulation of the fixing solution in the 5-mm-diameter glass tube with a Perista pump, followed by 1 h circulation of 500 mL distilled water.
5. Avoid an SDS equilibration step, because such a step easily spreads proteins out of the agarose gel.

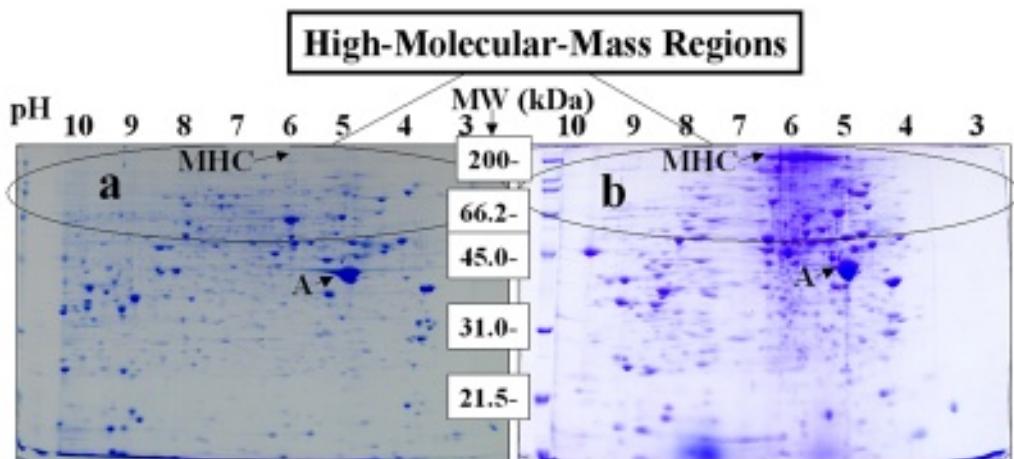


Fig. 1. Comparison of IPG-Dalt and agarose two-dimensional gel electrophoresis (2-DE). Tissues used are rat duodenum. (A) A 2-DE gel with immobilized pH gradients (IPGs) in the first dimension; (B) a 2-DE gel with agarose gels in the first dimension. Proteins loaded in weight on an agarose isoelectric focusing gel and an IPG gel were 740 µg. Note that spot densities are different in the high-molecular-mass regions. A = actin; MHC = myosin heavy chain. The gels were stained with PhastGel Blue R.

6. Thorough removal of Pharmalytes from first-dimensional agarose IEF gel is useful in two respects: one is to reduce spot deformations by Pharmalyte pH 8–10.5 in the basic pI (>8) and low-molecular-mass (<30 kDa) region of a 2-DE gel, and the other is to lower the background level of a Coomassie-stained 2-DE gel.

Acknowledgments

The authors thank Dr. Yoshio Kodera for helpful discussions. We also thank Ms. Kaori Dobashi for technical assistance. This work was supported by a Grant-in-Aid for Scientific Research on Priority Areas “Medical Genome Science” from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

1. O’Farrell, P. H. (1975) High resolution two-dimensional electrophoresis of proteins. *J. Biol. Chem.* **250**, 4007–4021.
2. Hirabayashi, T. (1981) Two-dimensional gel electrophoresis of chicken skeletal muscle proteins with agarose gels in the first dimension. *Anal. Biochem.* **117**, 443–451.
3. Hori, S., Sugiura, H., Shimizu, T., et al. (1989) Detection of dystrophin on two-dimensional gel electrophoresis. *Biochem. Biophys. Res. Commun.* **161**, 726–731.
4. Oh-Ishi, M. and Hirabayashi, T. (1989) Comparison of protein constituents between atria and ventricles from various vertebrates by two-dimensional gel electrophoresis. *Comp. Biochem. Physiol. B* **92**, 609–617.
5. Oh-Ishi, M., Satoh, M., and Maeda, T. (2000) Preparative two-dimensional gel electrophoresis with agarose gels in the first dimension for high molecular mass proteins. *Electrophoresis* **21**, 1653–1669.
6. Laemmli, U. K. (1970) Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **227**, 680–685.
7. Gorg, A., Postel, W., and Gunther, S. (1988) The current state of two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis* **9**, 531–546.

Using Ultra-Zoom Gels for High-Resolution Two-Dimensional Polyacrylamide Gel Electrophoresis

Sjouke Hoving, Hans Voshol, and Jan van Oostrum

1. Introduction

Two-dimensional polyacrylamide gel electrophoresis (2-D PAGE) is one of the core technologies—together with mass spectrometry—of proteome research. It is the only method currently available that is able to simultaneously separate the thousands of proteins found in biological samples. The method originates from the seminal work of O’Farrell and Klose in the 1970s (1,2). The main drawback of the original method was the poor reproducibility, due to the instability of the carrier ampholyte pH gradient in the first dimension. Insufficient sample loading capacity was another limitation. With the introduction of immobilized pH gradients (IPGs), the problems of gradient instability and loading capacity were largely overcome (3). The work of Görg and co-workers has further refined the technology with respect to the separation of proteins with extreme pI values (4). In short, the technology has made substantial improvements, but in many studies proteome analysis is carried out on single wide-range pH gels—for example, on pH 3.0–10.0 or on pH 4.0–7.0. While the first gradient lacks the high resolution (1500–2000 proteins per gel depending on the staining procedure) necessary to separate subtle differences between proteins, the latter does not cover the complete pH range. One possibility to obtain higher resolution and visualize more proteins is the use of a larger gel format, where up to 9000 proteins can be detected. However, this methodology makes use of carrier ampholyte-based IEF technology and large second-dimension gels. These gels are not commercially available (5). Another way of increasing the resolution is the separation of complex protein mixtures on a series of high-resolution “ultra-zoom” 2-D gels (Fig. 1). These multiple, partially overlapping narrow-range IPG strips combine high resolution with high sample loading, enabling the visualization of proteins of lower abundance (6,7). Whereas the ultra-zoom gels perform very well in the acidic pH range, care has to be taken when moving to higher pH. Especially above pH 7.0, the technique is not straightforward. On a complete series of overlapping 2-D gels it is possible to visualize over 10,000 unique protein spots (6–8). In this chapter, full details will be given on how to handle these ultra-zoom IPG strips. Different protocols are given for the use of acidic ultra-zoom IPGs and alkaline IPGs. For the alkaline gradients (IPG 6.2–8.2 and IPG 7.5–9.5), which are not commercially available, recipes are provided for their preparation. However, it is strongly advisable not to start a proteome analysis study on narrow-range strips in the alkaline range, but

From: *The Proteomics Protocols Handbook*
Edited by: J. M. Walker © Humana Press Inc., Totowa, NJ

first to optimize the sample preparation and the separation on wide-range IPGs. When the optimal conditions are obtained, narrow-range strips in the acidic pH range can be employed. Finally, the most common protocols for postseparation staining of 2-D gels are provided.

2. Materials

2.1. Equipment

This section describes the equipment that is used in our laboratory. However, most equipment is available from a variety of suppliers.

1. Flatbed electrophoresis system for isoelectric focusing (e.g., Multiphor II, Amersham Biosciences, Uppsala, Sweden). Although compact IEF systems, such as the IPGphor (Amersham Biosciences) or the Protean IEF cell (Bio-Rad Laboratories, Richmond, VA) are used in many laboratories, in our hands a conventional flatbed IEF system offers maximum reproducibility and flexibility in terms of sample loading.
2. Programmable Power Supply (e.g., EPS 3501 XL, Amersham Biosciences).
3. Multi-gel system for SDS-PAGE. In our laboratory the Iso-DALT system for SDS-PAGE (Amersham Biosciences) is used, which is not available anymore. Alternatives are, for example, the Ettan DALTwelve (Amersham Biosciences) or the Protean Plus Deodeca Cell (Bio-Rad Laboratories).
4. Thermostatic circulator (e.g., MultiTemp III, Amersham Biosciences).
5. Immobiline DryStrip Reswelling Tray (Amersham Biosciences).
6. Immobiline DryStrip Kit, including sample cups (Amersham Biosciences).
7. Rotary shaker.
8. Image capture devices (e.g., Personal Densitometer, Molecular Dynamics, Sunnyvale, CA, or FLA-3000R, Fuji, Tokyo, Japan).

2.2. Solutions and Reagents

Most reagents and ready-made solutions are available from a variety of suppliers. Here we indicate what is used in our laboratory. All solutions are prepared fresh before use, except where indicated.

1. The best quality of water available, e.g., from Milli-Q System (Millipore, Billerica, MA), or HPLC water (Fluka, Buchs, Switzerland).
2. Urea, thiourea, ethylenediaminetetraacetic acid (EDTA), and acrylamido buffer (pK 10.3) from Fluka (Buchs, Switzerland). CHAPS, dithiothreitol (DTT), iodoacetamide, and agarose from Sigma (St. Louis, MO). Pharmalytes, Immobilines, IPG buffers, and “CleanGel” electrode strips from Amersham Biosciences (Uppsala, Sweden). ProtoGel from National Diagnostics (Atlanta, GA). SDS, ammonium persulfate (APS), and TEMED from Bio-Rad Laboratories (Richmond, CA). GelBond PAG film from FMC (Rockland, ME). SYPRO Ruby from Molecular Probes (Eugene, OR). Coomassie Blue (Serva Blue G) from Serva (Heidelberg, Germany). All reagents were of the highest purity commercially available.
3. IPG strips, commercially available as pH 3.0–10.0, pH 3.0–10.0 NL, pH 4.0–7.0, pH 4.0–5.0, pH 4.5–5.5, pH 5.0–6.0, pH 5.5–6.7, pH 6.0–9.0, pH 6.0–11.0 (Amersham Biosciences), or pH 3.0–10.0, pH 3.0–10.0 NL, pH 4.0–7.0, pH 3.0–6.0, pH 5.0–8.0, pH 7.0–10.0, pH 3.9–5.1, pH 4.7–5.9, pH 5.5–6.7, and pH 6.3–8.3 (Bio-Rad Laboratories) (see **Note 1**).
4. Immobilines (pK 4.6, pK 6.2, pK 7.0, pK 8.5, and pK 9.3, Amersham Biosciences) and acrylamido buffers (pK 10.3, Fluka) to prepare custom made gradients.

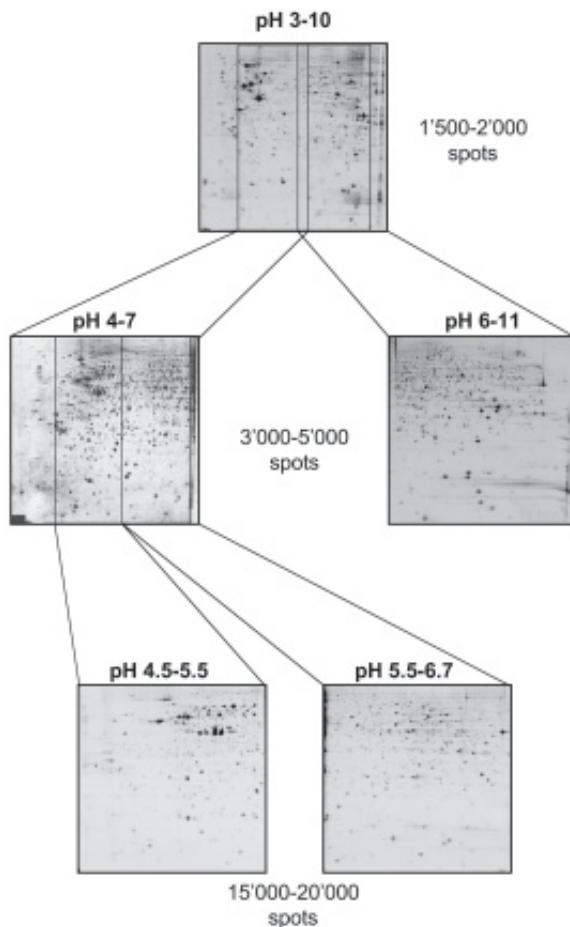


Fig. 1. The analysis depth of the proteome is limited when single wide-range pH gels such as pH 3.0–10.0 are used, because only about 1500 spots can be detected on such gels. The ultra-zoom gels, however, with multiple overlapping narrow-range immobilized pH gradient (IPG) strips for the first dimension, combine high resolution with high sample loading capacity to visualize more (low abundant) proteins per gel. This figure shows the separation of a protein extract of a pre-B lymphoma cell line on wide-range IPG (pH 3.0–10.0), two intermediate IPGs (pH 4.0–7.0 and pH 6.0–11.0), and two ultra-zoom IPGs (pH 4.5–5.5 and pH 5.5–6.7).

5. Sample buffer according to Rabilloud (9): 7 M urea, 2 M thiourea, 4% (w/v) CHAPS, 1% (w/v) DTT, 2% (v/v) Pharmalytes 3-10, cocktail of protease inhibitors (Complete, Roche, 1 tablet/50 mL solution). This sample buffer is stored at -80°C in small aliquots and should be thawed only once (see Note 2).
6. Rehydration solution according to Hoving et al. (7) for alkaline IPGs: 7 M urea, 2 M thiourea, 4% (w/v) CHAPS, 10% (v/v) iso-propanol, 5% (v/v) glycerol, 2.5% (w/v) DTT, 2% (v/v) IPG buffer 6-11, cocktail of protease inhibitors (Complete, Roche, 1 tablet/50 mL solution). This rehydration solution is stored at -80°C in 1 mL aliquots and should be thawed only once (see Note 3).
7. Equilibration solution: 50 μ M Tris-HCl (pH 8.8), 6 M urea, 30% (v/v) glycerol, 2% (w/v) SDS, and 0.01% (w/v) bromophenol blue.

8. Ready-made solutions for SDS-PAGE:
 - a. ProtoGel (30% (w/v) acrylamide, 0.8% (w/v) bisacrylamide stock solution (37.5:1) (National Diagnostics).
 - b. Gel buffer 1.5 M Tris-HCl, pH 8.8 (Bio-Rad Laboratories).
 - c. Running buffer (10X), final concentration 25 μ M Tris, 192 μ M glycine, 0.1% (w/v) SDS (Bio-Rad Laboratories).

3. Methods

Note: A very practical 2-D electrophoresis handbook (Product Code 80-6429-60) with many tips and tricks with respect to basic methodology can be downloaded for free at the following site: <http://www1.amershambiosciences.com/aptrix/upp00919.nsf/content/89432677DD1DB2FCC1256EB400418043?OpenDocument&hometitle=search>

3.1. Protein Sample Preparation

In general, protein samples were prepared from the 697 pre-B lymphoma cell line (available from the German Collection of Microorganisms and Cell Cultures, Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH, DSMZ No: ACC 42) (10), of which 10^7 cells correspond to approx 1 mg of extracted proteins. It is essential to wash the cells thoroughly (twice in PBS) in order to remove components of the growth medium (especially serum proteins) before collecting the cells. For the final cell pellet, all PBS is carefully removed with a fine pipet tip. Cell pellets (usually 10^8 cells per tube) are stored at -80°C . Human brain protein extracts and protein extracts from a wide variety of other organisms (for example, protein extracts from bacteria, yeast, and fruit flies) are best prepared by homogenizing the sample directly in the buffer described by Rabilloud (9) using a glass-Teflon homogenizer or a tip sonicator. The final protein concentration is approx 5–15 mg/mL.

1. Cells are lysed directly by resuspending the pellet rapidly in the sample buffer described by Rabilloud—see **Subheading 2.2.5.** (9)—followed by centrifugation for 10 min at 15,000 or 100,000g. It is critical to resuspend the cells rapidly to prevent proteolytic activity. The supernatant is applied to the strips, either by in-gel rehydration or by cup loading. Samples can be stored at -80°C (see **Note 4**).
2. Brain protein extracts are prepared by homogenizing the tissue directly in 8 vols of the same sample buffer (9) in a Potter-Elvehjem glass-Teflon homogenizer with eight strokes at 800 rpm on ice. Care has to be taken during the homogenizing not to cause foaming of the sample buffer. Before use, the samples are centrifuged for 15 min at 100,000g. Brain protein extracts have to be centrifuged at this speed to remove insoluble particles. Also in this case, the supernatant is applied to the strips either by in-gel rehydration or by cup loading. Samples can be stored at -80°C (see **Note 4**).

3.2. Immobilized pH Gradient: The First Dimension

Two-dimensional electrophoresis is performed according to standard protocols (11) with the following modifications.

1. Wide-range pH and ultra-zoom IPG strips in the acidic pH range are rehydrated in the sample buffer according to Rabilloud (9) with 0.5–2 mg protein (400 μ L solution per strip) in a reswelling cassette. Because of its simplicity, in-gel rehydration is the preferred method of protein application on the wide-range pH and acidic ultra-zoom gradients (12).

The maximum amount of protein that can be loaded on the ultra-zoom IPG strips is strongly dependent on the gradient. For example, the pH 4.5–5.5 strip is very robust, and up to 1.5 mg protein (brain extract) or even 3 mg (cell lysate) can be easily loaded on this particular gradient. Even much higher loadings of up to 15 mg protein by in-gel rehydration have been reported previously (12). On the other hand, the pH 5.5–6.7 gradient is much more sensitive to protein loading. It is noteworthy that it is not only the amount of sample or the protein content that influences the final 2-D PAGE results. At comparable protein loading, bacterial samples and cell lysates give much better results by 2-D PAGE than tissue extracts, such as from human brain (Fig. 2). This is mainly due to the much higher complexity of the tissue sample, with its different cell types and—especially in brain tissue—high lipid content.

2. The strips are overlaid with 2 mL paraffin oil to prevent urea crystallization. Rehydration of the strips should take at least 6 h, but recommended and most practical is rehydration overnight at room temperature. Some proteins, especially high-molecular-weight proteins, need more time to enter into the strip during the rehydration step.
3. After complete rehydration, the strips are aligned on a flatbed electrophoresis system, where the temperature of the cooling block is held constant at 20°C.
4. The Immobiline DryStrip Kit components such as humid paper wicks and electrodes are positioned and the strips are covered with paraffin oil (see Note 5).
5. Focusing is always started at 300 V to desalt the sample, and the voltage is slowly increased to 3500 V until a total of about 85 kWh is reached (see Table 1 for detailed running conditions) (see Note 6).

3.3. Immobilized Alkaline pH Gradient: The First Dimension

The conditions for running alkaline (zoom) IPG strips have to be changed compared to the acidic ultra-zoom IPGs. During isoelectric focusing at alkaline pH, water transport and migration of the reducing agent DTT takes place. To minimize these effects, a higher amount of DTT is used in the rehydration buffer, and DTT is constantly introduced at the cathode in combination with 10% isopropanol and 5% glycerol, resulting in adequate focusing at higher pH. In contrast with the acidic pH range, in the alkaline pH range, one pH unit IPGs of 180 mm length do not yield acceptable separations (Fig. 3). Therefore gradients spanning 2–3 pH units are used, e.g., pH 6.2–8.2, pH 7.5–9.5 (custom made strips), pH 6.0–9.0, pH 6.0–11.0 (Amersham Biosciences), or pH 6.3–8.3, pH 7.0–10.0 (Bio-Rad Laboratories). On the alkaline IPG strips, protein samples are always applied by cup loading at the anodic side and not by in-gel rehydration (Fig. 4). The maximum protein load on alkaline zoom gels is lower than on the acidic zoom gels. It is strongly suggested not to load more than 0.5 mg protein per cup to prevent streaking.

1. The IPG strip is rehydrated (without protein sample) with the rehydration buffer supplemented with 10% iso-propanol and 5% glycerol—see Subheading 2.2.6. (7)—in the reswelling cassette or between two glass plates at room temperature. Since no proteins are present during rehydration, the time can be as short as 6 h. Therefore, it is possible to start rehydration of the strips in the morning and start focusing the same day.
2. After complete rehydration, the IPG strips are placed on the flatbed electrophoresis system, where the temperature of the cooling block is held constant at 20°C.
3. The Immobiline DryStrip Kit components such as humid paper wicks, electrodes, and sample cups (at anode) are positioned and the strips are covered with paraffin oil (see Notes 5 and 7).

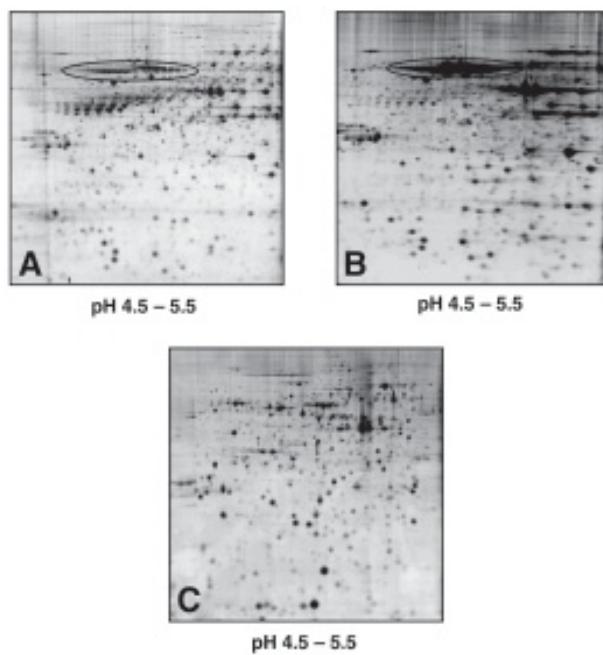


Fig. 2. (A) Human brain (Brodmann area 7) protein extract (750 μ g protein) on immobilized pH Gradient (IPG) 4.5–5.5 gel and (B) human brain (Brodmann area 7) protein extract (2 mg protein) on IPG 4.5–5.5 gel. The latter gel is clearly overloaded, resulting in heavy streaking at the high molecular weight and loss of resolution. (C) Pre-B lymphoma protein extract (1 mg) on IPG 4.5–5.5.

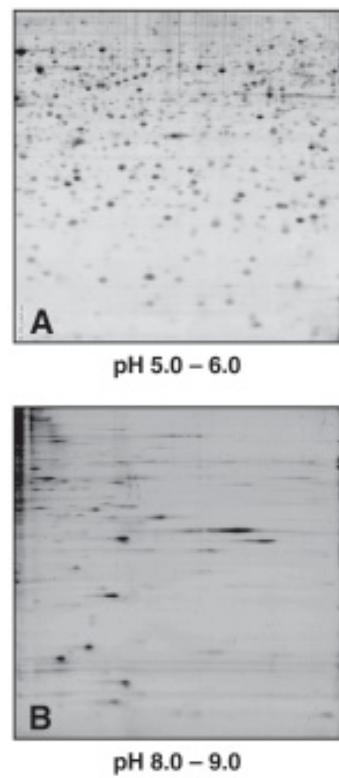


Table 1
Running Conditions for Immobilized pH Gradients (IPGs)
(all voltage gradients are linear)

Applicable for IPGs with pH 3.0–10.0, pH 3.0–10.0 NL, pH 4.0–7.0, pH 3.0–6.0, pH 5.0–8.0, pH 4.5–5.5, pH 5.0–6.0, pH 5.5–6.7, pH 3.9–5.1 and pH 4.7–5.9.

Vstart (V)	Vend (V)	Time (h)	Volt-hour product (Vh)
0	300	0.01	1.5
300	300	3	900
300	3500	5	9500
3500	3500	20	70000
	Total	28.01	80401.5

Vstart (V)	Vend (V)	Time (h)	Volt-hour product (Vh)
0	300	0.01	1.5
300	300	3	900
300	1400	6	5100
1400	1400	10	14000
1400	3500	3	7350
3500	3500	2	7000
	Total	24.01	34351.5

Vstart (V)	Vend (V)	Time (h)	Volt-hour product (Vh)
0	300	0.01	1.5
300	300	3	900
300	3500	5	9500
3500	3500	15	52,500
	Total	23.01	62,901.5

4. The sample is carefully pipetted under the oil (maximum 100 μ L sample solution per cup). Since the samples are prepared for all IPG gradients, only Pharmalytes 3–10 are present in the sample. In general it is not feasible to use different sample preparations for each gradient. No specific IPG buffers are added at this stage (see Note 8).
5. Because DTT will migrate towards pH 7.0, it is necessary to add a “CleanGel” paper wick at the cathodic side, immersed in a modified buffer containing 3.5% (w/v) DTT. At this point it is also possible to use the reducing agent hydroxylethyl disulphide (DeStreak, Amersham Biosciences) instead of using DTT (see Notes 3 and 9).
6. Focusing is always started at 300 V to allow the proteins to enter the gel, and the voltage is slowly increased until a total of about 34 kWh (IPG 6–9 and 6–11) or 64 kWh (IPG 6.2–8.2 and 7.5–9.5) is reached (see Table 1 for detailed running conditions) (see Note 6).

Fig. 3. (opposite page) Ultra-zoom gels of one pH unit. (A) Immobilized pH gradient (IPG) 5.0–6.0 with 1 mg of pre-B lymphoma protein extract and (B) IPG 8.0–9.0 with 0.5 mg of pre-B lymphoma protein extract. Whereas the acidic narrow gradient produces a good protein separation, the alkaline narrow gradient leads to heavy streaking.

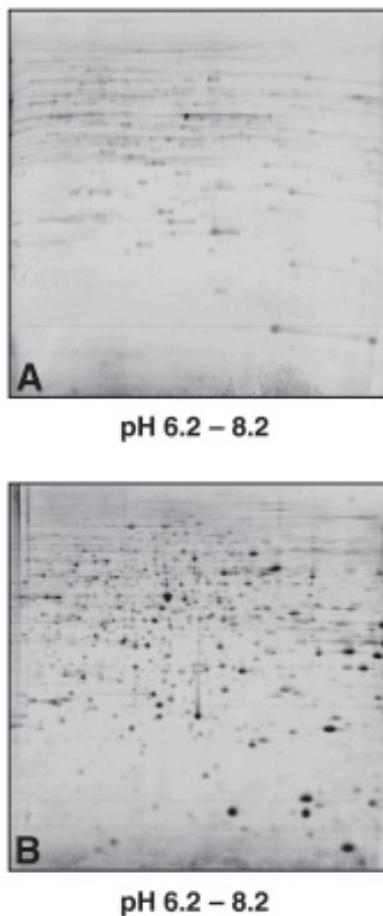


Fig. 4. (A) Sample application by in-gel rehydration, 400 µg of pre-B lymphoma protein extract and (B) sample application by cup-loading, 400 µg of pre-B lymphoma protein extract on an immobilized pH gradient (IPG) 6.2–8.2 strip. These strips were focused under the modified conditions for alkaline IPGs with DeStreak as reducing agent.

3.4. Preparation of Narrow-Range Immobilized Alkaline pH Gradients

This section describes briefly how custom-made alkaline IPGs are prepared. For detailed protocols on how to utilize the casting mould and the polymerization conditions, the reader is referred to the literature (13,14). Several computer software programs for pH gradient modeling, with routines optimizing the concentrations of the acrylamido buffers needed to give the required pH gradient, have been described (15,16). Our protocol for the IPG with a linear gradient of pH 6.2–8.2 and pH 7.5–9.5 was calculated using an in-house developed program, “Immobilizer.” **Table 2** shows the recipes for these gradients, resulting in approx 180-mm-long strips.

1. The two solutions—“heavy” and “light”—are prepared according to the recipe in **Table 2**. The catalysts APS and TEMED are subsequently added to the gradient mixer just before starting the casting (see **Note 10**).
2. The mixed solution is simply flowed into the casting mould by gravity and overlaid with some water-saturated iso-butanol.

Table 2**Casting of Alkaline Immobilized pH Gradient Linear Gradient pH 6.2–8.2 and pH 7.5–9.5**

Solutions ^a	pH 6.2 “heavy” solution	pH 8.2 “light” solution	pH 7.5 “heavy” solution	pH 9.5 “light” solution
Immobiline pK 4.6	286 µL	202 µL	193 µL	139 µL
Immobiline pK 6.2	190 µL	222 µL	176 µL	176 µL
Immobiline pK 7.0	146 µL	111 µL	107 µL	225 µL
Immobiline pK 8.5	56 µL	288 µL	163 µL	45 µL
Immobiline pK 9.3	—	—	11 µL	137 µL
Acrylamido buffer pK 10.3	—	—	—	89 µL
Acrylamide/ Bisacrylamide ^b	1.08 mL	1.08 mL	1.08 mL	1.08 mL
50% (v/v) Glycerol	3.47 mL	0.52 mL	3.47 mL	0.52 mL
Deionized water	1.25 mL	4.06 mL	1.25 mL	4.06 mL
TEMED	3 µL	3 µL	3 µL	3 µL
10% (w/v) Ammo- nium persulfate	15 µL	15 µL	15 µL	15 µL
Final volume	6.5 mL	6.5 mL	6.5 mL	6.5 mL

^aFor effective polymerization, the solutions are adjusted to pH 7.0 with 3 M acetic acid before adding the

3. The gradient is left for 5–10 min at room temperature to stabilize, before moving the cassette into a heating cabinet for polymerization for 1 h at 50°C.
4. The IPG plate is extensively washed with 1% (v/v) glycerol solution to remove any unpolymerized materials (see Note 11).
5. The IPG plate is then dried in a dust-free heating cabinet (with a ventilator), and strips (3-mm width) are cut using a typical photo paper cutter. Strips are stored at –20°C (shelf life approx 1 yr).

3.5. Equilibration of IPG Gels

After the isoelectric focusing step is completed, it is necessary to equilibrate the IPG strips in two steps prior to the second dimension. The equilibration buffer contains 6 M urea and 30% glycerol in order to diminish electro-endosmotic effects (17), which are responsible for poor transfer from first to second dimension. Further, the equilibration buffer contains 2% (w/v) SDS to make the proteins negatively charged for the SDS-PAGE. Loss of proteins during the equilibration step and subsequent transfer from the first to the second dimension are mainly due to proteins that are strongly absorbed to the IPG gel matrix, and due to wash-off effects. The majority of these proteins appear to be lost during the first minutes of equilibration, whereas protein losses during the second equilibration step are only minimal (17).

1. After focusing is complete, the paraffin oil is poured off. The equilibration of the IPG strips is performed in the DryStrip holder to minimize strip handling (see Note 12).
2. The strips are first equilibrated in 100 mL equilibration buffer supplemented with 2% (w/v) DTT (12–20 min) to reduce disulfide bonds.
3. In the second step, the strips are equilibrated in 100 mL equilibration buffer supplemented with 5% (w/v) iodoacetamide (6–20 min) to alkylate the formed sulphydryl groups. This is

to prevent re-oxidation of the sulphydryl groups during the second-dimension SDS-PAGE in order to prevent streaking of spots.

3.6. SDS-PAGE: The Second Dimension

When equilibration is completed, the strips are ready for the second dimension (SDS-PAGE). When more than one gel is to be run (which is always the case in a comparative proteomics study), only vertical systems are recommended. In general, three to five identical gels for each sample are prepared. Typically, electrophoresis chambers holding 10–12 large SDS-PAGE gels are used. Home-made gels are commonly used, but some ready-made gels are available on the market. In our hands, best results are obtained using 12% pore size gels (12% T; 2.6% C). These gels are the best compromise between ease of preparation and resolution. Several laboratories use gradient gels, which provide better resolution but are difficult to prepare reproducibly on a large scale. The Iso-DALT system (or similar systems currently on the market) is designed to allow a relatively large batch of slab gels to be run under identical conditions.

1. In the casting chamber, 22–24 slab gels are prepared. The glass cassettes (with fixed spacers—1.5 mm—and a hinge) are assembled in the chamber and separated by a plastic sheet to prevent the plates sticking to each other. Each gel is given a unique serial identity number by placing a piece of printed Whatman paper between the glass plates prior to casting.
2. In a vacuum flask, prepare 920 mL ProtoGel (30% (w/v) acrylamide, 0.8% (w/v) bisacrylamide solution (37.5:1), 575 mL 1.5 M Tris-HCl (pH 8.8), and 770 mL water.
3. The mixture is degassed for 10 min before the SDS (2.3 g in 35 mL water) and catalysts (700 mg APS and 300 μ L TEMED) are added.
4. The gel casting chamber is filled from the bottom to a height of about 2 cm below the top of the glass plates. The gels are carefully overlaid with 1.0–1.5 mL buffer-saturated isobutanol to allow for complete polymerization (at least 1 h). Gels can be stored up to 1 wk at 4°C (*see Note 13*).
5. Before use, polyacrylamide residues are cleaned off of the glass plates, and they are placed in racks before placing the equilibrated IPG strip on top of the gel.
6. In order to ensure good contact between the strip and the gel, an agarose solution is added. The agarose not only keeps the IPG strip in place, but also ensures good electrophoretic properties between the IPG strip and the gel. The agarose solution is kept at 65°C and added first on top of the gel. Immediately after this, the equilibrated strip is placed on the gel (*see Note 14*).
7. If desired, a molecular-weight marker can be run at the side (loaded through a small paper wick). Allow for 10 min to solidify the agarose before the gels are placed in the Iso-DALT gel chamber.
8. The current is set to 10 mA/gel initially to let the proteins enter the gel, and subsequently set to 20 mA/gel overnight at 15°C until the blue front reaches the end of the gel (*see Note 15*).

3.7. Staining and Storage of 2-D Gels

Proteins can be marked before electrophoresis by labeling methods such as DIGE (18,19), by incorporating radioactivity (^{32}P , ^{35}S , ^{14}C , and so on), or by several postseparation staining methods. Labeling methods have to be precisely optimized for each sample, whereas incorporating radioactivity is not applicable to postmortem tissue samples. After electrophoresis is completed, for postseparation staining the gels have to be fixed and the proteins visualized. Proteins can be stained with the fluores-

Table 3
Most Common Postseparation Protein Stains for 2-D Gels in Proteome Analysis

Staining*	Detection (ng)	Dynamic range	Sensitivity	Reproducibility	Cost	References
Coll. Coomassie Blue	15	+/-	+/-	++	++	21
Sypro Ruby	5	++	+	++	-	20
Silver (nitrate)	1-2	-	++	-	+	22

The detection limit is given per protein spot in the gel. (-) not good, (+/-) acceptable, (+) good, and (++) very good. For further explanation see text.

*For an accurate differential analysis on 2-D gels, it is essential that the staining procedure is both highly reproducible and quantitative. Fluorescent dyes for protein staining combine good reproducibility with high dynamic range for quantitation and are therefore the dyes of choice in 2-D analysis.

cent dye SYPRO Ruby (20), colloidal Coomassie blue (21), or with a silver stain that is compatible with mass spectrometry (22). For specific staining of phosphoproteins, a novel fluorescent dye Pro-Q Diamond has been described recently (23). **Table 3** summarizes the properties of each staining. The type of experiment determines which stain is used; for example, for quantitative comparison studies, fluorescent dyes are preferred, since the dynamic range of these stains cover a broad concentration range (approx 2 ng–1 µg protein per spot). **Table 4** provides detailed protocols for these protein stains. Gels are scanned on a fluorescence scanner (Fuji FLA-3000R) or on the Personal Densitometer (Molecular Dynamics), and image analysis is performed with Progenesis (Nonlinear Dynamics, Newcastle-upon-Tyne, UK). For a detailed discussion on image analysis, the reader is referred to the appropriate chapter of this volume. Gels can be stored in boxes in water at 4°C for months (see **Note 16**).

4. Notes

1. All strips described for these protocols have following dimensions: length 180 mm, width 3 mm, average thickness (after reswelling) approx 0.5 mm. Gel rehydration causes more swelling at the alkaline end of the strips. Other IPG lengths can be used, especially the mini format (7 cm) and the large format (24 cm). In both cases, mainly the sample loading conditions and focusing conditions need to be adjusted.
2. Never heat urea-containing solutions above 37°C to avoid carbamylation of proteins.
3. Instead of 2.5% (w/v) DTT, the reducing agent hydroxyethyl disulphide (DeStreak Reagent, Amersham Biosciences) can be added (12 µL/mL rehydration solution) to improve the resolution at alkaline pH.
4. At this stage it is important that the rehydration solution contain 2% (v/v) Pharmalytes 3–10 in order to prevent aggregation of DNA. Cells (e.g., 10⁸ cells) are best disrupted by rapid addition of 1 mL sample buffer (8) to the cell pellet. Nucleases (DNase, RNase, and benzonase) are commonly added as well, but their efficacy is questionable in this highly denaturizing buffer.
5. Because of the relative high sample loads on ultra-zoom IPGs, it is necessary to use paper wicks at the electrodes to collect proteins that are outside the pH range of interest. The paper wicks function as a desalting tool as well; it is possible to exchange the paper wicks during the IEF. The paraffin oil prevents drying out of the strip and crystallization of the urea/thiourea.

Table 4

Most Common Protocols for Protein Staining As Used in Our Laboratory
(Updated protocols for the fluorescent dyes can be found at <http://www.probes.com/lit>. For both colloidal Coomassie and silver staining, many variations on these protocols are used.)

Staining with the fluorescent dye SYPRO Ruby (17)		
Step	Solutions	Time
1. Fixing ^a	40% (v/v) ethanol, 10% (v/v) acetic acid	3 h
2. Washing	Distilled water	3 × 30 min
3. Staining	SYPRO Ruby ready-to-use solution (150 mL/gel)	O/N
4. Washing	Distilled water	2 × 30 min
Staining with colloidal Coomassie Blue (18)		
Step	Solutions	Time
1. Fixing	50% (v/v) ethanol, 3% (v/v) phosphoric acid	3 h
2. Washing	Distilled water	3 × 30 min
3. Staining—first step	34% (v/v) methanol, 3% (v/v) phosphoric acid, 17% (w/v) ammonium sulfate	1 h
4. Staining—second step ^b	Coomassie Blue G-250 (350 mg/L into solution 3A)	1–5 d ^c
5. Washing	Distilled water	3 × 30 min ^d
Staining with MS compatible silver (19)		
Step	Solutions	Time
1. Fixing	40% (v/v) ethanol, 10% (v/v) acetic acid	3 h
2. Washing	30% (v/v) ethanol	2 × 20 min
3. Washing	Distilled water	20 min
4. Sensitizing	0.02% (w/v) sodium thiosulfate	3 min
5. Washing	Distilled water	3 × 1 min
6. Staining	0.2% (w/v) silver nitrate	30–60 min
7. Washing	Distilled water	3 × 1 min
8. Developing ^e	3% (w/v) sodium carbonate, 0.05% (v/v) formaldehyde solution (37%)	2 × 5–10 min
9. Washing	Distilled water	1 min
10. Stopping ^f	1.5% (w/v) Na ₂ EDTA	10 min
11. Washing	Distilled water	3 × 10 min

^aFixing times are minimal for large-format gels (20 × 25 cm). Gels can be left overnight in fixing solution. Up to five gels can be processed simultaneously during the staining procedure.

^bCoomassie Blue G-250 (also available as Serva Blue G-250) is added as solid to the solution and immediately forms small colloidal particles. The gel boxes are sealed with tape to prevent evaporation.

^cAfter overnight staining, spots will be visible, but complete end-point staining will be reached after 4–5 d.

^dThe gels are sufficiently washed when no solid Coomassie particles are present anymore.

^eDuring the developing process, it is best to limit the number of gels at two during processing. As soon as the developing solution becomes yellow, it should be replaced with fresh solution. Development time is experimentally determined; the background should not be too high.

^fStopping the development in a Na₂ethylenediaminetetraacetic acid (EDTA) solution prevents the gel spots from bleaching from brown/black to yellow.

6. For the wide-range IPGs, such as pH 3.0–10.0, pH 3.0–10.0 NL, and pH 4.0–7.0, the last step of the voltage gradient might be shortened until a total volt-hour product of about 65–70 kWh. Running conditions will change with lower protein loadings. But the ultra-zoom IPG strips are generally used in conjunction with high protein loads.
7. The paraffin oil serves here as well to prevent carbon dioxide from entering the system. Carbon dioxide from the air can dissolve into the IPG gel, acting as a buffer with pK 6.3 and thus changing the pH gradient.
8. The protein concentration in the sample cup should not be too high (<10 mg/mL) to avoid protein precipitation at the entry point. If possible, the sample should be diluted with the sample buffer.
9. “Clean-Gel” paper wicks (Amersham Biosciences) were originally used for high protein loading, and consist of thick filter paper (24). When using the DeStreak reducing agent, it is not necessary to use the “Clean-Gel” paper wicks; in that case the normal humid paper wicks are sufficient.
10. It is important that fresh acrylamide/bisacrylamide solutions be used for the preparation of IPG strips (maximum 1 wk old).
11. Alkaline IPG strips can swell considerably; adding 1% (v/v) glycerol to the washing solution prevents excessive swelling of the gel.
12. Especially at this stage, the Multiphor II system with the Immobiline DryStrip Kit offers a great advantage over other systems, since the two equilibration steps can be performed in the same tray, without any strip handling. This avoids contamination or any other disruption of the strips. In our hands it is currently the best possible way for a semi-high-throughput running of 2-D gels.
13. Casting a large number of gels requires some practice. The amount of catalyst that is used is minimal to prevent the gels from polymerizing too rapidly, which leads to excessive heating of the casting chamber. Ideally, initial polymerization, which can be seen by the development of a distinct gel surface below the butanol layer, should take 30–60 min. Subsequently, gels can be washed, covered with gel buffer, and stored at room temperature. Overnight, residual polymerization will take place. The amounts of catalyst can be increased by 10% if necessary.
14. The agarose overlay is prepared in such a way that it remains fluid at relative low temperatures. In order to realize this, a mixture of low-melting agaroses is used (0.4% [w/v] standard low Mr from Bio-Rad Laboratories and 0.1% (w/v) type VII-A-low gelling temperature from Sigma, dissolved in running buffer). A trace of bromophenol blue can be added as tracking dye.
15. It is important to fill the Iso-DALT chamber with 20 L running buffer beforehand and start the cooling at 15°C. It will take a few hours until the temperature is reached.
16. Gels can be stored in the refrigerator or in a cold room at 4°C for months, provided that 0.02% (w/v) sodium azide is present to prevent bacterial and fungal growth. It is no problem to identify proteins from gel spots that were stored over a long period of time.

Acknowledgments

The authors wish to acknowledge and thank Alexandra Rüchti and Klaus Eichin for excellent technical assistance.

References

1. O’Farrell, P. H. (1975) High resolution two-dimensional electrophoresis of proteins. *J. Biol. Chem.* **250**, 4007–4021.

2. Klose, J. (1975) Protein mapping by combined isoelectric focusing and electrophoresis in mouse tissues. A novel approach to testing for induced point mutations in mammals. *Humangenetik* **26**, 231–243.
3. Bjellqvist, B., Ek, K., Righetti, P. G., et al. (1982) Isoelectric focusing in immobilized pH gradients: principle, methodology and some applications. *J. Biochem. Biophys. Methods* **6**, 317–339.
4. Görg, A., Obermaier, C., Boguth, G., Csordas, A., Diaz, J. J., and Madjar, J. J. (1997) Very alkaline immobilized pH gradients for two-dimensional electrophoresis of ribosomal and nuclear proteins. *Electrophoresis* **18**, 328–337.
5. Klose, J. (1999) Large-gel 2-D electrophoresis. *Methods in Molecular Biology* **112**, 147–172.
6. Hoving, S., Voshol, H., and van Oostrum, J. (2000) Towards high performance two-dimensional gel electrophoresis using ultrazoom gels. *Electrophoresis* **21**, 2617–2621.
7. Hoving, S., Gerrits, B., Voshol, H., Müller, D., Roberts, R. C., and van Oostrum, J. (2002) Preparative two-dimensional gel electrophoresis at alkaline pH using narrow range immobilized pH gradients. *Proteomics* **2**, 127–134.
8. Fey, S. J. and Mose Larsen, P. (2001) 2D or not 2D. Two-dimensional gel electrophoresis. *Curr. Opin. Chem. Biol.* **5**, 26–33.
9. Rabilloud, T. (1998) Use of thiourea to increase the solubility of membrane proteins in two-dimensional electrophoresis. *Electrophoresis* **19**, 758–760.
10. Findley, H. W., Cooper, M. D., Kim, T. H., Alvarado, C., and Ragab, A. H. (1982) Two new acute lymphoblastic leukemia cell lines with early B-cell phenotypes. *Blood* **60**, 1305–1309.
11. Görg, A., Obermaier, C., Boguth, G., et al. (2000) The current state of two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis* **21**, 1037–1053.
12. Sanchez, J. C., Rouge, V., Pisteur, M., et al. (1997) Improved and simplified in-gel sample application using reswelling of dry immobilized pH gradients. *Electrophoresis* **18**, 324–327.
13. Westermeier, R. *Electrophoresis in practice*, 3rd Edition, Wiley-VCH, Weinheim 2001, pp. 223–238.
14. Gianazza, E. (1999) Casting immobilized pH gradients (IPGs). In: Link, A. J. (ed), *2-D Proteome Analysis Protocols*. Humana, Totowa, NJ: 175–188.
15. Altland, K. (1990) IPGMAKER: a program for IBM-compatible personal computers to create and test recipes for immobilized pH gradients. *Electrophoresis* **11**, 140–147.
16. Giaffreda, E., Tonani, C., and Righetti, P. G. (1993) pH gradient simulator for electro-phoretic techniques in a windows environment. *J. Chromatogr.* **630**, 313–327.
17. Görg, A., Postel, W., and Günther, S. (1988) The current state of two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis* **9**, 531–546.
18. Unlu, M., Morgan, M. E., and Minden, J. S. (1997) Difference gel electrophoresis: a single gel method for detecting changes in protein extracts. *Electrophoresis* **18**, 2071–2077.
19. Yan, J. X., Devenish, A. T., Wait, R., Stone, T., Lewis, S. and Fowler, S. (2002) Fluorescence two-dimensional difference gel electrophoresis and mass spectrometry based proteomic analysis of *Escherichia coli*. *Proteomics* **2**, 1682–1698.
20. Lopez, M. F., Berggren, K., Chernokalskaya, E., Lazarev, A., Robinson, M., and Patton, W. F. (2000) A comparison of silver stain and SYPRO Ruby Protein Gel Stain with respect to protein detection in two-dimensional gels and identification by peptide mass profiling. *Electrophoresis* **21**, 3673–3683.
21. Neuhoff, V., Arold, N., Taube, D., and Ehrhardt, W. (1988) Improved staining of proteins in polyacrylamide gels including isoelectric focusing gels with clear background at nanogram sensitivity using Coomassie Brilliant Blue G-250 and R-250. *Electrophoresis* **9**, 255–262.

22. Blum, H., Beier, H., and Gross, H. J. (1987) Improved silver staining of plant proteins, RNA and DNA in polyacrylamide gels. *Electrophoresis* **8**, 93–99.
23. Steinberg, T. H., Agnew, B. J., Gee, K. R., et al. (2003) Global quantitative phosphoprotein analysis using Multiplexed Proteomics technology. *Proteomics* **3**, 1128–1144.
24. Sabourchi-Schütt, F., Astrom, J., Olsson, I., Eklund, A., Grunewald, J., and Bjellqvist, B. (2000) An Immobiline DryStrip application method enabling high-capacity two-dimensional gel electrophoresis. *Electrophoresis* **21**, 3649–3656.

NEpHGE and pI Strip Proteomic 2-D Gel Electrophoretic Mapping of Lipid-Rich Membranes

Steven E. Pfeiffer, Yoshihide Yamaguchi, Cecilia B. Marta,
Rashmi Bansal, and Christopher M. Taylor

1. Introduction

Two-dimensional gel electrophoresis (2-DE) has become a powerful and widely used technique for proteomic analyses. However, proteins that are highly basic (pI 8–12), suggesting interactions with acidic sulfo- and phospholipids (1), can become compacted at the edge of the gel. Even if the isoelectric focusing (IEF) pH gradient is extended to high pH, slightly basic proteins enter the gel but are not well resolved, and the pH gradient is not sufficiently extended to include highly basic proteins. Therefore, the use of nonequilibrium pH gradient electrophoresis (NEpHGE) may be recommended (1–3) (see Note 1). In the first dimension, positively charged, basic proteins move toward the negative end of the gel while the pI gradient is forming. Because running the gel to equilibrium would result in many highly basic proteins exiting from the basic end of the gel (negative electrode), the electrophoresis must be stopped at a critical point (thus, “nonequilibrium”).

Further, the limited ability of 2-DE to resolve transmembrane and glycosyl-phosphatidylinositol (GPI)-anchored proteins has slowed the identification of proteins from some biological samples, especially those high in lipid content, such as the myelin membrane. Therefore, as a prelude to an in-depth proteomic analysis of the complex community of myelin proteins (4–7) (see Notes 2–4), we have identified conditions for the improved separation and display of proteins by 2-DE using pI strips, including those that emphasize the high pI end of the scale. In particular, we have identified two detergents, the nonionic *n*-dodecyl β -D-maltoside (DM) and the zwitterionic amido-sulfobetaine 14 (ASB-14) (see Note 5), that are more effective in solubilizing myelin proteins than the commonly used zwitterionic detergent CHAPS (8). These detergents significantly enhance the solubility of both transmembrane and GPI-anchored myelin proteins, and enable their resolution by 2-DE. Our results demonstrate significant improvements in the resolution of myelin proteins by 2-DE, using an IEF buffer consisting of 7 M urea, 2 M thiourea, 100 mM dithiothreitol (DTT), 0.5% carrier ampholytes, and either 2% DM or 2% ASB-14 (8).

Here we present protocols for applying these two methods of 2-DE. These protocols have proven to be effective tools for the 2-DE analysis of lipid-rich samples such as

myelin, and they are expected to be more generally useful for the analysis of membrane-rich biological samples.

2. Materials

Reagents are obtained from Sigma (St. Louis, MO) unless otherwise indicated. All solutions are made in nanopure water.

2.1. NEpHGE Electrophoresis

1. Standard tube gel electrophoresis apparatus.
2. Gel tubes, 3.5 mm diameter \times 13 cm long.
3. Hypodermic syringe, 10 mL, with long thin needle.
4. Petri dishes, 60 mm diameter.
5. Gel solution (4%): The amount of gel solution indicated is sufficient for six tube gels of 3.5 mm inside diameter \times 13 cm long: urea, 9.2 M (5.5 g), 4% acrylamide-*bis* (19:1) (1.0 mL, 40% stock solution, BioRad Laboratories, Hercules, CA), 2% NP-40 (2.0 mL, 10% stock solution), AmpholineTM pH 7.0–9.0 (225 μ L) and Ampholyte pH 9.0–11.0 (25 μ L) (Sigma-Aldrich Chemical Co., St. Louis, MO, or other equivalent reagents such as Servalyt 9-11, Serva Electrophoresis GmbH, Heidelberg, Germany), water (2.5 mL), ammonium persulfate (20 μ L, 10% stock solution), *N,N,N',N'*-tetramethylethylenediamine (TEMED) (14 μ L). All components except ammonium persulfate (APS) and TEMED are mixed and the solution is de-gassed under vacuum.
6. Lysis buffer: urea, ultrapure grade, 8.6 M (2.59 g), 4% NP-40 (2.0 mL, 10% stock solution), DTT (Fisher Scientific, Fairlawn, CA), 100 mM (500 μ L, 1 M stock solution), Ampholyte pH 7.0–9.0 (300 μ L), H₂O (350 μ L)
7. Equilibrium buffer: 0.4 mM ethylenediaminetetraacetic acid (EDTA) (80 μ L, 500 mM [pH 8.0] stock solution), 10% glycerol (10 mL, 100% stock), 3% sodium dodecyl sulfate (SDS) (30 mL, 10% stock solution), 20 mM Tris-HCl (pH 8.8) (1.33 mL, 1.5 M stock solution), water to 100 mL.
8. 20 mM Sodium hydroxide.
9. 10 mM Phosphoric acid.

2.2. pI Strip 2-D Gel Electrophoresis

1. IPGphor IEF unit and a Hoefer DALT Vertical System (Amersham Biosciences, Piscataway, NJ), or equivalent apparatus.
2. Precast immobilized pH gradient (IPG) gel (IPG strip, pI 3–10 (or pI range of interest), 18 cm (Amersham Biosciences, Piscataway, NJ).
3. IPGphor strip holders, 18 cm (Amersham Biosciences, Piscataway, NJ).
4. Centrifuge for 24,000g sedimentation.
5. Bath sonicator.
6. Filter wicks (BioRad Laboratories, Hercules, CA).
7. Mineral oil.
8. Platform rocker.
9. Syringes, 20 mL, with needles.
10. Lysis buffer (per 1 mL; freshly made from stock solutions): 1% amidosulfobetaine-14 (ASB-14) (Calbiochem, La Jolla, CA) (100 μ L, 10% stock solution), 25 mM Tris (100 μ L, 250 mM stock solution), 5 mM EDTA (100 μ L, 50 mM stock solution), protease inhibitor cocktail (1 μ L leupeptin/aprotinin, 10 mg/mL stock solution, 5 μ L PMSF, 200 mM stock solution).

Table 1
Acrylamide Solutions

Acrylamide, %	7	8	9	10	11	12	13	14	15	16
Acrylamide Stock, 30%	156	178	200	223	244	267	289	311	333	356
Tris-HCl, 1.5 M, pH 8.8	167	167	167	167	167	167	167	167	167	167
Water	329	308	285	263	241	219	197	175	153	130
SDS, 10% stock	6.7	6.7	6.7	6.7	6.7	6.7	6.7	6.7	6.7	6.7
APS, 2.5% stock	6.7	6.7	6.7	6.7	6.7	6.7	6.7	6.7	6.7	6.7
TEMED, 10% stock	1.63	1.43	1.27	1.14	1.03	0.95	0.88	0.82	0.76	0.72

For 10 gels, in mL, to make 667 mL. APS, ammonium persulfate; SDS, sodium dodecyl sulfate.

Freshly prepared 2.5-3% APS is used; if 10% APS is used (such as for 1-D gels), the gel will polymerize before the pouring is completed. All components, except APS and TEMED, are mixed and the solution is de-gassed under vacuum.

11. Rehydration buffer (per 1 mL): 7 M urea, ultrapure grade (400 mg), 2 M thiourea (Fisher Scientific, Fair Lawn, NJ) (150 mg), 2% ASB-14 (200 μ L, 10% stock solution), 0.5% pI 3–10 carrier ampholytes (5 μ L of commercial stock, or other ampholytes for different pI ranges) (Amersham Biosciences, Piscataway, NJ), 0.001% bromophenol blue (1 μ L, 1% stock solution), 100 mM DTT, added just before use (15 mg). Urea, thiourea, and the carrier ampholytes can be combined and frozen, but not solutions of ASB-14.
12. Equilibration buffer (per 1 L): 50 mM Tris-HCl (pH 8.8) (33.5 mL, 1.5 M stock solution); 6 M urea (ultrapure grade, ICN, Aurora, OH) (360 g); 30% glycerol (300 mL, 100% stock); 2% SDS (20 g); 65 mM DTT (1%) (ultrapure grade, Fisher Scientific, Fairlawn, CA) (400 mg per 40 mL), added just before use; 110 mM iodoacetamide (2%) (simply a good molar excess), added just before use in the second equilibration step (200 mg per 10 mL). Store frozen without DTT at -20°C in 40-mL portions.
13. SDS electrophoresis buffer, 1X (freshly prepared): 25 mM Tris (60.5 g), 192 mM glycine (288 g), 0.1% SDS (20 g), water to 20 L.
14. Acrylamide solutions (see Table 1).
15. Displacement solution (freshly prepared): 1.5 M Tris-HCl (pH 8.8) (50 mL), glycerol (100 mL, 100% solution), water (50 mL), 0.001% bromophenol blue (200 μ L, 1% stock solution).
16. Gel overlay buffer: 1X SDS polyacrylamide gel electrophoresis (PAGE) buffer (15 mL), isopropyl alcohol (15 mL) (alternatively, use water-saturated *n*-butanol). The overlay buffer is drawn into two 20-mL syringes with needles.
17. Agarose sealing solution (freshly prepared): 0.5% agarose (500 mg), 1X electrophoresis buffer (100 mL), 0.001% bromophenol blue (200 μ L, 1% stock solution). The agarose solution is melted in a microwave just before adding the bromophenol blue, during the iodo-acetamide equilibration.

2.3. Transferring 2-D Gels onto PVDF

1. Transfer apparatus (see Subheading 2.2.1.).
2. Filter paper 20 \times 20 cm.
3. Poly(vinylidene difluoride) (PVDF) sheets, 20 \times 20 cm.
4. Transfer buffer, 10X (2 L): 25 mM Tris-HCl (60.5 g), 192 mM glycine (288 g), water to 2 L.
5. Transfer buffer, 1X (4 L): 25 mM Tris-HCl (12.1 g), 192 mM glycine (57.6 g), water to 4 L.
6. 100% Methanol, 2 L.

2.4. Silver Staining of Gels

1. Acid alcohol solution: ethanol (250 mL), acetic acid (50 mL), water (200 mL).
2. Glutaraldehyde solution: 0.5 *M* sodium acetate (20.5 g), 1% glutaraldehyde (10 mL, 50% stock), water to 500 mL.
3. Naphthalene disulfonic acid solution: 0.05% w/v 2,7-naphthalene-disulfonic acid (250 mg), water (500 mL).
4. Ammoniacal silver nitrate solution: silver nitrate solution (4 g in 20 mL water). Combine: ammonium hydroxide (5.7 mL, 29% solution) and sodium hydroxide (1.0 mL, 10 *N* stock solution). Add the ammonium/sodium hydroxide solution to the silver nitrate solution; a transient brown precipitate will form. Bring to 500 mL with water.
5. Formaldehyde solution: 0.01% w/v citric acid (50 mg); 0.1% formaldehyde (1.35 mL, 37% stock), water to 500 mL.
6. Glycerol solution: 5% glycerol (25 mL, 100% stock), water to 500 mL.

3. Methods

3.1. Nonequilibrium pH Gradient Electrophoresis (NEpHGE) of Myelin Proteins

The following procedure is derived from that of Yamaguchi and Pfeiffer (1), a method modified from that of Celis et al. (3), and Mozdzanowski et al. (9) (see Note 1).

1. The bottoms of the gel tubes (3.5 mm inside diameter \times 13 cm long) are sealed with parafilm and stood vertically.
2. Before polymerization, the gel mixture lacking ammonium persulfate and TEMED is degassed under vacuum for 5–10 min.
3. The ammonium persulfate and TEMED are added and the gel solution is poured into the tubes.
4. The gels are overlaid with 50 μ L of 8 *M* urea and allowed to polymerize for 2 h.
5. The 8 *M* urea overlay is removed and the tops of gels are equilibrated for 2 min in 50 μ L of lysis buffer.
6. The myelin protein sample is solubilized in lysis buffer at room temperature for 30 min.
7. The solubilized sample (50 μ L) is applied to the top of the NEpHGE gels and overlaid with 10 μ L 80% lysis buffer.
8. The lower chamber (+) of the electrophoresis apparatus is filled with 20 mM NaOH, and the upper chamber (−) is filled with 10 mM H₃PO₄.
9. Electrophoresis of the samples is carried out at 600 V for 2.5 h.
10. Each sample is removed from its tube to a 60 mm diameter plastic Petri dish, using a long, fine hypodermic needle attached to a syringe filled with water.
11. Each gel is incubated for 15 min at room temperature in 6 mL equilibrium buffer.
12. The liquid is removed and the gels are quickly frozen in Petri dishes sitting on dry ice. When frozen, the gels are wrapped in parafilm and stored at -80°C .
13. For the second dimension, electrophoresis (5–20% gradient gel, 16 cm \times 16 cm, 1 mm thick) is carried out using standard methods, such as those of Celis et al. (3), Mozdzanowski et al. (9), and Laemmli (10).

3.2. *pI* Strip 2-D Gel Electrophoresis of Myelin Proteins (5,7) (see Notes 2–4)

Directions are given for using an IPGphor IEF unit and a Hoefer DALT Vertical System (Amersham Biosciences, Piscataway, NJ). Modifications can be made as necessary for other similar types of apparatus.

An important note on cleanliness: To avoid keratin contamination, important for subsequent mass spectrometric analysis and minimal background during silver staining, all gel caster and apparatus components, including the glass plates, spacers, separator sheets, and gel tank, and a 1-L flask for mixing gel reagents, must be carefully hand-washed. Air-dry all parts (except tank) in a clean, dust-free environment.

3.2.1. Sample Preparation

1. Purified membrane proteins (generally 200–300 μg) (e.g., myelin [4], *see Notes 2 and 3*) are solubilized in 500 μL lysis buffer at 37°C for 30 min.
2. The material is then precipitated overnight with two volumes of 100% ethanol with 20 mM DTT (3 mg/mL) at –20°C.
3. The material is centrifuged at 24,000g for 25 min at 4°C.
4. The precipitated pellet is resuspended in 20 μL of 10% ASB-14 (*see Note 5*) and bath sonicated with 10 to 15 5-s pulses at moderate energy, or until suspended.
5. The samples are then brought up to 370 μL with rehydration buffer and incubated at room temperature for 1 h with occasional mixing.
6. The material is centrifuged at 14,000g for 10 min to clear the supernatant fraction of any remaining particulate material.

3.2.2. Rehydrating and Loading the IPGphor Strips With Sample

1. Each sample supernatant (350 μL each) is transferred to an 18-cm IPGphor strip holder.
2. A precast immobilized pH gradient gel (IPG strip), 18 cm, pI 3–10, is placed over each sample solution and covered with 3.5 mL of mineral oil.
3. The IPG strips are rehydrated actively at 30 V overnight (12–16 h) at 20°C using the IPGphor isoelectric focusing unit.
4. After rehydration, the IPG strips are briefly rinsed with nanopure water (to remove crystallized urea) and placed back into the holders.

3.2.3. Isoelectric Focusing

1. Hydrated filter wicks are placed between the IPG strips and the electrodes. The cathodic filter wick is rehydrated with 100 mM DTT (prepared from a 150 mg/mL [10X] frozen stock). The anodic filter wick is hydrated with water.
2. Proteins are separated in the first dimension at 20°C in the IPGphor unit as follows:
 - a. 200 V, 1 h.
 - b. 500 V, 1 h.
 - c. 1000 V, 1 h.
 - d. Ramped to 6000 V, 30 min.
 - e. Held at 6000 V for 20,000 Vh (for pI 3–10). (This is an experimentally derived parameter; for example, for pI 4–7, this can be increased to 25–30,000 Vh, depending on the protein load).
3. During IEF, the filter wicks are replaced every 6000 Vh.
4. Following IEF, the IPG strips are rinsed briefly with deionized water (do not blot!) and either frozen and stored at –80°C for later analysis, or equilibrated and used immediately to run the second-dimension SDS-PAGE (next subheading).

3.2.4. Equilibration of IPG Strips Prior to Running the Second Dimension

1. Equilibration buffer is thawed, and 400 mg of DTT are dissolved (65 mM [1%]) in each 40-mL aliquot.
2. The IPG strips are equilibrated in 10 mL of equilibration buffer containing 65 mM DTT on a rocking platform for 15 min.

3. The buffer is discarded, and the strips are similarly equilibrated for 15 min in 10 mL of equilibration buffer containing 110 mM (2%) iodo-acetamide (200 mg/10 mL).
4. Rinse the strips briefly in 1X electrophoresis buffer (simply dip them into the electrophoresis tank). The strips are now ready for further second-dimension electrophoresis (next subheading).

3.2.5. Preparing the Second-Dimension Gel

Prepare the gels in advance (**steps 1–8**) so that they are ready before equilibration (**Subheading 3.2.4.**).

1. The entire gel assembly (assembled per manufacturer's instructions) is placed into a large tray (to collect spillage after pouring), making sure that the assembly is level.
2. APS and TEMED are added to the degassed gel solution, and the solution is slowly poured through a funnel into the apparatus until the gel height is approx 80% to the top.
3. The funnel is removed when all of the acrylamide has entered the apparatus. This allows the displacement solution to flow down, causing the gel level to rise about another inch or so. The gel level should be approx 1 cm below the top of the glass plates. If the level isn't high enough, the remaining 50 mL of displacement solution can be added.
4. Each gel is overlaid with at least 2 mL of overlay buffer. It is best to pour the overlay buffer from first one side of the gel and then the other side; do not pour it all along the gel.
5. The gels are allowed to polymerize undisturbed for 20–30 min.
6. When polymerization is complete, the gel caster is disassembled, and each part is rinsed with deionized water.
7. The polymerized gels are rinsed, especially the wells, to remove isopropyl alcohol from the overlay buffer.
8. The gels are placed directly into the second-dimension electrophoresis tank previously filled with 1X electrophoresis buffer.
9. Each IPG strip is placed into the well of a gel with the acidic side facing the glass plate hinge.
10. Each strip is sealed in place with agarose sealing solution (poured slowly from one side, avoiding bubbles).
11. Electrophoresis is carried out at a constant current (approx 10–15 mA/gel), overnight (for a 10% gel; longer for 12%, less for 8%), at 20°C. At approx 12 mA/gel, the dye front will move at approx 1 cm/h; therefore, 10 gels at 120 mA will travel 18 cm in approx 18 h. If the migration is not complete the next morning, the voltage may be increased to 30 mA/gel.
12. The gels can now be stained with Coomassie blue or silver nitrate (**Subheading 3.4.**), or the proteins can be transferred onto PVDF membrane for immunoblotting (next subheading).

3.3. Transferring 2-D Gels onto PVDF

1. The tank used for the second-dimension SDS electrophoresis is emptied and the tank, electrodes, and gel holders are rinsed with deionized water.
2. The electrodes are placed back in the tank along with the transfer cassette holder apparatus.
3. The 16 L of water, 2 L of 10X transfer buffer, and 2 L of 100% methanol are added to the tank and mixed well.
4. The gel glass plates from the 2-DE are opened. The gels are cut at the edges of the IPG strip and at the electrophoresis front. In addition, a small diagonal indicator cut is made on the bottom, acidic side of the gel for orientation.
5. The 4 L of transfer buffer is poured into a large container for assembling the transfer cassettes according to the manufacturer's recommendations, using 20 × 20 cm sheets of filter paper and PVDF membrane. The assembled cassettes are placed into the tank.

6. The temperature of tank is set to 15°C, and the transfer of the proteins in the gels to the membrane is carried out overnight at 400 mA (≥ 14 h).
7. The membranes can now be used for immunoblotting, and so on.

3.4. Silver Staining (Ammoniacal Silver Nitrate) of 2-D Gels

Note: Silver nitrate is extremely sensitive and dirty. Be *very* careful in weighing and pouring solutions. After staining, wash both the inside and the outside of trays very carefully with deionized water to avoid tracking silver.

The following procedure is for a $20 \times 20 \times 1.5$ mm gel. Use 500 mL of each solution per gel. All steps are carried out using nanopure water.

1. The gels are fixed in acid alcohol solution for 10 min (the gels can be left in this solution for many days).
2. The gels are soaked in glutaraldehyde solution for 10 min.
3. The gels are washed twice in water for 10 min each time.
4. The gels are soaked in naphthalene disulfonic acid solution for 15 min.
5. The gels are washed three times with water for 10 min each wash.
6. The gels are stained with ammoniacal silver nitrate solution for 30 min.
7. After staining, the gels are washed four times with water for 4 min each wash.
8. The gels are developed with formaldehyde solution for 2–5 min, depending on the protein load and background levels.
9. When a slight yellow background stain appears, development is stopped by adding 30 mL of 100% glacial acetic acid. The gels are then rocked gently for 15 min.
10. The gels are washed twice in water for 10 min each wash.
11. The gels are stored in glycerol solution.
12. The gels are scanned in a flatbed scanner at 300 dpi, grayscale.

4. Notes

1. Application to developmental analysis. Using NEPHGE 2-DE, we have analyzed metabolically radiolabeled oligodendrocytes in culture at specific stages of the developmental lineage, and demonstrated the developmental regulation of a number of basic membrane proteins (1). We have used similar procedures to analyze the developmental regulation of the expression of small GTP-binding proteins (11) and a myelin protein residing in glycosphingolipid-cholesterol microdomains (“lipid rafts”) (12).
2. The methods presented here using precast IPGphor strips have been designed to significantly enhance the entry/focusing of transmembrane and GPI-anchored proteins by 2-DE (5,8). Although these studies were aimed at analyses of the unusually lipid-rich myelin membrane, it is expected that they will prove more generally useful for the analysis of proteins in a variety of membrane-rich samples.
3. The myelin proteome. Using these methods for high-resolution, two-dimensional gel electrophoresis, coupled with mass spectrometry and immunoblotting, we have developed an extensive proteomic map of proteins present in central and peripheral nervous system myelin, identifying 98 proteins corresponding to at least 130 of the approx 200 spots on the map (5). This proteomic map has been applied to analyses of the localization and function of selected proteins, providing a powerful tool to investigate the diverse functions of myelin, and it suggests a paradigm for similar studies in other systems, including glycosphingolipid-cholesterol microdomains (lipid rafts) (13,14).
4. Application to signal transduction and clinical disease. We have recently applied a proteomic strategy to a novel signaling mechanism in mature oligodendrocytes, with

important implications for demyelinating disease and nerve regeneration in the central nervous system (6,7). Antibody cross-linking of myelin oligodendrocyte glycoprotein (MOG; a molecule strongly implicated in multiple sclerosis) on oligodendrocytes in culture rapidly leads to changes in the phosphorylation state of specific proteins, followed by major changes in cell cytoarchitecture. Whereas the low quantities of proteins available from the primary cultures precluded direct mass spectrometric analysis, often we were able to identify these proteins by comparing the silver-stained oligodendrocyte gels with the myelin 2-D proteomic map. Conversely, the strategy also has often worked in reverse, identifying myelin proteins from previously identified proteins on the oligodendrocyte map.

5. Many detergents have been synthesized recently that have enhanced the 2DE separation of hydrophobic proteins, including ASB-14 (used in the present protocol [8]) and C8Ø (15). As a general rule, optimum sample preparation must be determined empirically for each unique tissue. It is likely that sequential extraction under conditions of increasing stringency will be required to identify all the proteins in a sample.

Acknowledgments

This work was supported by National Institutes of Health grants NS10861 (SEP/RB), NS41078 (SEP), and NS45440 (CMT), and National Multiple Sclerosis Society grants RG2181 (SEP) and FG1423 (CBM).

References

1. Yamaguchi, Y. and Pfeiffer S. E. (1999) Highly basic myelin and oligodendrocyte proteins analyzed by NEPHGE two dimensional gel electrophoresis: Recognition of novel developmentally regulated proteins. *J. Neurosci. Res.* **56**, 199–205.
2. O'Farrell, P. Z., Goodman, H. M., and O'Farrell, P. H. (1977) High resolution two-dimensional electrophoresis of basic as well as acidic proteins. *Cell* **12**, 1133–1142.
3. Celis, J. E., Ratz, G., Basse, B., Lauridsen, J. B., and Celis, A. High resolution two-dimensional gel electrophoresis of proteins: isoelectric focusing (IEF) and non equilibrium pH gradient electrophoresis (NEPHGE). Internet web site: <http://proteomics.concer.dk/>
4. Menon, K., Rasband, M. N., Taylor, C. M. Brophy P., Bansal, R., and Pfeiffer, S. E. (2003) The Myelin-Axolemmal Complex: biochemical dissection and the role of galactosphingolipids. *J. Neurochem.* **87**, 995–1009.
5. Taylor, C. M., Marta, C. B., Claycomb, R. J., et al. (2004) Proteomic mapping provides powerful insights into functional myelin biology. *Proc. Natl. Acad. Sci. USA* **101**, 4643–4648.
6. Marta, C. B., Taylor, C. M., Coetzee, T., et al. (2003) Antibody Crosslinking of myelin oligodendrocyte glycoprotein leads to its rapid repartitioning into detergent insoluble fractions and altered protein phosphorylation and cell morphology. *J. Neurosci.* **23**, 5461–5471.
7. Marta, C. B., Taylor, C. M., Cheng S., Quarles, R., Bansal, R., and Pfeiffer, S. E. (2004) Myelin associated glycoprotein cross-linking triggers its partitioning into lipid rafts, specific signaling events and cytoskeletal rearrangements in oligodendrocytes. *Neuron Glia Biology* **1**, 35–46.
8. Taylor, C. M. and Pfeiffer, S. E. (2003) Enhanced resolution of glycosylphosphatidylinositol-anchored and transmembrane proteins from the lipid-rich myelin membrane by two-dimensional gel electrophoresis. *Proteomics* **3**, 1303–1312.
9. Mozdzanowski, J., Speicher, D., and Harper, S. (1995) Two-dimensional electrophoresis. In: Colligan, J. E., Dunn, B. M., Ploegh, H. L., Speicher, D. W., and Wingfield, P. T. (eds), *Current Protocols in Protein Science*. New York: John Wiley & Sons, pp. 10.4.1–30.

10. Laemmli, U. K. (1970) Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **22**, 680–685.
11. Hüber, L. A., Madison, D. L., Simons, K., and Pfeiffer, S. E. (1994) Myelin membrane biogenesis by oligodendrocytes: developmental regulation of low-molecular weight GTP-binding proteins. *FEBS Lett.* **347**, 273–278.
12. Kim, T., Fiedler, K., Madison, D. L., Krueger, W. H., and Pfeiffer, S. E. (1995) Cloning and characterization of MVP17: a developmentally regulated myelin protein in oligodendrocytes. *J. Neurosci. Res.* **42**, 413–422.
13. Taylor, C. M., Marta, C. B., Bansal, R., and Pfeiffer, S. E. (2004) The transport, assembly and function of myelin lipids. In: Lazzarini, R. (ed) *Myelin Biology and Disorders, Vol. I*, New York: Academic Press, pp. 57–88.
14. Schafer, D. P., Bansal, R., Hedstrom, K. L., Pfeiffer, S. E., and Rasband, M. N. (2004) Does paranode formation and maintenance require partitioning of Neurofascin 155 into lipid rafts? *J. Neurosci.* **24**, 3176–3185.
15. Chevallet, M., Santoni, V., Poinas, A., et al. (1998) New zwitterionic detergents improve the analysis of membrane proteins by two-dimensional electrophoresis. *Electrophoresis* **19**, 1901–1909.

Silver Staining of 2-D Gels

Julia Poland, Thierry Rabilloud, and Pranav Sinha

1. Introduction

1.1. Staining Methods Used in Proteomics: An Overview

Proteomics, in its commonly used meaning, refers to the detection of differentially expressed proteins between control and experimental samples. The basic tools include two-dimensional gel electrophoresis (2-DE) for protein fractionation followed by mass spectrometry (MS) of excised protein spots for protein identification. After separation, proteins need to be detected on the gel, meaning the spot pattern has to be distinguished from the surrounding background. This is generally performed directly in the polyacrylamide gel by binding a dye molecule to the proteins.

One of the current challenges in the field of proteomics is the development of highly sensitive protein-staining methods compatible with sophisticated identification techniques such as matrix-assisted laser desorption/ionization (MALDI) and tandem electrospray ionization (ESI). Several conventional staining methods available for spot detection on 2-D gels do not possess both qualities. Protocols using Coomassie brilliant blue, for instance, are highly compatible with MS but are not at all sensitive. On the other hand, silver-staining methods have raised the detection limit to the nanogram range, but protein identification of excised spots is often an obstacle that cannot be overcome easily. In particular, very sensitive alkaline silver methods using glutaraldehyde as a sensitizer are not compatible with MS.

In the past, an attempt to solve these problems has been the combination of silver and Coomassie staining, making use of the advantages of both approaches: in a first step, analytical 2-DE is carried out using a highly sensitive silver-staining protocol for image analysis of the gels. A second experimental run comprises semipreparative 2-DE utilizing a Coomassie blue protocol for spot excision and subsequent protein identification (see **Note 1**). This combined approach must fulfill a basic condition: an adequate amount of sample must be available (up to 1 mg for each Coomassie gel + at least 100 µg for each silver gel). Additionally, the approach is both time-consuming and expensive, because twice the amount of gels is required.

In order to avoid the complexity and costs of this approach, researchers often use Coomassie staining for gel analysis and spot excision from a single gel, thus accepting the low sensitivity of the stain, resulting in a relatively small number of spots.

Another solution is the use of fluorescence staining methods now available. The introduction of fluorescence-based technologies (e.g., SYPRO Ruby stain; see Chapters

20–23) provides simple, highly sensitive, and MS-compatible staining methods, but these methods require special, expensive hardware and software (e.g., fluorescence scanner), and are thus not generally utilizable, especially for smaller laboratories. Nevertheless, technologies like 2-D differential in-gel electrophoresis (2D DIGE; Chapter 24), with the possibility of simultaneously separating up to three samples on a single 2-D gel, are very good methods for present-day proteomics.

A summary showing features of selected staining methods used for proteome analysis is given in **Table 1**.

Recently, a breakthrough in proteome analysis has been reached by the introduction and optimization of silver-staining protocols that show good compatibility with mass spectrometry. Several methods have been published, mainly based on already existing silver-staining protocols (1–3).

Because silver staining is a useful, sensitive, non-radioactive method for permanently staining proteins in polyacrylamide gels, requiring relatively inexpensive equipment and reagents, this chapter deals with this topic, with the stress on MS-compatible staining methods.

1.2. Silver Staining: Principles

The staining procedure is performed in several individual steps, which can be subdivided into five main phases (4):

1. Fixation: interfering substances present on sodium dodecyl sulfate (SDS)-polyacrylamide gel electrophoresis (PAGE) gels showing a high affinity to silver (e.g., SDS, Tris, chloride) have to be removed to avoid an interfering background.
2. Sensitization: sensitivity enhancers (e.g., formaldehyde, glutaraldehyde, naphthalene disulfonic acid, sodium thiosulfate, potassium tetrathionate) can be used between fixation and silver impregnation. These substances bind to proteins, resulting in increased silver binding (naphthalene disulfonic acid) or forming of metallic silver (glutaraldehyde, formaldehyde) or silver sulfide (potassium tetrathionate).
3. Silver impregnation: the gel is soaked in a staining solution containing soluble silver ions; the silvering agent is either silver nitrate (acidic methods) or silver ammonia/silver diamine (basic methods).
4. Development: development is achieved by treatment with a reductant, generally a dilute formaldehyde solution. Reduction of soluble silver ions to metallic silver (insoluble and visible) is a strong autocatalytic process that is primed by sensitivity enhancers. A combination of factors contribute to this autocatalytic character of metal ion reduction, resulting in the extraordinary high sensitivity of silver staining. A way to further increase sensitivity is the use of thiosulfate, which sequesters silver ions weakly bound to the matrix and thereby reduces development of the background. Thiosulfate can be added to the development bath or can be integrated in the gel, if self-cast gels are used.
5. Stopping: the reaction is stopped to avoid over-development, usually at the point of incipient background development.

A detailed description of the physico-chemical principles underlying the mechanisms of silver staining can be found in reviews by Rabilloud et al. (4,5).

1.3. Silver-Staining Protocols Compatible With Mass Spectrometry

During the last few years, there has been an increasing development and optimization of silver-staining methods used for detection of proteins prior to in-gel digestion

Table 1
Comparison of Selected Staining Methods Used for 2-D Gels

Staining method	Technology	Detection limit	MS compatibility
Coomassie Brilliant Blue	Absorbance	50–100 ng	YES
Colloidal Coomassie	Absorbance	10–20 ng	YES
Silver nitrate (acidic methods)	Absorbance	5–10 ng	YES (without sensitizer glutaraldehyde)
Silver ammonia/ silver-amine complex (basic methods)	Absorbance	<1 ng	NO
SYPRO® Ruby	Fluorescence	1 ng	YES

and MS identification. The first landmark was the introduction of a stain compatible with MALDI-time-of-flight (TOF) by Shevchenko et al., shortly followed by others (1–3). These protocols show some disadvantages over the existing silver-staining methods used only for spot detection, but the increasing need of high sensitivity for studying low-abundance proteins, as well as rapid protein identification after fractionation, makes this new direction absolutely indispensable for up-to-date proteomics.

For MS, specific reagents such as glutaraldehyde have to be omitted from the protocol because the proteins must not be modified. This can reduce the sensitivity of the stain somewhat, resulting in a decreased number of spots on the gel; nevertheless, the published protocols show satisfying results comparable to other silver stains.

The main problem of these protocols is that exceptional cleanliness must be practiced after the second dimension to receive acceptable MS spectra. Not only must reagent and water quality be very high, but gel handling itself must be reduced to a minimum to avoid keratin contamination (see Note 2). Keratins from the skin and air present on the excised gel plug increase the keratin:protein ratio, making successful protein identification impossible. After excision, gel pieces should be destained by silver oxidation, because metallic silver interferes with MS. This is another step that prolongs the procedure. Another complication is the need for rapid progress after spot detection. Spots should be excised directly after staining and rapidly subjected to further processing (6). Storage of gels and excised gel plugs can impair successful identification.

2. Materials

2.1. Staining Protocol by Heukeshoven and Dernick

1. Fixing solution: 50% (v/v) ethanol, 10% (v/v) acetic acid.
2. Sensitizing solution: 0.5 M sodium acetate, 0.2% (w/v) sodium thiosulfate pentahydrate, 30% (v/v) ethanol, 0.5% (v/v) glutardialdehyde.
3. Staining solution: 0.1% (w/v) silver nitrate, 0.01% (v/v) formaldehyde.
4. Incubating solution: 2.5% (w/v) sodium carbonate.
5. Developing solution: 2.5% (w/v) sodium carbonate, 0.01% (v/v) formaldehyde, 0.05% (w/v) sodium bicarbonate, 0.02% (w/v) thimerosal.
6. Stop solution: 0.05 M ethylenediaminetetraacetic acid (EDTA) disodium salt.

2.2. Staining Protocol by Hochstrasser

1. Fixing solution I: 40% (v/v) ethanol, 10% (v/v) acetic acid.
2. Fixing solution II: 5% (v/v) ethanol, 5% (v/v) acetic acid.
3. Sensitizing solution I: 0.5 M sodium acetate, 1% (v/v) glutardialdehyde.
4. Sensitizing solution II: 0.05% (w/v) naphthalene-disulfonic acid.
5. Staining solution: 0.8% (w/v) silver nitrate, 0.34% (v/v) ammonia, 0.2% (v/v) 10 N NaOH (see Note 4).
6. Developing solution: 0.01% (w/v) citric acid, 0.1% (v/v) formaldehyde.
7. Stop solution: 5% (w/v) Tris, 2% (v/v) acetic acid.

2.3. Staining Protocol by Sinha et al.

1. Fixing solution: 30% (v/v) ethanol, 10% (v/v) acetic acid.
2. Sensitizing solution: 0.3% (w/v) potassium tetrathionate, 0.5 M potassium acetate, 30% (v/v) ethanol.
3. Staining solution: 0.2% (w/v) silver nitrate.
4. Developing solution: 3% (w/v) potassium carbonate, 0.0125% (v/v) sodium thiosulfate (10%), 0.03% (v/v) formaldehyde (37%).
5. Stop solution: 4% (w/v) Tris, 2% (v/v) acetic acid.

2.4. Staining Protocol by Yan et al.

1. Fixing solution: 40% (v/v) methanol, 10% (v/v) acetic acid.
2. Sensitizing solution: 30% (v/v) methanol, 0.2% (w/v) sodium thiosulfate pentahydrate, 0.5 M sodium acetate.
3. Staining solution: 0.1% (w/v) silver nitrate.
4. Developing solution: 2.5% (w/v) sodium carbonate, 0.01% (v/v) formaldehyde.
5. Stop solution: 0.05 M EDTA disodium salt.

2.5. Equipment

1. 24 × 20 cm gels: large, covered plastic boxes (not less than 30 × 35 cm); minigels: covered plastic/glass boxes.
2. Shaker.
3. Staining apparatus (optional).

3. Methods

In this section, selected protocols for silver staining of 2-D gels are given in detail. The first part (**Subheading 3.1.**) deals with conventional silver-staining methods that are ultrasensitive but not compatible with mass spectrometry. In the second part (**Subheading 3.2.**), two examples for mass spectrometry-compatible silver staining are presented. Since silver staining is a multistep procedure that requires proper fixation and exchange of substances from the liquid into the gel phase, the reliable protocols are rather long. But you will be rewarded by reproducible gels with a clear spot pattern.

All solutions specified in the following should be prepared fresh.

3.1. Conventional Staining Methods

3.1.1. Silver Staining Protocol by Heukeshoven and Dernick (Modified)

This protocol published by Heukeshoven and Dernick (7) has been modified by Klose and Kobalz (8) and represents a reliable method of very good sensitivity. The protocol uses silver nitrate for impregnation and thus belongs to the acidic methods.

The sequence of the procedure is as follows (use approx 2.5 L of each solution for simultaneous staining of 10 large gels [20 × 25 cm]); for solutions (see **Subheading 2.1.**), all steps have to be performed with shaking:

1. Fix the gel in fixing solution for at least 2 h or overnight.
2. Sensitize for 2 h in sensitizing solution.
3. Wash the gel using Milli-Q water, 2 × 20 min.
4. Impregnate with silver staining solution for 30 min.
5. Wash the gel using Milli-Q water for a few seconds (see **Note 4**).
6. Incubate the gel prior to development in incubating solution for 1 min (see **Note 4**).
7. Develop the gel for 5–20 min in developing solution.
8. Stop the reaction using stop solution with shaking for approx 15 min.

3.1.2. Silver Staining Protocol by Hochstrasser

This method (<http://us.expasy.org/ch2d/protocols/protocols.fm4.html>) is representative of basic silver staining methods and offers extremely high sensitivity. The sequence of the procedure is as follows (use approx 2.5 L of each solution for simultaneous staining of 10 large gels [20 × 25cm]); for solutions see **Subheading 2.2.**; all steps have to be performed with shaking:

1. Fix the gel in fixing solution I for 1 h.
2. Fix the gel in fixing solution II for at least 2 h or overnight.
3. Wash the gel using Milli-Q water for 5 min.
4. Sensitize for 2 h in sensitizing solution I.
5. Wash the gel using Milli-Q water, 3 × 10 min.
6. Sensitize for 2 h in sensitizing solution II, 2 × 30 min.
7. Wash the gel using Milli-Q water, 4 × 4 min.
8. Impregnate with silver staining solution for 30 min.
9. Wash the gel using Milli-Q water, 4 × 4 min.
10. Develop the gel for 5–10 min in developing solution (see **Note 5**).
11. Stop the reaction using stop solution with shaking for approx 15 min.

3.2. MALDI-TOF-Compatible Staining Methods

3.2.1. Silver-Staining Protocol by Sinha *et al.*

Our protocol has been designed to combine high sensitivity, minimal protein-to-protein variation, and high reproducibility, with the possibility of identifying excised protein spots using MALDI-TOF MS (3). A MALDI-compatible silver stain is especially recommended, if the amount of protein extract is limited. With this method, only the surface of the proteins is stained, and protein loads as small as 100 µg per gel can be used. Development can be extended to very long periods, making the procedure extremely reproducible from batch to batch. Take care that stringent conditions of cleanliness are maintained to reduce the risk of keratin contamination (see also **Note 2**)!

The sequence of the procedure is as follows (use approx 2.5 L of each solution for simultaneous staining of 10 large gels [20 × 25 cm]); all steps have to be performed with shaking:

1. Fix the gel in fixing solution for at least 1 h, change the solution, and fix overnight or incubate the gel in fixing solution 4 × 30 min.
2. Sensitize for 45 min in sensitizing solution.
3. Wash the gel using Milli-Q water, 6 × 10 min.

4. Impregnate with silver staining solution for 30 min.
5. Wash the gel using Milli-Q water for a maximum of 15 s (see **Note 6**).
6. Develop the gel for 30–40 min in developing solution.
7. Stop the reaction using stop solution with shaking for approx 45 min.
8. Wash the gel using Milli-Q water for 2 × 30 min.

3.2.2. Silver Staining Protocol by Yan et al.

This method is a modification of the commercial kit Silver Stain PlusOne (Amersham Biosciences), which is based on the original silver staining protocol by Heukeshoven and Dernick (2). By omitting glutaraldehyde in the sensitization step and formaldehyde in the silver impregnation step, the staining protocol is compatible with MALDI and ESI.

The sequence of the procedure is as follows (use approx 2.5 L of each solution for simultaneous staining of 10 large gels [20 × 25cm]); for solutions, see **Subheading 2.1.4.**; all steps have to be performed with shaking:

1. Fix the gel in fixing solution for 2 × 15 min.
2. Sensitize for 2 h in sensitizing solution.
3. Wash the gel using Milli-Q water, 3 × 5 min.
4. Impregnate with silver staining solution for 30 min.
5. Wash the gel using Milli-Q water for 2 × 1 min.
6. Develop the gel for 5–20 min in developing solution.
7. Stop the reaction using stop solution with shaking for approx 15 min.
8. Wash the gel using Milli-Q water for 3 × 5 min.

3.2.3. Silver Oxidation of Excised Protein Spots

Removal of silver ions prior to tryptic digestion is recommended for successful protein identification (see also **Note 7**). The method described by Gharadaghi et al. (9) is as follows:

1. Add a small volume of a 1:1 solution potassium ferricyanide (30 mM) and sodium thiosulphate (100 mM) to the gel piece.
2. Stir frequently until the dark color disappears.
3. Wash three to five times with Milli-Q water to stop the reaction.
4. Discard the water and cover the gel piece with 200 mM ammonium bicarbonate buffer for about 20 min.
5. Wash with water.
6. Discard water and cover the gel piece with acetonitrile.

4. Notes

1. Apart from low sensitivity and the need to apply a high amount of protein extract, preparative Coomassie-stained gels may cause some problems: certain protein spots do not bind Coomassie brilliant blue, making spot cutting impossible. Note that the spot pattern of Coomassie-stained gels may be different from the one of silver-stained gels.
2. The use of a staining apparatus can help to ensure clean working conditions. We use a homemade staining chamber that allows simultaneous staining of up to 10 gels (3). Gels are placed on Plexiglas grids that can be slid into position like a drawer and clamped with screws on each side of the basic inner chamber to prevent the gels from falling off the platform during shaking. The outer chamber of the apparatus is equipped with an outlet whose diameter allows 8 L to be drained in 1 min. The main body of the staining chamber

can be connected to an orbital shaker. Staining chambers are also commercially available (e.g., Bio-Rad).

3. The impregnation solution should be freshly prepared. To prepare 750 mL of this solution, 6 g of silver nitrate is dissolved in 30 mL Milli-Q water, which is slowly mixed into a solution containing 160 mL Milli-Q water, 10 mL concentrated ammonia (25%), and 1.5 mL of sodium hydroxide (10 N). A transient brown precipitate might form. After it clears, water is added to give the final volume (<http://us.expasy.org/ch2d/protocols/protocols.fm4.html>).
4. If high-throughput staining (e.g., 10 gels per box) is performed without the aid of a staining apparatus, handling of the two short steps prior to development can be simplified in the following way:
 - Take three extra boxes; fill the first with Milli-Q water, the second with incubating solution, and the third with developing solution.
 - Place the box with the developing solution (after formaldehyde has been added) on the shaker.
 - Start with the first gel, take it out of the impregnation solution, rinse it a few seconds in the first box (washing step), subsequently shake it manually for approx 1 min in the second box (incubation step), and finally put it into the third box (development step) on the shaker; set an alarm clock (count up).
 - Proceed with the other gels in the same way.
 - Stop the development at the point when a slight background stain appears on the first gel (usually after 15–20 min); begin with the first gel to ensure equal staining of all gels.Individual gel handling at this phase of staining ensures that gels are soaked properly in each solution despite the short duration of the washing and incubation steps.
5. With this method, development time can be less than specified above. Sometimes, the reaction has to be stopped after 2–3 min to avoid over-development of the background. Although this protocol is one of the most sensitive silver staining procedures available, the Heukeshoven staining method (development 15–20 min) should be preferred if high-throughput analysis with reproducible results from gel to gel (and between experiments) is required.
6. Wash gels individually during the step prior to development and the development step itself to get cleanly stained gels.
7. It has been recently shown that the formaldehyde present in the developer is responsible for most of the residual interference between silver staining and mass spectrometry (6). Consequently, the time elapsed between the end of the silver staining process and the destaining process is also a critical parameter for the quality of the resulting MS spectra (6). Thus, best results are obtained when the destaining is performed on the same day as silver staining.

References

1. Shevchenko, A., Wilm, M., Vorm, O., and Mann, M. (1996) Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels. *Anal. Chem.* **68**, 850–858.
2. Yan, J. X., Wait, R., Berkelman, T., et al. (2000) A modified silver staining protocol for visualization of proteins compatible with matrix-assisted laser desorption/ionization and electrospray ionization–mass spectrometry. *Electrophoresis* **21**, 3666–3672.
3. Sinha, P., Poland, J., Schnolzer, M., and Rabilloud, T. (2001) A new silver staining apparatus and procedure for matrix-assisted laser desorption/ionization-time of flight analysis of proteins after two-dimensional electrophoresis. *Proteomics* **1**, 835–840.
4. Rabilloud, T., Vuillard, L., Gilly, C., and Lawrence, J. J. (1994) Silver-staining of proteins in polyacrylamide gels: a general overview. *Cell. Mol. Biol. (Noisy-le-grand)* **40**, 57–75.

5. Rabilloud, T. (1990) Mechanisms of protein silver staining in polyacrylamide gels: a 10-year synthesis. *Electrophoresis* **11**, 785–794.
6. Richert, S., Luche, S., Chevallet, M., Van Dorsselaer, A., Leize-Wagner, E., and Rabilloud, T. (2004) About the mechanism of interference of silver staining with peptide mass spectrometry. *Proteomics* **4**, 909–916.
7. Heukeshoven, J. and Dernick, R. (1985) Simplified method for silver staining of proteins in polyacrylamide gels and the mechanism of silver staining. *Electrophoresis* **6**, 103–112.
8. Klose, J. and Kobalz, U. (1995) Two-dimensional electrophoresis of proteins: an updated protocol and implications for a functional analysis of the genome. *Electrophoresis* **16**, 1034–1059.
9. Gharahdaghi, F., Weinberg, C. R., Meagher, D. A., Imai, B. S., and Mische, S. M. (1999) Mass spectrometric identification of proteins from silver-stained polyacrylamide gel: a method for the removal of silver ions to enhance sensitivity. *Electrophoresis* **20**, 601–605.

Zn²⁺ Reverse Staining Technique

Carlos Fernandez-Patron

1. Introduction

With the advent of proteomics, many previously unidentified proteins will be isolated in gels for subsequent structural and functional characterization. It is, therefore, important that methods are available for detecting these proteins with minimal risk of modification. This chapter describes a reverse staining technique that facilitates the sensitive detection of unmodified proteins (1–9).

The Zn²⁺ reverse staining technique exploits the ability of biopolymers (in particular, proteins and protein-sodium dodecyl sulfate [SDS] complexes) to bind Zn²⁺ (9–11) and that of imidazole (ImH, C₃N₃H₄, *see Note 1*) to react with unbound Zn²⁺ to produce insoluble zinc imidazolate (ZnIm₂). Deposition of ZnIm₂ along the gel surface results in the formation of a deep white-stained background, against which unstained biopolymer bands contrast (as biopolymer bands bind Zn²⁺, they locally inhibit deposition of ZnIm₂ and are not stained; *see Notes 2 and 3*).

With regard to protein analysis, reverse staining is a very gentle method of detection that does not require strong acid solutions, organic dyes, chemical modifiers, or protein sensitizers. Protein interactions with Zn²⁺, established during the reverse staining step, are abrogated by metal chelation. Therefore, the risk of protein modification during reverse staining is minimal and reverse-stained proteins can be efficiently eluted and used in biological and enzymatic assays (*see Note 4*) (7–9,15). Proteins can also be processed for microsequencing or mass spectrometric analysis at any time after detection (*see Note 5*) (5–9,13,14).

Another benefit of the reverse staining technique is its high sensitivity of detection (approx 10 ng protein/band in SDS-polyacrylamide gel electrophoresis [PAGE] gels), which is greater than that of the Coomassie blue stain (approx 100 ng protein/band) and approaches the sensitivity of the silver stains (1–10 ng protein/band). Indeed, reverse staining often reveals many proteins that display low affinity for Coomassie blue and are thus not detected with this stain (*see Note 6*) (11).

Other advantages of the reverse staining technique include the speed of the procedure (approx 15 min), significantly faster than the Coomassie blue (>1 h) and silver (>2 h) stains. The reverse-stained gels can be kept in water for several hours to years without loss of image or sensitivity of detection. Similar to other stains, reverse-stained patterns can be analyzed densitometrically, and a gel toning procedure has been recently developed to preserve the image upon gel drying (12).

The Zn^{2+} reverse staining technique can be applied to detect virtually any gel-separated biopolymer that binds Zn^{2+} (i.e., proteins/peptides, glycolipids, oligonucleotides, and their multimolecular complexes) (see ref. 9 and references therein). Moreover, these biopolymers are detected regardless of the electrophoresis system used for their separation (see Notes 7–9) (9), which was not possible with the previously developed metal salt stains (1–3). Therefore, Zn^{2+} reverse staining is a widely applicable detection method.

The SDS-PAGE method of protein electrophoresis is probably the most popular. Thus, the author chose to describe a standard reverse staining method that works very well with SDS-PAGE. The detection of proteins, peptides, and their complexes with glycolipids in less commonly used gel electrophoresis systems is addressed as well (see Notes 7–9).

2. Materials

Reagent- and analytical-grade zinc sulfate, imidazole, acetic acid, and sodium carbonate are obtained from Sigma (St. Louis, MO).

1. Equilibration solution (1X): 0.2 M imidazole, 0.1% (w/v) SDS (see Note 1).
 2. Developer (1X): 0.3 M zinc sulfate.
 3. Storage solution for reverse-stained gels (1X): 0.5 % (w/v) sodium carbonate.
- All solutions are prepared as 10X concentrated stocks, stored at room temperature, and diluted (1:10) in distilled water, to yield the working concentration (1X) just before use.

3. Methods

3.1. Polyacrylamide Gel Electrophoresis

SDS-PAGE is conducted following the method of Laemmli (16). Native PAGE is conducted following the protocol of Laemmli, except that SDS is not included in the gel and electrophoresis solutions. Conventional agarose gel electrophoresis is conducted in 0.8% agarose gels, and using Tris-acetate, pH 8.0, as gel and running buffer (17).

3.2. Standard Reverse Staining of SDS-PAGE and Native PAGE Gels

The following reverse staining method detects proteins in standard polyacrylamide gels (see Note 2, **Fig. 1**). All incubations are performed under continuous gentle agitation in a plastic or glass tray with a transparent bottom. The volume of the corresponding staining/storage solutions must be enough to cover the gel (typically 50 mL for one mini-gel [10 cm × 7 cm × 0.75 mm]).

1. Following electrophoresis, the gel is incubated for 15 min in the equilibration solution (see Note 2).
2. To develop the electropherogram, the imidazole solution is discarded and the gel soaked for 30–40 s in developer solution. Caution: This step must not be extended longer than 45 s or band overstaining and loss of the image will occur. Overstaining is prevented by pouring off the developer solution and rinsing the gel three to five times (approx 1 min) in excess water (see Note 3).

At this point, the reverse-stained gel can be photographed (**Fig. 1**). Photographic recording is best conducted with the gel placed on a glass plate held a few centimeters above a black underground and under lateral illumination.

2D-PAGE gel

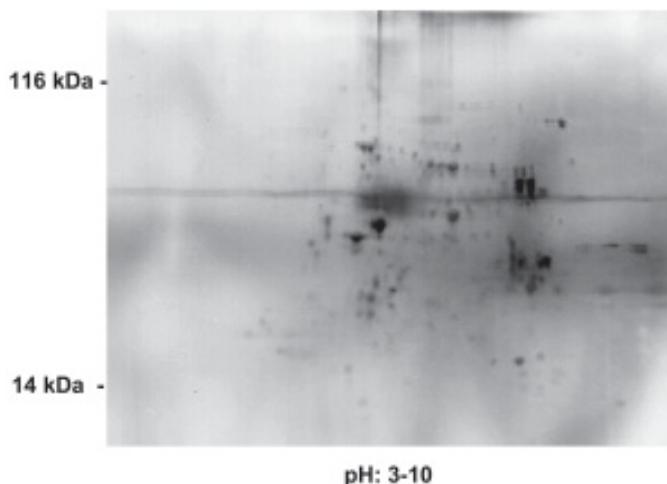


Fig. 1. Reverse staining of rat brain homogenate proteins (70 µg load) after 2-D polyacrylamide gel electrophoresis. Proteins in a wide range of molecular weights and isoelectric points are detected as transparent spots. These spots contain protein-sodium dodecyl sulfate complexes that bind Zn²⁺ and thereby inhibit the precipitation of ZnIm₂ locally.

While SDS-PAGE is a popular, high-resolution method for separating complex protein mixtures, sometimes it is desired to avoid either protein denaturation or disruption of macromolecular complexes during electrophoresis. In this case, the proteins are separated in the absence of SDS (native PAGE or agarose gel electrophoresis). Therefore, it is desirable to avoid the use of SDS during the reverse staining step. Two procedures have been developed to resolve this problem (*see Notes 8 and 9, Figs. 2 and 3*).

4. Notes

1. Imidazole is a five-membered heterocyclic ring containing a tertiary (pyridine) nitrogen at position 3, and a secondary (pyrrole) nitrogen at position 1. It is a monoacidic base whose basic nature is due to the ability of pyridine nitrogen to accept a proton. The pyrrole nitrogen can lose the hydrogen atom, producing imidazolate anion (Im⁻), at high pH values (pKa approx 14.2). However, deprotonation of imidazole's pyrrolic nitrogen may also occur at lower pH values, upon complexation of Zn²⁺ at the pyridinium nitrogen (*see ref. 9 and references therein*). As a result, ZnIm₂ can form and precipitate at pH > 6.2. Upon treatment of a polyacrylamide or agarose electrophoresis gel with salts of zinc(II) and imidazole, a complex system is generated, due to the presence of amide groups in the polyacrylamide matrix and sulfate groups in the agarose gel, as well as Tris, glycine, dodecyl sulfate, hydroxyle, and carbonate anions in the electrophoresis buffers. These groups can coordinate with Zn(II), act as counteranions in the complexes of zinc with imidazole, and lead to the formation of complex salts and hydroxides. Diffusion phenomena (reflected in the times required for optimal gel equilibration and development during reverse staining) also critically influence the reverse-staining reactions between Zn²⁺ and imidazole (*see Notes 2 and 3*). Nevertheless, when the protocol described in this chapter is followed, ZnIm₂ is the major component of the precipitate that stains the gels treated with zinc sulfate and imidazole (**9**).

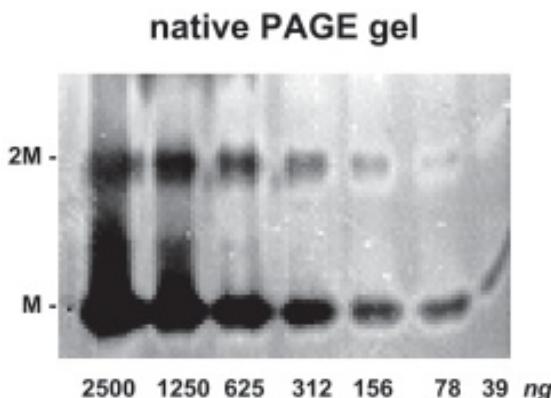


Fig. 2. Reverse staining of serial dilutions of human serum albumin (HSA) after native polyacrylamide gel electrophoresis. HSA migrates, yielding two main bands corresponding to its monomer (M) and its dimer (2M), and is detected under native conditions (no sodium dodecyl sulfate) due to its natural ability to complex with Zn^{2+} .

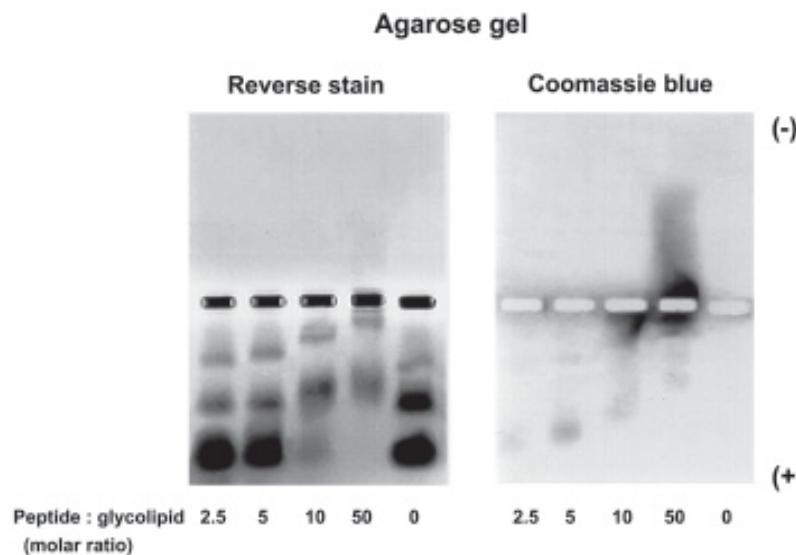


Fig. 3. Reverse staining of complexes between a synthetic cationic peptide with lipopolysaccharide binding properties and the glycolipid of a *N. meningitidis* strain. Formation of complexes was promoted by incubating ($37^{\circ}C$) peptide and glycolipid for 30 min in 25 mM Tris-HCl, 0.1% Triton X-100, pH 8.0, at indicated peptide:glycolipid molar ratios. Reverse staining revealed bands of both peptide-glycolipid complexes and uncomplexed glycolipid. Coomassie blue stained mainly the uncomplexed peptide migrating towards the negative (-) electrode. Coomassie blue failed to stain the glycolipid bands and stained peptide-glycolipid complexes very weakly.

2. This reverse staining protocol is optimized for use with any standard PAGE system, regardless of whether PAGE is the first (1-D PAGE) or the second (2-D PAGE) dimension in the separation strategy. The equilibration step assures that proteins in the gel are all uniformly coated with SDS. Therefore, protein's ability to bind Zn^{2+} is modulated by the protein-SDS complex, and limits of detection (approx 10 ng protein/band) are similar for

gels cast with and without SDS (i.e., SDS-PAGE and native PAGE, respectively). The larger the gel thickness or acrylamide concentration, the longer the equilibration step. A 15-min-long equilibration is enough for gels with =15% acrylamide and =1 mm thickness. Preparative gels are often as thick as 3–5 mm; these gels should be equilibrated for 30–60 min. Insufficient equilibration may result in faint reverse-stained patterns, which may fade upon prolonged storage.

3. Development time must be between 30 and 45 s, as insufficient development results in pale background staining and overdevelopment causes overstaining. An overstained gel can be re-stained. For this, the gel is treated with 10 mM ethylenediaminetetraacetic acid (EDTA) or 100 mM glycine for 5–10 min to redissolve the white ZnIm₂ precipitate that has deposited on the gel surface, rinsed in water (30 s), and, finally, soaked in the storage sodium carbonate solution. If the reverse-stained pattern is not restored in approx 5 min, the reverse-staining procedure can be repeated as indicated in Methods. Usually, the reverse-stained pattern will be restored during the equilibration step due to the precipitation of traces of Zn²⁺ already present in the gel with the imidazole from the equilibration solution. If the above suggestions do not lead to a homogenous reverse-stained pattern of suitable quality, the gel can be “positively” re-stained with Coomassie blue or silver (5). In this case, it is recommended that the reverse-stained gel be treated with EDTA (50 mM, pH 8.0; 30 min incubation) to free proteins of Zn²⁺. Then, the gel can be processed with Coomassie blue or silver stains.
4. Protein elution from reverse-stained gels can be performed following any conventional method, such as electroelution or passive elution; however, a highly efficient procedure has been developed (7,8). The reverse-stained band of interest is excised, placed in a (1 mL) plastic vial, and incubated with EDTA (50 mM, 2 × 5 min, and 10 mM, 1 × 5 min) to chelate protein-bound zinc ions. Supplementation of EDTA with nonionic detergents is optional but convenient if protein is to be in-gel refolded; e.g., Triton X-100 was useful when refolding proteins that were to be bioassayed (9,15). EDTA (or EDTA-detergent mixture) is replaced by an appropriate assay buffer (e.g., phosphate saline solution), in which the protein band is equilibrated. Finally, the band is homogenized and protein is eluted into an appropriate volume of the assay buffer (7,8,14). The slurry is centrifuged and the supernatant filtered to collect a clean, transparent solution of protein ready for subsequent analysis (7–9,14,15).
5. An important application of reverse staining is in structural/functional proteomics (13,14). Identification of proteins separated by electrophoresis is a prerequisite to the construction of protein databases in proteome projects. Matrix-assisted laser desorption/ionization mass spectrometry (MALDI-MS) has a sensitivity for peptide detection in the lower fmol range. In principle, this sensitivity should be sufficient for an analysis of small amounts of proteins in silver-stained gels. However, a variety of known factors (e.g., chemical sensitizers such as glutaraldehyde) and other unknown factors modify the silver-stained proteins, leading to low sequence coverage (13). Low-abundance proteins have been successfully identified after ZnIm₂ reverse staining (13,14).
6. Reverse staining of gels that have already been stained with Coomassie blue reveals proteins undetected with Coomassie blue, thus improving detection. Double staining can enhance sensitivity of detection in silver-stained gels as well. Before subjecting a Coomassie- or silver-stained gel to double staining, the gel should be rinsed in water (2–3 × 10 min). This step assures a substantial removal of acetic acid from Coomassie and silver stains. Then, the gel is subjected to the standard reverse-staining protocol. The resultant double-stained patterns consist of the previously seen Coomassie blue (brown, in the case of silver) stained bands and new unstained (reverse-stained) bands that contrast against the deep-white, ZnIm₂-stained background.

7. Reverse staining can be used in combination with other specific protein stains to indicate the presence of both protein and glycolipid. As with proteins, the biological properties of the presumed glycolipid bands are testable in a functional experiment following elution from the reverse-stained gel (9,18). After the electrophoresis step, the gel is rinsed for 30 min (2 × 15 min) in aqueous solution of methanol (40%, v/v). This step presumably removes excess SDS from glycolipid molecules that co-migrate with the proteins in SDS-PAGE, making the glycolipid adopt a conformation with high affinity for Zn²⁺. Next, the gel is reverse stained (Methods). Similar to proteins, glycolipid bands show up transparent and unstained. Sensitivity of detection is approx 10 ng glycolipid/band. If the same or a parallel gel is treated with Coomassie blue (silver) stain, that does not detect glycolipids; protein and glycolipid bands can be distinguished by their distinct staining and migration properties (9,18). A replicate protein sample treated with proteinase K is a useful control, as this protease digests protein but not glycolipid.
8. An alternative method of precipitating ZnIm₂ facilitates the reverse staining of native PAGE gels without the use of SDS (9). A quick and a slow reverse-staining procedure were implemented (9). The quick version makes use of the characteristic neutral to basic pH of the gels immediately after electrophoresis. When the gel is incubated (3–6 min) in a slightly acidic solution containing zinc sulfate and imidazole (pre-mixed to yield the molar ratio Zn(II):ImH = 15 mM:30 mM, pH approx 5.0–5.5), the gel stains negatively as ZnIm₂ deposits along its surface. Blotchy deposits are prevented by intense agitation and re-solubilized by lowering the pH of the zinc-imidazole solution; this also provides a basis for a slow version of this method. In the slow version (Fig. 2), the gel is incubated (10–20 min) in a solution of zinc sulfate (15 mM) and imidazole (30 mM), adjusted to pH 4.0. No precipitation occurs at this pH. The solution is poured off, and the gel is rinsed with water (30 s); the electropherogram is developed by incubating the gel for approx 5 min in 1% sodium carbonate. Sodium carbonate increases the pH first at the gel surface. Therefore, Zn²⁺ and imidazole, already present in the gel, react at the gel surface to form ZnIm₂. Again, protein bands are not stained, as they complex with Zn²⁺ and locally prevent the precipitation of ZnIm₂.
9. Many proteins and peptides, as well as glycolipids and their complexes with certain proteins/peptides, separate well on agarose gels under non-denaturing conditions (9). To reverse stain an agarose gel (9), following electrophoresis, the gel is incubated for 25–30 min in a zinc sulfate–imidazole solution (Zn(II):ImH = 15 mM:30 mM, adjusted to pH 5.0 with glacial acetic acid). During this step, Zn²⁺ and imidazole diffuse into the agarose matrix. The gel is then rinsed in water for 5–8 min to remove any excess of the staining reagents from the gel surface. The reverse-stained electropherogram is developed by incubating the gel for 5–8 min in 1% Na₂CO₃ (Fig. 3). Of note, due to poorly understood factors, agarose gels are not as amenable to reverse staining as polyacrylamide gels. Patterns of positive and negative bands are often seen in agarose gels.

Acknowledgments

The author expresses his gratitude to his former colleagues, Drs. L. Castellanos-Serra and E. Hardy from the Centre for Genetic Engineering and Biotechnology (CIGB), Havana, Cuba, for their contributions to the development of the reverse-staining concept, and to the CIGB for early support. Currently, the author is a Heart and Stroke Foundation of Canada Scholar, funded by grants from Natural Sciences and Engineering Council, Heart and Stroke Foundation of Canada, and the Canadian Institutes of Health Research.

References

1. Lee, C., Levin, A., and Branton, D. (1987) Copper staining: a five-minute protein stain for sodium dodecyl sulfate-polyacrylamide gels. *Anal. Biochem.* **166**, 308–312.
2. Dzandu, J. K., Johnson, J. F., and Wise, G. E. (1988) Sodium dodecyl sulfate-gel electrophoresis: staining of polypeptides using heavy metal salts. *Anal. Biochem.* **174**, 157–167.
3. Adams, L. D. and Weaver, K. M. (1990) Detection and recovery of proteins from gels following zinc chloride staining. *Appl. Theor. Electrophor.* **1**, 279–282.
4. Fernandez-Patron, C., and Castellanos-Serra, L. (1990) Eight International Conference on Methods in Protein Sequence Analysis, Kiruna, Sweden, July 1–6, Abstract booklet.
5. Fernandez-Patron, C., Castellanos-Serra, L., and Rodriguez, P. (1992) Reverse staining of sodium dodecyl sulfate polyacrylamide gels by imidazole-zinc salts: sensitive detection of unmodified proteins. *Biotechniques* **12**, 564–573.
6. Fernandez-Patron, C., Calero, M., Collazo, P.R., et al. (1995) Protein reverse staining: high-efficiency microanalysis of unmodified proteins detected on electrophoresis gels. *Anal. Biochem.* **224**, 203–211.
7. Castellanos-Serra, L. R., Fernandez-Patron, C., Hardy, E., and Huerta, V. (1996) A procedure for protein elution from reverse-stained polyacrylamide gels applicable at the low picomole level: An alternative route to the preparation of low abundance proteins for microanalysis. *Electrophoresis* **17**, 1564–1572.
8. Castellanos-Serra, L. R., Fernandez-Patron, C., Hardy, E., Santana, H., and Huerta, V. (1997) High yield elution of proteins from sodium dodecyl sulfate-polyacrylamide gels at the low-picomole level: Application to N-terminal sequencing of a scarce protein and to in-solution biological activity analysis of on-gel renatured proteins. *J. Protein Chem.* **16**, 415–419.
9. Fernandez-Patron, C., Castellanos-Serra, L., Hardy, E., et al. (1998) Understanding the mechanism of the zinc-ion stains of biomacromolecules in electrophoresis gels: generalization of the reverse-staining technique. *Electrophoresis* **19**, 2398–2406.
10. Kaim W. and Schwederki, B. (eds) (1994) *Bioinorganic Chemistry: Inorganic Elements in the Chemistry of Life*, John Wiley & Sons, Chichester-New York-Brisbane-Toronto-Singapore.
11. Fernandez-Patron, C., Hardy, E., Sosa, A., Seoane, J., and Castellanos, L. (1995) Double staining of coomassie blue-stained polyacrylamide gels by imidazole-sodium dodecyl sulfate-zinc reverse staining: sensitive detection of coomassie blue-undetected proteins. *Anal. Biochem.* **224**, 263–269.
12. Ferreras, M., Gavilanes, J. G., and Garcia-Segura, J. M. (1993) A permanent Zn²⁺ reverse staining method for the detection and quantification of proteins in polyacrylamide gels. *Anal. Biochem.* **213**, 206–212.
13. Scheler, C., Lamer, S., Pan, Z., Li, X. P., Salnikow, J., and Jungblut, P. (1998) Peptide mass fingerprint sequence coverage from differently stained proteins on two-dimensional electrophoresis patterns by matrix assisted laser desorption/ionization-mass spectrometry (MALDI-MS). *Electrophoresis* **19**, 918–927.
14. Castellanos-Serra, L., Proenza, W., Huerta, V., Moritz, R. L., and Simpson, R. J. (1999) Proteome analysis of polyacrylamide gel-separated proteins visualized by reversible negative staining using imidazole-zinc salts. *Electrophoresis* **20**, 732–737.
15. Hardy, E., Santana, H., Sosa, A., Hernandez, L., Fernandez-Patron, C., and Castellanos-Serra, L. (1996) Recovery of biologically active proteins detected with imidazole-sodium dodecyl sulfate-zinc (reverse stain) on sodium dodecyl sulfate gels. *Anal. Biochem.* **240**, 150–152.

16. Laemmli, U. K. (1970) Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **227**, 680–685.
17. Sambrook, J., Fritsch, E. F., and Maniatis, T. (eds) (1982) *Molecular Cloning. A Laboratory Manual*, Cold Spring Harbor Press, Cold Spring, NY.
18. Hardy, E., Pupo, E., Castellanos-Serra, L., Reyes, J., and Fernandez-Patron, C. (1997) Sensitive reverse staining of bacterial lipopolysaccharides on polyacrylamide gels by using zinc and imidazole salts. *Anal. Biochem.* **244**, 28–32.

Multiplexed Proteomics Technology for the Fluorescence Detection of Glycoprotein Levels and Protein Expression Levels Using Pro-Q® Emerald and SYPRO® Ruby Dyes

Birte Schulenberg and Wayne F. Patton

1. Introduction

Protein glycosylation is increasingly being recognized as one of the most prominent posttranslational modifications associated with malignant transformation and tumorigenesis as well as cell differentiation. The multiplexed proteomics platform is a new technology that permits quantitative, multicolor fluorescence detection of proteins in gels and on Western blots. This methodology allows the sequential determination of both protein expression level changes as well as altered glycosylation patterns within a single experiment. The linear responses of the fluorescent dyes allow rigorous quantitation of changes in protein expression using advanced image analysis software over a broad 3-log linear dynamic range. Global analysis of glycosylation changes can be expanded by dichromatic lectin-based profiling methods for rapidly categorizing glycan branching structure as well as deglycosylation studies employing exo- and endoglycosidases (1,2). The basic principles and methodologies for the multiplexed proteomics technology are explained in this chapter.

2. Materials

All fluorescent stains were obtained from Molecular Probes, Inc., Eugene, OR.

PNGase F was obtained from New England Biolabs. The volumes used for staining large-format 2-D gels are typically 500 mL/gel, except for the staining step, which requires a minimum of 330 mL/gel (see protocol below). For smaller gels, a volume of stain that is roughly ten times the volume of the gel should be employed (see Note 1).

2.1. Deglycosylation of Samples Using PNGase F

This step is not a part of the staining protocol and is required only if a distinction between *N*- and *O*-linked glycosylation is needed. For further explanation please see Subheading 3.1. The following PNGase F enzyme was used successfully in our experiments. It also includes the buffers needed for the digest.

1. PNGase F (New England Biolabs, cat. no. P 0704S).
2. 10X glycoprotein denaturing buffer: 5% sodium dodecyl sulfate [SDS], 10% β -mercaptoethanol.

3. 10X reaction buffer: 500 mM sodium phosphate, pH 7.5.
4. 10% Nonidet P-40.

2.2. Glycoprotein Detection Using Pro-Q® Emerald 300 Glycoprotein Gel Stain

1. Pro-Q Emerald 300 glycoprotein gel stain kit (2) (cat. no. M-33307, Molecular Probes, Eugene, OR). The kit contains all the components for oxidation and staining using the Multiplexed Proteomics technology. Included in the kit:
 - a. Pro-Q Emerald 300 reagent, component A, enough for 1 L of total staining solution, which can be used to stain up to three large 2-D gels (20 × 20 cm).
 - b. Pro-Q Emerald 300 staining buffer (component B), 1L.
 - c. Periodic acid, 10 g (component C).
 - d. SYPRO® Ruby protein gel stain.
2. Fix solution for Pro-Q Emerald dye staining: 50% methanol, 10% acetic acid in dH₂O.
3. Wash solution for Pro-Q Emerald dye staining: 3% acetic acid in dH₂O.
4. Carbohydrate oxidation solution: 1% periodic acid in 3% acetic acid. Add 1L 3% acetic acid to component C and mix until completely dissolved.
5. Pro-Q Emerald staining solution: Add 24 mL of dimethylformamide (DMF) to the Pro-Q Emerald 300 reagent (component A) and mix gently until completely dissolved. DMSO can be used instead of DMF, although the gel background might be a little bit higher. Unused solution should be stored in the freezer (-20°C). Each large-format 2-D gel is stained with 330 mL (absolute minimum in order to obtain high-quality staining) of diluted Pro-Q Emerald 300 dye.
6. Wash solution for SYPRO Ruby protein gel stain: 10% methanol (or ethanol), 7% acetic acid.

2.3. Glycoprotein Detection Using Pro-Q Emerald 488 Glycoprotein Gel Stain

Using this dye, it will be possible to analyze the gels with a laser-based imaging system. However, it is important to note that the sensitivity will be 10 times less than that of Pro-Q Emerald 300 dye, which is the dye of choice for 2-D gels. The following buffers and solutions are needed:

1. Pro-Q Emerald 488 glycoprotein gel stain kit (cat. no. M-21875, Molecular Probes, Eugene, OR).
 - a. Pro-Q Emerald 488, 10 vials, component A, enough for 250 mL of total staining solution.
 - b. Pro-Q Emerald 488 staining buffer (component B), 250 mL.
 - c. Periodic acid, 2.5 g (component C).

Also needed are the solutions in **Subheading 2.1., items 1–6**, except that the Pro-Q Emerald 488 dye is used in this case. **Subheading 2.1., item 4** requires only 250 mL of 3% acetic acid to make up a 1% periodic acid, 3% acetic acid oxidation solution. If a large 2-D gel is to be oxidized, add 500 mL of 3% acetic acid in order to cover the whole gel. One vial of reagent should be dissolved in 0.5 mL of DMSO or DMF. Do not freeze unused dye stocks—discard them.

2.4. Lectin Detection of Glycoproteins

Lectins can be used to distinguish between different branching structures of carbohydrates present on proteins. Because lectins are a very diverse family, only two differ-

ent ones will be described in this chapter. The protocol for blotting detection should be easily adaptable to others, however.

1. Concanavalin A conjugated to alkaline phosphatase (Pro-Q Glycoprotein Blot Stain Kit no. 1 with concanavalin A and DDAO phosphate, cat. no. P-21870, Molecular Probes, Eugene, OR). Make a stock solution of 2 mg/mL in dH₂O.
2. Ricin conjugated to alkaline phosphatase (EY Labs, cat. no. LA-2002-1, San Mateo, CA). Make a stock solution of 2 mg/mL in dH₂O. Care should be exercised with the ricin conjugate, as it is a toxin and should be handled with gloves at all times.
3. DDAO phosphate (cat. no. D-6487, Molecular Probes, Eugene, OR).
4. Wash buffer for blots: 50 mM Tris-HCl, 150 mM NaCl, pH 7.5.
5. Blocking buffer: 50 mM Tris-HCl, 150 mM NaCl, 0.2% Tween-20, 0.25% Mowiol 4-88, pH 7.5. Mowiol 4-88 was obtained from Hoechst-Celanese Corporation, Charlotte, NC (cat. no. 50661910). It is easiest if a 10% Mowiol stock solution is prepared in 60°C water and then diluted into the buffer to the final concentration.
6. Incubation buffer: wash buffer plus 1 mM CaCl₂, 0.5 mM MgCl₂.
7. DDAO phosphate reaction buffer: 10 mM Tris, 1 mM MgCl₂, pH 9.5.
8. DDAO phosphate stock solution: 1.25 mg/mL in DMF, store at -20°C. This is stable for at least 6 mo. When the solution turns blue, the substrate has broken down and can no longer be used.

2.5. Image Analysis

Compugen's Z3 software (Tel Aviv, Israel), Decodon's Delta2D software (Greifswald, Germany), or PDQuest (BioRad, Hercules, CA) are some of the most often used software packages used for multiplexed proteomics analysis.

3. Methods

3.1. PNGase F Treatment

This step aids in the identification of glycoprotein but is not a requirement for the staining. PNGase F is a widely used endoglycosidase that cleaves oligosaccharides that are *N*-linked on asparagine residues. The deglycosylation can be used as an additional control for determining the glycosylation status of a protein, and allows one to distinguish between *O*- and *N*-linked oligosaccharides. The only known inhibitor of that cleavage function is α (1-3)-linked fucose, which can be found in bromelin and horseradish peroxidase. Best activity is achieved at pH 8.6, but the enzyme can be used at a range of 7.5-9.5 with 20% maximal loss in activity. The following protocol was used for most applications.

1. Dilute the protein mixture to 1 mg/mL in 1X glycoprotein denaturing buffer.
2. Heat to 95°C for 5 min.
3. Cool down the solution to room temperature.
4. Dilute the sample twofold by adding 1% Nonidet P-40 final concentration, 1X reaction buffer, and dH₂O to make up the final volume. Example: To 10 μ L protein sample add 6 μ L dH₂O, 2 μ L 10% Nonidet P-40, and 2 μ L 10X reaction buffer.
5. Add the recommended amount of PNGase F for the protein amount to be deglycosylated (this is lot and manufacturer dependent).
6. Incubate at 37°C for 2 h to overnight.

3.2. Staining of Glycoproteins

The Pro-Q Emerald 300 or 488 glycoprotein gel stain kits provide an easy method for detecting glycoproteins. Pro-Q Emerald 300 dye is excited at 280–300 nm and emits maximally at 530 nm. If Pro-Q Emerald 488 glycoprotein Gel Stain Kit is utilized, the dye is maximally excited with a 473 or 488 nm laser. For emission, a 520 nm bandpass or longpass emission filter can be used. Even though the Pro-Q Emerald 488 dye is the glycoprotein stain of choice when imaging with a laser scanner system, sensitivity is 10 times lower than with the Pro-Q Emerald 300 dye. It also requires more washes than the Pro-Q Emerald 300 dye, so care should be employed during the wash steps. For 2-D gels, the Pro-Q Emerald 300 dye is strongly recommended (*see Notes 2 and 3*).

1. Fix the gel. Fix all gels before staining, 1 × 30 min followed by an incubation at room temperature with gentle agitation (e.g., on an orbital shaker at 50 rpm) overnight. This is to ensure that all sodium dodecyl sulfate is washed out of the gel (*see Note 4*). For large 2-D gels, use 500 mL of fix solution.
2. Wash the gel. Incubate the gel in 500 mL of wash solution with gentle agitation for 15 min. Repeat this step two to three times for large gels.
3. Oxidize the carbohydrates. Incubate the gel in 500 mL of carbohydrate oxidation solution with gentle agitation for 60 min.
4. Wash the gel. Incubate the gel in 500 mL of wash solution with gentle agitation for 20 min. Repeat this step three times more for large 2-D gels. *Important:* If the Pro-Q Emerald 488 dye is used, add two more wash steps for 20 min each in order to remove the periodic acid solution thoroughly. Otherwise the staining will either fail or be very dim and unspecific.
5. Pro-Q Emerald staining solution. Dilute the Pro-Q Emerald reagent (made in **Subheading 2.2., step 5**) 50-fold into Pro-Q Emerald staining buffer (provided in the kit). Large 2-D gels require 330 mL of staining solution as a minimum.
6. Staining the gel. Incubate the gel in the dark in 330 mL of Pro-Q Emerald staining solution while gently agitating for 2.5–3 h. Staining overnight is not recommended, because it increases background and unspecific staining.
7. Reducing the background staining. For Pro-Q Emerald 300 dye staining, incubate the gel in 500–700 mL of wash solution at room temperature for 20 min. Repeat this wash once for a total of two washes. The wash step should not exceed 2 h, as the staining intensity will decrease and bands might be lost. For the Pro-Q Emerald 488 dye staining, one wash for 20 min followed by three washes for 45 min each is recommended.
8. Imaging. If the Pro-Q Emerald 300 dye was used for staining, image the gels on a CCD camera-based ultraviolet (UV) transillumination system using a 520-nm emission filter. Typically, optimal visualization of the signal requires up to 10-s exposures (*see Note 5*). If the Pro-Q Emerald 488 dye was used for staining, the gels can be imaged on either a 473 nm or 488 nm containing laser-based gel scanner system. A 520-nm bandpass or longpass filter should be employed. *Important:* Image all gels before continuing with a general protein stain (including SYPRO Ruby dye).

3.3. Staining of Total Protein Using SYPRO Ruby Protein Gel Stain

Glycoprotein staining can be easily followed by SYPRO Ruby protein gel stain (3) and does not require another fixation step between the stains.

1. Staining. Pour 500 mL of stain on a large 2-D gel (50 mL for minigels) and stain the gel overnight.
2. Transfer the gel to a clean staining dish to minimize speckling.

3. Wash. Wash the gel in 7% acetic acid, 10% methanol, 2×30 min to reduce the background (see Note 6).
4. Wash the gel in dH₂O before imaging it, in order to prevent possible corrosion of the imaging machines.
5. Imaging. SYPRO Ruby dye stained proteins may be detected using a UV transilluminator equipped with a 520-nm longpass filter or a 555-nm bandpass filter. If a laser-based gel scanner system is used, the dye should be excited with a 473- or 488-nm laser. Filters compatible with the emission include 520- or 580-nm longpass filters as well as a 555-nm bandpass filter.

3.4. Western Blot Detection of Glycoproteins Using Lectins

Lectins are known to have different affinities for carbohydrate branching structures. The two lectins used for our studies are concanavalin A (Con A) and ricin. Con A has a high affinity for high mannose or hybrid-type glycans as well as complex-type mono- and bi-antennary complexes. Ricin preferentially binds tri-, tetra-, and penta-antennary structure of the complex type. Both are specific for *N*-linked glycan structures.

1. Wash the blot: After electrotransfer of the proteins onto PVDF membrane, the membrane should be dried in order to reduce the background staining. Wash the blot with wash buffer 3×10 min.
2. Blocking: Incubate the blot in blocking buffer for 1–2 h or until the membrane is completely wet.
3. Reaction with lectin-alkaline phosphatase conjugate: Dilute the lectin-alkaline phosphatase conjugate (Con A AP or ricin AP) stock solution 2000-fold into incubation buffer (prepared in Subheading 2.3., step 6) for a final concentration of 1 μ g/mL. Incubate in the diluted lectin conjugate for 1 h at room temperature.
4. Washing the blot: Wash off the excess lectin with two to three washes in blocking buffer for 10 min each, followed by another two washes in wash buffer for 2×10 min. The last washes remove the Tween-20, which would inhibit the alkaline phosphatase reaction.
5. Reaction with DDAO phosphate: Dilute the DDAO phosphate stock 1000-fold into the DDAO phosphate reaction buffer. Pipet a sufficient amount of diluted detection solution on top of the membrane (3–5 mL/minigel blot) to cover the whole area. Incubate for 15–30 min. Air-dry the membrane before imaging.
6. Image the blot: The DDAO phosphate can be imaged on a laser scanner equipped with a 633- or 635-nm laser, because it has an excitation maximum of 646 nm and an emission maximum of 659 nm. **Figure 1** shows a result obtained from human liver tumor lysate run on 2-D gels after gel staining as well as Western blotting results to show the complementary data that can be obtained by the multiplexed proteomics approach described in this chapter.

3.5. 2-D Gel Analysis Using Z3 Software

Image analysis with Z3 software works well with images in TIFF format that do not exceed 4.5 Mb in file size. Version 2.0 of the software allows for multiple gels (more than two) to be analyzed, whereas the older version 1.5 works only with two gels at a time. The newest release is Z3000, which allows for high throughput 2-D data analysis and automation. The software serves as a useful visual tool for quick analysis of gels, as well as quantitation of spots. When pairs of gels are compared, each gel is assigned a color (purple for the comparative image and green for the reference image). After superimposing the images, all spots that are present in both gels appear black or gray.

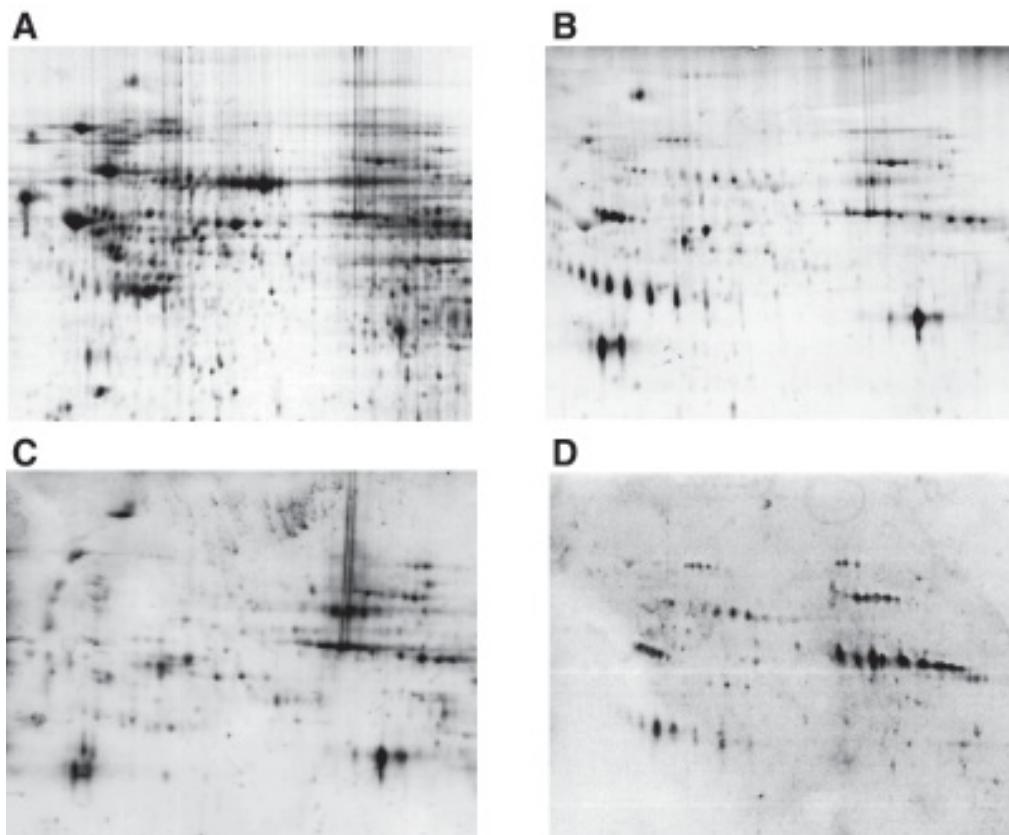


Fig. 1. Results obtained after gel staining as well as Western blotting of human liver tumor lysate on 2-D gels. (A) Total protein profile obtained with SYPRO Ruby protein gel stain. (B) Glycoprotein staining of the same gel as in A using Pro-Q Emerald 300 dye. (C) Result obtained by Western blotting using alkaline phosphatase conjugated concanavalin A. (D) Result obtained by Western blotting using alkaline phosphatase conjugated ricin lectin.

Differences appear purple or green. This allows for a quick visual analysis of 2-D gels and ready recognition of the appearance and the disappearance of proteins. For more detailed analysis, especially for protein differential expression, all spots may be quantitated automatically by the software and analyzed in a table format. However, it should be pointed out that the Z3 software expresses spot volumes in parts per million (ppm). This means the software takes the total area of the spots on a gel and then expresses a single spot relative to the total spot volume on a gel. If a project requires different protein loads on gels (e.g., gel no. 2 is loaded with twice as much material as gel no. 1), Z3 software should be used only for the detection of novel spots. The following paragraphs briefly outline the workflow for the analysis of two gels. Prior manipulation of images in graphics programs should be avoided (even with respect to size and grayscale) unless absolutely necessary, since this might result in a loss of data.

1. Open the reference-image first (green), then the comparative image (purple).
2. Detect the spots on both gels until enough features are indicated.

3. Edit the spots.
4. Overlay the gels. It is important that the gels have the same overall dimensions at this point. If the stains were imaged on different instruments, it might be necessary to adjust the image size in the Z3 program.
5. Add anchor points and register images until a satisfactory overlay is achieved. This can be done either manually or automatically. It generally works better if three to five anchors are placed manually before an automatic registration is attempted.
6. Match the spots (fully automatic).
7. Analyze the data.

Z3 version 2.0 compares images also in pairs, but assigns one master image that all others are compared to. The software guides the user through the steps and is very user friendly. If the registered image is to be used in a publication or for demonstration purposes, there are two ways to print that image. One way is to save the image in a GIF format directly from the Z3 software package. The image does lose some resolution, but most of the time can be made into a publication-quality picture using standalone graphics programs. The other way is to print screen and paste the image directly into a dedicated graphics program, which preserves the original image resolution. If specific glycoproteins are to be readily recognized by eye from a profile of total protein, it seems to be easier to color the specific features (in this case the glycoproteins) in purple and the general stain (in this case the SYPRO Ruby dye) in green. The spot tables that contain all the data analysis results can be exported into spreadsheet programs (like Microsoft Excel) in a text file format.

3.6. Image Analysis Using Delta2D Version 3.0

Decodon's Delta 2D software allows the same kind of analysis as Z3 software does, with the capability to generate virtual overlays of the gels and perform data analysis in spreadsheet format. However, there are some major differences between the programs worth noting. Pictures that have been taken with different imaging programs have to be set to the same resolution before opening them in Delta2D (e.g., 100–300 dpi). It is not possible to resize single images in the Delta2D software as it is in the Z3 package. Another minor difference is the way the programs deal with opening and closing images. Z3 software opens and closes images as most conventional graphics programs do. The Delta2D program only opens images and does not have a closing function. So, if one wants to open a different master image, just open a new picture and it will replace the older one. A major difference with Z3 software is that the spot finding and editing is performed after matching the gels and not before. The overall workflow in Delta2D software is as follows:

1. Open the master image.
2. Open the sample image.
3. Place anchor points (always place the anchor first on the master image and then on the sample image. The anchor point is displayed with the color of the image it has to be placed on).
4. Perform either an automatic warp (rubber sheeting) or, if required, perform a global warp first. Add more anchor points if needed in specific areas of the gel and then warp the image exactly. It is always possible to remove single anchors too. Coloring of spots in the master and sample images can be changed very easily by going to <color scheme> and loading a different default from the scroll-down menu. This allows one to switch between

black or white background, and so on.

5. After overlaying the images successfully, spot detection and editing is performed, as well as all the data analysis.

Delta2D software allows the user to switch between a ppm-based spot quantitation or the spot volume, depending upon the project. If there is a large difference in the spot intensities between gels (e.g., one gel was loaded with twice as much material as the other) it is advisable to use spot volume, whereas if 70–90% of the spots should remain constant, the ppm value will be the more useful tool.

4. Notes

1. The recommended staining volume for a 0.75 mm thick, 8 × 10 cm mini-gel is 50–100 mL.
2. Polypropylene dishes such as Rubbermaid containers are best suited for large-format 2-D gels.
3. All gels should be agitated on an orbital shaker at 50 rpm. Faster speeds usually result in damage of the gels.
4. If unspecific staining of nonglycosylated proteins is obtained, the fixation period in 50% methanol, 7% acetic acid was not sufficient. In this case, it might be necessary to perform a 60-min fix in methanol/acetic acid first to remove the bulk of the SDS, and then immerse the gels overnight in the fixation buffer.
5. If spots (or bands) appear to have a nonfluorescent center and only the perimeter is brightly stained, this is an indication of overloading of the gel. Lower the load or, if very low-abundance proteins are to be visualized, simply do not use this spot for quantitation. However, in the case of heavily overloaded gels, the unspecific staining will increase in signal intensity too.
6. The SYPRO Ruby protein dye-stained gels can be washed with either ethanol- or methanol-containing wash solution (3). Ethanol has the advantage of being more environmentally friendly, but might take a little bit more time (45 min instead of 30).

Acknowledgments

The authors would like to thank Dr. Tom Steinberg and Courtenay Hart for intellectual contributions to this project. This work was supported by a grant from the National Cancer Institute (grant number R33CA093292) awarded to Molecular Probes, Inc. Pro-Q and SYPRO are registered trademarks of Molecular Probes, Inc.

References

1. Schulenberg, B., Beecham, J. M., and Patton, W. F. (2003) Mapping glycosylation changes related to cancer using the multiplexed proteomics technology platform: a protein differential display approach. *J. Chromatog. B* **293**, 127–139.
2. Steinberg, T. H., Pretty On Top, K., Berggren, K. N., et al. (2001) Rapid and simple single nanogram detection of glycoproteins in polyacrylamide gels and on electroblots. *Proteomics* **7**, 841–855.
3. Berggren, K. N., Schulenberg, B., Lopez, M. F., et al. (2002) An improved formulation of SYPRO Ruby protein gel stain: comparison with the original formulation and with ruthenium II tris (bathophenanthroline disulfonate) formulation. *Proteomics* **5**, 486–498.

Multiplexed Proteomics Technology for the Fluorescence Detection of Phosphorylation and Protein Expression Levels Using Pro-Q® Diamond and SYPRO® Ruby Dyes

**Birte Schulenberg, Terrie Goodman, Thomas H. Steinberg
and Wayne F. Patton**

1. Introduction

The reversible phosphorylation of serine, threonine, and tyrosine residues is arguably one of the most important covalent posttranslational modifications regulating the functional status of proteins in eukaryotic organisms. Tools and techniques for determining the phosphorylation status of proteins and peptides thus play a prominent role in the investigation of diverse biological phenomena, including signal transduction, cell division, cell motility, apoptosis, metabolism, differentiation, gene regulation, and carcinogenesis. Typically, cells or isolated proteins are labeled with ^{32}P or ^{33}P prior to gel electrophoresis using protein kinases (1). The phosphoproteins are then usually detected by autoradiography, using film or storage phosphor imaging screens. In vitro radiolabeling provides a measure only of the phosphate groups attached during the actual labeling. No information is provided with respect to the physiological phosphorylation status of the proteins (2). Radiolabeling phosphoproteins also presents safety issues associated with handling the material and contamination of instrumentation (3).

Alternatively, phosphoproteins may be detected using standard Western blotting procedures (4,5). Although specific antibodies for phosphotyrosine residues are commercially available, antibodies that recognize phosphoserine or phosphothreonine residues are typically sensitive to amino acid sequence context and do not provide universal detection of proteins posttranslationally phosphorylated at these sites.

Pro-Q® Diamond phosphoprotein gel stain provides a novel fluorescence-based, in-gel detection system suitable for monitoring the phosphorylation status of proteins (6). Gels are fixed and then stained by a simple incubation in the dye solution. After staining, gels are destained and visualized using any of a variety of laser-based gel scanners or CCD/UV transilluminator systems. Pro-Q Diamond dye can detect between 1 ng (of a multiply phosphorylated protein) and 10 ng (of a singly phosphorylated protein) in gels. The stain is fully compatible with matrix-assisted laser desorption time-of-flight mass spectrometry, thus facilitating protein identification after gel electrophoresis. The dye can be used in combination with a total protein stain, such as SYPRO® Ruby protein gel stain, allowing protein phosphorylation levels and expression levels to be

monitored in the same gel. This chapter presents the basic protocol to achieve sensitive detection of phosphoproteins and total protein staining in 1- and 2-D gels.

2. Materials

Chemicals or buffers needed and not provided in the kit include: 1 *M* sodium acetate buffer (pH 4.0), filtered with an 0.45 μ m filter; acetonitrile (high-performance liquid chromatography [HPLC] grade); methanol (HPLC grade); and glacial acetic acid.

2.1. Gel Staining With Pro-Q Diamond Phosphoprotein Gel Stain

1. Multiplexed Proteomics phosphoprotein gel stain kit (Molecular Probes, Inc. cat. no. M-33305), containing 1L Pro-Q Diamond dye staining solution and 1 L SYPRO Ruby protein gel stain.
2. Fixative: 50% methanol, 10% acetic acid.
3. Destain/wash: 20% acetonitrile, 50 mM sodium acetate (pH 4.0) (see **Note 1**).
4. dH₂O.

2.2. Poststaining With SYPRO Ruby Protein Gel Stain

1. SYPRO Ruby protein gel stain (Molecular Probes, Inc., cat. no. S-12000, included in the kit), 1L.
2. Destain/wash: 10% methanol, 7% acetic acid.
3. dH₂O.

2.3. Gel Imaging and Data Analysis

1. A laser-based gel scanner equipped with a 532 to 560-nm laser and a 555-nm bandpass filter or a 580-nm longpass filter is optimal for visualizing the Pro-Q Diamond dye signal. However, an ultraviolet (UV) transilluminator equipped with similar emission filters may be used as well.
2. Software packages that are available for data analysis on 2-D gels include: Delta2D (Decodon, Greifswald, Germany), Z3 (Compugen, Tel Aviv, Israel), or Progenesis (Perkin Elmer, Boston, MA).

2.4. In-Gel Dephosphorylation by β -Elimination

Chemicals needed for this procedure include barium hydroxide (Sigma, St Louis, MO) and argon gas. The following volumes are for the treatment of one 6 \times 9 cm minigel.

1. Make a fresh saturated Ba(OH)₂ solution (0.155 *M*) as follows: de-gas 200 mL of dH₂O with argon for 5–10 min to remove all atmospheric carbon dioxide, which would result in the formation of insoluble barium carbonate later on. Make up 20 mL of a 155 mM BaOH solution in the degassed dH₂O. Stir for 20 min, avoiding introduction of air (cover the beaker with parafilm and replace the air with argon gas).

2.5. In Vitro Dephosphorylation

In addition to the β -elimination protocol, there is the possibility of dephosphorylating a sample enzymatically using phosphatases. The following reagents are optimized for the use of calf-intestine alkaline phosphatase.

1. Alkaline phosphatase (Sigma, cat. no. P-3681)
2. 1 *M* Tris-HCl buffer, pH 8.6–9.5.

3. 1 M MgCl₂.
4. 10% Sodium dodecyl sulfate (SDS) (optional).

2.6. *In Vitro* Phosphorylation

1. Protein kinase A, catalytic subunit (Sigma, cat. no. P-2645).
2. Protease inhibitor: 1 mg/mL leupeptin in dH₂O, 1 mg/mL pepstatin in ethanol, and 300 mM phenylmethyl sulfonamide (PMSF) in ethanol.
3. 200 mM ATP (pH 7.5).
4. 100 mM cAMP.
5. 1 M NaF in dH₂O.

3. Methods

In general, gel electrophoresis should be performed according to standard procedures (7,8). Minimal staining volumes for typical gel sizes are the following: 50 mL, for 8 cm × 10 cm × 0.75 mm gels (mini-gels); 330 mL, for 16 cm × 20 cm × 1 mm gels; 500 mL, for 20 cm × 20 cm × 1 mm gels; or approx 10 times the volume of the gel for other gel sizes.

Incubations are performed with continuous agitation on an orbital shaker at 50 rpm.

3.1. Staining Phosphoproteins In-Gel Using Pro-Q Diamond Dye

1. Fixing gels: After the electrophoresis is completed, gels are fixed in 50% methanol, 10% acetic acid. An overnight fix is recommended for large-format 2-D gels, with a 1-h fixation first and a second fixation step overnight to ensure the removal of SDS. Minigels can be fixed for 2 × 45–60 min, or overnight for convenience.
2. Washing gels: Incubate the gels two to three times (three times is recommended for large-format 2-D gels) in dH₂O for 15 min each step to remove residual acetic acid and methanol, which will interfere with the staining.
3. Staining: Incubate the gel with gentle agitation in the Pro-Q Diamond staining solution for 90–120 min. A 120-minute incubation is recommended for large-format gels. Overnight staining will only increase nonspecific staining as well as the gel background.
4. Destaining: To remove the background fluorescence, the gels are washed with either the proprietary destain solution (cat. no. 33310) or a 20% acetonitrile, 50 mM sodium acetate solution (pH 4.0) for 3 × 30–45 min each. On large gels, it might be necessary to include one more wash step for 30 min to obtain satisfactory results. Overnight destaining can be done, but will result in signal loss, whereas the background remains largely unchanged.
5. Optional wash: In order to reduce possible corrosion on imagers due to the destain solution, it is advisable to briefly wash the gels in dH₂O (10–20 min).
6. Imaging: the Pro-Q Diamond stain has an excitation maximum at 550 nm and an emission maximum at 580 nm. Stained gels are imaged best using a laser-based gel scanner with an excitation source of 532–560 nm. Emission is best captured using a 580-nm longpass filter or a 605-nm bandpass filter (see **Fig. 1A** for a typical result obtained with a laser-based gel scanner). If a Polaroid system is used, a 300-nm excitation source is recommended. The images can be documented on Polaroid film (667 black and white) using an exposure time of 2–5 s at an f-stop of 4.5, and a SYPRO Ruby photographic filter (Molecular Probes cat. no. S-6656). For CCD-based camera systems, use a filter that closely resembles the emission characteristics of the Pro-Q Diamond dye (e.g., a 600-nm bandpass filter). Exposure times of up to 40 s might be necessary to obtain the optimal grayscale coverage.

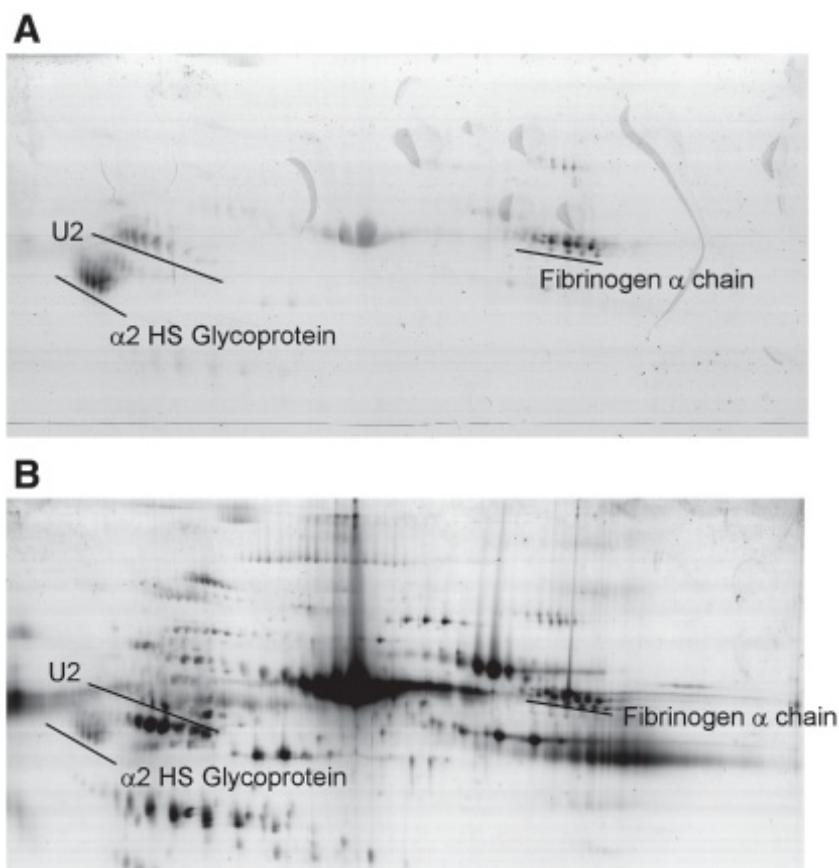


Fig. 1. Multiplexed proteomics staining of human plasma, separated on a 3.0–10.0 NL immobilized pH gradient strip (18 cm). (A) Image obtained after staining with Pro-Q® Diamond protein gel stain. (B) Image obtained after post-staining with SYPRO® Ruby protein gel stain. The three major phosphoproteins present are indicated.

3.2. Poststaining Using SYPRO Ruby Total Protein Stain

1. Staining: the gels can be directly stained with SYPRO Ruby protein gel stain solution without repeating any fixation. Staining overnight is recommended to obtain the best signal possible.
2. Destaining: the background fluorescence should be decreased by a brief wash for 2 × 15 min (30 min for large-format gels) in 10% methanol, 7% acetic acid.
3. Optional wash: a brief wash in dH₂O should be performed again (see Subheading 3.1., step 5).
4. Imaging: SYPRO Ruby can be imaged on either a UV-based gel imaging system or a laser-based scanner. SYPRO Ruby can be maximally excited with a 473- or 488-nm laser. Emission filters used include 530- or 580-nm longpass filters as well as a 555-nm bandpass filter (see Fig 1B for a typical result obtained after SYPRO Ruby poststaining). If a Polaroid-based camera system is used, gels can be photographed at an f-stop of 4.5 using a 1- to 2-s exposure time and a SYPRO Ruby photographic filter. For CCD-based camera systems, use a filter that closely resembles the emission characteristics of the dye.

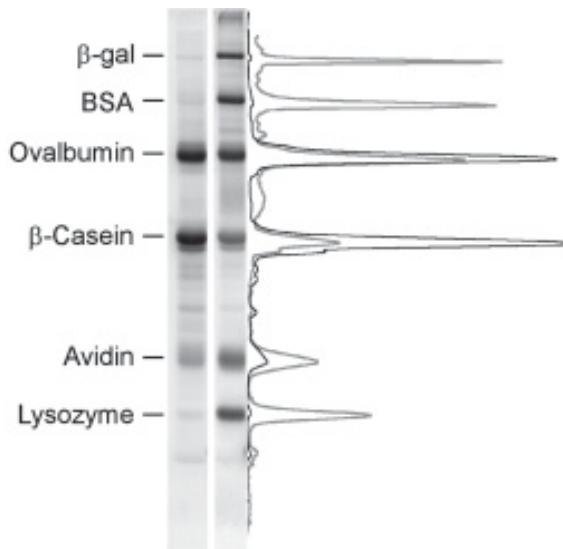


Fig. 2. Analysis of a 1-D gel using Image Gauge software. Peppermint-stick markers were separated on a 12% acrylamide gel (1000 ng per protein) and stained with Pro-Q® Diamond dye (D) followed by SYPRO® Ruby protein gel stain (S). The signal intensity profiles were plotted (Image Gauge, Fuji Film, Tokyo, Japan) for each stain and overlaid using Adobe Photoshop 7.0. The light gray trace reflects the profile of the SYPRO Ruby dye stained gel lane, and the black trace indicates the Pro-Q Diamond dye intensities. Phosphoproteins present in the marker include β -casein (five phosphates) and ovalbumin (two phosphates). All other proteins are not phosphorylated and serve as a negative control. These include β -gal (β -galactosidase), BSA (bovine serum albumin), avidin, and lysozyme.

3.3. Multiplexed Proteomics Analysis

1. Signal quantitation: Signals obtained after Pro-Q Diamond dye staining, as well as signals from total protein staining, can be quantitated using standard imaging software specific for the imaging platform used (e.g., Quantity One, BioRad, Hercules CA; Image Gauge from Fuji Film, Tokyo, Japan; or Scion Image [NIH Image for MacIntosh systems]). In addition, it is recommended to include two to four nonphosphorylated proteins as negative controls as well as two to three different phosphoproteins (with different levels of phosphate) as positive controls (see Note 2 and Fig. 2).
2. Semi-quantitative analysis: after quantitation of all proteins of interest and the controls, a ratio can be calculated of Pro-Q Diamond dye signal/SYPRO Ruby dye signal (9). Phosphoproteins will have a much higher ratio than nonphosphorylated proteins. This is an extremely helpful approach in order to distinguish between specific and unspecific signals, especially in lysates containing a variety of proteins present at different amounts. The two to three different phosphoproteins included in the controls will span a range of ratios, depending on the number of phosphates per protein. A protein with eight phosphates (α -casein) will have a much higher ratio than pepsin (one phosphate), but both are specific. These ratios should be higher than the calculated ratios for the nonphosphorylated protein controls (e.g., BSA, β -galactosidase, or carbonic anhydrase).
3. Qualitative analysis: In order to identify the phosphoproteins present in the total protein profile, any software capable of overlaying two images can be employed (e.g., Compugen's

Z3 (10), Tel Aviv, Israel; Decodon's Delta2D, Greifswald, Germany; or Adobe Photoshop, San Jose, CA). A brief overview is given in Chapter 20, and so shall not be repeated at this point.

3.4. In-Gel Dephosphorylation

After completing the electrophoresis, the gels are fixed for 2 \times 30 min (up to overnight) in 50% methanol, 10% acetic acid.

1. Dilute the saturated barium hydroxide solution to 50 mM final concentration with de-gassed dH₂O.
2. Pour 50–60 mL of the diluted solution onto the gel and incubate at 50°C for 30–50 min under argon gas (see Note 3).
3. To stop the reaction, add 2–4 mL of glacial acetic acid to reduce the pH to 4.0 in the barium hydroxide solution.
4. Wash 3 \times 10 min in 100 mL dH₂O (see Note 4).
5. Proceed with the Pro-Q Diamond dye staining protocol starting at Subheading 3.1., step 1.
6. Comparison of a gel treated with chemical dephosphorylation with an untreated control gel reveals proteins phosphorylated on serine and threonine residues. Phospho-tyrosines are resistant to β -elimination.

3.5. In Vitro Dephosphorylation

If a more gentle dephosphorylation protocol is needed, the alkaline phosphatase treatment should be preferred. The sample should not contain phosphatase inhibitors. If inhibitors were used during preparation, a protein precipitation step should be performed prior to alkaline phosphatase treatment. All cited amounts mentioned can be scaled up or down as needed, but the ratio of phosphatase to protein sample should be kept constant.

1. To 30 μ g of sample, add 25 mM Tris-HCl (pH 8.6–9.5) and 5 mM MgCl₂.
2. Add 0.5 μ g of alkaline phosphatase.
3. Incubate for 1–2 h at 37°C.
4. Stop the reaction by adding urea or SDS sample buffer.
5. Visualize the changes in phosphorylation using Pro-Q Diamond dye as described in Subheadings 3.1.–3.3.

If the phosphate was not removed by the above protocol, it might be necessary to add 1% SDS to make the phosphate residue more accessible to the alkaline phosphatase, which will be active in SDS.

3.6. In Vitro Phosphorylation

In some cases it might be desirable not only to study the phosphorylation present in a given sample after purification, but also to find proteins that can be phosphorylated in vitro by kinase reactions. This will result in information about a kinase motif contained within a protein, as well as information about potential regulatory pathways that may impinge on the target protein.

The variety of kinases commercially available (Upstate Biotechnology, Panvera, or New England Biolabs are good sources) has increased tremendously and cannot be covered in this chapter. The following protocol is an example describing a widely used kinase (protein kinase A).

1. Dilute the protein to 0.5–1 mg/mL in 50 mM Tris-HCl (pH 7.5).
2. Add 10 mM MgCl₂, 1 mM ATP (pH 7.5), 1 µg/mL pepstatin, 1 µg/mL leupeptin, 1 mM PMSF, 20 mM NaF, 20 µg/mL oligomycin, and 20 U of the active site of phosphoprotein kinase A (PKA). Oligomycin is an inhibitor of the ATP synthase, which would hydrolyze the added ATP very quickly. If a maximum of kinase activity is desired, 60 µM cAMP could be added in addition to activate kinases that are already present in the sample.
3. Incubate at 30° C for 30 min.
4. Stop the reaction by adding 1-D or 2-D sample buffer.
5. Visualize the changes in phosphorylation using Pro-Q Diamond dye as described in **Sub-headings 3.1.–3.3.**

3.7. MS Analysis of Phosphoproteins

The method for an MS analysis using matrix-assisted laser desorption time-of-flight (MALDI-TOF) is described in Chapter 32, and so shall not be repeated here.

4. Notes

1. Alternatively, an optimized destain solution can be purchased from Molecular Probes, Inc. (cat. no. P-33310). The protocol is the same for either destain.
2. Molecular-weight standards can be used as controls. Peppermint-stick markers contain a blend of phosphorylated and non-phosphorylated proteins for that purpose (Molecular Probes, Inc., cat. no. P-33350).
3. This step works best in a closed container that is flushed with argon gas.
4. The gels will have a white precipitate (barium carbonate) on top, which can be washed off to a certain extent. In addition, the gels will swell after the washes to double their normal size. The β-elimination procedure will also increase the background fluorescence after Pro-Q Diamond staining.

Acknowledgments

The authors would like to thank Jill Hendrickson for valuable contributions to the phosphoprotein project. This work was supported by a grant from the National Cancer Institute (grant number R33CA093292) awarded to Molecular Probes, Inc. Pro-Q and SYPRO are registered trademarks of Molecular Probes, Inc.

References

1. Guy, G., Philip, R., and Tan, Y. (1994) Analysis of cellular phosphoproteins by two dimensional gel electrophoresis: applications for cell signaling in normal and cancer cells. *Electrophoresis* **15**, 417–440.
2. Wind, M., Edler, M., Jakubowski, N., Linscheid, M., Wesch, H., and Lehmann, W. (2001) Analysis of protein phosphorylation by capillary liquid chromatography coupled to element mass spectrometry with 31P detection and to electrospray mass spectrometry. *Anal. Chem.* **73**, 29–35.
3. Conrads, T., Issaq, H., and Veenstra, T. (2002) New tools for quantitative phosphoproteome analysis. *Biochem. Biophys. Res. Commun.* **290**, 885–890.
4. Moore, D. and Sefton, B. (1995) In: *Current Protocols in Molecular Biology*, Vol. 3 pp. 18.0.3–18.4.5, Wiley, NY.
5. Kaufmann, H., Bailey, J., and Fussenegger, M. (2001) Use of antibodies for detection of phosphorylated proteins separated by two-dimensional gel electrophoresis. *Proteomics* **1**, 194–199.

6. Steinberg, T. H., Agnew, B. J., Gee, K. R., et al. (2003) Global quantitative phosphoprotein analysis using Multiplexed Proteomics technology. *Proteomics* **3**, 1128–1144
7. Laemmli, U. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **227**, 680–685.
8. O'Farrell, P. (1975) High resolution two-dimensional electrophoresis of proteins. *J. Biol. Chem.* **250**, 4007–4021.
9. Schulenberg, B., Aggeler, R., Beechem, J. M., and Patton, W. F. (2003) Analysis of steady-state protein phosphorylation in mitochondria using a novel fluorescent phosphosensor dye. *J. Biol. Chem.* **278**, 27,251–27,255
10. Smilansky, Z. (2001) Automatic registration for images of two-dimensional protein gels. *Electrophoresis* **22**, 1616–1626.

Sensitive Quantitative Fluorescence Detection of Proteins in Gels Using SYPRO® Ruby Protein Gel Stain

Birte Schulenberg, Nancy Ahnert, and Wayne F. Patton

1. Introduction

The fluorescent, noncovalent staining method, using SYPRO® Ruby dye resembles colorimetric staining procedures, such as colloidal Coomassie Blue dye staining. The dye provides noncovalent staining and is applied postelectrophoretically, which means that no alteration in the migration of proteins during electrophoresis can occur. The SYPRO Ruby dye-based staining method is also compatible with downstream applications such as Edman-based protein sequencing and peptide mass profiling by mass spectrometry-based methods.

SYPRO Ruby protein gel stain differs from the other SYPRO dyes in that it does not bind through intercalation into sodium dodecyl sulfate (SDS) micelles (1–4). The binding mechanism is actually quite similar to Coomassie Brilliant Blue stain, being via direct electrostatic interaction with basic amino acid residues (3). The staining procedure is simple and does not involve a lot of different solutions like silver staining (which can involve up to 10). This makes the dye ideal for high-throughput gel staining and large-scale proteomics applications. The stain is as sensitive as the best silver-staining methods available, and superior to them in terms of ease of use, linear dynamic range, and compatibility with downstream microchemical characterization techniques (5–7). SYPRO Ruby protein gel stain has exceptional photostability, allowing long exposure times for maximum sensitivity, and can be imaged on UV as well as laser-based gel scanners.

This chapter presents protocols for staining of 1- and 2-D gels with SYPRO Ruby protein gel stain. Methods for identification of proteins from stained gels are also presented.

2. Materials

Materials needed include standard equipment for 1- or 2-D gel electrophoresis, an orbital shaker, and a gel imaging system (ultraviolet [UV] or laser-based). Dishes for staining have to be cleaned and rinsed thoroughly before use to remove any detergent residue, which would interfere with the staining and result in higher background fluorescence (see Note 1).

2.1. Gel Staining With Recommended Procedure

1. Fixative for gels: 50% methanol (or ethanol), 10% acetic acid (*see Note 2*).
2. SYPRO Ruby protein gel stain (Molecular Probes, Inc., cat. no. S-12000), 1L.
3. Destain/wash: 10% methanol, 7% acetic acid.

2.2. Gel Staining With Reused SYPRO Ruby Protein Gel Stain

1. All solutions are the same as in **Subheading 2.1.** except that the dye is collected in a dark plastic bottle (HDPE) and stored until the next use.

2.3. Gel Staining With Diluted SYPRO Ruby Dye

1. Dilute the dye up to threefold maximum in dH₂O before use.
2. All other solutions are the same as in **Subheading 2.1.**

2.4. In-Gel Trypsin Digestion After SYPRO Ruby Dye Staining

1. Wash solutions: (a) 50% acetonitrile, 50 mM ammonium bicarbonate (pH 8.3); (b) 0.1% trifluoroacetic acid (TFA); (c) 100% acetonitrile.
2. Reduction/alkylation (if it was not performed already in the sample preparation): (a) 20 mM dithiothreitol (DTT) in 100 mM ammonium bicarbonate; (b) 100 mM iodoacetamide, 50 mM ammonium bicarbonate; (c) 50% acetonitrile, 100 mM ammonium bicarbonate (pH 8.3).
3. Trypsin digest: (a) 0.05 mg/mL trypsin (modified) in 50 mM ammonium bicarbonate, 10% acetonitrile (keep on ice until used); (b) 50 mM ammonium bicarbonate, 10% acetonitrile.
4. Extraction of peptides: (a) 0.1% TFA; (b) 30% acetonitrile, 0.1% TFA; (c) 60% acetonitrile, 0.1% TFA.
5. Spotting of peptides on a matrix-assisted laser desorption/ionization (MALDI) steel target: (a) 20 mg/mL nitrocellulose in acetone, vortex for 5 min; (b) 40 mg/mL α -cyanohydroxycinnamic acid in acetone, vortex for 5 min (this is a saturated solution, and not everything will be dissolved); (c) isopropanol; (d) 10% acetonitrile, 0.1% trifluoroacetic acid (TFA); (e) 3% formic acid.

3. Methods

In general, gel electrophoresis should be performed according to standard procedures (8,9). The stain is compatible with all other known gel chemistries as well (Tris/acetate, Tris/tricine, NUPAGE, and so on *see Note 3*). Minimal staining volumes for typical gel sizes are the following: 50 mL, for 8 cm \times 10 cm \times 0.75 mm gels (minigels); 330 mL, for 16 cm \times 20 cm \times 1 mm gels; 500 mL, for 20 cm \times 20 cm \times 1 mm gels; or approx 10 times the volume of the gel for other gel sizes.

Incubations are performed with continuous agitation on an orbital shaker at 50 rpm.

3.1. Standard Staining Protocol

The following protocol has been optimized for maximum sensitivity (1–2 ng) and linear dynamic range (from 1 to 1000 ng) on 1- and 2-D gels using the minimum volumes of stain possible.

1. *Fixation of gels:* Place minigel into 50–100 mL of fixative made in **Subheading 2.1., step 1.** Leave in the fix for 15 min and repeat this step once for optimal results. The fixation results in an increased signal intensity as well as better signal to background. If a large 2-D gel is to be stained, fix for 2 \times 30 min in 500 mL per step.

2. *Staining:* Stain the gel in SYPRO Ruby protein gel stain (50 mL for a minigel, 500 mL for a large 2-D) overnight (see Note 4).
3. *Destain/wash:* The next day, place the gel in destain solution (50 mL for minigels, 500 mL for large 2-D) to reduce the background (see Note 5). It is helpful to place the gels into a clean container for this step to reduce possible speckling. The gels should be washed for 2 × 15–30 min (2 × 30–60 min for large gels).
4. *Optional wash:* In order to reduce possible corrosion on imagers due to the destain solution, it is advisable to briefly wash the gels in dH₂O to remove the methanol/acetic acid solution.

3.2. Imaging of SYPRO Ruby Protein Gel Stain

The gels can be imaged on a standard 300-nm UV or a blue-light transilluminator (see Note 6). Gels may also be visualized using various laser-based scanners containing either a 473-nm (SHG) laser or a 488-nm argon-ion laser. Alternatively, use a xenon arc lamp, blue fluorescent light, or blue light-emitting diode (LED) source. Gels may be imaged using a Polaroid or CCD camera. Use Polaroid 667 black-and-white print film and the SYPRO protein gel stain photographic filter (Molecular Probes, Inc., cat. no. S-6656). Exposure times vary with the intensity of the illumination source; for an f-stop of 4.5, a 1–2 s exposure should give satisfactory results. Examples for SYPRO Ruby dye staining on 1- and 2-D gels can be seen in Figs. 1 and 2.

3.3. Dilution and Reuse of SYPRO Ruby Protein Gel Stain

In order to conserve stain, it might seem desirable to dilute the stain or use it more than once. This is not going to result in the optimal staining, and will decrease signal intensity as well as compromise the linear dynamic range. Nevertheless, it can be feasible for certain experiments, as outlined below.

1. *Dilution of stain:* SYPRO Ruby protein gel stain can be diluted with dH₂O as described in ref. 10. A dilution above 1:2 is not advisable due to too much loss in signal. Even with a 1:2 dilution, the signal is reduced roughly sixfold. At any dilution, the staining intensity is not linear with the protein load anymore and cannot be used for quantitation. However, the limit of detection is unchanged compared to the standard protocol.
2. *Reuse of stain:* SYPRO Ruby protein gel stain can be reused up to two times, with a signal reduction of up to 2.5 times. The stain should be stored in the dark between uses. As with dilution, the linear dynamic range is impacted, but full sensitivity is retained even in the third use.

Of the two methods, reuse or dilution, reuse will give the better results overall and should be used preferentially. On 2-D gels, neither of the above is advisable, since it will not be possible to quantitate results anymore.

3.4. In-Gel Trypsin Digestion From SYPRO Ruby Dye Stained Gels

1. Cut the band of interest with a clean razorblade (see Note 7) and place into a clean 1.5-mL tube. For large spots, cutting off a 1-mL pipet tip approx 1–3 mm from the tip to obtain the needed diameter works well. For smaller spots, 0.2-mL tips might be advisable. All tips should be autoclaved before use, and dust free. The cut spot can be removed with a small wire. Commercially available manual spot and band cutters are more convenient for larger sample numbers (available from The Gel Company, San Francisco, CA). Care should be taken at every step from running the gel to the MALDI analysis to minimize keratin contamination. Gloves are recommended for all procedures.

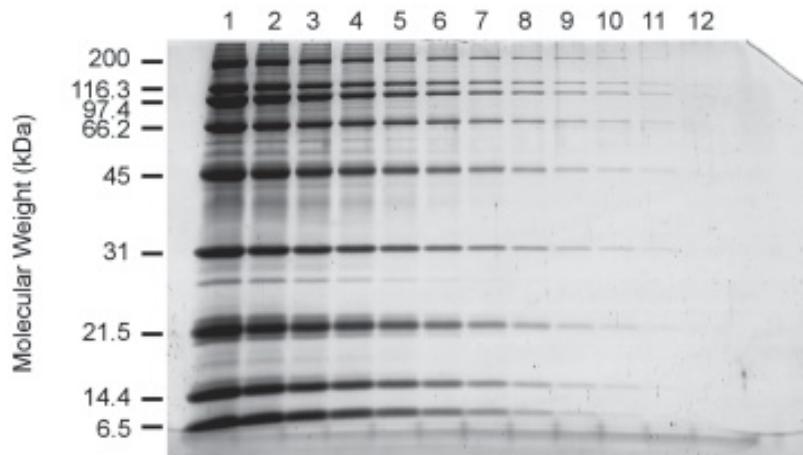


Fig. 1. —Sodium dodecyl sulfate (SDS)-polyacrylamide gel electrophoresis (PAGE) gel stained with SYPRO® Ruby protein gel stain. Broad range molecular-weight markers were loaded as a twofold dilutions series starting at a 1000 ng in the first lane and ending at 0.5 ng in lane 12. SYPRO Ruby protein gel stain readily detects 1 ng of protein.

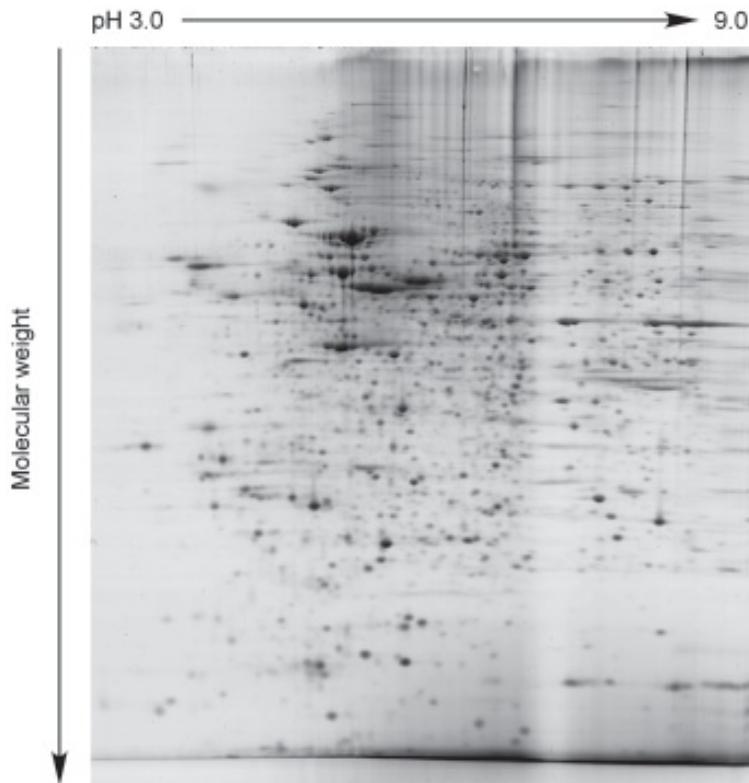


Fig. 2. Two-dimensional gel stained with SYPRO Ruby protein gel stain. 150 μ g of a Jurkat cell extract was loaded on the first-dimension gel (carrier ampholyte gel). A 12.5% sodium dodecyl sulfate (SDS)-polyacrylamide gel electrophoresis (PAGE) gel was used for the second dimension. More than 1000 different protein species are detectable on the gel.

2. Cut the gel into 1-mm² pieces and wash with 100 μ L 50% acetonitrile, 50 mM ammonium bicarbonate for 30 min. Wash another 30 min with 0.1% TFA. Repeat the 50% acetonitrile, 50 mM ammonium bicarbonate wash and then dehydrate the pieces in 100% acetonitrile for 5–10 min. Remove the acetonitrile and air-dry the gel.
3. Add 50 μ L fresh DTT solution and incubate at 56°C for 1 h. Remove the DTT and add 50 μ L iodoacetamide solution for 30 min at room temperature in the dark. To remove the iodoacetamide solution, wash the pieces with 100 μ L, 100 mM ammonium bicarbonate for 15 min. Repeat twice. Wash once with 50% acetonitrile, 50 mM ammonium bicarbonate for another 15 min before dehydrating the gel pieces in 100% acetonitrile again.
4. Add 10 μ L freshly made trypsin solution to each piece processed and allow the solution to soak into the gel before proceeding (10 min). Fully re-swell the pieces by adding 20 μ L 50 mM ammonium bicarbonate, 10% acetonitrile. Incubate the trypsin at 37°C overnight.
5. Remove the supernatant the next morning and transfer to a clean 0.5-mL tube. Add 50 μ L 0.1% TFA to the gel pieces and extract the peptides for 30 min by mixing on a vortex. Remove the supernatant and combine with the first one. Add 50 μ L 30% acetonitrile, 0.1% TFA to the gel pieces for 30 min and combine the supernatant again with the first two. Add 50 μ L 60% acetonitrile, 0.1% TFA to the gel pieces and repeat the extraction step. All supernatants combined should be dried down using a Speed-Vac concentrator (Thermo Savant, Milford, MA).
6. Resuspend the extracted peptides after drying in 10 μ L 10% acetonitrile, 0.1% TFA. Wait 30 min and vortex heavily before spotting on the MALDI steel target. For the matrix solution, mix one part nitrocellulose solution with one part isopropanol and two parts matrix solution. Spot 1 μ L of peptide sample and mix on the plate with 1 μ L of matrix solution immediately. Dry the spot. Wash the spot with 3% formic acid and dry again before analysis.

4. Notes

1. Staining containers: Polypropylene dishes, such as Rubbermaid Servin' Savers (Wooster, OH), are the optimal containers for staining because the high-density plastic adsorbs only a minimal amount of the dye. If containers have been used for staining before, rinse the containers with ethanol to remove residual dye. For small gels, circular staining dishes provide the best fluid dynamics on orbital shakers, resulting in less dye aggregation and better staining. For large-format 2-D gels, a convenient staining option uses the Clearview 3 Drawer Organizer (Sterilite, cat. no. 1790, Townsend, MA). This polypropylene box provides a convenient format for staining three gels per unit with up to three units stacked, and is available at department stores. Glass dishes are not recommended, as they have a tendency to bind dye.
2. Fixation: A variety of fixes will result in the same overall staining results, but the 50% methanol (or ethanol) 10% acetic acid has been found to work very reliably in 1- as well as 2-D gels. The following fixatives can be combined with SYPRO Ruby dye staining but might result in elevated background signal: 10% methanol/7% acetic acid, 25% ethanol/12.5% trichloroacetic acid, 40% ethanol/10% acetic acid.
3. Plastic-backed gels: The polyester backing on some precast gels is highly fluorescent. For maximum sensitivity using a UV transilluminator, the gel should be placed polyacrylamide side down and an emission filter, such as the SYPRO protein gel stain photographic filter (S-6656), used to screen out the blue fluorescence of the plastic. The use of a blue-light transilluminator or laser scanner will reduce the amount of fluorescence from the plastic backing so that the gel may be placed polyester side down.
4. Staining times can be varied between 4 h to overnight. However, it should be noted that decreasing the staining times to even 7 h results in a 50% decrease in overall brightness.

5. The destain solution can be replaced by water washes only, but the background will be higher than with the destain.
6. Place the gel directly on the imagers; do not use plastic wrap or plastic backing. It is important to clean the surface of the imagers after each use with deionized water and possibly ethanol and a soft tissue paper (like Kimwipes). Otherwise, fluorescent dyes, such as SYPRO Ruby dye, will accumulate on the glass surface and cause a high background fluorescence.
7. Even if bands are not to be processed immediately, the gel pieces should be cut out and placed in the freezer to prevent possible degradation and further contamination.

Acknowledgments

The authors would like to thank Kiera Berggren, Terrie Goodman, and Dr. Tom Steinberg for intellectual contributions to this project.

References

1. Berggren, K., Chernokalskaya, E., Steinberg, T., et al. (2000) Background-free, high-sensitivity staining of proteins in one- and two-dimensional sodium dodecyl sulfate-polyacrylamide gels using a luminescent ruthenium complex. *Electrophoresis* **21**, 2509–2521.
2. Berggren, K., Schulenberg, B., Lopez, M., et al. (2002) An improved formulation of SYPRO Ruby protein gel stain: Comparison with the original formulation and with a ruthenium II tris (bathophenanthroline disulfonate) formulation. *Proteomics* **2**, 486–498.
3. Patton, W. (2000) A thousand points of light; the application of fluorescence detection technologies to two-dimensional gel electrophoresis and proteomics. *Electrophoresis* **21**, 1123–1144.
4. Patton, W. (2002) Detection technologies in proteome analysis. *J. Chromatog. B* **771**, 3–31.
5. Lopez, M., Berggren, K., Chernokalskaya, E., Lazarev, A., Robinson, M., and Patton, W. (2000) A comparison of silver stain and SYPRO Ruby Protein Gel Stain with respect to protein detection in two-dimensional gels and identification by peptide mass profiling. *Electrophoresis* **21**, 3673–3683.
6. Nishihara, J. and Champion, K. (2002) Quantitative evaluation of proteins in one- and two-dimensional polyacrylamide gels using a fluorescent stain. *Electrophoresis* **23**, 2203–2215.
7. Gerner, C., Vejda, S., Gelmann, D., et al. (2002) Concomitant determination of absolute values of cellular protein amounts, synthesis rates, and turnover rates by quantitative proteome profiling. *Mol. Cell Proteomics* **1**, 528–537.
8. Laemmli, U. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **227**, 680–685.
9. O'Farrell, P. (1975) High resolution two-dimensional electrophoresis of proteins. *J. Biol. Chem.* **250**, 4007–4021.
10. Krieg, R. C., Paweletz, C. F., Liotta, L. A., and Petricoin, E. F., 3rd. (2003) Dilution of protein gel stain results in retention of staining capacity. *Biotechniques* **35**, 376–378.

Rapid, Sensitive Detection of Proteins in Minigels With Fluorescent Dyes

Coomassie Fluor Orange, SYPRO® Orange, SYPRO Red, and SYPRO Tangerine Protein Gel Stains

Thomas H. Steinberg, Courtenay R. Hart, and Wayne F. Patton

1. Introduction

The advent of polyacrylamide gel electrophoresis (PAGE) in the minigel format (i.e., gel dimensions in the range of 6×9 cm or 8×8 cm \times 0.75 to 1.5 mm thick), the widespread use of precast minigels, and the commercialization of colloidal Coomassie® Blue G-250 staining formulations have raised expectations for rapid, sensitive post-electrophoresis staining methods. The proprietary commercially available fluorescent stains Coomassie Fluor™ Orange protein gel stain, SYPRO® Orange dye, SYPRO Red dye, and SYPRO Tangerine dye (Molecular Probes, Inc., Eugene, OR) offer sensitivity and ease of use that rival or surpass colloidal colorimetric staining. These stains are characterized by high sensitivity, a 3-log linear dynamic range, and a simple staining protocol, allowing rapid, accurate quantitation of proteins in gels. Staining is reversible and can be followed by staining with Pro-Q® Diamond phosphoprotein stain or Pro-Q Emerald glycoprotein stain for detection of posttranslational modifications. Staining can also be followed by detection with other total protein stains, including SYPRO Ruby stain or colorimetric stains such as Coomassie Blue or silver. A common feature of these fluorescent protein gel stains is a bimodal excitation spectrum, with maxima in the ultraviolet (UV) and visible ranges, resulting in versatility of use on a broad range of imaging platforms.

These dyes bind to proteins in gels noncovalently, through interaction with sodium dodecyl sulfate (SDS) micelles (1–4). The basis of detection is twofold: (1) the dyes preferentially bind to hydrophobic detergent–protein complexes and (2) the transition from an aqueous staining solution to the hydrophobic environment results in a several-hundred-fold fluorescence signal enhancement. Protein quantitation with fluorophores of this type is generally more reliable than that achieved with fluorophores that label primary amines alone (5). Because proteins are not covalently modified with dye molecules and staining is performed postelectrophoretically, no alteration in the migration of proteins during electrophoresis occurs. These fluorescence-based staining methods are also highly compatible with downstream microchemical characterization methods such as Edman-based protein sequencing and peptide mass profiling by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS).

From: *The Proteomics Protocols Handbook*
Edited by: J. M. Walker © Humana Press Inc., Totowa, NJ

Coomassie Fluor Orange, SYPRO Orange, and SYPRO Red dyes are chemically similar. Coomassie Fluor Orange and SYPRO Orange protein gel stains are brighter than SYPRO Red protein gel stain, which shows lower background fluorescence in gels. SYPRO Tangerine stain is distinguished from these products on the basis of staining diluent. In contrast to SYPRO Orange and SYPRO Red protein gel stains, which depend upon dilute acetic acid as a staining diluent, staining with SYPRO Tangerine dye may be performed in a nonfixative solution that permits subsequent electroblotting, electroelution, or zymography (in-gel detection of enzyme activity). Detection sensitivity for SYPRO Orange and SYPRO Red stains has been reported to be as low as 1–2 ng, but these results depend upon electrophoresis in nonstandard detergent conditions (0.05% SDS) with 0.75 mm thick gels (1). With standard conditions (0.1% SDS) detection sensitivity with the SYPRO Orange, Red, and Tangerine dyes and with Coomassie Fluor Orange stain is commonly 4 to 10 ng in a 1 mm thick gel, comparable to rapid silver-staining methods and generally superior to colloidal Coomassie Blue staining methods ([1–6]; C. Hart and T. Steinberg, unpublished data). In contrast to many silver-staining and reverse-staining methods, these dyes do not stain nucleic acids or bacterial lipopolysaccharides to a significant extent (1–4). These protein gel stains are not suitable for staining proteins on blotting membrane or in isoelectric focusing (IEF) gels, and they show reduced sensitivity when staining proteins in 2-D gels.

The following protocols will describe procedures to obtain optimal staining following the preparation and running of SDS-polyacrylamide gels. An outstanding feature of these gel stains is that the staining protocol is truly simple. In general, gels are put immediately into staining solution and washed briefly to reduce background before image acquisition. Detection sensitivity of 2–10 ng per band can be consistently obtained within 1 to 2 h after electrophoresis. Staining is also as stable as it is simple: gels can be stained overnight or over the weekend, allowing the investigator good time flexibility. By contrast, the simplest “easy to use” Coomassie Blue staining protocols require several wash steps before a staining step of at least 1 h, followed by at least hours of destaining to obtain the lowest background. Even so, the signal-to-background ratio with the fluorescent stains remains superior (Fig. 1).

2. Materials

2.1. Equipment

1. Staining dishes: Staining containers can be polypropylene, polystyrene, acrylic, or glass. It is preferable to use tightly covered containers, especially with formulations that contain acetic acid or other volatile organic solvents. The container size should be appropriate for the gel size and anticipated staining volume. A polypropylene container with a tight, leakproof lid, such as a commonly available food container, is an excellent staining dish. Staining can also be done in large, polystyrene weigh boats commonly used in laboratories.
2. Imaging and viewing devices: As a consequence of their bimodal excitation spectra, these protein gel stains are efficiently excited by UV and by visible illumination. Thus, stained gels can be viewed and photographed with a standard laboratory UV transilluminator and Polaroid film, or imaged with CCD camera-based UV transillumination or xenon-arc lamp systems, in conjunction with the proper filters. Their various excitation maxima in the visible spectrum make these dyes suitable for use with a variety of laser-based scanning-imaging systems. Coomassie Fluor Orange stain, SYPRO Orange stain (excitation/emission: approx 470/569 nm) and SYPRO Tangerine stain (490/640 nm) are optimal for argon

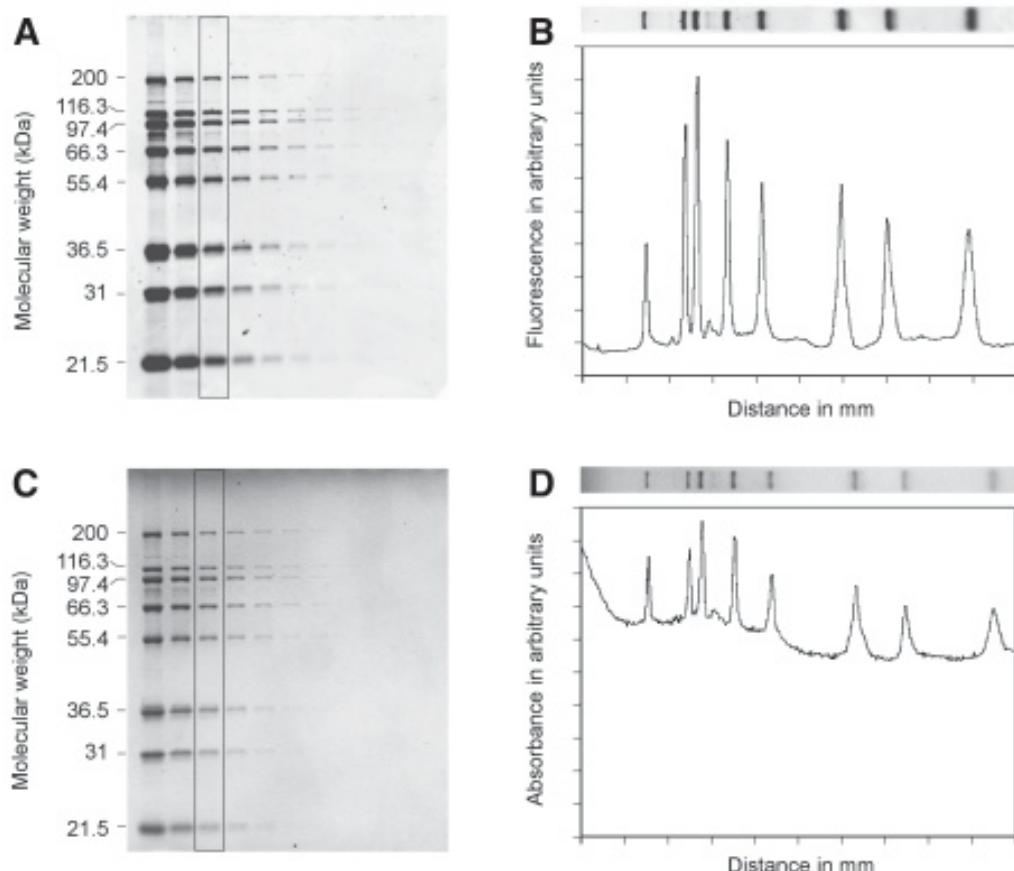


Fig. 1. Comparison of staining obtained with Coomassie Fluor Orange protein gel stain (Molecular Probes, Eugene, OR) and Simply Blue™ Safestain (Invitrogen Corp., Carlsbad, CA), a Coomassie Blue G-250 formulation. A twofold dilution series of molecular-weight markers (Mark 12, Invitrogen) was separated by electrophoresis on 12% Tris-glycine gels (Invitrogen) and stained with Coomassie Fluor Orange stain. (A) gel was placed in 75 mL stain for 85–90 min, washed with two 5- to 10-min changes of water (100 mL), and imaged. (B) Fluorescence profile of lane 3, panel A. (C) The other gel was stained with SimplyBlue Safestain: gel was washed with three 5-min changes of water (100 mL), placed in 25 mL stain for 70 min, washed with two changes of water, 60 min per wash, and imaged. (D) densitometry profile of lane 3, panel C. CCD image acquisition was with the Fluor-S Max (BioRad, Hercules, CA) with the manufacturer's recommended settings for (A) SYPRO Orange, (B) Coomassie Blue. For display purposes, the fluorescent image is presented in negative mode. Grayscales were adjusted with Adobe Photoshop 7.0.1. Profiles were obtained with ImageGauge (Fuji Instruments, Tokyo, Japan). Lane 1: 5 μ L undiluted markers; lanes 2–10, twofold dilution series, 5 μ L per lane. Proteins are, in order of decreasing MW: myosin, β -galactosidase, phosphorylase B, BSA, glutamic dehydrogenase, lactate dehydrogenase, carbonic anhydrase, and soybean trypsin inhibitor.

ion or second-harmonic generation (SHG) laser-based instruments, while SYPRO Red stain (excitation/emission: approx 547/631 nm) is optimal for green He-Ne or Nd:YAG lasers.

3. Film and filters for UV photography: Use Polaroid 667 black-and-white print film and the SYPRO protein gel stain photographic filter (Molecular Probes, Inc., cat. no. S-6656). This filter is optimal for use with all of the total protein fluorescent stains. A Kodak Wratten no. 9 filter is also suitable.

2.2. Reagents

1. Coomassie Fluor Orange protein gel stain (Molecular Probes, Inc., cat. no. C-33250, C-33251) is a 1X, ready-to-use dilute acetic acid solution, containing dye. In appearance it is a clear colorless liquid, and it is stable at room temperature.
2. SYPRO Orange protein gel stain (Molecular Probes, Inc., cat. no. S-06650, S-06651) is provided as 5000X concentrated solutions in dimethylsulfoxide (DMSO), either as a single vial containing 500 μ L of stock solution or as a set of 10 vials, each containing 50 μ L of stock solution. In each case, enough reagent is supplied to prepare a total of 2.5 L of working stain solution, which is sufficient to stain 30 to 50 polyacrylamide minigels. Before opening the vial, warm it to room temperature to avoid water condensation and subsequent precipitation. After thawing completely, briefly centrifuge the vial in a microcentrifuge to deposit the DMSO solution at the bottom of the vial.
3. SYPRO Red protein gel stain (Molecular Probes, Inc., cat. no. S-06653, S-06654) is provided as a 5000X concentrated solution in DMSO, either as a single vial containing 500 μ L of stock solution or as a set of 10 vials, each containing 50 μ L of stock solution. This is handled as for SYPRO Orange dye.
4. SYPRO Tangerine protein gel stain (Molecular Probes, Inc., cat. no. S-12010) is provided as a 5000X concentrated solution in DMSO as a single vial containing 500 μ L of stock solution. This is handled as for SYPRO Orange dye.
5. Deionized water (18 $M\Omega/cm$).
6. Glacial acetic acid: This is used to prepare acetic acid solutions (7.5%) for staining diluent for SYPRO Orange, Red, or Tangerine dyes.
7. Buffer salt solution: 50 mM phosphate buffer, 150 mM NaCl (pH 7.0) for nonfixative staining with SYPRO Tangerine dye.
8. Methanol: This is used to prepare aqueous methanol (30 to 50%) if complete destaining of protein bands is desired.
9. 0.1% Tween-20. This can be used for complete destaining of protein bands.
10. 10% SDS stock solution. This can be used for staining proteins in gels where SDS is not present.

3. Methods

3.1. General Practice

Gel electrophoresis should be performed according to standard procedures (7,8). These stains are also compatible with proprietary electrophoresis gel and buffer systems such as the NuPAGE® System (Invitrogen, Carlsbad, CA). Because the dyes intercalate into detergent micelles, the SDS front at the bottom of the gel stains very heavily, resulting in an extremely bright, broad band that may interfere with imaging. Moreover, a large SDS front combined with a low staining volume may sequester enough dye to reduce staining intensity or significantly increase gel background. It is advantageous to run the SDS front off the gel or to remove the section of gel containing the SDS front. Colored stains and marker dyes, as well as commercially prestained protein markers, may interfere with staining and quench fluorescence.

It is critical that staining be done in clean containers. Residual detergent or dye from previous usage will interfere with proper staining. If a container is to be reused, it

should be cleaned thoroughly with ethyl alcohol, followed by rinsing with deionized water before each use, and it is preferable to have dedicated containers for each type of stain. Although the stains described in this chapter are generally photostable, staining dishes should be covered to protect the fluorescent dyes from photobleaching by ambient light.

Cleanliness and good care of the imaging surfaces is essential. Dust particles, surface scratches, and residual stain will convert a carefully prepared, elegantly stained gel into an object of frustration. After imaging, the surface should be washed with ethyl alcohol followed by a water rinse to prevent accumulation of fluorescent dye on the glass surface. Washing and drying should be with soft, lint- and dust-free cloths or laboratory delicate-task wipers such as Kimwipes®, or an equivalent. It is generally prudent to clean the surface prior to photography or image acquisition. If good quantitation is desired, it is also useful to image the blank surface to assess any unnoticed dye contamination or surface defect.

3.2. Protocols for Coomassie Fluor Orange Protein Gel Stain

3.2.1. Rapid Staining of SDS Gels After Electrophoresis

1. After electrophoresis, combine the gel with the Coomassie Fluor Orange protein gel staining solution (**Subheading 2.2., step 1**) in a clean staining dish (**Subheading 2.1., step 1**). A standard 1-mm minigel requires 50 to 100 mL of stain, or 10 to 20 gel volumes. Staining is at room temperature on a rocker or orbital shaker with gentle agitation—e.g., 50 rpm. Staining can be monitored with a handheld UV light or by placing the staining dish on a UV transilluminator, or with a blue-light transilluminator with an orange filter. Stained bands (100 to 1000 ng protein) are apparent within 10 to 15 min, against an orange background. Progression of staining involves both an increase in signal from the protein bands and a large decrease in background as free SDS diffuses out of the gel. Staining is essentially complete in 45 to 60 min, with continuing improvement (due to decreasing background) over the next hour. Staining can be accelerated by briefly microwaving (*ca.* 1 min) the staining solution containing the gel. The gel can remain in the staining solution for several hours to overnight with no decrease in signal. Gels may be stored in the staining solution for several days to several weeks with only two- to fourfold loss of sensitivity.
2. The gel is removed from stain and washed with two changes of water (50 to 100 mL), 5 min per wash. Loss of signal may occur with longer wash times.
3. The gel is photographed or an image obtained (**Subheading 3.2.3.**).

3.2.2. Staining Native Gels or Fixed Gels

Gels that have been fixed with alcohol to remove SDS, or “native” gels lacking SDS, may be stained by adding SDS to the staining solution to a final concentration of 0.005% and incubating for several hours to overnight. Native gels may also be stained by incubating the gel in 0.1% SDS for several hours and then staining with Coomassie Fluor Orange stain.

3.2.3. Viewing, Photographing, and Imaging Stained Gels

Detection with UV light is best done with a 300-nm transilluminator. Place the gel directly on the transilluminator. When viewed on a UV transilluminator, protein bands stained with Coomassie Fluor Orange dye are bright orange. Do not use plastic wraps or plastic backing, since these contribute significantly to background fluorescence, interfering with detection sensitivity. Use Polaroid 667 black-and-white print film and

the SYPRO protein gel stain photographic filter (Molecular Probes, Inc., cat. no. S-6656). This filter is optimal for use with all of the total protein fluorescent stains. Exposure times vary with the intensity of the illumination source; for an f-stop of 4.5, use 2 to 5 s exposure, which is easily done by multiple 1-s exposures if the camera does not have a timer. The longer exposure times increase sensitivity, detecting bands not visible to the eye. Maximum detection sensitivity is achieved when the gel background in the photograph appears to be very light gray. The choice of photographic filter is important. Standard ethidium bromide filters may be used, but exposure times may have to be lengthened. Some filters are autofluorescent in UV light, giving rise to increased background and diminished detection sensitivity (3). A blue-light transilluminator (e.g., Dark Reader™ transilluminator, Clare Chemical Research, Boulder, CO) is also an excellent excitation source for viewing and photographing proteins stained with Coomassie Fluor Orange dye.

A CCD camera-based image analysis system can gather quantitative information that will allow comparison of fluorescence intensities between different bands or spots, adding to the utility of these stains. Images are best obtained by digitizing at least 1024 × 1024 pixels resolution with 12- or 16-bit grayscale levels per pixel. Filter sets for each system are different, and manufacturer's recommendations should be heeded. For Coomassie Fluor Orange stain, any longpass filter with a cutoff greater than 520 nm is suitable, and good results are obtained with 600-nm bandpass filters.

Detection with laser scanners also gives excellent detection sensitivity. Excitation can be with 473-nm SHG laser, 488-nm argon ion, or 473-nm He-Ne laser-based imaging systems, paired with suitable emission filters suggested by the manufacturers.

3.2.4. Destaining Gels

Gels may be completely destained by washing in 30 to 50% aqueous methanol for 1 h to overnight, or in 0.1% Tween-20 for several hours to overnight.

3.3. Protocols for SYPRO Orange and SYPRO Red Protein Gel Stains

3.3.1. Rapid Staining of SDS Gels After Electrophoresis

1. Prepare the staining solution by diluting the stock dye solutions 1:5000 into dilute acetic acid solution and mixing. We recommend 7.5% acetic acid, but concentrations may range from 2 to 10%. The staining solution will be a clear, colorless liquid and is stable at room temperature.
2. Proceed as for Coomassie Fluor Orange staining steps 1 to 3 (Subheading 3.2.1.). For SYPRO Orange dye, staining can be monitored as with Coomassie Fluor Orange stain. SYPRO Red dye staining can be monitored with UV light.

3.3.2. Staining Proteins During Electrophoresis in SDS-PAGE

SYPRO Orange or SYPRO Red dye may be added directly to the cathode running buffer to stain proteins during electrophoresis. The dye intercalates into the SDS micelles and the SDS-protein complexes, but does not interfere with protein migration. After electrophoresis, the gel is washed to decrease background signal resulting from highly fluorescent SDS-dye complexes in the gel matrix.

1. Dilute the concentrated dye stock directly into the cathode running buffer and mix thoroughly. Detection sensitivity with this method, using a 1:10,000 dilution of stain into cathode buffer followed by one wash in 7.5% acetic acid, has been reported to be four- to

eightfold poorer than with post-electrophoresis staining (1). Detection may be substantially improved, however, by using the dye at 1:3300 dilution, with two 15-min destaining water washes before imaging (C. Hart, unpublished data) or by two 15-min dilute acetic acid washes followed by a 5-min water wash. Destaining can be accelerated by microwaving the gel in the wash solution.

2. Electrophoresis is done by standard procedures.
3. After electrophoresis, the gels are washed with dilute acetic acid or with water.
4. Images may be taken (**Subheading 3.2.4.**) within 30 to 40 min.

By this method colorimetric staining is observed at high protein loads. With SYPRO Orange stain, orange protein bands can be observed during electrophoresis, visible to the eye down to 300–400 ng; SYPRO Red stain gives purplish bands, visible down to approx 200 ng. Fluorescent images taken immediately after electrophoresis with no destaining can reveal protein bands down to 50 ng, but image quality is improved and detection limits are increased 10-fold after washing.

3.3.3. Staining Native Gels or Fixed Gels

Gels that have been fixed with methanol to remove SDS, or native gels, can be stained by the methods described above for Coomassie Fluor Orange stain (**Subheading 3.2.2.**).

3.3.4 Viewing, Photographing, and Imaging Stained Gels

Gels stained with SYPRO Orange dye may be viewed, photographed, and imaged in the same way as with Coomassie Fluor Orange stain (**Subheading 3.2.3.**). When viewed on a UV transilluminator, protein bands with SYPRO Orange stain are bright orange; SYPRO Red dye stained proteins are distinctly red.

Gels stained with SYPRO Red dye require longer exposure times with UV transillumination (3 to 8 s). For UV transilluminators, the SYPRO protein gel stain photographic filter is optimal. For detection in CCD-based systems, the manufacturer's recommendations should be followed. SYPRO Red dye is not suitable for detection with blue light transilluminators. For laser scanners, excitation for SYPRO Red stain is with green He-Ne or Nd:YAG (532 nm) lasers, paired with the manufacturer's suggested emission filter.

3.3.5. Destaining

Gels can be destained by the methods in **Subheading 3.2.4.**

3.4. Protocols With SYPRO Tangerine Protein Gel Stain

Coomassie Fluor Orange, SYPRO Orange, and SYPRO Red are optimal with an acetic acid diluent, making them less suitable for applications involving electroblotting, electroelution, or measuring enzyme activity. SYPRO Tangerine protein gel stain is a versatile stain for detecting proteins separated by SDS-PAGE (4). Staining is performed in a nonfixative solution that permits subsequent electroblotting, electroelution, or detection of enzyme activity. Proteins stained without fixation can be used for further analysis by zymography (in-gel enzyme activity assay), provided SDS does not inactivate the protein of interest. Stained proteins can also be easily eluted from gels and used for further analysis. If protein fixation is preferred, the dye can be used with 7% acetic acid. In this case, however, one should expect slightly higher background staining than with Coomassie Fluor Orange, SYPRO Orange, and SYPRO Red stains.

3.4.1. Nonfixative Staining With SYPRO Tangerine Dye

1. Prepare the staining solution by diluting the stock SYPRO Tangerine reagent (**Subheading 2.2., step 4**) 1:5000 in an appropriate buffer, as detailed below, and mixing. The staining solution will be a clear, very pale yellow-orange liquid.
If the proteins are to be used for electroelution, electroblotting, or zymography, dilute the stock solution into 50 mM phosphate, 150 mM NaCl, pH 7.0. If no fixative is used before or during staining, some diffusion of the protein bands may occur, especially for smaller proteins. Alternatively, use one of a wide range of buffers that are compatible with the stain. These include: formate, pH 4.0; citrate, pH 4.5; acetate, pH 5.0; MES, pH 6.0; imidazole, pH 7.0; HEPES, pH 7.5; Tris acetate, pH 8.0; Tris-HCl, pH 8.5; Tris borate, 20 mM ethylenediaminetetraacetic acid (EDTA), pH 9.0; and bicarbonate, pH 10.0. Buffers should be prepared as 50–100 mM solutions containing 150 mM NaCl. The stock dye solution may also be diluted directly into 150 mM NaCl.
2. Proceed as for Coomassie Fluor Orange stain protocol (**Subheading 3.2.1., steps 1–3**). Staining can be monitored as with Coomassie Fluor Orange stain.
If the gel is to be used for zymography, the poststaining washes should be in buffer compatible with the intended procedure. If the proteins are to be transferred to a blot, incubate the gel in Western blotting buffer containing 0.1% SDS after imaging the gel. The SDS is not absolutely required, but it helps in the transfer of some proteins to the blot.

3.4.2. Photographing and Imaging Stained Gels

Gels stained with SYPRO Tangerine dye may be viewed, photographed, and imaged in the same way as with Coomassie Fluor Orange stain (**Subheading 3.2.3.**). SYPRO Tangerine dye has a relatively broad emission spectrum, so it appears to the eye as orange-red, despite its emission maximum being red-shifted relative to SYPRO Red dye. As a result of the broad emission peak, emission filters suitable for ethidium bromide or SYPRO Red dye can also be used.

3.4.3. Destaining

Gels can be destained by the methods in **Subheading 3.2.4.**

References

1. Steinberg, T., Jones, L., Haugland, R., and Singer, V. (1996a) SYPRO Orange and SYPRO Red protein gel stains: one-step fluorescent staining of denaturing gels for detection of nanogram levels of protein. *Anal. Biochem.* **239**, 223–237.
2. Steinberg, T., Haugland, R., Singer V., and Jones, L. (1996b) Applications of SYPRO Orange and SYPRO Red protein gel stains. *Anal. Biochem.* **239**, 238–245.
3. Steinberg, T., White, H., and Singer, V. (1997). Optimal filter combinations for photographing SYPRO Orange or SYPRO Red dye-stained gels. *Anal. Biochem.* **248**, 168–172.
4. Steinberg, T., Lauber, W., Berggren, K., Kemper, C., Yue, S., and Patton, W. (2000) Fluorescence detection of proteins in SDS-polyacrylamide gels using environmentally benign, non-fixative, saline solution. *Electrophoresis* **21**, 497–508.
5. Patton, W. (2002) Detection technologies in proteome analysis. *J. Chromatog B* **771**, 3–31.
6. Patton, W. and Beechem, J. (2002) Rainbow's end: the quest for multiplexed fluorescence quantitative analysis in proteomics. *Curr. Opin. Chem. Biol.* **6**, 63–69.
7. Laemmli, U. (1970) Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **227**, 680–685.
8. Schagger, H. and Von Jagow, G. (1987) Tricine-sodium dodecyl sulphate-polyacrylamide gel electrophoresis for the separation of proteins in the range from 1 to 100 kDa. *Anal. Biochem.* **166**, 368–379.

Differential In-Gel Electrophoresis in a High-Throughput Environment

Richard I. Somiari, Stephen Russell, Stella B. Somiari, Anthony G. Sullivan, Darrell L. Ellsworth, Henry Brzeski, and Craig D. Shriver

1. Introduction

Proteomics deals with the large-scale analysis of proteins expressed by the genome of an organism. In general, the aim is to search for qualitative and quantitative changes in protein expression that occur as a function of development, disease, environmental insults, or treatment. Because of the complexity of the proteome of most organisms, the complete and rapid mapping of protein expression changes and characterization of posttranslational modifications is challenging and will require an analytical effort that exceeds the technology and throughput that exist in most standard biochemistry laboratories (1). While novel automated methods for rapid separation and analysis of proteins are being evaluated in many laboratories, it is widely accepted that two-dimensional gel electrophoresis (2-DE) is still the benchmark for large-scale separation of complex protein mixtures. The availability of standard reagents and equipment, and development of many modifications of the 2-DE process, have led to remarkable improvements in reproducibility (2,3). However, it is still difficult to fully duplicate the pattern of protein expression using conventional 2-DE methods (4,5), particularly when conducting multiple comparative and quantitative proteomics experiments in a high-throughput research environment where reproducibility and speed are critically important parameters.

The relatively new variant of 2-DE, differential in-gel electrophoresis (DIGE), promises to significantly improve the speed, reproducibility, and sensitivity of 2-DE-based experiments and performance of large-scale comparative proteomics studies using automated and semi-automated platforms. The DIGE concept involves the use of spectrally matched and resolvable fluorescent dyes to differentially label two to three protein samples prior to two-dimensional polyacrylamide gel electrophoresis (PAGE) on a single gel. The 2-DE gel pattern is then visualized by sequential imaging of the gel with a fluorescent scanner. Quantitative analysis of differential protein expression using the 2-D DIGE technology and appropriate software is fast and relatively more accurate than conventional 2-DE because it is based on the relative fluorescent intensities captured from a single gel, and errors attributable to gel-gel variations are nonexistent or minimal.

Two DIGE procedures, “minimal” and “saturation” labeling, will be described in this chapter. In the minimal labeling procedure, two to three protein samples are separately labeled covalently with Cy2, Cy3, or Cy5 fluorescent dyes. The minimal labeling dyes contain an NHS ester active group that binds covalently to the lysine residue in proteins via an amide linkage. The protein-to-dye ratio is deliberately kept high (>95%) to ensure that only a single lysine residue on each protein is labeled in the minimal labeling protocol. The saturation labeling procedure is specifically developed for the analysis of scarce protein samples. The saturation labeling dyes (CyDye DIGE Fluor Cy3 and CyDye DIGE Fluor Cy5) contain maleimide reactive groups that covalently bind to the cysteine residues on proteins via a thioether linkage. The aim is to label all available cysteine residues.

In a high-throughput proteomics research environment, it is important to have a robust tracking system to permit the tracking of each sample, all proteins of interest, and all proteins that will be identified by mass spectrometry. Before proteins quantified by 2-D DIGE are analyzed by mass spectrometry, they must be picked from the gels, destained, digested, and extracted from the gel (6). Many reports have demonstrated that the DIGE technology affords great sensitivity and reproducibility (7–9), and because it is also amenable to automation (1), it is the variant of 2-D electrophoresis that we have found to be the most suitable for large-scale comparative and quantitative proteomics in a high-throughput biomedical research environment. This chapter covers the standard 2-D DIGE process and provides a brief introduction to the analysis of DIGE gels, and preparation of 2-D DIGE gels for automated picking of protein spots using a spot picker.

2. Materials

The materials listed below are required to successfully perform the GE Healthcare Biosciences minimal and saturation labeling 2-D DIGE procedure using matched and spectrally resolved cyanine dyes. Although there are a few reports indicating that other fluorescent dyes—e.g., AlexaTM dyes (Molecular Probes, Inc.)—can be used (5), the GE Healthcare DIGE package represents the most advanced and well-integrated DIGE package that is currently available commercially and that has been evaluated in many laboratories (1,4,7–9).

2.1. Minimal Labeling

1. Reagent-grade water.
2. Standard cell wash buffer: 5 mM magnesium acetate, 10 mM Tris-HCl (pH 8.0). Store at 2–8°C and discard after 1 mo storage (see Note 1).
3. Lysis buffer: 7 M urea, 2 M thiourea, 4% CHAPS, 1% NP-40, 5 mM magnesium acetate, 30 mM Tris-HCl (pH 8.5), at 4°C. Store in aliquots at –20°C.
4. Mechanical homogenizer.
5. Microcentrifuge pestle with motor tool (Kimble/Kontes, Vineland, NJ).
6. pH indicator strips pH 2.0–9.0 (EM Science, Gibbstown, NJ).
7. Sodium hydroxide (NaOH), 50 mM solution.
8. Coomassie Plus protein assay reagent kit (Pierce Chemical Co., Rockford, IL).
9. Albumin Standard concentrate (Pierce Chemical Co., Rockford, IL).
10. Clear bottom 96-well microtiter plates.
11. Microtiter plate reader capable of reading absorbance at 595 nm.

12. 99.8% anhydrous dimethylformamide (DMF). Must be less than 3 mo old from date of opening.
13. Cy2, Cy3, and Cy5 minimal label dyes (GE Healthcare Biosciences, Piscataway, NJ).
14. Lysine (Aldrich Chemical Co., Milwaukee, WI).
15. DeStreak reagent (GE Healthcare Biosciences).
16. Immobilized pH gradient (IPG) buffer (GE Healthcare Biosciences).
17. Rehydration buffer: 7 *M* urea, 2 *M* thiourea, 4% CHAPS, 1% NP-40, 10% isopropanol, 5% glycerol, 1.2% DeStreak reagent, 0.5% IPG buffer, trace amount of bromophenol blue.
18. 2X rehydration buffer: 7 *M* urea, 2 *M* thiourea, 4% CHAPS, 1% NP-40, 10% isopropanol, 5% glycerol, 2.4% DeStreak reagent, 1.0% IPG buffer.
19. Bromophenol blue.
20. Immobiline DryStrips isoelectric focusing (IEF) gel strips (GE Healthcare Biosciences).
21. DryStrip cover fluid (GE Healthcare Biosciences).
22. Strip holders (GE Healthcare Biosciences).
23. IPGphor IEF unit (GE Healthcare Biosciences).
24. Sodium dodecyl sulfate (SDS) equilibration buffer: 6 *M* urea, 30% glycerol, 2% SDS, 50mM Tris-HCl (pH 8.8).
25. Equilibration buffer A: 1X SDS equilibration buffer, 0.5% dithiothreitol.
26. Equilibration buffer B: 1X SDS equilibration buffer, 4.5% iodoacetamide.
27. SDS-PAGE running buffer: 25 mM Tris, 192 mM glycine, 0.2% SDS.
28. Agarose sealing solution: 1X SDS-PAGE running buffer, 0.5% low melting point agarose, small amount of bromophenol blue.
29. 40% Acrylamide/bis solution (GE Healthcare Biosciences).
30. Bind-Silane (GE Healthcare Biosciences).
31. Low-fluorescence glass gel plates (GE Healthcare Biosciences).
32. Gel casting equipment.
33. TEMED (GE Healthcare Biosciences).
34. Ammonium persulfate (GE Healthcare Biosciences).
35. 1.5 *M* Tris-HCl, pH 8.8.
36. SDS-PAGE equipment (GE Healthcare Biosciences).
37. Gel fix solution: 10% methanol, 7% acetic acid.
38. Deep Purple Total Protein Stain (GE Healthcare Biosciences).
39. Development solution: 0.1% ammonium hydroxide.
40. Stabilization solution: 0 .75% acetic acid.
41. DeconTM (branded ContradTM in the US).

2.2. Saturation Labeling

1. Reagent-grade water.
2. Cell lysis buffer: 7 *M* urea, 2 *M* thiourea, 4% (w/v) CHAPS, 30 mM Tris-HCl (pH to 8.0 with 1.0 *M* HCl). Aliquot and store at -15°C to -30°C. Discard after 3 mo.
3. Tris-(2-carboxyethyl) phosphine hydrochloride (TCEP) (Pierce Chemical Co., Rockland, IL).
4. Cy3 and Cy5 maleimide dye (GE Healthcare Biosciences).
5. IPG buffer (GE Healthcare Biosciences).
6. 1X rehydration buffer: 7 *M* urea, 2 *M* thiourea, 4% CHAPS, 1% NP-40, 10% isopropanol, 5% glycerol, 130 mM dithiothreitol (DTT), 1.0% IPG buffer, trace amount of bromophenol blue. Add IPG buffer and DTT immediately before use.
7. 2X rehydration buffer: 7 *M* urea, 2 *M* thiourea, 4% CHAPS, 1% NP-40, 10% isopropanol, 5% glycerol, 130 mM DTT, 2.0% IPG buffer.
8. Bromophenol blue.

9. Immobiline DryStrip IEF gel strips (GE Healthcare Biosciences).
10. DryStrip cover fluid (GE Healthcare Biosciences).
11. Strip holders (GE Healthcare Biosciences).
12. IPGphor IEF unit (GE Healthcare Biosciences).
13. SDS equilibration buffer: 6 M urea, 30% glycerol, 2% SDS, 50 mM Tris-HCl (pH 8.8).
14. Equilibration buffer A: 1X SDS equilibration buffer, 0.5% DTT.
15. SDS-PAGE running buffer: 25 mM Tris, 192 mM glycine, 0.2% SDS.
16. Agarose sealing solution: 1X SDS-PAGE running buffer, 0.5% low-melting-point agarose, small amount of bromophenol blue.
17. 40% Acrylamide/bis solution (GE Healthcare Biosciences).
18. Bind-Silane (GE Healthcare Biosciences).
19. Low-fluorescence glass gel plates (GE Healthcare Biosciences).
20. Gel casting equipment.
21. TEMED.
22. 10% Solution of ammonium persulfate.
23. 1.5 M Tris-HCl, pH 8.8.
24. SDS-PAGE equipment.
25. Gel fix solution: 10% methanol, 7% acetic acid.
26. Deep Purple Total Protein Stain (GE Healthcare Biosciences).
27. Development solution: 0.1% ammonium hydroxide.
28. Stabilization solution: 0.75% acetic acid.
29. Decon.

2.3. Quantitation of Protein Concentration

30. Coomassie Plus protein assay reagent or equivalent.
31. Spectrophotometer.

2.4. Isoelectric Focusing

1. Rehydration buffer: 7 M urea, 2 M thiourea, 4% (w/v) CHAPS, 1% (w/v) PharmalyteTM, broad range pH 3.0–10.0, 13 mM DTT. Prepare fresh by adding DTT (1 mg/mL) and Pharmalytes to 1X sample buffer. Use immediately and discard any unused material.
2. IPG strips (ImmobilineTM DryStrip IPG strips, GE Healthcare Biosciences).
3. Reswelling tray (ImmobilineTM DryStrip reswelling tray, GE Healthcare Biosciences).
4. Cup loading strip holders (EttanTM IPGphorTM cup loading strip holder, GE Healthcare Biosciences).
5. DryStrip cover fluid (GE Healthcare Biosciences).
6. IEF and electrode strips (Ettan IPGphor IEF system, GE Healthcare Biosciences).
7. IEF electrode strips (GE Healthcare Biosciences).

2.5. SDS-PAGE Separation

1. Equilibration tubes or 9-cm diameter Petri dishes.
2. Equilibration buffer: 6 M urea, 0.1 M Tris (pH 8.0), 30% (v/v) glycerol, 2% (w/v) SDS, 0.5% (w/v) DTT. A DTT-free stock can be prepared and stored at room temperature for 6 mo. DTT is then added immediately before use and any unused material discarded.
3. 12.5% acrylamide gel: 281 mL acrylamide/bis 40% (v/v), 225 mL Tris (1.5 M pH 8.8), 9 mL 10% (w/v) SDS, 9 mL 10% (w/v) ammonium persulfate (freshly prepared on day of use), 1.24 mL 10% (v/v) TEMED. Make up to 900 mL with distilled water. 900 mL is sufficient to prepare 14 24 cm × 24 cm gels.
4. Low fluorescence glass plates (GE Healthcare Biosciences).

5. Gel caster (Ettan DALT gel caster, GE Healthcare Biosciences).
6. Displacement solution: 375 mM Tris-HCl (pH 8.8), 50% glycerol, bromophenol blue (2 mg/100 mL). Prepare fresh and use immediately. Discard unused portion.
7. Agarose overlay solution: 0.5% low-melting-point (LMP) agarose preparation, 0.1% (w/v) bromophenol blue in 1X SDS electrophoresis running buffer (see **item 9**). Stable for 1 mo at room temperature.
8. Water-saturated butanol: Add 50 mL water to 50 mL butan-2-ol until two layers are visible. Stable for 6 mo at room temperature.
9. 1X SDS electrophoresis running buffer: 25 mM Tris, 192 mM glycine, 0.2% (w/v) SDS. Store at room temperature.
10. Electrophoresis system: e.g., Hoefer™ SE600 Ruby Gel system, Ettan DALTtwelve gel system, Ettan DALTsix gel system, or equivalent system.
11. Bind-Silane solution: 100 μ L PlusOne Bind-Silane (code 17-1330-01). Add to 80 mL ethanol, 2 mL glacial acetic acid, and 18 mL water (GE Healthcare Biosciences).

2.6. Imaging and Image Analysis

1. Fluorescent gel imager (e.g., Typhoon™ 9000 series Variable Mode Imager, GE Healthcare Biosciences, or equivalent).
2. Image analysis software (DeCyder™ Differential Analysis Software, GE Healthcare Biosciences).

2.7. Post-DIGE Staining for Spot Picking and Mass Spectrometric Analysis

1. SYPRO® Ruby stain (Molecular Probes, Inc.).
2. SYPRO Ruby gel fixation solution: 30% (v/v) methanol and 7.5% (v/v) acetic acid in distilled water. Store at room temperature.
3. SYPRO Ruby gel destain solution: 10% (v/v) methanol and 6% (v/v) acetic acid in distilled water. Store at room temperature.

3. Methods

The methods described below outline the 2-D DIGE protocols for (1) the minimal labeling, (2) saturation labeling, (3) DIGE image analysis, and (4) protein spot processing; these have been successfully used in the authors' laboratory for 2-D DIGE analysis of protein isolated from human tissue, human cell lines, and cells procured by laser microdissection. Pay particular attention to the protocol because some of the methods used in preparation of protein samples for classical 2-D electrophoresis may not be compatible with DIGE. Also, note that there are some differences between the minimal and saturation labeling procedures.

3.1. Experimental Design

An experimental design needs to be carefully developed before carrying out 2-D DIGE experiments. The power of 2-D DIGE is fully realized when an in-gel internal standard is included. The in-gel standard is created by combining aliquots of all the biological samples in the experiment and labeling this mixture with one of the CyDye fluores (usually Cy2). This in-gel standard is then run on every single gel along with each individual sample. An example of the experimental design we have used successfully for analysis of human breast biopsies is shown in **Table 1**. The six-gel and three-gel experimental designs shown in **Table 1** will permit the comparison of protein spot

Table 1
Experimental Design for Two-Color (A) and Three-Color (B) Two-Dimensional Differential In-Gel Electrophoresis

A. Two-color analysis of protein from six samples (6-gel experiment)		
Gel	CyDyes	
Number	Cy3 (test)	Cy2 (Internal standard)
1	50 µg Sample 1	50 µg (8.3 µg each of Sample 1–6)
2	50 µg Sample 2	50 µg (8.3 µg each of Sample 1–6)
3	50 µg Sample 3	50 µg (8.3 µg each of Sample 1–6)
4	50 µg Sample 4	50 µg (8.3 µg each of Sample 1–6)
5	50 µg Sample 5	50 µg (8.3 µg each of Sample 1–6)
6	50 µg Sample 6	50 µg (8.3 µg each of Sample 1–6)

B. Three-color analysis of protein from six samples (3-gel experiment)*			
Gel	CyDyes		
Number	Cy3	Cy5	Cy2
1	50 µg Sample 1	50 µg Sample 2	50 µg (8.3 µg each of Sample 1–6)
2	50 µg Sample 3	50 µg Sample 4	50 µg (8.3 µg each of Sample 1–6)
3	50 µg Sample 5	50 µg Sample 6	50 µg (8.3 µg each of Sample 1–6)

*Using three colors reduces the number of gels required by 50%.

volumes across a range of samples and gels. The example given will allow (1) intra-gel co-detection of sample and internal standard protein spots and (2) inter-gel matching of internal standard samples across all the gels in the experiment.

3.2. Sample Preparation for 2-D DIGE

3.2.1. Cells

1. Harvest cells by centrifugation at 4°C in a suitable centrifuge.
2. Pour off growth medium without disturbing the cell pellet.
3. Wash cells by re-suspending the pellet in 1 mL of standard cell wash buffer in a microfuge.
4. Pellet in a benchtop centrifuge at 12,000g for 4 min at 4°C.
5. Carefully remove and discard the supernatant.
6. Re-suspend cell pellet in 1 mL of standard cell wash buffer.
7. Repeat **steps 4–6** three times.
8. Make sure that all standard cell wash buffer is removed after washing of cells.
9. Re-suspend the washed cell pellet in 500 µL of standard lysis buffer.
10. Homogenize with a mechanical homogenizer and leave on ice for 30 min, vortexing three to four times during this incubation. Take care not to heat the sample during homogenization.
11. Centrifuge the cell homogenate at 15,000g for 5 min at 4°C.
12. Re-homogenize the pellet using a microcentrifuge tube pestle with motor.
13. Centrifuge the cell homogenate at 15,000g for 10 min at 4°C.
14. Carefully draw off the supernatant with a Pasteur or adjustable pipet and place in a clean tube.

15. Check to insure the pH of the lysate is between 8.0 and 9.0 by spotting 1 μ L of lysate on the indicator pads of the pH indicator strips (there are three pads). If the pH is too low, it can be raised by careful addition of 50 mM NaOH.
16. Place protein extract on ice and use immediately or store at -20°C to -80°C until needed.

3.2.2. Tissue

1. Place frozen tissue in a container big enough for the tissue and suitable for homogenization using a mechanical homogenizer.
2. Add approx 3 vols of ice-cold lysis buffer.
3. Rapidly homogenize for 3–5 s using a suitable laboratory homogenizer, set on high speed (see Note 2).
4. Place homogenate on ice immediately after homogenization. Step 2 can be repeated up to three times. The generator on the homogenizer should be of sufficient size to handle the piece(s) of tissue.
5. Transfer the tissue homogenate to 1.5-mL microcentrifuge tube(s) and incubate on ice for 30 min. Vortex the homogenate three to four times during this incubation period.
6. Centrifuge the tissue homogenate at 15,000g for 5 min at 4°C .
7. Re-homogenize the pellet using a motorized microcentrifuge tube pestle.
8. Centrifuge the tissue homogenate at 15,000g for 10 min at 4°C .
9. Carefully draw off the supernatant and place in a clean tube.
10. Check to insure the pH of the lysate is between 8.0 and 9.0 by spotting 1 μ L of lysate on the indicator pads pH indicator strips (there are three pads). If the pH is too low, it can be raised by careful additions of 50 mM NaOH.

3.3. Quantitation of Total Protein Concentration

It is absolutely essential that equal total protein concentrations be used in DIGE experiments. The method chosen for the determination of total protein concentration must be sensitive, reproducible, and compatible with 2-D DIGE buffer systems. The method must be compatible with detergents and thiourea, if these are present. The method described below is used routinely in the authors' laboratory for determining the total protein concentration of samples before 2-D DIGE.

1. Bring the Coomassie Plus dye reagent to room temperature and mix by inverting the tube several times.
2. Make dilutions of human serum albumin standard to obtain known concentrations (e.g., 0.125, 0.25, 0.5, 1.0, and 1.5 mg/mL). Include lysis buffer in the standards so the final concentration of lysis buffer in each assay well is identical.
3. Pipet 10 μ L of the standards into microtiter plate wells, in duplicate.
4. For blank wells, pipet water with identical concentration of lysis buffer included. Prepare duplicate wells for the blank.
5. For sample wells add an amount of the lysate to duplicate wells; 1–2 μ L should be enough to be in the range of the standard curve. Add enough water to the sample wells to bring the volume up to 10 μ L.
6. Pipet into all wells 300 μ L of the Coomassie Plus dye reagent and incubate at room temperature for 5 min.
7. Read the absorbance at 595 nm in a plate reader and calculate the total protein concentration based on the albumin standard curve.

Table 2
Examples of Dilutions of CyDye Used for Differential In-Gel Electrophoresis

Vol. of stock solution (μL)	Vol. of DMF (μL)	Total vol. (μL)	Conc. CyDye (pmol/μL)
1	4	5	200
2	3	5	400
2	2	4	500
1	—	1	1000

3.4. Differential Labeling of Protein Samples With CyDye DIGE Fluor

3.4.1. Minimal Labeling Protocol

The appropriate amount of CyDye fluor required for optimal labeling of specific protein samples will have to be determined individually. In general, the recommended ratio of fluor Cydye to protein is 400 pmol/50 μg. Examples of the Cydye concentrations that have been used are shown in **Table 2**.

The protocol described below has been found adequate for proteins isolated from human cell lines, biopsy specimens, and whole tissue. It has also been successfully used to study proteins isolated from cells procured by laser-assisted microdissection.

1. Re-suspend the Cydye as recommended by the manufacturer and dilute the stock Cydye 2.5 times with DMF. This makes a 400pmole/μL concentration of Cydye. Cydye fluor working solution is stable for only 1 wk at –20°C.
2. Transfer 50 μg of protein lysate to a clean microcentrifuge tube, and add 1 μL of diluted Cydye (400 pmol) (*see Note 3*).
3. Place tube on ice in an ice bucket, cover from light, and incubate for 30 min in the dark.
4. Add 10 mM lysine to stop the reaction. For every 1 μL of dye used, add 1 μL of 10 mM lysine.
5. Vortex, briefly centrifuge, and incubate the tube on ice, protected from light, for 10 min.
6. Determine the total volume in the tube and add an equal volume of 2X rehydration buffer. Vortex to mix, and centrifuge briefly.
7. Use immediately or store at –70°C, in the dark. Samples can be stored for 3 mo.

3.4.2. Saturation Labeling Protocol

The appropriate amount of CyDye fluor required for optimal labeling of specific protein samples must be determined for each batch of proteins. The protocol described below has been found adequate for proteins isolated from human cell lines, biopsy specimens, and whole-tissue samples. It has also been successfully used to study proteins isolated from cells procured by laser-assisted microdissection. As little as 5 μg of total protein can be labeled and analyzed.

1. Re-suspend the Cydye for saturation labeling as recommended by the manufacturer (GE Healthcare Inc.).
2. Transfer as little as 5 μg of protein lysate to a clean microcentrifuge tube, and add 1 μL of 2 mM TCEP (*see Note 4*).
3. Mix well using a pipet.
4. Briefly centrifuge tube and incubate at 37°C for 1 h. Protect the tube from light.
5. Add the Cydye to the tube. The amount of dye must be added at a 1:2 ratio to the TCEP (e.g., use 2 μL of Cydye when 1 μL of TCEP is used). Mix well using a pipet.

6. Centrifuge the tube briefly and incubate at 37°C. Protect the tube from light.
7. Determine the volume in the tube and stop the reaction by adding an equal volume of 2X rehydration buffer; mix well with a pipet.
8. The labeled samples can be used immediately or stored at -70°C, in the dark for up to 3 mo.

3.5. 2-D DIGE

3.5.1. IEF of CyDye Labeled Proteins—First Dimension

2-D DIGE is different from the standard 2D-PAGE because you can mix up to three differentially labeled protein samples before running on a single first- and second-dimension gel. The method described is for performing IEF with IPG strips using the Ettan IPGphor IEF system. Rehydration of the IPG strip can be done in the presence or absence of the labeled protein.

1. Choose an IEF strip length and range that is appropriate for the sample to be analyzed (see **Note 5**).
2. Combine the labeled samples in one tube. Samples that are to be combined must be labeled with different CyDye fluors.
3. Perform IEF using the recommended protocol for the chosen IEF strip.

3.5.2. PAGE of CyDye Labeled Proteins—Second Dimension

To identify and pick proteins of interest from 2-D DIGE gels using automatic spot pickers, e.g., Ettan Spot Picker (GE Healthcare Biosciences), the gels must be cast with two reference markers attached to the low-fluorescent glass plate, and the gel must be bound to the glass plate with Bind-Silane to prevent shifting and deformation during picking. Following the steps below carefully will enhance reproducibility. Focused strip must be equilibrated prior to running in the second dimension.

3.5.2.1. PREPARATION OF DIGE SECOND-DIMENSION GELS FOR AUTOMATED SPOT PICKING

1. Clean low-fluorescence glass plates thoroughly and wash each plate in 1% Decon with a sponge to completely remove residual gel.
2. Soak plates in 1% Decon overnight and wash with a sponge.
3. Rinse plate with distilled water and leave to soak in 1% HCl (v/v) for 1 h.
4. Wash plate in 1% Decon and rinse with Milli-Q water.
5. Dry plates with lint-free tissue and protect from dust.
6. Coat the surface of the plate with 2–4 mL of Bind-Silane working solution and wipe with a lint-free tissue until dry.
7. Cover plate with a lint-free tissue and leave on bench for 1–1.5 h to dry (not leaving Bind Silane to dry before casting could create problems, because BindSilane will evaporate off the treated plate, coat the facing plate, and cause the gel to stick to the coating plate when it sets).
8. Attach reference markers on the treated glass plates. The markers should be placed where they will not interfere with the resolved proteins in the gel, and the spacers should not cover the markers.
9. Pour gels and allow to solidify before use.

3.5.2.2. PERFORMING SECOND-DIMENSION DIGE

1. Equilibrate strips in equilibration buffer A for 15 min and in equilibration buffer B for 15 min, with gentle agitation during both steps. Note: If the focused strip contains samples that were labeled with saturation dye, they should be equilibrated for 15 min in equilibration buffer A only.

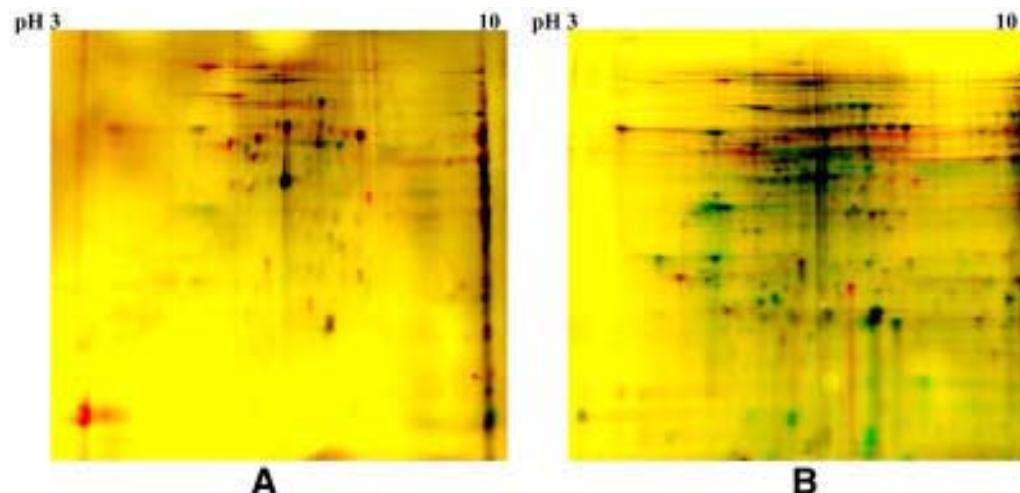


Fig. 1. Representative two-dimensional difference gel electrophoresis (2D-DIGE) images captured with the Typhoon 9400 Variable Mode Imager. The gel images show the signals generated following the labeling and analysis of 100 μ g of total protein isolated from breast biopsies using (A) the 2D-DIGE “minimal” labeling protocol or (B) the 2D-DIGE “saturation” labeling protocol. Note that more candidate protein spots are visible on gel “B.”

2. Carefully place equilibrated strips on top of the second-dimension gel and cover strip with melted agarose sealing solution.
3. Perform electrophoresis overnight (approx 16 h at 15°C and 2 W/gel). The best resolution in the second dimension is achieved if electrophoresis is performed overnight.

3.6. Image Capture, Spot Matching, and Spot Processing

The separated proteins are visualized by imaging on a fluorescence imager. In our high-throughput proteomics laboratory, we use the Typhoon 9400 Variable Mode digital imager for capturing of gel images after 2-D DIGE (Fig. 1). The protein spots are matched with the DIA and BVA modules of DeCyder, the gels are stained with SYPRO Ruby, a pick list is generated, and the protein spots are picked and processed using spot pickers and digesters or a fully automated spot-handling workstation (1). The use of DeCyder software, the Ettan spot picker, Ettan Digester (GE Healthcare Biosciences), or Ettan automated spot-handling workstation are beyond the scope of this protocol. Only brief overviews are presented. Detailed protocols can be found in the manuals accompanying each product.

3.6.1. Scanning of 2-D DIGE Gels

1. Carefully transfer gel to a suitable fluorescent scanner, e.g., the Typhoon Variable Mode Imager (GE Healthcare Inc.). Protect from light to avoid bleaching of dyes.
2. Obtain two scans of each gel (e.g., Cy2 and Cy3) if a two-color experiment was performed, and three scans of each gel (Cy2, Cy3, and Cy5) if a three-color experiment was performed. The excitation/emission wavelengths for Cy2, Cy3, and Cy5 are 488 nm/520 nm, 532 nm/580 nm, and 633 nm/670 nm, respectively. Use the protocol recommended for the scanner to position, select resolution (pixel size), and set focal plane.
3. Process the acquired image with image analysis software, e.g., DeCyder.

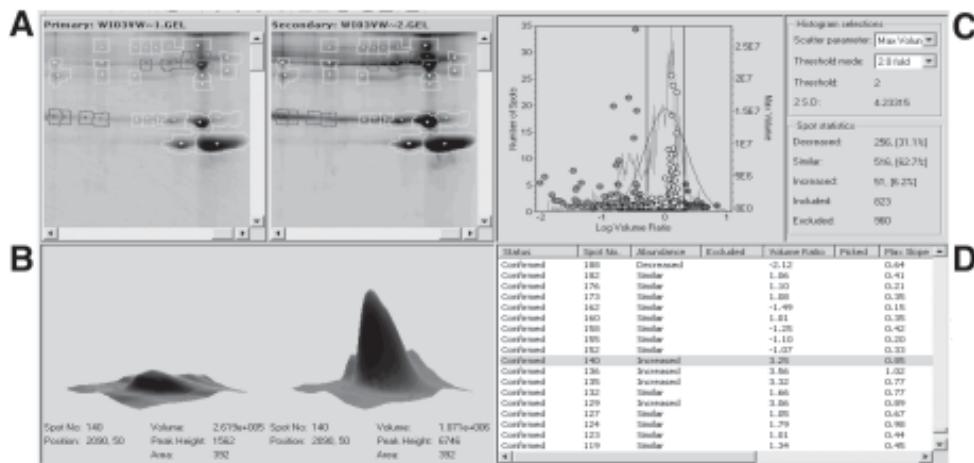


Fig. 2. Screen shot showing the DeCyder Difference In gel Analysis (DIA) workspace divided into four windows. (A) Image view: Cy3 (*primary window*) and Cy5 (*secondary window*) images generated by scanning a single DIGE gel at two wavelengths, (B) 3-D view: Simulated 3-D representation of the gel region shown in panel "A." (C) Histogram view: Shows the graphical representation of the data associated with the spot population, parameters used for generation of histogram and spots statistics window. In this example, the spots within the window defined by the two vertical lines sandwiching the peak (*twofold difference*) occur at similar levels between sample 1 (*primary*) and sample 2 (*secondary*), spots on the left of the window are more abundant in sample 1 and spots on the right of the window are more abundant in sample 2. Out of the 823-candidate proteins considered 516 (62.7%) were similar and 51 (6.2%) occurred at higher levels in sample 2. (D) Table view: Shows the tabulated data associated with the selected spot shown in panel "A" and "B."

3.6.2. Spot Matching

1. Use software capable of processing multiple 2-D images from a single and multiple gels. The most robust software currently available for DIGE is DeCyder. The co-detection algorithms in DeCyder permit comparison of the protein abundance of each sample to the reference, and generation of a 3-D image (Fig. 2).
2. Using the software, generate the ratio of (Cy3:Cy2) and (Cy5:Cy2) from all the gels, so that cross-sample comparison of protein abundance between samples on different gels can be performed.
3. Using the BVA module in DeCyder, automatically match the position of each protein spot on all the gels to a master gel and plot the relative abundance of each protein against the normalized internal standard (Fig. 3).

3.6.3. Staining With SYPRO Ruby, Generating Picking List, Picking and Processing Protein Spots

The picking and identification of protein spots by mass spectrometry after 2-D DIGE requires assigning of proteins for picking on the analyzed DIGE gels, spot detection on the preparative gel (gel prepared for picking and stained with SYPRO Ruby), identification of reference markers on gels, and exporting the picking list to the spot picker. All spots detected can be assigned for picking, or spots can be selected for picking based on the experimental design or specific properties of each spot.

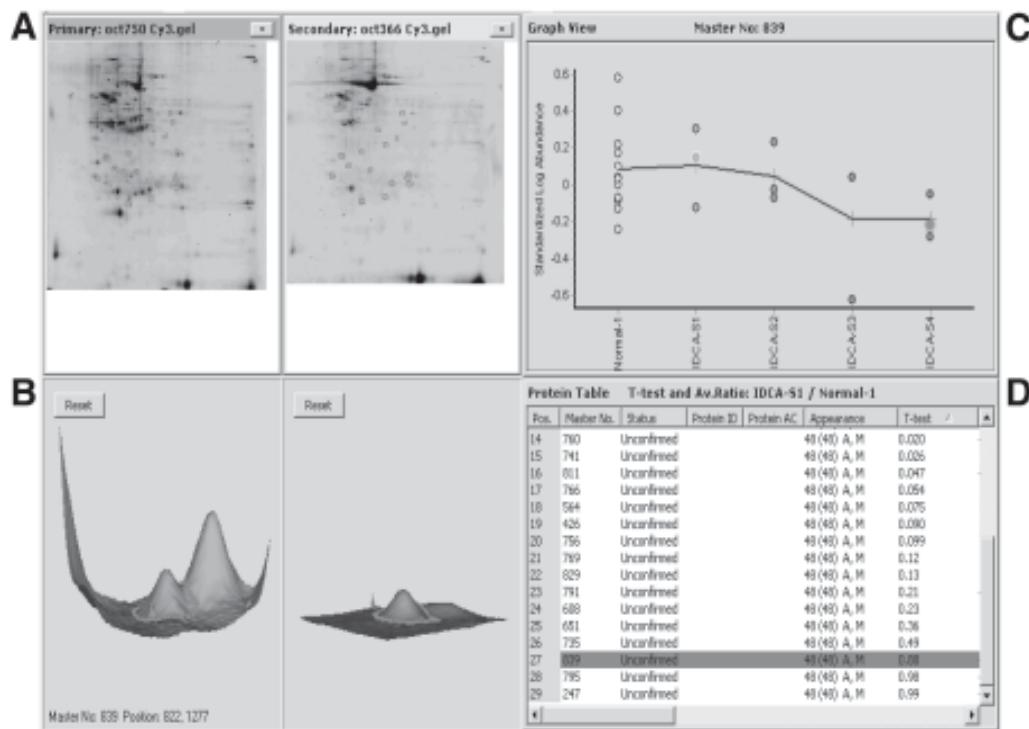


Fig. 3. Screen shot showing the DeCyder Biological Variation Analysis (BVA) workspace, divided into four windows. (A) Image view: Cy3 (primary window) and Cy5 (secondary window) images generated by scanning a single DIGE gel at two wavelengths, (B) 3-D view: Simulated 3-D representation of the gel region shown in panel "A." (C) Graphical view: Shows the graphical representation of the data associated with spot population. In this example, the relative abundance of protein number 839 in "normal" breast tissue and stage 1 to stage 4 IDCA of the breast is shown. In this example, there were twelve normal samples and three each for stage 1–4 IDCA. The line trace is the mean. (D) Table view: Shows the tabulated data associated with the selected spot shown in panel "A" and "B."

3.6.3.1. SYPRO RUBY STAINING OF 2-D DIGE GEELS

To pick and identify proteins after 2-D DIGE, the gels must be stained with a fluorescent dye like SYPRO Ruby to ascertain that the majority of the unlabeled spots are detected and picked. All staining should be done in the containers suitable for SYPRO Ruby. Gels must be protected from light during the entire process to prevent bleaching of the dye.

1. Remove the top glass plate. The top is the plate that was not treated with Bind-Silane.
2. Place the DIGE gel, still attached to the bottom glass plate, into a suitable container (e.g., polypropylene, polycarbonate, or polyvinyl chloride tray).
3. Add SYPRO Ruby gel fix solution and incubate for at least 2 h on a shaking platform.
4. Pour off gel fixing solution.
5. Add SYPRO Ruby stain. Add enough stain to completely cover gel, protect from light, and incubate for at least 5 h or overnight with gentle shaking.
6. Pour off stain and wash gel with SYPRO Ruby gel destain solution for 1–2 h.

7. Briefly rinse gel with reagent-grade water.
8. Dry the back of the glass plate.
9. Place gel, glass side down, onto a clean, dust-free surface.
10. Wet the bottom edge of the gel with distilled water.
11. Place a clean, low-fluorescence glass plate on the gel, taking care not to damage gel or form bubbles.
12. Pick the gel up and drain excess water carefully.
13. Dry the outside of the glass plates.
14. Image the gel with the appropriate filter set and exposure times.
15. After scanning, remove the top plate and store the gel in the gel fixing solution.

3.6.3.2. GENERATING PICKING LIST

Suitable software is needed for assigning of spots for picking. In our laboratory, we use the DIA and BVA modules of DeCyder for 2-D DIGE gel analysis and generation of the pick list. The DIA module is used for assigning spots for picking in small-scale experiments utilizing two samples. All spots can be picked by selecting a “Pick All” check box in DeCyder DIA workspace. The BVA module of DeCyder is used for assigning protein spots for picking based on the statistical information generated from the BVA workspace.

1. Upload a pair of analysis maps into the DIA module of DeCyder.
2. Identify the reference markers on the preparative gel.
3. Select the “Pick All” check box to pick all protein spots detected, or select spots for picking using the “Pick Filter” function.
4. The protein spots showing differences in expression can be picked on the basis of their physical properties—e.g., area, volume, and peak height.
5. Assign the analysis gel as the master gel, and then assign this master gel as the pick spot map.
6. From the spot map table select the preparative gel as the primary image and generate the picking list.
7. Transfer the picking list to the preparative gel.
8. Transfer the list to the spot picker and pick spots into microtiter plates. Use the protocol recommended for the spot picker and spot digester.

3.6.3.3. PROCESSING PROTEIN SPOTS

After identification of all protein spots of interest, the spots have to be excised from the gel manually or automatically with robotic spot pickers. In our laboratory, we use the Ettan spot picker for spot picking and the Ettan Digester for spot digestion, or the Spot Handling Workstation for fully automated spot picking, digestion, and matrix-assisted laser desorption/ionization (MALDI) target slide preparation (1). The procedure must be performed in a clean environment to prevent the entry of contaminating proteins, particularly keratin, into the protein identification workflow process at this stage (see **Note 6**).

4. Notes

1. A cell wash buffer should not contain primary amines, because primary amines such as ampholytes will compete with the proteins for fluors, thereby reducing the total number of proteins labeled.
2. Protein breakdown must be prevented. Homogenize the tissue immediately in lysis buffer and keep tube on ice all the time. When using a motorized homogenizer, make sure that the sample is not heated during the process by placing on ice intermittently.

Table 3
Recommended Amounts of TCEP and Dye Required for Labeling Optimization*

Gel	2 mM TCEP (μL)	TCEP (nmol)	2 mM Dye (μL)	Dye (nmol)
1	0.5	1	1	2
2	0.75	1.5	1.5	3
3	1	2	2	4
4	1.25	2.5	2.5	5
5	1.5	3	3	6
6	2	4	4	8

*Reproduced from the GE Healthcare saturation labeling protocol handbook.

3. The recommended protein concentration of the sample is between 5 and 10 mg/mL; however, samples containing 1mg/mL have been successfully labeled and used for 2-D DIGE.
4. For the best results, the amount of TCEP and CyDye DIGE used needs to be determined and optimized for each protein type being analyzed, or when a nonstandard cell lysis buffer is to be used. The recommended amounts are shown in **Table 3**.
5. Note that the IEF parameters may be different for different sample types and different batches of the same sample type.
6. Although keratins are a major contaminant, and attempts should be made to prevent their introduction into the workflow process, note that cytoskeletal keratins (especially the Type 1, e.g., the 40 kDa KRT19, also known as CK19, an epithelial biomarker) show differential expression as a function of disease. Many other keratins are also within the resolving limits of 2-D PAGE, and so when detected may not be a contaminant. Caution should, however, be used when interpreting results if it is observed that a batch of test and reference samples processed in the same way and separated in the same run all show elevated and/or comparable levels of keratin.

Acknowledgments

We acknowledge the research funds and support received from the US Department of Defense and the Henry Jackson Foundation for the Advancement of Military Medicine, Rockville, MD, for the Clinical Breast Care Project.

References

1. Somiari, R. I., Sullivan, A., Russell, S., et al. (2003) High throughput proteomic analysis of infiltrating ductal carcinoma of the breast. *Proteomics* **10** (3), 1863–1873.
2. Bjellqvist, B., Sanchez, J. C., Pasquali, C., et al. (1993) Micropreparative two-dimensional electrophoresis allowing the separation of samples containing milligram amounts of proteins. *Electrophoresis* **14**, 1375–1378.
3. Rabilloud, T. (1994) Two-dimensional electrophoresis of basic proteins with equilibrium isoelectric focusing in carrier ampholyte-pH gradients *Electrophoresis* **15**, 278–282.
4. Zhou, G., Li, H., DeCamp, D., et al. (2002) 2D differential in-gel electrophoresis for the identification of esophageal scans cell cancer-specific protein markers. *Mol. Cell. Proteomics* **1.2**, 117–124
5. Von Eggeling, F., Gawriljuk, A., Fiedler, W., et al. (2001) Fluorescent dual colour 2D-protein gel electrophoresis for rapid detection of differences in protein pattern with standard image analysis software. *Inter. J. Mol. Med.* **8**, 373–377.

6. Westermeier, R. and Naven, T. (2002) Proteomics in practice: A laboratory manual of protein analysis. Wiley-VCH Verlag-GmbH, Weinheim, p. 316.
7. Unlu, M., Morgan, M. E., and Minden, J. S. (1997) Difference gel electrophoresis: a single gel method for detecting changes in protein extracts. *Electrophoresis* **18**, 2071–2077.
8. Tonge, R., Shaw, J., Middleton, B., et al. (2001) Validation and development of fluorescence two-dimensional differential gel electrophoresis proteomics technology. *Proteomics* **1**(3), 377–396.
9. Gharbi, S., Gaffney, P., Yang, A., et al. (2002) Evaluation of two-dimensional differential gel electrophoresis for proteomic expression analysis of a model breast cancer cell system. *Mol. Cell. Proteomics* **1**(2), 91–98.

Statistical Analysis of 2-D Gel Patterns

Françoise Seillier-Moiseiwitsch

1. Introduction

This chapter reviews the analytical methods implemented in software packages such as MELANIE (1–3) and HERMeS (4–8). Let $I(x,y)$ denote the two-dimensional image, and by convention, the larger* $I(x,y)$ is, the darker the pixel is.

2. State-of-the-Art Analytical Methods

2.1. Filtering Gel Images

To reduce the high-frequency background noise, the signal is extracted by applying a smoothing filter. The most popular filters are

1. *Gaussian smoothing*, which convolves the image with the operator

$$\frac{1}{16} \begin{pmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{pmatrix},$$

2. *diffusion smoothing*, that is,

$$\begin{aligned} I^{(t+1)}(x,y) &= \frac{1}{2} (I^{(t+1)}(x-1, y) + I^{(t)}(x+1, y)) \\ I^{(t+2)}(x,y) &= \frac{1}{2} (I^{(t+1)}(x-1, y) + I^{(t+2)}(x+1, y)) \\ I^{(t+3)}(x,y) &= \frac{1}{2} (I^{(t+3)}(x, y-1) + I^{(t+2)}(x, y+1)) \\ I^{(t+4)}(x,y) &= \frac{1}{2} (I^{(t+3)}(x, y-1) + I^{(t+4)}(x, y+1)), \end{aligned}$$

3. *polynomial smoothing*, where the pixel intensities in a small area (e.g., 3×3 , 7×7) are approximated by a second-degree polynomial function in x and y ,

*In particular, these analytical methods are used to filter gel images, detect spots, remove background noise, quantify spots, align images, match spots, create synthetic gels, and identify differential expression patterns. Wavelet methods are described and applied to analyze 2-D gel images.

4. *adaptive smoothing*, that is,

$$I^{(t+1)}(x, y) = \frac{1}{N^{(t)}} \sum_{i=-1}^1 \sum_{j=-1}^1 I^{(t)}(x+i, y+j) w^{(t)}(x+i, y+j)$$

with

$$w^{(t)}(x, y) = \exp\left(-\frac{(d^{(t)}(x, y))^2}{2 K^2}\right), \quad N^{(t)} = \sum_{i=-1}^1 \sum_{j=-1}^1 w^{(t)}(x+i, y+j), \quad d^{(t)}(x, y) = \sqrt{G_x^2 + G_y^2}$$

where G_x and G_y are the gradients along the x - and y -axis, respectively (3).

Gaussian deconvolution is an alternative approach to remove noise and blur (22). Each spot is modeled as

$$I(x, y) = \sum_{k=-m}^m A(x+k, y) g_k + e(x, y) \text{ with } A(x+k, y) \geq 0 \text{ and } g_k = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{k^2}{2\sigma^2}\right)$$

where m is the integer part of 3σ and $e(x, y)$ represents random noise. Estimates are obtained via the constrained least-squares procedure. This approach tends to be overly sensitive to noise and to oversplit spots (9).

2.2. Spot Detection

For automatic spot detection, nonparametric procedures, based either on second derivatives (2,3) or on mathematical morphology (22), are utilized.

Let $\mathbf{p} = (x, y)$ be a point on the image, S_i a spot, and T the saturation threshold

$$\max(I) - \frac{100 - \text{saturation}}{100} (\max(I) - \min(I))$$

where $0 \leq \text{saturation} \leq 100$ ($\text{saturation} = 100$ when no pixel is saturated), and $\Delta I(\mathbf{p})$ the Laplacian

$$-\left(\frac{\partial^2}{\partial x^2} I(\mathbf{p}) + \frac{\partial^2}{\partial y^2} I(\mathbf{p})\right).$$

Is \mathbf{p} part of the spot S_i ? Select thresholds l, r, c (i.e., small positive constants). If $I(\mathbf{p}) < T$,

$$\mathbf{p} \in S_i \Leftrightarrow \min\left(\frac{\partial^2}{\partial x^2} I(\mathbf{p}) - r, \frac{\partial^2}{\partial y^2} I(\mathbf{p}) - c\right) > 0$$

when $-\Delta I(\mathbf{p}) - l \geq 0$. If $I(\mathbf{p}) > T$,

$$\mathbf{p} \in S_i \Leftrightarrow \min\left(\frac{\partial^2}{\partial x^2} I(\mathbf{p}), \frac{\partial^2}{\partial y^2} I(\mathbf{p})\right) > 0.$$

Small values of l allow the detection of as many spots as possible, while high values only yield dark spots with flat spots being ignored. High values of r and c help separate spots that are in close proximity of each other and to eliminate streaks. The algorithm identifies

the spots by searching for the most negative values of the Laplacian and the two second derivatives (2,3). The Laplacian is indeed most negative at local peaks, while at the inflection points between unresolved spots the minimum value of the two second derivatives is negative.

With mathematical morphology, one can study characteristics of objects by investigating whether a standard shape fits into them. In this context, one searches for elevations relative to the local background brightness, and constructs an image based on the heights of these elevations. This is achieved by subtracting the *closing* of the image from the original image. This is the so-called *top-hat transform*, which in essence assesses whether a cylinder of a chosen radius fits into the elevations. To obtain the closing of the image, one first replaces each pixel value $I(x,y)$ with the local minimum intensity in a disk around each pixel, and then one replaces the resulting pixel value $I'(x,y)$ with a local maximum intensity, that is,

$$I''(x,y) = \max_{k,l} I'(x+k, y+l)$$

where $\sqrt{k^2 + l^2} \leq R$ (the disc radius) and $I'(x,y) = \min_{k,l} I(x+k, y+l)$.

In this closing, only pixels within elevations narrower than the chosen disc size have changed their values from the original image, and thus will show when the two images are subtracted. The radius R is selected to be the smallest value so that the disk is larger than the smallest spot (22). Shapes (or *structuring elements*) other than disks can be considered (e.g., spheres [10]).

Alternatively, instead of a fixed structuring element, one can look for all h -domes (4,11,12). An h -dome is a connected region of pixels with intensity above h and greater than any pixel bordering the h -dome. These h -domes are not constrained by size. Algorithms for searching for these regions are more complex than the top-hat transform. The choice of h is crucial: if it is too small, background streaks will be recovered rather than spots, and, if it is too large, narrow peaks due to high-amplitude noise will be selected. Raising h stepwise allows the resolution of overlapping spots (11).

2.3. Background Filtering

The smooth background noise, which consists of vertical and horizontal streaks, is removed, either by subtracting the global minimum pixel value from all pixels or by estimating the background outside the spots with a third-order polynomial function (3). Because the background varies significantly across the image, a single threshold tends to work poorly: when it is set too high, faint spots are lost and, when it is set too low, high background is regarded as signal (11).

Mathematical morphology has also been utilized to remove the streaks (4,10). One subtracts from the original image its closing with respect to two structuring elements, one vertical and one horizontal bar of one-pixel width, the lengths of which are slightly greater than those of the vertical and horizontal extents of the largest spots.

2.4. Spot Quantification

Spot characteristics are estimated by fitting two-dimensional Gaussian curves via the least-squares method:

$$g(x, y) = A \exp \left\{ \left(\frac{x - x_c}{\sigma_x} \right)^2 + \left(\frac{y - y_c}{\sigma_y} \right)^2 \right\} + B$$

with A representing the amplitude, (x_c, y_c) the center, the σ 's the spread along the principal axes and B the background level (3,13,14). Spot models based on two half-Gaussian curves have also been utilized (4). However, many spots are not Gaussian in nature because of several factors: local inhomogeneity within the acrylamide, overloading of the sample within the gel, adsorption of some proteins onto the acrylamide matrix, failure of some polypeptides to focus in the first dimension, and tendency for chemically distinct but barely resolved proteins to displace each other (9).

Spot characteristics are thus better estimated directly:

$$\text{area} = \text{AREA} = \text{number of pixels} \times \text{pixel area}$$

$$\text{optical density} = \text{OD} = \max_{x, y \in \text{spot}} I(x, y)$$

$$\text{percent optical density} = \% \text{OD} = 100 \times \frac{\text{OD}}{\sum_{s=1}^n \text{OD}_s}$$

$$\text{volume} = \text{VOL} = \sum_{x, y \in \text{spot}} I(x, y)$$

$$\text{percent volume} = \% \text{VOL} = 100 \times \frac{\text{VOL}}{\sum_{s=1}^n \text{VOL}_s}$$

where OD_s and VOL_s are, respectively, the optical density and the volume of spot s in a gel containing n spots.

2.5. Image Alignment

Gels are aligned via polynomial image warping. Identify landmarks (or control points) on each image and choose one gel as the reference gel. The alignment algorithm attempts to superimpose these landmarks by stretching and shrinking the images. Let (x, y) be the pixel coordinates in the original image and $[u(x), v(y)]$ those in the warped image. The latter are first-, second- or third-order univariable polynomials or their inverses. Estimate the parameters of these polynomial functions via the least-squares criterion by summing over the landmarks. Specifically, let there be M landmarks on each gel. The parameters are obtained by minimizing, for instance, if $M \geq 4$,

$$\sum_{i=1}^M (u_i - (a_0 + a_1 x_i + a_2 x_i^2 + a_3 x_i^3))^2 \quad \text{and} \quad \sum_{i=1}^M (v_i - (b_0 + b_1 y_i + b_2 y_i^2 + b_3 y_i^3))^2$$

where (u_i, v_i) refers to landmark i on the reference gel and (x_i, y_i) to its position on another gel. The value M determines the order of the polynomial: a polynomial of degree n requires at least $n + 1$ landmarks.

2.6. Spot Matching

Local gel-to-gel variations make it impossible to utilize a single transformation to map the spots from one gel to another. One approach is to divide the image into a number of small rectangular regions and to select, in each segment, 3 or more evenly spaced spots as reference points (11). These reference points serve to compute a transformation that maps spot centers from one film to another. Spots are considered matched if the transformed spot center from one gel and the corresponding spot center on the other gel are within 0.8 mm [a slightly more stringent criterion of 0.7 mm has also been utilized (15)]. This procedure works best for the area defined by the reference points: spots located at the edges of the rectangular regions can be poorly matched. As a remedy, the following steps are added to the procedure: triangles of nearby matched spots are considered on both images, and the above algorithm is applied to the yet unmatched spots within these triangles.

Alternatively, for each spot on a gel, consider a cluster of neighboring spots (3). The central spot is regarded as the primary spot, and the surrounding spots as secondary spots. A spot belongs to a cluster if its centroid is inside a circle of fixed radius. This radius depends on the image dimension, number of spots on a gel, and minimum number of spots in the cluster. The clusters are characterized by polar coordinates centered at the primary spot. First, match the clusters with highest-intensity primary spots. Compare clusters via a probabilistic similarity measure. The probability that the next random hit falls within a cluster where $m - 1$ spots have been matched is given by

$$p_m = \frac{A_s - A_{m-1}}{A_c - A_{m-1}}$$

where A_s is the sum of the secondary areas in the cluster, A_c the total area within the boundary of the cluster, and A_{m-1} the total area of matched spots. If N stands for the number of spots in one cluster,

Prob(at least m spots are matched in N trials)

$$= \sum_{h=m}^N \binom{N}{h} \prod_{i=1}^h p_i \prod_{i=h+1}^N (1-p_i) \approx \sum_{h=m}^N \binom{N}{h} p_G^h (1-p_G)^{N-h} \text{ where } p_G \left(\prod_{i=1}^m p_i \right)^{1/m}.$$

That spots be reliably matched is of paramount importance in the creation of representative images and in subsequent pattern-recognition analyses. Proceed with a consistency check for possible mismatching:

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} t_x \\ t_y \end{pmatrix} + \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

by ensuring that $L = AD - BC \approx 1$ (rotation). For each primary cluster, estimate parameters from each set of 3 matched spots. When $L = 1.0 \pm 0.25$ and the rotation angle is ± 10 degrees, the pairing is declared suitable. To project the remaining spots in the clusters, estimate the rotation parameters A, \dots, D by the least-squares method from good matchings in the two clusters.

Artificial-intelligence methodology has also been used to match spot lists (6). Because they are not based on geometrical considerations, they should be able to cope

better with discontinuous gel distortions. Spot clusters are described via the angles and distances between any two spots in the cluster. Distances are divided into 3 classes and angles into 16 classes. Measurements are coded via their class identifiers. Spots are then matched via syntactic pattern-recognition techniques. Heuristic rules are imposed to limit the number of searches. Isolated spots tend to be problematic for this approach.

2.7. Creating Synthetic Gels

To obtain a master image from at least three pairwise matched gels, first select a reference gel (3). Check that the spots on the reference gel are well matched to spots on two other gels. These form triangles of matched spots, that is, the *starting groups*. Extend the starting groups by adding spots using the connectivity test: a spot must be matched with at least one other spot in the initial group. When all spots on the reference gel have been considered, create additional groups with the spots on the second gel that are not part of a group. Repeat with the other gels.

The synthetic gel contains the same number of spots as there are determined groups (these are the representative spots). The position of a spot on the synthetic gel is taken from the reference gel if the group has a spot on the reference gel. Otherwise, one translates the coordinates of the closest spot by considering a set of neighbors that have representatives on the reference gel. The intensity of the master spot is the average over the spots in the group. Its shape is the shape of the spot in the group that is closest in area to the average of the group (3).

To ensure the reliability of the master gel, some investigators compute an overlap measure between each master spot and the corresponding spots (considered one at a time) on the aligned gels (23). This in effect assesses the quality of the matches on which the master gel relies. The overlap measure is simply the value of a Gaussian function evaluated at the physical distance of the spot centers. This Gaussian function is chosen to have height of 1.0 and width depending on the matching criterion: for instance, 0.7 mm if spots on different aligned gels are considered matched when they are within 0.7 mm (15).

2.8. Pattern Recognition

For pattern-recognition purposes, only spots that yield highly reliable features on the master image are considered in the analyses. For instance, the overlap measure between the master spot and a corresponding object on one of the aligned gels (cf. **Subheading 2.7.**) needs to be above a specific threshold (typically 0.5) and to exceed 90% of the largest overlap of the master spot with the original spots or of the overlap of the gel spot with all the master spots (23).

To find significant protein patterns associated with a specific disease, investigators have first recourse to principal-component analysis (23) or correspondence analysis and factor analysis to reduce dimensionality (1–3). This requires the computation of the normalized observation table so that the columns have mean 0 and variance 1. In factor analysis, the eigenvalues and eigenvectors of the covariance matrix are extracted to determine a factorial space (usually of dimension between 1 and 3). The gels are projected as points onto the factorial space. Spots can also be mapped onto this space so that characteristic spots can be identified: spots fall within the cluster of gels they typify.

These authors then apply clustering algorithms to the transformed data. The difficulty here is to define a meaningful distance metric. With principal-component analysis, one candidate is the Euclidean distance in the transformed space after weighing each coordinate by the percentage of the total variance represented by the corresponding principal component (23). The usual hierarchical clustering procedures, based on complete or single or average linkage, are utilized (1-3,16,23). A heuristic clustering algorithm has also been proposed (1,2). Suppose that n gels are to be classified into k classes. Select k gels by maximizing the Euclidean distance between them. These define k classes. A heuristic search is then performed: each of the remaining $n - k$ gels is included into one class and class descriptions are formulated. Iterate this process by choosing one gel per class, excluding the first k gels, to form new classes and repeating the previous step. This process continues until the classification converges.

One is in effect searching for protein patterns that best distinguish two groups of images (one from a “disease” group and one from a “control” group). Classification procedures would be best suited for this purpose. In actuality, patterns are identified, ignoring the associated outcomes, and then the inferred patterns are reconciled to the known groups.

3. Brief Introduction to Wavelets

Wavelets are building-block functions like *sine* and *cosine* functions in the Fourier transform (17). They oscillate about 0 and dampen to 0. This localization in time or space renders them highly versatile to model signals with nonsmooth features or that vary over time or space. The *father wavelet* or *scaling function* ϕ represents smooth, low-frequency components while the *mother wavelet* ψ represents detail, high-frequency components:

$$\int_{-\infty}^{+\infty} \phi(t) dt = 1 \quad \text{and} \quad \int_{-\infty}^{+\infty} \psi(t) dt = 0.$$

A number of orthogonal wavelet families have been constructed; for instance, Haar wavelets (symmetric square waves with compact support), daublets (continuous waves with compact support, d2 to d20 in S-plus [18]), symmlets (nearly symmetric waves with compact support, s4 to s20 in S-plus [18]).

Through a multiresolution analysis, one obtains fine to coarse resolution (scale) components of the signal, that is, for a one-dimensional signal,

$$f(t) = \sum_k s_{j,k} \phi_{j,k}(t) + \sum_k d_{j,k} \varphi_{j,k}(t) + \sum_k d_{j-1,k} \varphi_{j-1,k}(t) + \dots + \sum_k d_{1,k} \varphi_{1,k}(t)$$

where J is the number of multiresolution components considered. The functions $\phi_{j,k}(t)$ and $\psi_{j,k}(t)$ are generated from ϕ and ψ by scaling and translation, that is,

$$\phi_{j,k}(t) = 2^{\frac{-j}{2}} \phi(2^{-j}t - k) = 2^{\frac{-j}{2}} \phi\left(\frac{t - 2^j k}{2^j}\right) \quad \text{and} \quad \psi_{j,k}(t) = 2^{\frac{-j}{2}} \psi(2^{-j}t - k) = 2^{\frac{-j}{2}} \psi\left(\frac{t - 2^j k}{2^j}\right).$$

The scale/dilation factor 2^j affects the width of $\phi_{j,k}(t)$ and $\psi_{j,k}(t)$. The translation/location parameter $2^j k$ is coupled to the scale factor: as the support of $\phi_{j,k}(t)$ and $\psi_{j,k}(t)$ gets wider the translation steps become larger. As 2^j increases, $\phi_{j,k}(t)$ and $\psi_{j,k}(t)$ become

shorter and more spread out. Finally, $s_{J,k}$, $d_{J,k}$, ..., $d_{I,k}$ are the wavelet-transform coefficients:

$$\begin{aligned} \text{scaling function coefficients} \quad s_{J,k} &= \int_{-\infty}^{+\infty} f(t) \phi_{J,k}(t) dt \\ \text{wavelet coefficients} \quad d_{J,k} &= \int_{-\infty}^{+\infty} f(t) \varphi_{J,k}(t) dt \end{aligned}$$

The $\phi_{j,k}(t)$'s and $\psi_{j,k}(t)$'s form an orthogonal basis:

$$\int_{-\infty}^{+\infty} \phi_{J,k}(t) \phi_{J,k'}(t) dt = \delta_{k,k'}, \quad \int_{-\infty}^{+\infty} \varphi_{J,k}(t) \phi_{J,k'}(t) dt = 0, \quad \int_{-\infty}^{+\infty} \varphi_{J,k}(t) \varphi_{J,k'}(t) dt = \delta_{k,k'}$$

where $\delta_{i,j} = 1$ if $i = j$ and 0 if $i \neq j$.

The discrete wavelet transform \mathbf{W} for the discrete signal $\mathbf{f} = (f_1, f_2, \dots, f_n)'$ is defined as

$$\mathbf{w} = \mathbf{W} \mathbf{f} \quad \text{where} \quad \mathbf{w}' = (\mathbf{s}_J' \ \mathbf{d}_J' \ \mathbf{d}_{J-1}' \ \dots \ \mathbf{d}_1')$$

with

$$\begin{aligned} \mathbf{s}_J &= (s_{J,1}, s_{J,2}, \dots, s_{J,n/2^J})', \quad \mathbf{d}_J = (d_{J,1}, d_{J,2}, \dots, d_{J,n/2^J})', \quad \mathbf{d}_{J-1} = (d_{J-1,1}, d_{J-1,2}, \dots, d_{J-1,n/2^{J-1}})', \dots, \\ \mathbf{d}_1 &= (d_{1,1}, d_{1,2}, \dots, d_{1,n/2})'. \end{aligned}$$

Each of the so-called crystals \mathbf{s}_J , \mathbf{d}_J , \mathbf{d}_{J-1} , ..., \mathbf{d}_1 contains the coefficients corresponding to a set of translated wavelet functions. In the multiresolution analysis,

$$f(t) \approx S_J(t) + D_J(t) + D_{J-1}(t) + \dots + D_1(t),$$

the smooth and detail signals are represented, respectively, by

$$S_J(t) = \sum_k s_{J,k} \phi_{J,k}(t) \quad \text{and} \quad D_j(t) = \sum_k s_{j,k} \varphi_{j,k}(t) \quad j = 1, \dots, J.$$

To compress an image, one utilizes a two-dimensional wavelet family

$$\begin{aligned} \Phi(x,y) &= \phi_h(x) \times \phi_v(y) = \text{horizontal father} \times \text{vertical father} \\ \Psi^v(x,y) &= \psi_h(x) \times \phi_v(y) = \text{horizontal mother} \times \text{vertical father} \\ \Psi^h(x,y) &= \phi_h(x) \times \psi_v(y) = \text{horizontal father} \times \text{vertical mother} \\ \Psi^d(x,y) &= \psi_h(x) \times \psi_v(y) = \text{horizontal mother} \times \text{vertical mother}. \end{aligned}$$

The father wavelet Φ deals with the smooth aspect and the mother wavelets Ψ deal with the details in the vertical (Ψ^v), horizontal (Ψ^h), and diagonal (Ψ^d) dimensions. (Figure 1 shows the diagonal s8 wavelet.)

The two-dimensional wavelet approximation is then

$$\begin{aligned} F(x,y) \approx \sum_{m,n} s_{J,m,n} \Phi_{J,m,n}(x,y) + \sum_{j=1}^J \sum_{m,n} v_{j,m,n} \Psi_{j,m,n}^v(x,y) + \sum_{j=1}^J \sum_{m,n} h_{j,m,n} \Psi_{j,m,n}^h(x,y) \\ + \sum_{j=1}^J \sum_{m,n} d_{j,m,n} \Psi_{j,m,n}^d(x,y) \end{aligned}$$

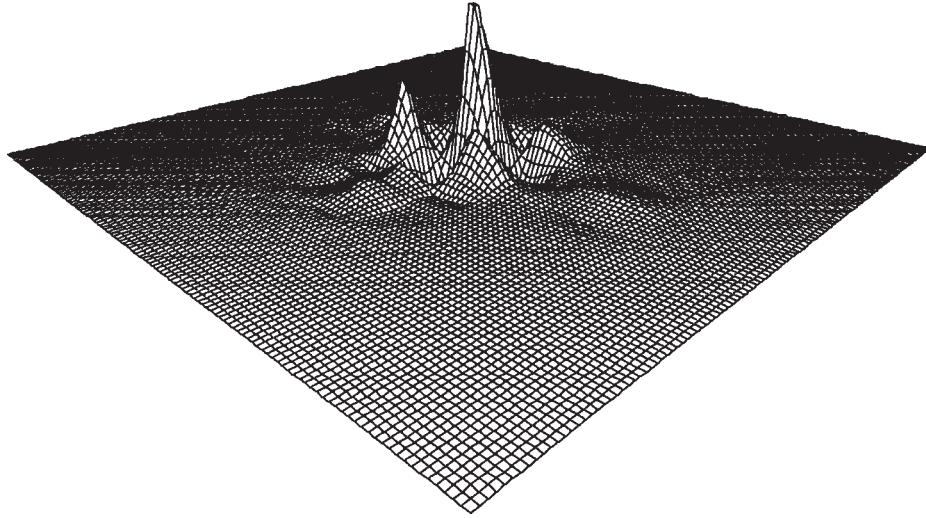


Fig. 1. s8 diagonal mother wavelet.

with

$$\begin{aligned}
 \Phi_{j,m,n}(x, y) &= 2^{-j} \Phi(2^{-j} x - m, 2^{-j} y - n), \Psi_{j,m,n}^v(x, y) = 2^{-j} \Psi^v(2^{-j} x - m, 2^{-j} y - n), \\
 \Psi_{j,m,n}^h(x, y) &= \Psi^h(2^{-j} x - m, 2^{-j} y - n), \Psi_{j,m,n}^d(x, y) = \Psi^d(2^{-j} x - m, 2^{-j} y - n), \\
 s_{j,m,n} &= \int \int \Phi_{j,m,n}(x, y) F(x, y) dx dy, v_{j,m,n} = \int \int \Psi_{j,m,n}^v(x, y) F(x, y) dx dy, \\
 h_{j,m,n} &= \int \int \Psi_{j,m,n}^h(x, y) F(x, y) dx dy, d_{j,m,n} = \int \int \Psi_{j,m,n}^d(x, y) F(x, y) dx dy.
 \end{aligned}$$

The two-dimensional discrete wavelet transform maps an $m \times n$ discrete image to $m \times n$ matrix of wavelet coefficients $\mathbf{w}_{m,n}$. In S-plus (18), $\mathbf{w}_{m,n}$ is decomposed into submatrices with coefficients for different multiresolution levels:

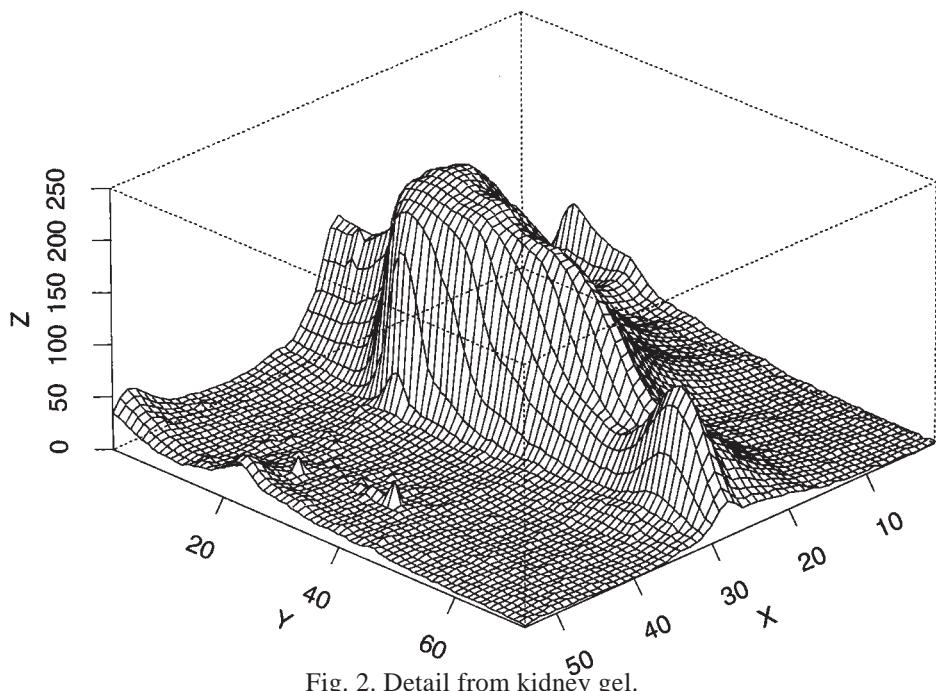
$\mathbf{sJ} - \mathbf{sJ}$	with coefficients $s_{j,m,n}$	for the smooth part
$\mathbf{d1} - \mathbf{s1}, \dots, \mathbf{dJ} - \mathbf{sJ}$	with coefficients $v_{j,m,n}$	for the vertical detail
$\mathbf{s1} - \mathbf{d1}, \dots, \mathbf{sJ} - \mathbf{dJ}$	with coefficients $h_{j,m,n}$	for the horizontal detail
$\mathbf{d1} - \mathbf{d1}, \dots, \mathbf{dJ} - \mathbf{dJ}$	with coefficients $d_{j,m,n}$	for the diagonal detail.

Hence, in the multiresolution analysis,

$$F(x, y) \approx S_j(x, y) + \sum_{j=1}^J D_j^v(x, y) + \sum_{j=1}^J D_j^h(x, y) + \sum_{j=1}^J D_j^d(x, y),$$

where

$$\begin{aligned}
 S_j(x, y) &= \sum_{m,n} s_{j,m,n} \Phi_{m,n}(x, y), \quad D_j^h(x, y) = \sum_{m,n} h_{j,m,n} \Psi_{m,n}^h(x, y), \\
 D_j^v(x, y) &= \sum_{m,n} v_{j,m,n} \Psi_{m,n}^v(x, y), \quad D_j^d(x, y) = \sum_{m,n} d_{j,m,n} \Psi_{m,n}^d(x, y).
 \end{aligned}$$



4. Wavelets for Two-Dimensional Electrophoretic Data

Few statistical techniques can cope with the high dimensionality of 2D-PAGE data. Hence, for analytical reasons, the information is often reduced to the volumes of a manageable set of selected spots. This prohibits exploratory investigations of the data for the purpose of formulating testable hypotheses. We explored the possibility of fitting Gaussian curves. We selected a number of spots from a gel and assessed via statistical tests that we would not be justified in assuming that their shape is Gaussian (as is clearly evidenced by [Fig. 2](#)). We turned to wavelets for their versatility in representing irregular signals.

With this methodology, much effort is needed to identify the most suitable representation. Once the coefficients are selected, mainstream techniques can be applied to investigate the scientific questions of interest. We now review a few of these issues.

4.1. What Is the Most Suitable Wavelet Family to Represent Gels?

We considered Haar wavelets, daublets, and symmlets. All seemed suitable with the Haar family giving slightly worse results ([Figs. 3–6](#)). Even with an image corrupted by noise (Gaussian or Poisson), the reconstruction was highly successful ([Figs. 7–9](#)).

4.2. What Multiresolution Level Should One Select?

A high multiresolution level is not necessary: most of the information is contained in the smoother crystals. The percentage of coefficients one wishes to retain is the more pertinent issue and depends on one's threshold for the volume of significant spots ([Figs. 10–13](#)).

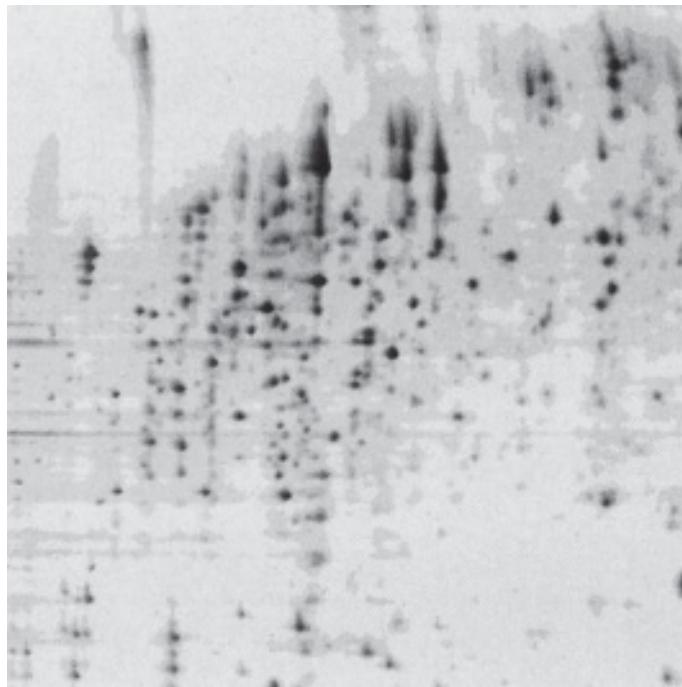


Fig. 3. Cropped image of kidney gel (512×512).

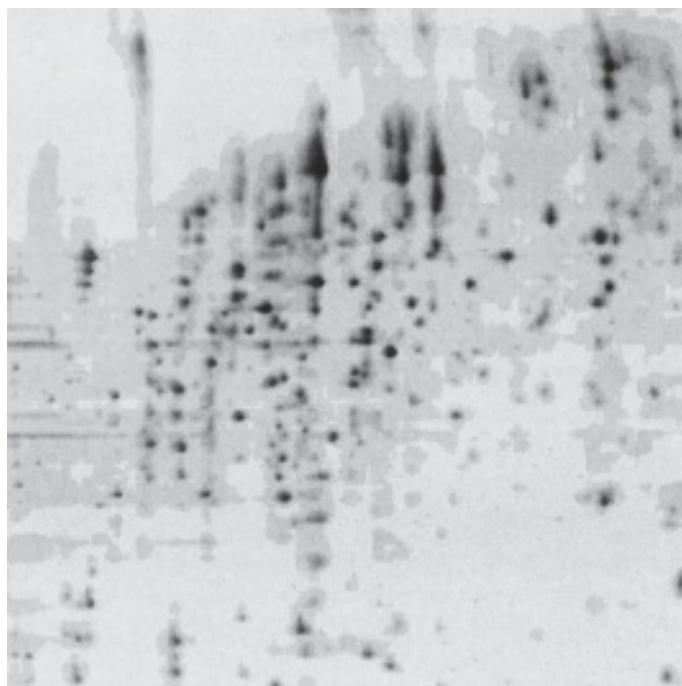


Fig. 4. Reconstruction of the cropped image of a kidney gel with daublets d4 at multiresolution level 4 and the largest 5% of the coefficients.

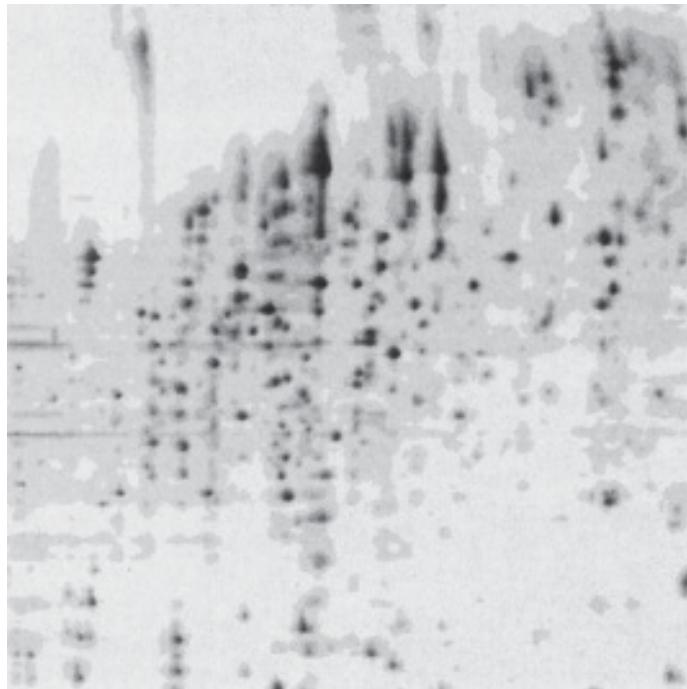


Fig. 5. Reconstruction of the cropped image of a kidney gel with symmlets s8 at multiresolution level 4 and the largest 5% of the coefficients.

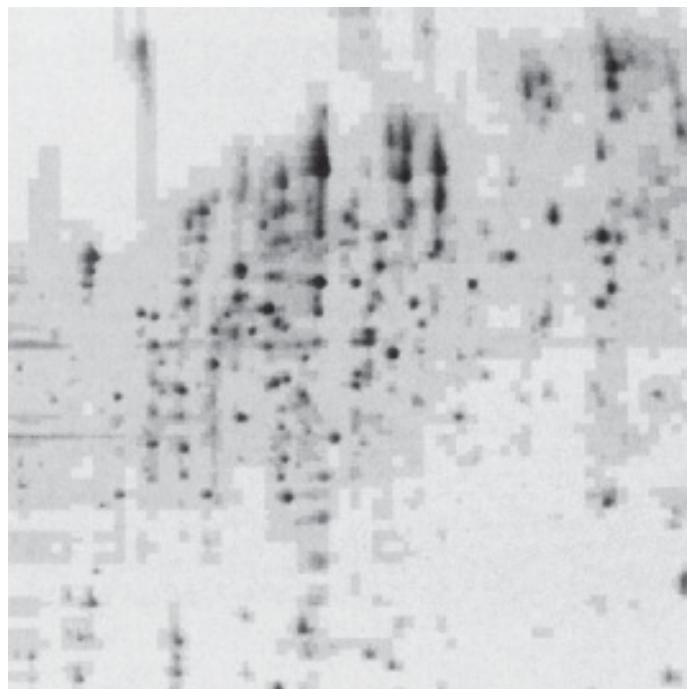


Fig. 6. Reconstruction of the cropped image of a kidney gel with Haar wavelets at multiresolution level 4 and the largest 5% of the coefficients.

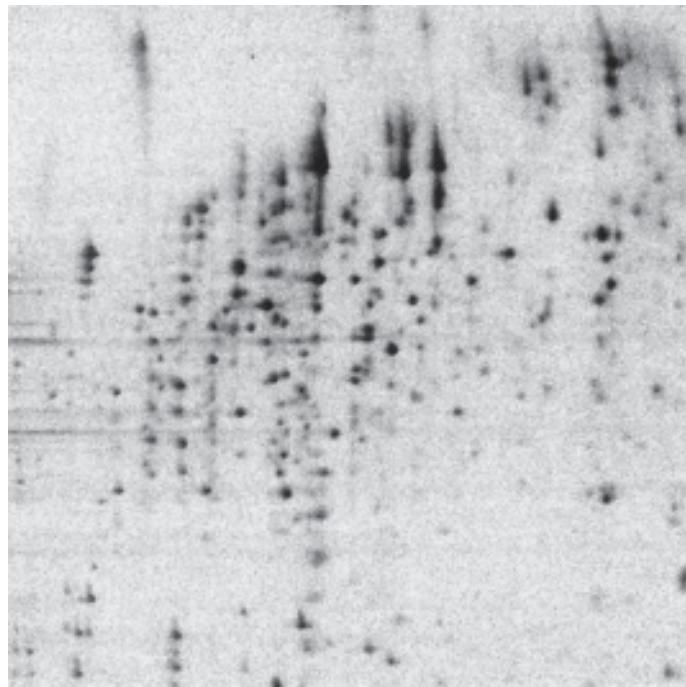


Fig. 7. Cropped gel image with added Gaussian noise (with variance 100).

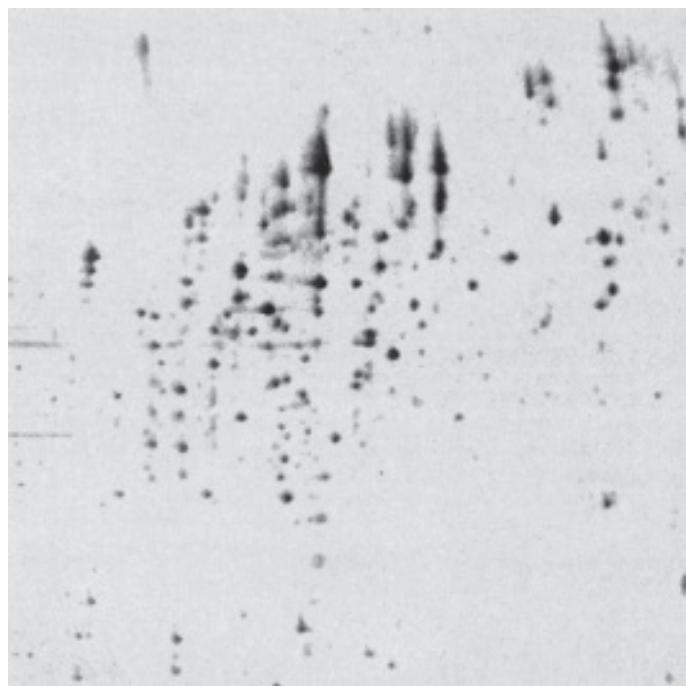


Fig. 8. Reconstruction of noisy gel with symmlets s8 at multiresolution level 1 and the largest 1% of the coefficients.

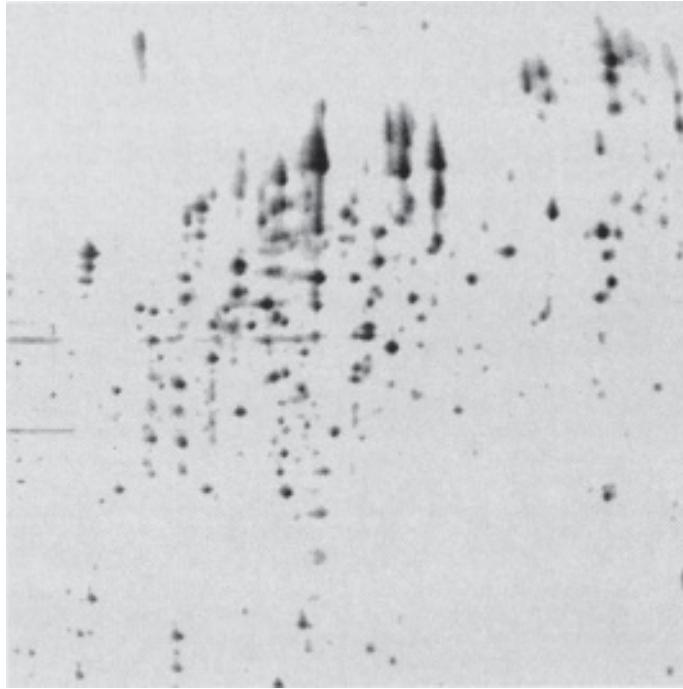


Fig. 9. Reconstruction of cropped image of kidney gel with symmlets s8 at multiresolution level 1 and the largest 5% of the coefficients.

4.3. How Does One Remove the Noise?

In the WaveShrink algorithm of S-plus (35), one applies the wavelet transform with J levels then shrinks the detail coefficients

$$\tilde{\mathbf{d}}_1 = \delta_{\lambda_1 \sigma_1}(\mathbf{d}_1), \dots, \tilde{\mathbf{d}}_J = \delta_{\lambda_J \sigma_J}(\mathbf{d}_J)$$

and reconstructs the image using $\mathbf{d}_1, \dots, \mathbf{d}_J, \mathbf{s}_J$. Shrinkage is performed using the so-called soft or hard shrinkage functions

$$\delta_\gamma^S(x) = \begin{cases} 0 & \text{if } |x| \leq \gamma \\ \text{sign}(x)(|x| - \gamma) & \text{if } |x| > \gamma \end{cases}, \quad \delta_\gamma^H(x) = \begin{cases} 0 & \text{if } |x| \leq \gamma \\ x & \text{if } |x| > \gamma \end{cases}.$$

For the thresholds λ_j , one can select the so-called universal value

$$\lambda_j = \sqrt{2 \log n}$$

where n is the sample size. Alternatively, the value which minimizes the upper bound of the asymptotic risk (minimax) will result in less smoothing as it is always smaller than the universal threshold (19). Finally, we also considered Stein's unbiased risk estimator (SURE) which is adapted to each multiresolution level; the threshold for \mathbf{d}_j with K coefficients is

$$\lambda_j = \arg \min_{t \geq 0} \text{SURE}(\mathbf{d}_j, t) \text{ where } \text{SURE}(\mathbf{d}_j, t) = K - 2 \sum_{k=1}^K 1_{[d_{j,k} \leq t \sigma_j]} + \sum_{k=1}^K \min((d_{j,k} \sigma_j)^2, t^2)$$

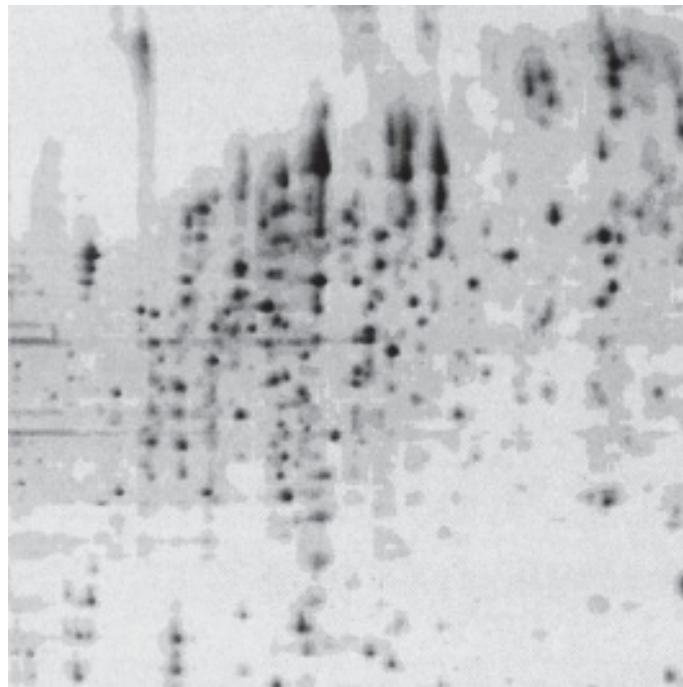


Fig. 10. Reconstruction of cropped image of kidney gel with daublets d4 at multiresolution level 3 and the largest 5% of the coefficients.

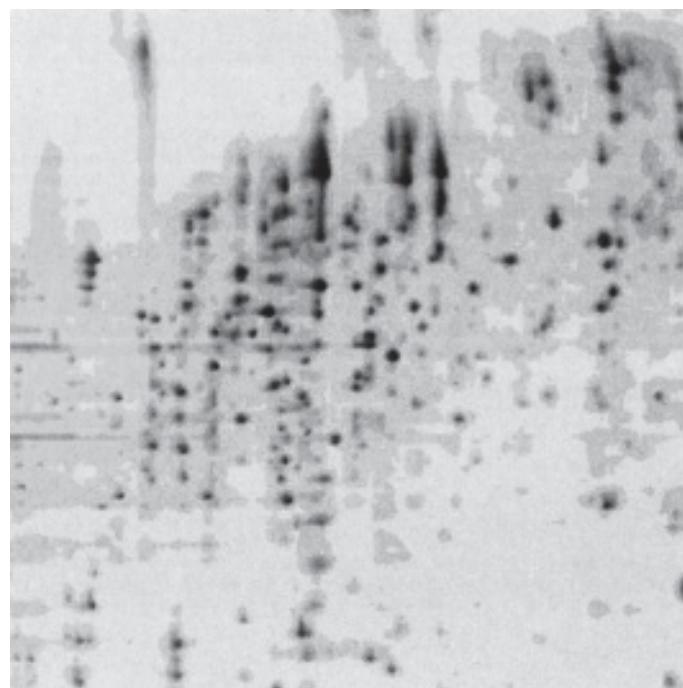


Fig. 11. Reconstruction of cropped image of kidney gel with daublets d4 at multiresolution level 5 and the largest 5% of the coefficients.

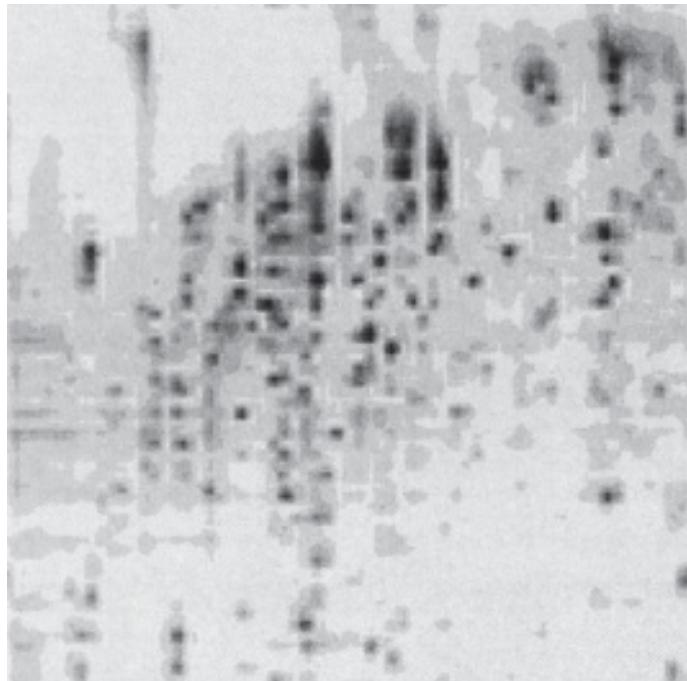


Fig. 12. Reconstruction of cropped image of kidney gel with daublets d4 at multiresolution level 3 and the largest 1% of the coefficients.

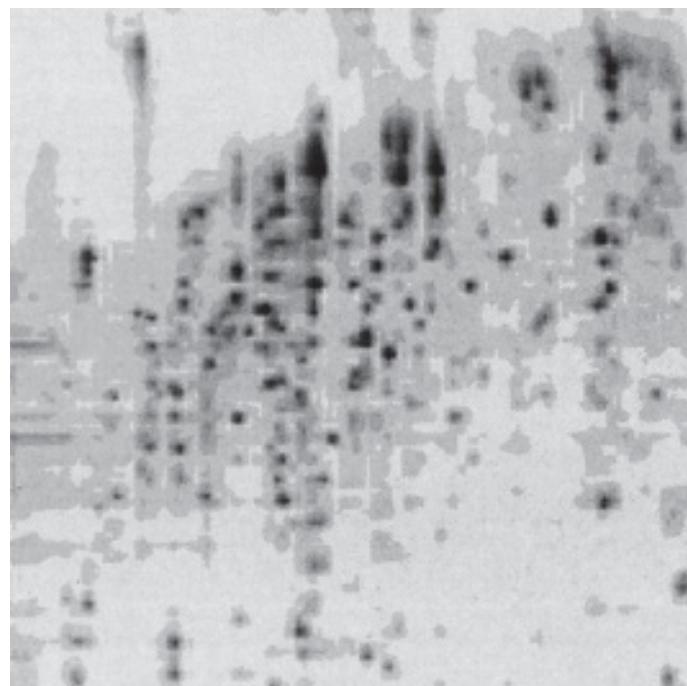


Fig. 13. Reconstruction of cropped image of kidney gel with daublets d4 at multiresolution level 5 and the largest 1% of the coefficients.

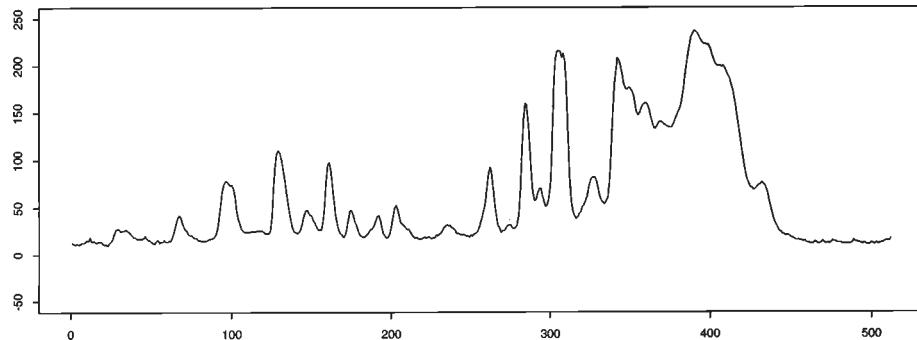


Fig. 14. Vertical slice of cropped image of kidney gel.

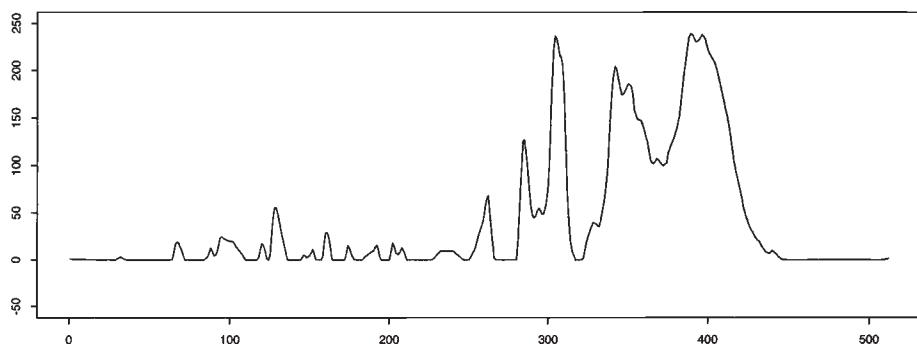


Fig. 15. Slice after hard shrinkage with universal threshold, estimating the scale of the noise separately for each crystal (multiresolution level 4).

(20,21). To estimate the scale of the noise, that is, σ_j , one can rely either on the crystal corresponding to the finest detail or on all the detail crystals, or one can consider each crystal in turn, that is,

$$\tilde{\sigma}_j(\mathbf{d}_1), \dots, \tilde{\sigma}_j(\mathbf{d}_1, \dots, \mathbf{d}_j), \tilde{\sigma}_j(\mathbf{d}_j).$$

These algorithms yield rather poor results with the 2D-PAGE data (Figs. 14–16). Indeed, they are really smoothing techniques that are suitable to reduce highly localized and peaked noise. Here, the noise takes the form of streaks. All combinations we have tried result in the removal of important features.

We devised a hybrid procedure that seems to work well: hardshrinkage is utilized on the level 1 coefficients (these all tend to be very small) and the SJ - SJ crystal is multiplied by a constant between 0 and 1 (Fig. 17). This constant depends on the level of detail to be retained.

This routine is currently being optimized with respect to a biologically relevant objective criterion that involves the size of the spots being ignored.

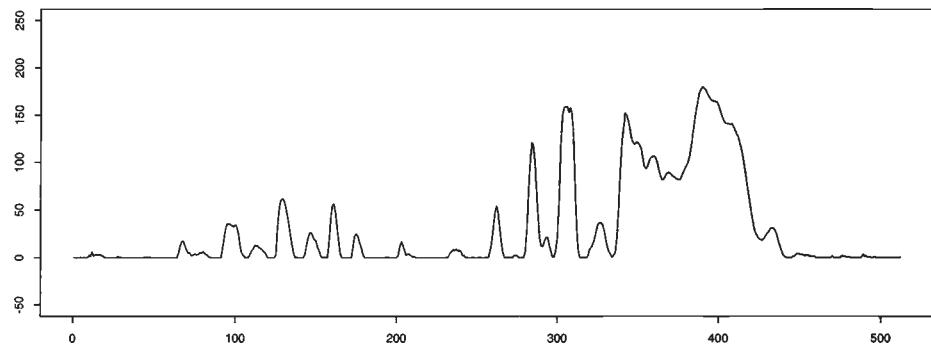


Fig. 16. Slice after soft shrinkage of the SJ - SJ crystal with universal threshold, estimating the scale of the noise from the SJ - SJ crystal (multiresolution level 4).

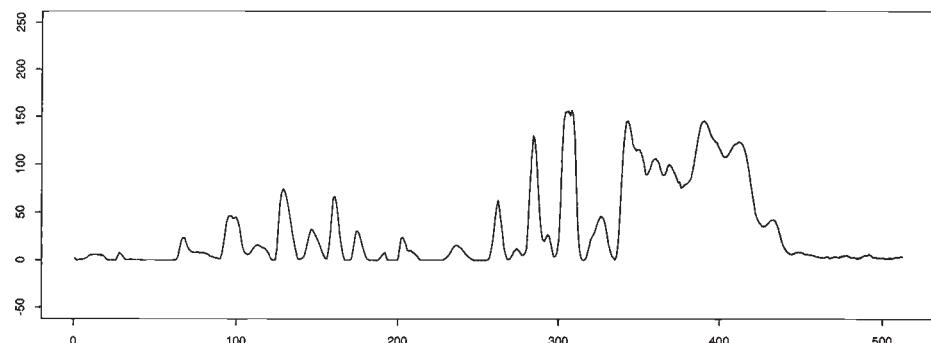


Fig. 17. Slice after hybrid shrinkage with the universal threshold and the multiplier set to 0.25 (multiresolution level 4).

4.4. How Does One Create a Master Gel?

Assume that the gels have been aligned. Wavelet coefficients are obtained for each of them. The synthetic gel is constructed by averaging the coefficients or by taking their median values. Variability in this construct can easily be computed.

4.5. How Does One Find Specific Protein Patterns for a Disease?

An analysis of the wavelet coefficients, for gels from diseased and control samples, based on classification and regression trees (CART), will highlight relevant clusters that best discriminate between the two groups. This has the advantage of considering both the location and the intensity of the spots simultaneously.

5. Conclusion

Electrophoresis has developed over the past 60 yr from a crude method able only to distinguish between very specific one-dimensional changes in experimental protocols to a highly complex technique. It is now possible not only to separate the genomic fingerprint of samples but also their proteome. While the technology has developed at an ever-increasing rate, the statistical techniques necessary to analyze such complex data structures has been left wanting. We have outlined some of the new methodolo-

gies that are currently available to take full advantage of the technology that is now in common usage in molecular biology laboratories.

References

1. Appel, R., Hochstrasser, D. F., Roch, C., Funk, M., Muller, A. F., and Pellegrini, C. (1988) Automatic classification of two-dimensional gel electrophoresis pictures by heuristic clustering analysis: a step toward machine learning. *Electrophoresis* **9**, 136–142.
2. Appel, R., Hochstrasser, D. F., Funk, M., Vargas, J. R., Pellegrini, C., Muller, A. F., and Scherrer, J. R. (1991) The MELANIE project: from a biopsy to automatic protein map interpretation by computer. *Electrophoresis* **12**, 722–735.
3. Appel, R., Palagi, P. M., Walther, D., Vargas, J. R., Sanchez, J. C., Ravier, F., et al. (1997) MELANIE II — A third-generation software package for analysis of two-dimensional electrophoresis images: I. Features and user interface. *Electrophoresis* **18**, 2724–2734.
4. Vincens, P. (1986) HERMeS: a second generation approach to the automatic analysis of two-dimensional electrophoresis gels. Part II: Spot detection and integration. *Electrophoresis* **7**, 357–367.
5. Vincens, P., Paris, N., Pujol, J. L., Gaboriaud, C., Rabilloud, T., Pennetier, J. L., et al. (1986) HERMeS: a second generation approach to the automatic analysis of two-dimensional electrophoresis gels. Part I: Data acquisition. *Electrophoresis* **7**, 347–356.
6. Vincens, P. and Tarroux, P. (1987) HERMeS: a second generation approach to the automatic analysis of two-dimensional electrophoresis gels. Part III: Spot list matching. *Electrophoresis* **8**, 100–107.
7. Vincens, P. and Tarroux, P. (1987) HERMeS: a second generation approach to the automatic analysis of two-dimensional electrophoresis gels. Part IV: Data base organization and management. *Electrophoresis* **8**, 173–186.
8. Tarroux, P., Vincens, P., and Rabilloud, T. (1987) HERMeS: a second generation approach to the automatic analysis of two-dimensional electrophoresis gels. Part V: Data analysis. *Electrophoresis* **8**, 187–199.
9. Miller, M. J., Vo, P. K., Nielsen, C., Geiduschek, E. P., and Xuong, N. H. (1982) Computer analysis of two-dimensional gels: semi-automatic matching. *Clin. Chem.* **28**, 867–875.
10. Skolnick, M. M., Sternberg, S. R., and Neel, J. V. (1982) Computer programs for adapting two-dimensional gels to the study of mutation. *Clin. Chem.* **28**, 969–978.
11. Vo, K. P., Miller, M. J., Geiduschek, E. P., Nielsen, C., Olson, A., and Xuong, N. H. (1981) Computer analysis of two-dimensional gels. *Analyt. Biochem.* **112**, 258–271.
12. Vincens, P. (1993) Morphological grayscale reconstruction in image analysis. *IEEE Trans. Image Proc.* **2**, 176–201.
13. Lutin, K. W. A., Kyle, C. F., and Freeman, J. A. (1978) Quantitation of brain proteins by computer-analyzed two dimensions electrophoresis, in *Electrophoresis '78* (Catsimpoles, ed.), *Developments in Biochemistry*, vol. 2, Elsevier, NY, pp. 93–106.
14. Garrels, J. (1979) Two-dimensional gel electrophoresis and computer analysis of proteins synthesized by clonal cell lines. *J. Biol. Chem.* **254**, 7961–7977.
15. Taylor, J., Anderson, N. L., and Anderson, N. G. (1981) A computerized system for matching and stretching two-dimensional gel patterns represented by parameter lists, in *Electrophoresis '81* (Allen, R. A. and Arnoud, P., eds.), W de Gruyter, NY, pp. 383–400.
16. Tarroux, P. (1983) Analysis of protein patterns during differentiation using 2-D electrophoresis and computer multidimensional classification. *Electrophoresis* **4**, 63–70.
17. Daubechies, I. (1992) Ten Lectures on Wavelets. Society for Industrial and Applied Mathematics, Philadelphia, PA.
18. S-Plus (2000) Data Analysis Products Division, MathSoft, Seattle, WA.

19. Donoho, D. L. and Johnstone, I. M. (1994) Ideal spatial adaptation via wavelet shrinkage. *Biometrika* **81**, 425–455.
20. Donoho, D. L. and Johnstone, I. M. (1995) Adapting to unknown smoothness via wavelet shrinkage. *J. Am. Statist. Assoc.* **90**, 1200–1224.
21. Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D. (1995) Wavelet shrinkage: asymptopia? *J. Roy. Statist. Soc. Ser. B* **57**, 301–369.
22. Horgan, G. W. and Glasbey, C. A. (1995) Uses of digital image analysis in electrophoresis. *Electrophoresis* **16**, 298–305.
23. Anderson, N. L., Hofmann, J. P., Gemmell, A., and Taylor, J. (1984) Global approaches to quantitative analysis of gene-expression patterns observed by use of two-dimensional gel electrophoresis. *Clin. Chem.* **30**, 2031–2036.

2-DE Databases on the World Wide Web

Christine Hoogland, Khaled Mostaguir, and Ron D. Appel

1. Introduction

With the development of the Internet, a growing number of two-dimensional gel electrophoresis (2-DE) databases become available (50 in January 2004, for a total of 270 image maps). By linking the two components constituting 2-DE databases—images and protein information—the active hypertext links provide a powerful tool for data integration, in addition to navigation from one database to another. This chapter shows how to remotely retrieve data from these databases over the Internet.

2. Materials

1. Computer connected to the Internet, such an Apple Macintosh®, a Windows-based personal computer (PC), or a Unix X-Windows system.
2. Worldwide web (WWW) browser such as Mozilla, Netscape® Navigator, or Microsoft® Internet Explorer. Free copies of WWW browsers can generally be downloaded from Internet. *See*, for example, <http://www.mozilla.org/>.
3. Optionally, ImageMaster® 2D Platinum, Amersham Biosciences. A free viewer can be downloaded from Internet (<http://www.expasy.org/melanie/>).

3. Methods

There are basically two ways of retrieving data from a 2-DE database over the Internet: querying and browsing the database using a WWW browser, or directly linking spot data on a gel image within a 2-D analysis software such as ImageMaster 2D Platinum to corresponding entries in the remote database. The latter method also requires a WWW browser.

3.1. Querying a 2-DE Database

An up-to-date list of available 2-DE databases is given in the WORLD-2DPAGE index (URL: <http://www.expasy.org/ch2d/2d-index.html>), referencing more than 270 maps. Moreover, an additional document describes how to link to specific entries of these 2-DE databases (URL: <http://www.expasy.org/ch2d/2d-access.html>). Currently, these federated 2-DE databases (**1**) can be accessed by different ways (*see Note 1*):

1. Remotely queried on the WWW.
2. Attainable through a search in the Swiss-Prot protein sequence database.
3. Linked to other 2-DE databases through hypertext links.

4. Queried graphically by clicking on a spot in a 2-DE gel image.
5. Directly reached from within 2-D analysis software.

To query a 2-DE database over the Internet, run your favorite WWW browser and perform the following operations:

1. Go to WORLD-2DPAGE. Choose Open Location in menu File and enter the following URL: <http://www.expasy.org/ch2d/2d-index.html>. Click OK. The WORLD-2DPAGE document will be retrieved from the ExPASy WWW server (2,3).
2. Go to 2-DE Database. In the list of federated 2-DE databases, select the database of your choice and click on it. The database's home page will be requested from the corresponding Web server. If you have selected the SWISS-2DPAGE database (4), then you may choose among the following options to query the database (see Note 1 for other 2-DE databases):
 - by keyword (description, accession number, author, spot serial number, full text search);
 - by clicking on a spot;
 - using SRS, the Sequence Retrieval System, for complex queries;
 - retrieve protein list for a reference map.

3.1.1. Query by Keyword (Description, Accession Number, Author, Spot Serial Number, Full Text Search)

1. Select option. In the section named "Access to SWISS-2DPAGE," click on the keyword option you want to search, currently by description (any word in the DE, OS, GN, and ID lines), by accession number (AC lines), by author (RA lines), by spot serial number (2D and 1D lines), by full text search. This requests the corresponding keyword search page from ExPASy.
2. Enter keyword. Once the keyword search page is displayed on your screen, enter a keyword of your choice (or partial word) in the designated field (see Note 2). For example, you may select the full text search and enter the terms *heat shock* AND { *hsp70* OR *hsc70* } in order to retrieve all SWISS-2DPAGE entries containing the terms *heat shock* and either *hsp70* or *hsc70* anywhere in the text. Click the return key. The browser will send the query to ExPASy and will then display the result.
3. Select database entry. The result page shows all entries in the database that match the queried keyword(s). The entries are listed by their short name with full description including gene name and species (Fig. 1). Select one of them and click on it. For example, click on HS7C_HUMAN. This will retrieve the complete entry from SWISS-2DPAGE. Various types of information are shown in this document, such as general information about the entry (protein name, accession number in the database, full description, gene name, organism origin, and taxonomic information), bibliographical references, as well as active cross-references to other related databases and active icons to available SWISS-2DPAGE master images (Figs. 2, 3).
4. Compute pI/Mr. To obtain the theoretical pI and Mr of the protein, click on Compute the theoretical pI/Mw (see Note 3).
5. Get gel image with spot. To get the full-size image of one of the available master protein maps, and to see the exact location of the protein, click on the corresponding icon in the section called "2-D PAGE maps for identified proteins." For example, you may select the NUCLEOLI_HELA_2D_HUMAN map.
6. Get gel image without spot. To retrieve just the theoretical region, in which you may expect to find the protein on any of the master gels the protein has not been identified in, click on the corresponding icon in the section called "2-D PAGE maps for unidentified proteins." For example, click on CSF.

ExPASy Home page Site Map Search ExPASy Contact us SWISS-2DPAGE

Hosted by SIB Switzerland. Mirror sites: Australia Bolivia Canada China Korea Taiwan USA

Search swiss-2dpage for heat shock AND {hsp70 OR hsc70} Go Clear

Search
[\[by description\]](#)
[\[by accession number\]](#)
[\[by clicking on a spot\]](#)
[\[by author\]](#)
[\[by serial number\]](#)
[\[by full text search\] >](#)
[\[SRS\]](#)

Search in SWISS-2DPAGE for: heat shock AND {hsp70 OR hsc70}

(Release 16 and updates up to 17-Nov-2003)

Enter search keywords:

 Prefix and append wildcard *** to words.

By default, this search engine searches for complete words only. If you did not find what you expected, and would try to do a substring match, you should perform a new search and select 'prefix and append wildcard to words'.

- Number of documents found in SWISS-2DPAGE: 8
- Note that the selected sequences can be saved to a file to be later retrieved; to do so, go to the [bottom](#) of this page.
- For more directed searches, you can use the Sequence Retrieval System [SRS](#)

DNAK_ECOLI (P04475)
Chaperone protein dnaK (Heat shock protein 70) (Heat shock 70 kDa protein) (HSP70). (GENE: DNAK OR GRPF OR GROP OR BEG OR B0014 OR C0019 OR Z0014 OR EC80014) - *Escherichia coli*

H104_YEAST (P31539)
Heat shock protein 104. (GENE: HSP104 OR YLL026W OR L0948) - *Saccharomyces cerevisiae* (Baker's yeast)

HS71_ARATH (P22953)
Heat shock cognate 70 kDa protein 1 (Hsc70.1). (GENE: HSC70-1 OR HSP70-1 OR AT5G02500 OR T22P11.90) - *Arabidopsis thaliana* (Mouse-ear cress)

HS71_HUMAN (P08107)
Heat shock 70 kDa protein 1 (HSP70.1) (HSP70-1/HSP70-2). (GENE: (HSPA1A OR HSPA1) AND HSPA1B) - *Homo sapiens* (Human)

HS73_ARATH (D65719)
Heat shock cognate 70 kDa protein 3 (Hsc70.3). (GENE: HSC70-3 OR HSP70-3 OR AT3G09440 OR F11FB OR F3L24.33) - *Arabidopsis thaliana* (Mouse-ear cress)

HS7C_DICDI (P36415)
Heat shock cognate protein (Aginactin). (GENE: HSPB OR HSC70) - *Dicytostelium discoideum* (Slime mold)

HS7C_HUMAN (P11142)
Heat shock cognate 71 kDa protein. (GENE: HSPB8 OR HSPA10 OR HSC70 OR HSP73) - *Homo sapiens* (Human)

HS7C_MOUSE (P08109)
Heat shock cognate 71 kDa protein. (GENE: HSPB8 OR HSC70 OR HSC73) - *Mus musculus* (Mouse)

If you would like to retrieve all the SWISS-2DPAGE entries contained in this list, you can enter a file name. These entries will then be saved to a file under this name in the directory [outgoing](#) of the ExPASy anonymous ftp server, from where you can download it. (Please note that this temporary file will only be kept for 1 week.)

File name:
 or

Fig. 1. Entries in SWISS-2DPAGE that match the query “heat shock AND {hsp70 OR hsc70}” anywhere in the text. The protein entries information can be saved in a temporary file, to be downloaded later. Quick jumps to other SWISS-2DPAGE search engines are also provided.

7. Access related databases. In the cross-reference section, you may click on any of the links to retrieve corresponding entries, either from the Swiss-Prot protein sequence database (5) (from which you have access to numerous other databases (6)) or from other 2-DE databases. For example, from the SWISS-2DPAGE HS7C_HUMAN entry, you may directly reach the related entry in HSC-2DPAGE, PHCI-2DPAGE, and OGP-WWW 2D federated databases.

3.1.2. Query by Clicking on a Spot

This method lets you choose one of the master gel images that are available on the server, display the gel image and select one protein entry by clicking on the corresponding spot:



General information about the entry

[View entry in original SWISS-2DPAGE format](#)

Entry name **HS7C_HUMAN**

Primary accession number **P11142**

Entered in SWISS-2DPAGE in Release 09, January 1999

Last modified in Release 16, May 2003

Name and origin of the protein

Description Heat shock cognate 71 kDa protein.

Gene name(s) HSPAB OR HSPA10 OR HSC70 OR HSP73

From Homo sapiens (Human). [TaxID: 9606]

Taxonomy Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.

References

[1] MAPPING ON GEL.
Deruelle-Annessi I., Sanchez J.-C., Hoogland C., Rouge V., Blinz P.-A., Appel R.D., Hochstrasser D.F.; Submitted (JAN-1999) to the SWISS-2DPAGE database.

[2] MAPPING ON GEL.
MEDLINE=20529951; PubMed=11079567; [NCBI, ExPASy, EBI, Israel, Japan]
Jung E., Hoogland C., Chiappe D., Sanchez J.-C., Hochstrasser D.F.;
"The establishment of a human liver nuclei 2-DE reference map.";
Electrophoresis 21:3483-3487(2000).

[3] MAPPING ON GEL.
PubMed=12429849; [NCBI, ExPASy, EBI]
Scherl A., Coute Y., Deon C., Calle A., Kindbeiter K., Sanchez J.-C., Greco A., Hochstrasser D.F., Diaz J.-J.;
"Functional proteomic analysis of human nucleolus.";
Mol. Biol. Cell. 13:4100-4109(2002).

2D PAGE maps for identified proteins

[Compute the theoretical pI/Mw](#)

[How to interpret a protein map](#)

Fig. 2. Top of the HS7C_HUMAN entry (Nice View) in SWISS-2DPAGE (showing the protein short name, accession number, creation and modification date, description, gene name, organism and taxonomy, as well as bibliographical references).

1. Select option. In the section named Access to SWISS-2DPAGE, click on "by clicking on a spot." This requests the SWISS-2DPAGE map selection page from ExPASy.
2. Select master map. Once the SWISS-2DPAGE map selection page is displayed on your screen, select the master reference protein map of your choice by clicking on the respective icon. For example, you may click on the *Arabidopsis thaliana* icon in order to retrieve the image of the *Arabidopsis thaliana* master gel.
3. Select a spot. On the master image, all identified proteins are marked by a small, red cross. Among the marked spots, click on the protein of your choice. The related entry from SWISS-2DPAGE will be requested from ExPASy and displayed on the screen.
4. Access related data. Access additional data as outlined above in **Subheading 3.1.1., steps 4-7.**

3.1.3. Query Using SRS

1. Select option. In the section named Access to SWISS-2DPAGE, click on "SRS, using the Sequence Retrieval System." This requests the SRS search page from ExPASy.

2D PAGE maps for identified proteins

Compute the theoretical pI/Mw

How to interpret a protein map

Colorectal
adenocarcinoma cell
line (DL-1)

MAP LOCATIONS:

- SPOT 2D-001DK7: pI=5.18, Mw=69584

PEPTIDE MASSES:

- SPOT 2D-001DK7: 1253.53; 1481.7; 1487.59; 1981.83; TRYPSIN.

- MAPPING: MASS FINGERPRINTING [1].

Soluble nuclear
proteins and matrix
from liver tissue

MAP LOCATIONS:

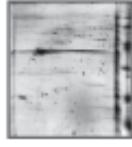
- SPOT 2D-001J00: pI=5.38, Mw=67421

- SPOT 2D-001J03: pI=5.30, Mw=67421

PEPTIDE MASSES:

- SPOT 2D-001J00: 861.415; 1074.472; 1081.553; 1199.682; 1228.633; 1235.634; 1251.64; 1253.63; 1426.68; 1481.81; 1487.708; 1659.848; 1691.739; 1773.9; 1982.006; 2774.294; TRYPSIN.
- SPOT 2D-001J03: 861.437; 882.408; 993.495; 1074.491; 1081.569; 1175.629; 1199.693; 1228.643; 1235.622; 1251.612; 1253.635; 1303.613; 1319.643; 1410.693; 1426.687; 1481.828; 1487.733; 1616.792; 1632.807; 1649.822; 1653.857; 1659.888; 1665.848; 1691.752; 1982.007; 2260.15; TRYPSIN.

- MAPPING: MASS FINGERPRINTING [2].

2D-PAGE of nucleolar
proteins from Human
HeLa cells

MAP LOCATIONS:

- SPOT 2D-001VGX: pI=5.50, Mw=70432

PEPTIDE MASSES:

- SPOT 2D-001VGX: 1081.7; 1197.8; 1199.8; 1228.8; 1251.8; 1253.8; 1426.8; 1480.9; 1487.9; 1632.9; 1666.0; 1691.9; 1746.0; 1788.1; 1982.2; 2263.4; 2697.6; 2774.6; TRYPSIN.

PEPTIDE SEQUENCES:

- SPOT 2D-001VGX: (R)GTLDPEVKE(A),312-319; (K)DAGTIAGLNVLRI(I),160-171; (K)VEIILANDQGNR(I),26-36; (K)NSLESYAFNMaxK(A),540-550; (R)RFDDAVVQSDMaxK(H),77-88; (R)TPSYVAFTDTER(L),37-49; (K)SFYPEEVSSMaxV(LTK)(M),113-128; (R)IIINEPTAAIAAYGLDKK(V),172-188; (K)TVTNAVTVTPAYFNDSQR(Q),138-155; (K)GPAVGIDLGTTYSCVGVFQHGK(V),4-25.

- MAPPING: PEPTIDE MASS FINGERPRINTING AND TANDEM MASS SPECTROMETRY [3].

Copyright

This SWISS-2DPAGE entry is copyright the Swiss Institute of Bioinformatics. There are no restrictions on its use by non-profit institutions as long as its content is in no way modified and this statement is not removed. Usage by and for commercial entities requires a license agreement (See <http://www.isb-sib.ch/announce/> or send an email to license@isb-sib.ch).

Cross-references

SWISS-PROT	P11142 ; HS7C_HUMAN.
HSC-2DPAGE	P11142 ; HUMAN.
PHC1-2DPAGE	P11142 ; HS7C_HUMAN.
OGP-WWW	P11142 ; -.

2D PAGE maps for unidentified proteins

- How to interpret a protein map

Fig. 3. Second part of HS7C_HUMAN entry (Nice View) in SWISS-2DPAGE (with for each reference map, active icons to the full size master images, spots location, experimental data, mapping procedure, as well as links to Swiss-Prot and other two-dimensional gel electrophoresis databases).

2. Select database. Press the Start button to “Start a new SRS session.” Click on the Swiss-Prot and TREMBL check boxes to deselect them. Click on SWISS-2DPAGE check box to select it. Press the Continue button.
3. Enter keyword. Once the “SRS: Query Form Page” is displayed on your screen, enter one keyword of your choice in the first field (see Note 4). For example, you may enter the

keyword *polymerase* in order to retrieve all SWISS-2DPAGE entries containing the word *polymerase* anywhere in the text. Press the Do query button. The browser will send the query to ExPASy and will then display the result.

4. Select database entry. The result page shows all entries in the database that reference the given keyword. The entries are listed by their short name as described in **Subheading 3.1.1., step 3.**
5. Access related data. Access additional data as outlined above in **Subheading 3.1.1., steps 4–7.**

3.1.4. Retrieve Protein List for a Reference Map

1. Select option. In the section named Access to SWISS-2DPAGE, click on “retrieve all the protein entries identified on a given reference map.” This requests the page from ExPASy that allows getting protein list for a map.
2. Select reference map. Select in the list the reference map for which you want the list of identified proteins. You may choose to sort the protein list according to their accession number or short name. Click the button “click here to compute the list.” The browser will send the query to ExPASy and will then display the result (**Fig. 4**).
3. Select database entry. The result page shows all identified proteins on the chosen reference map. The entries are listed with general information (gene name, protein description, accession number) and all 2-DE information (spot serial number, experimental pI and Mw –read from gel-, mapping procedure, references).
4. Access related data. Access additional data as outlined above in **Subheading 3.1.1., steps 4–7.**

3.2. Linking Spot Data to a 2-DE Database

Linking spot data to a 2-DE database allows you to automatically retrieve data from a 2-DE database while analyzing 2-D gel images using the ImageMaster 2D Platinum analysis software (*see Note 5*). Open or create an image gel with at least one label with a valid accession number, as described in Chapter 27, then follow these steps:

1. Setting the browser. You have to select your favorite browser. Choose Edit → Settings from the menu and select the desired software in the Internet tab.
2. Setting the database. You have to define the database Web server and database query engine that allow querying the remote database. The database should be a federated 2-DE database (**1**). An example of such a database is SWISS-2DPAGE. Choose Edit → Settings → Categories → AC and enter the URL in the External engine field. For SWISS-2DPAGE on ExPASy (this is the default), enter: <http://www.expasy.org/cgi-bin/nice2dpage.pl>?
3. Querying the database. To query the remote database, select one feature (spot) that contains a label with a valid accession number (AC). When double-clicking on the label, the ImageMaster 2D Platinum analysis software will automatically launch your favorite browser with a query consisting of the concatenation of the HTTP address, query engine, and accession number. As a result, the entry for the protein with the given accession number will be displayed.
4. Access related data. Access additional data as outlined above in **Subheading 3.1.1., steps 4–7.**

4. Notes

1. Other 2-DE databases may have slightly different interfaces, but all 2-DE databases that are federated (**1**) provide a means to query the database by keyword search, to link to other databases, and to access protein data by clicking on a spot on the gel image or directly

Protein list for: *Arabidopsis thaliana* (ARABIDOPSIS)
Sorted by: Accession Number

(Release 16 and updates up to 17-Nov-2003)

Gene Name	Protein Description	SWISS-2DPAGE Serial Number	SWISS-2DPAGE Access Number	Identification method	Exp. pI	Exp. Mw	References
RBC1	Ribulose bisphosphate carboxylase large chain (EC 4.1.1.39) (RuBisCO large subunit)	2D-001K6Q	003042	PMF, Gm	5.71	50664	1, 2
		2D-001K74			5.64	51074	
		2D-001K7A			5.78	51284	
		2D-001K9S			6.15	47034	
		2D-001K8M			6.10	43852	
		2D-001KHU			6.02	33897	
		2D-001KJ2			5.92	32050	
		2D-001KGP			5.55	30287	
		2D-001KL6			5.79	29705	
		2D-001KLE			5.79	29370	
CRT1	Catreticulin 1	2D-001K5F	004151	Gm	4.41	55005	1, 2
AT3G16470	Myrosinase binding protein-like At3g16470	2D-001KAA	004309	PMF	5.22	46361	2
AT2G47470	Probable protein disulfide isomerase AB (EC 5.3.4.1) (P5)	2D-001KEG	022263	PMF	8.12	39238	2
CIMS	5-methyltetrahydropteroylglutamate-homocysteine methyltransferase (EC 2.1.1.14) (Vitamin-B12-independent methionine synthase isozyme) (Cobalamin-independent methionine synthase isozyme)	2D-001KDN	050008	Gm	8.15	78088	1, 2
CPN21	20 kDa chaperonin, chloroplast (Protein Cpn21) (Chloroplast protein Cpn10) (Chloroplast chaperonin 10) (Ch-CPN10) (Chaperonin 20)	2D-001KMY	085282	PMF	5.25	27447	2
GDCST	Aminomethyltransferase, mitochondrial (EC 2.1.2.10) (Glycine cleavage system T protein) (GCVT)	2D-001KCA	085396	PMF	6.83	42694	2
HSC70-3	Heat shock cognate 70 kDa protein 3 (Hsc70.3)	2D-001K1H	085719	PMF, Gm	4.92	64425	1, 2
AT2G37880	Protein At2g37880, chloroplast	2D-001KXM	080934	PMF	5.26	30338	2
AT2G38330	Myrosinase binding protein-like At2g38330	2D-001KRW	080948	PMF	6.23	46937	2
VSP2	Vegetative storage protein 2	2D-001KV	082122	PMF	6.47	29848	2
RCA	Ribulose bisphosphate carboxylase/oxygenase activase, chloroplast (RuBisCO activase) (RA)	2D-001K9K	P10895	PMF	5.01	47034	2
CHLI	Magnesium-chelatase subunit chl, chloroplast (Mg-protoporphyrin IX chelatase) (Protein C5/CH-42)	2D-001KCZ	P16127	PMF	8.33	41395	2

Fig. 4. List of SWISS-2DPAGE entries identified on the *Arabidopsis thaliana* master gel, presented in a table with general (gene name, protein description, accession number) and experimental information (spot number, pI, Mw, identification method).

- from a 2-D PAGE analysis software package such as the one discussed in Chapter 27. There are currently approx 40 databases that are either fully or partially federated.
- With the full text search, you may use AND, OR, and NOT to restrict your search, with braces (i.e., { and }) to specify the order. Note that parentheses (i.e., (and)) may be part of certain words in SWISS-2DPAGE (e.g., nad(+)) and can therefore not be used to group your search expressions. If words are given without any boolean operator (AND, OR, NOT), these words will be searched as adjacent words in the database. With the search by description, the boolean operators are not required. Thus, entering more than one keyword in that case will return entries with both the individual queries.
 - Protein pI is calculated using pK values of amino acids described in (7,8), which were defined by examining polypeptide migration between pH 4.5 and 7.3 in an immobilized pH gradient gel environment with 9.2 M and 9.8 M urea at 15°C or 25°C. Prediction of protein pI for highly basic proteins is yet to be studied, and it is possible that current pI predictions may not be adequate for this purpose.

4. The Sequence Retrieval System can be used for more directed or combined searches. For example, with SRS you may search for SWISS-2DPAGE entries that contain the word *reductase* in the Description field AND *Homo sapiens* in the Organism field. The Boolean operators AND, OR, and BUTNOT are available from the drop-down list box adjacent to the Do query button for combined searches.
5. For querying a database using a WWW browser as described in **Subheading 3.1.** (“Querying a 2-DE database”), the ImageMaster 2D Platinum software is not necessary. You may access the 2-DE databases using just the browser software. However, to link spot data to 2-DE databases as presented in **Subheading 3.2.** (“Linking spot data to a 2-DE database”), both ImageMaster 2D Platinum and a WWW browser are mandatory.

References

1. Appel, R. D., Bairoch, A., Sanchez, J. C., et al. (1996) Federated 2-DE database: a simple means of publishing 2-DE data. *Electrophoresis* **17**, 540–546.
2. Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R., and Bairoch, A. (2003) ExPASy the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.* **31**, 3784–3788.
3. Hoogland, C., Sanchez, J. C., Walther, D., et al. (1999) Two-dimensional electrophoresis resources available from ExPASy. *Electrophoresis* **20**, 3568–3571.
4. Hoogland, C., Mostaguir, K., Sanchez, J. C., Hochstrasser, D. F., and Appel, R. D. (2000) SWISS-2DPAGE, ten years later. *Proteomics* **4**, 2352–2356.
5. Boeckmann, B., Bairoch, A., Apweiler, R., et al. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 365–370.
6. Gasteiger, E., Jung, E., and Bairoch, A. (2001) SWISS-PROT: connecting biological knowledge via a protein database. *Curr. Issues Mol. Biol.* **3**, 47–55.
7. Bjellqvist, B., Hughes, G. J., Pasquali, C., et al. (1993) The focusing positions of polypeptides in immobilized pH gradients can be predicted from their amino acid sequences. *Electrophoresis* **14**, 1023–1031.
8. Bjellqvist, B., Basse, B., Olsen, E., and Celis, J. E. (1994) Reference points for comparisons of two-dimensional maps of proteins from different human cell types defined in a pH scale where isoelectric points correlate with polypeptide compositions. *Electrophoresis* **15**, 529–539.

Computer Analysis of 2-D Images

Patricia M. Palagi, Daniel Walther, Gérard Bouchet, Sonja Voordijk, and Ron D. Appel

1. Introduction

Indeed, the development of two-dimensional (2-D) electrophoresis technology has provided an important research tool for comparative proteomics. A number of other technologies for protein separation have also been investigated in the last few years, such as protein chips, liquid chromatography, and high-pressure liquid chromatography, to name just a few. However, 2-D gels are still considered the workhorse among research laboratories willing to accurately track changes in protein expression and to achieve differential proteomics.

The unlimited possibilities of 2-D applications, such as finding targets for drug treatments or biomarkers for medical diagnostics and prognostics, have been a major incentive to the development of specialized software systems. Their major utilities are to accurately detect and quantify protein spots over the gels, to find corresponding proteins across gels, to highlight significant protein expression changes, and to statistically validate the variability found.

Even though 2-D image-analysis tools have been significantly improved in terms of speed, accuracy, and throughput, they need to be continuously updated and enhanced to keep up with the development of related proteomics technologies. Coupling of 2-D and mass-spectrometry techniques, sample prefractionation, and high throughput, for instance, necessarily increase the number of gels to be analyzed (1). As a consequence, automated analysis of gel images with no human intervention would be greatly appreciated, to cite just one of many other challenges that the new generation of image-analysis tools are facing now. High financial and manpower investments are needed to develop those tools, which explains the lack of completely free versions and the existence of several commercial software packages (2–8). Although these packages differ in various aspects, such as the graphical interface, the file format, or the sequence of steps needed to fully process the gels, most of them have the functions necessary to perform the procedures described in this chapter. The logic behind the sequential operations performed in a 2-D gel analysis, necessary to find out differently expressed proteins, is explained hereafter with the assistance of the Melanie software version 4 (9,10).

2. Materials

2.1. Software

The Melanie software (version 4) was developed and is continuously updated at the Swiss Institute of Bioinformatics. It is marketed under the name ImageMaster™ 2D Platinum by GE Healthcare in collaboration with GeneBio* (1). A demonstration version of the program and support documentation are freely available from GeneBio's website (www.genebio.com). Melanie Viewer, a reduced version of this software, can be freely downloaded from the ExPASy server (<http://www.expasy.org>), and most of the procedures described hereafter may be done with the Viewer.

2.2. Hardware

The Melanie software runs on any of the current Windows operating systems (98, ME, NT, 2000, XP).

The minimum recommended virtual memory is 256 MB, which is enough to open and process a large number of gels.

2.3. Images

Before 2-D gels can be treated by image analysis software, they first have to be digitized into an adapted data format. Gel images may be produced with a variety of scan devices, including flatbed document scanners, camera systems, densitometers, phosphor imagers, and fluorescence scanners. The default output format for most imaging equipment, and certainly the most appropriate for further analysis by 2-D software, is TIFF (Tag Image File Format, Aldus Corporation). This is the recommended format for use with Melanie, although the software can read some other file types.

It is important to save images with sufficient resolution and depth, as these influence the amount of visible detail in the image. Taking into consideration the currently available computer systems, an image resolution of about 200 dpi and depth of 16 bits is a good compromise.

3. Methods

3.1. Melanie's Interface

Melanie deals with defined objects—the gels and their components: spots, annotations, labels, and regions. Its display is divided in six parts: the Menu bar, the Toolbar, the Reference gel zone, the Hidden gels zone, the Gel display zone, and the Status bar. The objects may be displayed or hidden, selected or unselected, and processed using the options from the Menu bar and the Toolbar. The Toolbar has the following buttons: the Workspace tool (to organize gels into projects), the Hand tool (to move gels), the Magnify tool (to zoom gels), the Region tool (to select rectangular regions over gels), the Spot tool (to select and edit spots), and the Annotation tool (to add and select annotations). Drag a gel to the Reference gel zone and it will become the reference gel (see Note 1). Drag a gel to the Hidden zone and the gel will disappear from the Display zone, even though it is open. The Status bar indicates the number of gels, spots, and

*Melanie 4 is currently also available from Bruker Daltonics, integrated with their PROTEINEER spII spot picking robot.

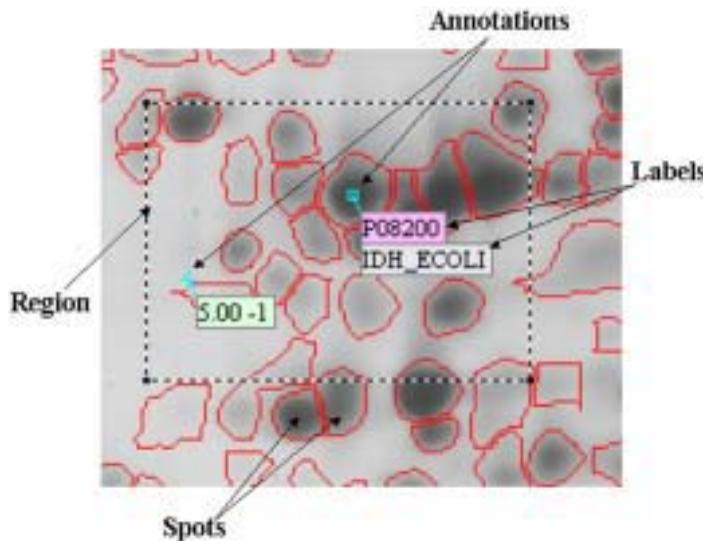


Fig. 1. Part of a gel image with a defined region, spots, annotations, and labels.

annotations selected. **Figures 1** and **2** briefly illustrate Melanie's objects, display, tools, and operators necessary to start working with it.

3.2. Setting Up an Experiment

1. Open the gel images with the Import function to start a work session with Melanie. If necessary, choose a reduction factor to modify the image dimensions of large scanned images.
2. Click on the Melanie Workspace icon in the toolbar to display the Melanie Workspace window and create a new workspace (see **Note 2**).
3. Inside the workspace, create a new project with a chosen name—for example, “Disease A Experiment.”
4. Select the control gels and add a new class to the project with the name “Control.” The selected gels will immediately be allocated to the new class.
5. Repeat the same procedure for the disease gels and create a new class called “Disease.” This means that the gels are now organized into two populations: “Control” and “Disease.” If necessary, repeat the same procedure for all the other classes of the experiment.
6. Choose the best representative gel among all gels and set it as the reference gel.
7. Position the mouse cursor on the class names and right-click to open the classes and make these settings active.
8. Click on the Save icon on the Workspace window to save the new workspace and the projects.

3.3. Detecting and Quantifying Spots

The elementary component of a gel image is the spot—a shape that can be automatically detected by a spot-detection algorithm or manually adjusted by the user. It delimits a region in the gel where a protein or a mixture of proteins is present. In Melanie, each spot in a gel has an associated spot ID (a unique sequential number) automatically assigned to it when it is created. Besides, a spot is automatically quantified—i.e., its

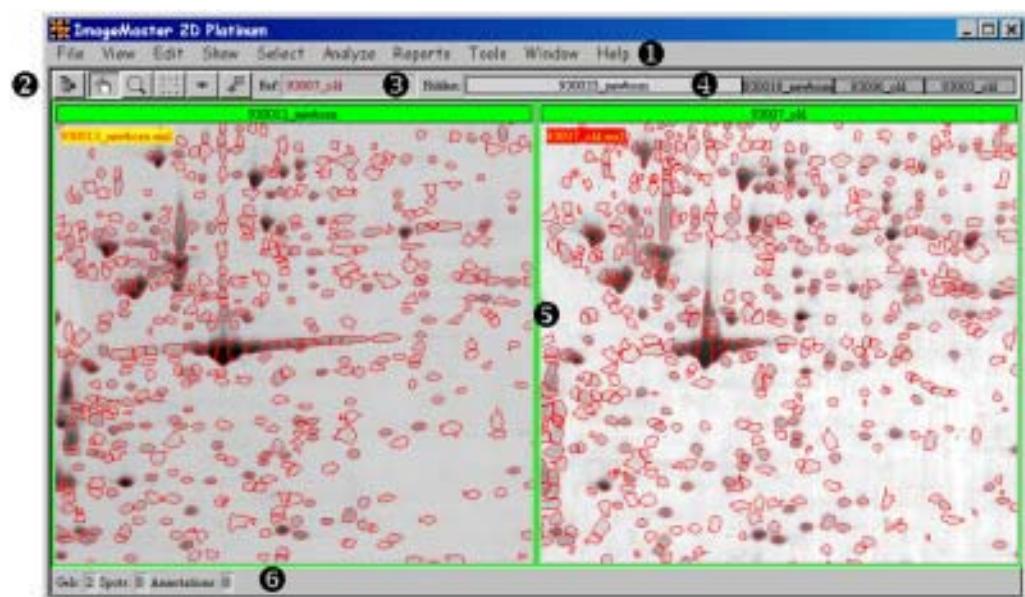


Fig. 2. The Melanie program window. (1) Menu bar, (2) toolbar, (3) reference gel zone, (4) hidden gels zone, (5) gel display zone, and (6) status bar.

optical density, area, and volume are computed. Follow the next steps to detect spots automatically:

1. Use the Hand tool to move the gels to an area of interest.
2. Select the gels from which spots will be detected.
3. Select the Region tool and draw a region on one or more selected gels in zones with representative spots.
4. Choose *Edit* → *Spots* → *Auto-Detect* from the menu. The *Detect Spots* window appears on the screen, and the spots in the active regions in selected gels are detected with the default parameters.
5. Adjust the detection parameters if necessary. The default parameters are optimized to typical sodium dodecyl sulfate (SDS)-polyacrylamide gel electrophoresis (PAGE); however, refining the saliency parameter may be indicated. Each change in one of the spot-detection parameters is immediately reflected in the selected region helping to choose the parameters. Use also the cursor information window to help find optimal values for the spot filtering.
6. Click on “OK” to detect all spots in the selected gels using the parameter values that have been set. The spot shapes will be displayed over the gels (*see Note 3*).
7. Choose *File* → *Save* → *Spots* to save the result of the detection operation.

3.4. Matching Gels

Once spots have been detected, the next essential step is to match gel images, i.e., to find the corresponding protein spots in different gels. A basic gel-matching algorithm compares two gel images to find Pairs of related spots—in other words, spots describing the same protein in both gels.

In Melanie, matching two gels means finding all the pairs between spots of the two gels. Matching several gels means picking out a reference gel, then successively match-

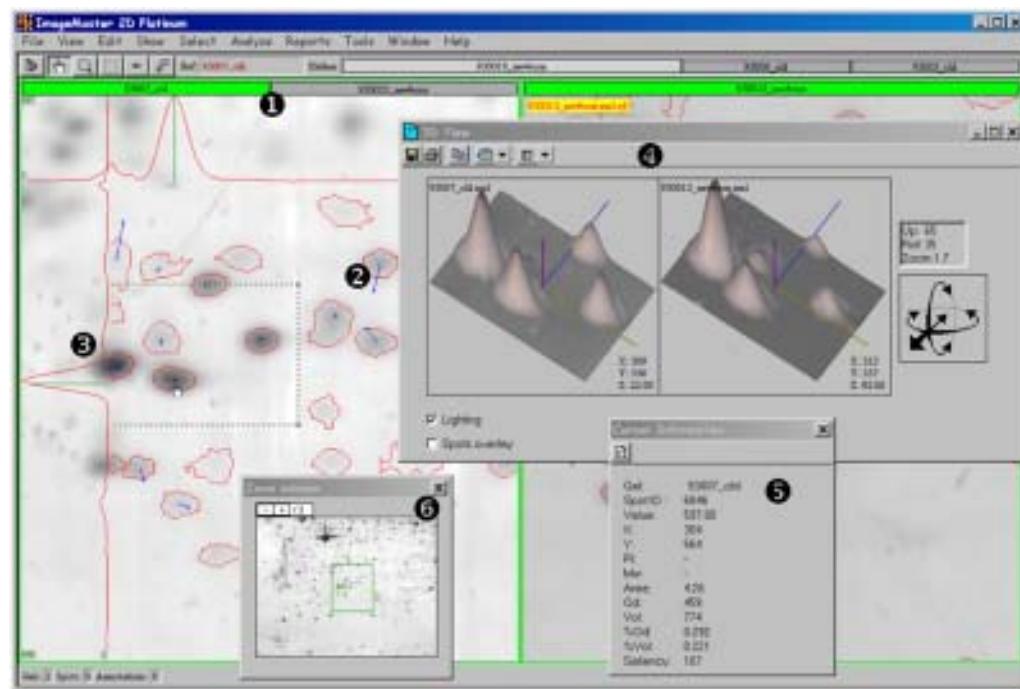


Fig. 3. Two gels are in stack mode (1), gel 930033_newbord is behind gel 93007_old, and the small vectors (2) represent the paired spots between them. The curves (3) in the vertical (left) and horizontal (top) directions represent the profile of the spot under the cursor. A 3D view (4) of selected regions of two gels is displayed. Only one region is seen on the image. The Cursor information window (5) displays the values of the spot under the cursor. The Zoom window (6) localizes the visible part of the gel under the cursor on a larger area.

ing each gel with the reference gel. In this way, spots in all gels may be compared with spots in the reference gel (*see Note 4*).

To match two or more gels automatically:

1. Select the gels to be matched (including the reference gel).
2. Choose **Edit** → **Pairs** → **Auto-Match Gels** from the menu.
3. Set the reference gel for matching in the pop-up dialog box, that is, specify to which image the other gels should be matched, and click **OK**.
4. All selected gels are matched with the chosen reference gel.
5. Choose **File** → **Save** → **Pairs** to save the result of the matching operation.

When matching is completed, Melanie displays the total number of pairs found (*see Note 5*). A report with this information can also be displayed through the **Matches Report** window.

3.5. Displaying Information

3.5.1. Pile Gels

When many gels are open at the same time, or to visualize some of Melanie's operation results, it is helpful to stack gels, i.e., display one gel on top of the others, thus creating a pile of gels. To stack two or more gels, select them and drag them onto one of the display cells (**Fig. 3**).

3.5.2. Show Pairs

When gels are in stack mode, the pairs between the Stack Reference and the other gels of the stack may be displayed in the form of blue vectors linking the locations of paired spots.

1. Stack two or more gels.
2. Select the Stack Reference (for comparison with other gels).
3. Make sure the gels were matched with the Stack Reference.
4. Choose View → Stack → Show Pairs from the menu.
5. Blue vectors are displayed between paired spots in the front gel and the Stack Reference.
6. You may swap the front gel and Stack Reference with Ctrl+F, and vice versa.
7. Click on any other gel in the stack to bring it to the top and compare it with the Stack Reference.
8. Choose View → Stack → Hide Pairs to hide the pair vectors.

3.5.3. View Profiles

The Profile function can help to determine whether spots should be split, or they are saturated, or they have some other problems (see Note 6).

1. Select the gels for which you would like to view profiles.
2. Choose File → Raw Image → Load (see Note 7).
3. Select View → Profile → Show from the menu.
4. Hold the mouse cursor over the desired gel. The horizontal and vertical profiles at the cursor position will be displayed.
5. Once you do not need the profile feature any more, choose View → Profile → Hide from the menu, and unload the raw image data.

3.5.4. Show 3-D View

Looking at the 3-D view of a gel region may help examine the intensity variations in a gel (Fig. 3).

1. Select a region in one or several gels with the Region tool (see Note 8).
2. Select the gels for which you would like to display a 3-D view.
3. Choose Reports → 3-D View from the menu.

Melanie displays a 3-D View for each active region or area around selected spots.

3.5.5. Show Spot Information

The Cursor Information window may be used to display information on pixels (X and Y coordinates, pI and Mw estimates, raw pixel values) and spots, if spots have been detected (Fig. 3).

1. Choose Window → Information from the menu, and the Cursor Information window will appear on the screen.
2. Move the cursor over the pixel of interest. The Cursor Information window displays the desired information (see Note 9). If the Value field shows unloaded, the raw pixel values are unavailable. In this case, choose File → Raw Image → Load to load the image data of the gels.

3.5.6. Zoom Window

The Zoom Window can be used to enlarge parts of gels—that is, to focus on the region under the cursor, or to localize the visible part of the gel on a larger gel area (Fig. 3). To zoom a region under the cursor:

1. Choose Window → Zoom from the menu.
2. Set the desired magnification by pressing the + and – buttons located at the top of the Zoom Window.
3. Move the cursor over the region to have the focus on. The Zoom Window will show an enlarged view of the region under the cursor.

3.5.7. Annotating Spots

Individual pixels and spots in a gel image may be indicated by annotations. In Melanie, annotations are active elements in the gel analyses, and they can have several different purposes—for example, they can be used to calibrate and match gels, or to mark spots with their own information, such as protein name, accession number, and so on. Annotations may also be used to mark spots with common characteristics, thus creating subsets.

An annotation is defined by a set of labels, and each label belongs to a predefined or user-defined category (see **Note 10**). To create an annotation:

1. Activate the Annotation tool.
2. Double-click on the pixel or spot in the gel where the annotation should be located.
3. In the pop-up window, enter the name of a new category or choose one of the existing categories by clicking on its name.
4. When a new category is created, the Create Category window appears.
5. Type the desired label in the next dialog box, and click OK.

The annotation is created and its label is displayed on the gel.

3.6. Finding Protein Variations

Once a number of gels have been matched to a reference gel, the analysis of changes in protein expression over this set of gels may be performed (see **Note 11**). There are numerous ways of finding variations in protein expression with Melanie. Two of them are explained in the following subheadings.

3.6.1. Analysis With Classes Set

In case the populations of gels are known—for example, when comparing gels of control tissues against disease-affected tissue samples—the analysis is based on this classification. Assume, for example, that the classes are defined as described in **Sub-heading 3.2**. The next step is to find out which are the characteristic spots of each class, i.e., the proteins that are differently expressed (**Fig. 4**). The class spot values may be described by statistical and overlapping values, such as mean, standard deviation, gap, ratio, and normalization. One possible way to investigate groups of spots according to these descriptors is:

1. Select the gels and then select the groups to be studied with Select → Groups → All.
2. Choose Analyze → Classes Report.
3. In the pop-up list, select the % Vol value type to be displayed.
4. Accept the default statistics (Mean 100% and Mean squared deviation 100%) in the subsequent dialog box.
5. Change the Displayed value at the top of the Classes Report window for Ratio, rank the report in descending order (by clicking on the column headers), and select the rows showing a ratio from the highest value to two.
6. Create a new report by choosing the Report from selection option in the Classes Report window.



Fig. 4. The rows of gels correspond to two different populations. The group of spots across the gels, marked with the annotation “Validated,” has been found through the analysis proposed **Subheading 3.6.1**. The histogram at the right shows that this protein is differently expressed in both classes, i.e., the protein is over-expressed in the class at the bottom.

7. In the new window report, change the Displayed value for Gap. Rank the gap values in descending order. Select all rows from the Gap report and create a Classes + Groups Histograms.
8. Use the created reports and the Select on Gels function on the reports to verify the pertinence of the given results. Use the green arrows on the report menu to select rows one by one.
9. In the Gap report, create an annotation of category “Set” with name Validated and type Boolean (*see Note 12*). If the results are reliable, select the field Validated in the Gap report.
10. When finished, select all rows in the Gap report, refine the selection using the column “Validated” with value 1, and re-select the spots on the gels and on the displayed reports with the function “Select on gels + reports.” Keep the selected spots and read **Subheading 3.7**.

3.6.2. Blindly Classifying Gels

If the populations of gels are not known, the heuristic clustering procedure may create classes of gels and highlight significant groups of differently expressed proteins (*see Note 13*).

1. Select the gels and then select the groups to be studied with Select → Groups → All.
2. Set the reference gel, if it has not already been selected.
3. Choose Analyze → Heuristic Clustering → Do Clustering.
4. In the pop-up list, select the % Vol value type to be displayed.

5. In the subsequent dialog box, set the Maximum number of classes to the desired number, and accept the proposed Sensitivity parameter.

When the heuristic clustering process is finished, Melanie automatically assigns default classes to the gels included in the analysis and displays a Classes Histograms window for the groups that have been found characteristic for one of the classes. Keep the selected spots proposed by Melanie and read **Subheading 3.7.**

3.7. Validating Data

To help investigate the significance of the resulting spot groups, Melanie provides three statistical tests: two-sample t, Mann-Whitney U, and Kolmogorov-Smirnov tests. Their principle is to calculate the probability of observing data sampled from populations with different means by chance or by fact, and consequently to find out the groups of spots that could possibly differentiate classes of gels. In order to get the statistical test values:

1. Groups of spots have to be selected. If the spots have not been selected through the procedures suggested in **Subheading 3.6.**, select the groups of spots by taking Select → Groups → All.
2. Choose Analyze → Statistical Tests.
3. In the pop-up list, select the % Vol value type.
4. Choose one or more of the statistical tests among the two-sample t, Mann-Whitney U, and Kolmogorov-Smirnov test values to be displayed.
5. Sort, for example, the t-test values in the report in descending order by clicking on the column header on the top of the Statistical Tests Report window.
6. Re-select the first 30 groups that have the highest t-test value to concentrate the analysis on the most significant spot differences between classes.
7. Click the Classes+Groups Histograms button at the top of the Statistical Tests Report window and then on each histogram to check the obtained results.
8. To obtain another view of the results, click the Classes+Groups Histograms button at the top of the Classes Histograms window and then on each histogram to check the obtained results (**Fig. 4**).
9. Mark the resulting spots with labels from the “Set:” category and name t-test (see **Note 12**).

4. Notes

1. A reference gel is the gel chosen to be the representative of all gels and make the connection among them.
2. It is highly recommended to define a workspace, since that facilitates the organization of the gel experiments and it helps to work in a personalized environment. The workspace holds information on the relationships between gels such as their organization into populations (classes) and their reference gel. The workspace allows organizing gels into projects that reflect the structure and design of the experimental studies, facilitating the subsequent work.
3. Once spots have been detected, the amount of protein present in each spot is automatically computed. Among the measured quantitative protein values, the most often used in the analyses is the relative volume (% Vol). It is a normalized value that remains relatively independent of uninteresting variations between gels, particularly those caused by varying experimental conditions. This measure takes into account variations due to protein loading and staining by considering the total volume over all the spots in the gel.

4. All spots in selected gels that are paired with a given spot in the reference gel form a group. A spot group is the basic element for analyzing spot variations across gels, for producing reports and histograms, as well as for performing statistical and clustering analysis. Moreover, when several gels have been matched to a given reference gel, this reference gel provides a unique numbering scheme for spots across all gels. Indeed, each paired spot in a gel image may be associated with the corresponding Spot ID in the reference gel. The Spot ID in the reference gel is then called the Group ID.
5. When the gels are very distorted or different, automatic matching may not be effective. In this case, the matching procedure may be enhanced with the addition of well-placed landmarks. Place two or three landmarks on clearly corresponding spots. Protein variants definitely should not be used as landmarks. Landmarks should be placed on small, sharp spots, rather than on large, diffuse ones (to reduce the error on the position). The choice of landmarks or initial pairs is very important to obtain good matching results. During matching with Melanie version 4.0, landmarks essentially correct global deformations of gels. Therefore, it is recommended not to put landmarks on spots in locally distorted regions, because this can worsen the matching results around such regions.
6. Sometimes spots may have a so-called donut structure, with low intensities in their center compared to their borders. It is very important to identify such problems, as they will lead to incorrect spot quantitation.
7. Whatever the image depth, Melanie re-maps the original gray values to an image with 256 gray levels. Thus, for normal viewing, the raw pixel values of the gel do not have to be kept in memory. This corresponds to the default-display, Unloaded, mode in Melanie. For some operations, such as spot detection, Melanie automatically loads the raw pixel values and immediately unloads them again when the task is done, thus freeing the unused memory. Viewing the Profile of gels or displaying the raw pixel values in the Cursor Information window are actions that need images in this Loaded mode.
8. The Shift key may be used to select the same region in all your open gels, or just delimit a particular region for each gel. Alternatively, just one or several spots in each gel may be selected. In such a view, the *x* and *y* axes, as usual, represent the PI and MW values, whereas the pixel intensity is plotted along the third dimension (*z*-axis). The resulting image shows a peak for each protein spot, with a peak height that is proportional to the spot intensity. It can be rotated in any direction to view the interesting spot(s) from all sides, thus facilitating spot editing or matching decisions.
9. Click on the Settings icon to choose the information to be displayed. In the pop-up dialog box, select the attributes to be shown in the Cursor Information window from the Hidden list (using the Shift or Ctrl keys for multiple selections), and transfer them to the Visible list using the >> button. Similarly, you may remove attributes from the information window by selecting them in the Visible list and clicking on the << button. The selected attributes will be displayed in the Cursor Information window.
10. Annotations offer also the possibility to link and associate gel objects to external query engines or data sources of any format (text, html, spreadsheet, multimedia, 2DE database entry), located locally or on the Internet. Among the available predefined annotation categories, three are worth mentioning. The Ac category is provided to hold the protein's accession number (AC) taken from a user-selected database—for example, SWISS-2DPAGE (11) or Swiss-Prot (12), and can be the link to Melanie's remote database query engine. When such a link is defined, a double-click on a label of this category displays the corresponding protein entry in the chosen database with the selected browser software. A Landmark is a predefined category of annotations used to mark pixels or spots in the gels as reference points, for the calculation of corresponding locations between gels. Two

annotations are considered to refer to the same point in different gels when they hold identical labels (names).

11. Based on the explained procedures, groups composed of spots whose quantification values are unusual may be located. The detected variations can result from protein expression changes among gels, or can be due to an inadequate detection or matching operation. Therefore, the analyses are not only useful for investigating the extracted data, but also for controlling them.
12. The pre-defined category Set: is used to mark spots with common properties by indicating that they belong to a set. The labels in such a category do not contain specific information. They only display the name of the set to which they belong. To create a set: Select one or many spots. Choose Edit → Annotations → Add Labels from the menu. In the pop-up window, click on the category called Set: and add a keyword, which will be the name of the set. For instance, in the specific case of this chapter, the final category name might indicate Set: Validated. A label containing the name of the set (e.g., Validated) will be attached to the selected spots.
13. Heuristic clustering is used to describe the characteristic spots of 2-D gels while using heuristics to speed up the search. The algorithm provides the possibility of blindly classifying similar gels into two or more classes. Depending on your computer, this could take time, especially when the maximum number of classes is high and many spots and gels were selected. The Sensitivity parameter determines the number of characteristic spots. This defines the gap between two classes (the difference between the highest spot value in one class and the lowest value in a second class). The smaller this parameter, the smaller the accepted difference between the class intervals in a group.

References

1. Voordijk, S., Walther, D., Bouchet, G., and Appel, R. D. (2003) Image Analysis Tools in Proteomics. In *Encyclopedia of the Human Genome*, Macmillan Publishers Ltd, Nature Publishing Group, London, pp. 404–410.
2. <http://www.amershambiosciences.com> (ImageMasterTM 2D Platinum and DeCyder software)
3. <http://www.decodon.com> (Delta2D software)
4. <http://www.proteomesystems.com> (ImagepIQ software)
5. <http://www.bio-rad.com> (PDQuest software)
6. <http://www.nonlinear.com> (Progenesis and Phoretix software)
7. <http://www.proteomweaver.com> (ProteomWeaver software)
8. <http://www.2dgels.com> (Z3 and Z4000 software)
9. Appel, R. D., Palagi, P. M., Walther, D., et al. (1997) Melanie II: a third generation software package for analysis of two-dimensional electrophoresis images—I. Features and user interface. *Electrophoresis* **18**, 2724–2734.
10. Appel R. D., Vargas J. R., Palagi P. M., Walther D., and Hochstrasser D. F. (1997) Melanie II: a third generation software package for analysis of two-dimensional electrophoresis images—II. Algorithms. *Electrophoresis* **18**, 2735–2748.
11. Hoogland, C., Sanchez, J.-C., Tonella, L., et al. (2000) The 1999 SWISS-2DPAGE database update. *Nucleic Acids Res.* **28**, 286–288.
12. Boeckmann, B., Bairoch, A., Apweiler, R., et al. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 365–370.

Comparing 2-D Electrophoretic Gels Across Internet Databases

An Open Source Application

Peter F. Lemkin, Gregory C. Thornwall, and Jai Evans

1. Introduction

In (1–4) and in the previous edition of this book (5), we described a Web-based computer-assisted visual method called Flicker for comparing two-dimensional (2-D) protein gel images across the Internet using a Java applet. We originally used the flicker method in the GELLAB 2-D gel analysis system (6–9). The applet (primarily for Web-based analysis using a Web browser) has since been converted to a Java stand-alone application that can run on a user's computer. The new application also can access images from the Web but can now more easily access the user's data on their local computer. The new version is available as open source on <http://open2dprot.sourceforge.net/Flicker>, where the executable program and the source code are available. Some of the code was derived from the old Flicker applet program and some from the MicroArray Explorer (10)—an open-source data-mining tool for microarray analysis (<http://maexplorer.sourceforge.net/>).

Because Flicker now can run on a user's computer, this gives them the ability to perform real-time comparisons of 2-D-gel image data with gel images residing on remote Internet databases on the Internet, or on the user's local file system, or a combination of both sources.

This approach may be useful for comparing similar protein samples created in different laboratories to help putatively identify or suggest possible protein spot identification. The gels should be run under similar pH and molecular-weight ranges if possible. Although available for over three decades, 2-D polyacrylamide gel electrophoresis (2-D PAGE) is still routinely used (11), even considering the now common use of mass spectrometry (12–17) and recently protein arrays (18) for protein identification.

Recent advances, such as isoelectric focusing (IEF) “zoom” fractionation gels (19) that divide the protein sample by pH range or immunoaffinity subtraction with liquid chromatography (LC) (15), greatly increase the resolution and numbers of spots able to be discriminated by subsequent 2-D gel electrophoresis. Another increasingly common image comparison technique uses two to six cyanine dyes using dye multiplexing to label multiple control and experimental samples run in the same gel, such as Amersham's differential in-gel electrophoresis (DIGE) (20), and scanned with systems

like Perkin Elmer's ProExpress (21). Multiple scans of the same gel using different color filters can then be color mapped to see the contributions of the different samples. This is useful if one has control over the experimental design when determining the reference gel, set of control gels, and experimental gels. However, it does not solve the problem of trying to putatively compare one's own sample against an Internet reference gel where they have identified protein spots.

A number of 2-D-gel databases that contain gel images are available on the Web for various types of tissues. Proteins are identified for some of the spots in a subset of these databases. Both WORLD-2DPAGE and 2D Hunt on the <http://www.expasy.org/> server can be used to find Web URL addresses for a number of 2-D protein gel databases. Google searches are also used, and we link to these sites in the Help menu. Many of these databases contain 2-D-gel images with identified proteins. Some of these databases let you identify spots in their gels by clicking on a spot in their gel image shown in your Web browser. It then queries their Web server database to determine whether the spot you pointed to is in that database and report its identity if found. These "clickable" 2-D-gel map images are often published using a common federated database (DB) schema (22–23). One of the more interesting databases is SWISS-2DPAGE (22–25), accessible from the Expasy site. It has a large number of tissues with over 30 gel databases, including a wide range of human tissues, mouse, *Escherichia coli*, *aribidopsis*, *dictyostelium*, and yeast. Their site also has a series of IEF zoom fractionated gels for *E. coli*.

We have incorporated links to these SWISS-2DPAGE database gel-map images so they may be loaded and accessed directly from Flicker after having putatively matched a spot in your gel with one in the SWISS-2DPAGE gels. If you have loaded one of these active gel-map images in Flicker and enabled the database access, then clicking on a spot in that image will pop up a Web page as it tries to look up the spot in the SWISS-2DPAGE database. If a SWISS-2DPAGE data entry exists for the spot coordinates you have selected, then it will report the corresponding protein; if not, it will tell you it can't be found. Access to PIR UniProt, iProClass, and iProLink (<http://pir.georgetown.edu>) is also available.

By comparing one's own experimental 2-D-gel image data with gel images of similar biological material from these Internet reference databases, it may be possible to use the spots in these reference gels to suggest the putative identification of apparently corresponding spots in your gels. The image analysis method described here allows scientists to more easily collaborate and compare their gel image data over the Web.

1.1. How Can We Compare Two Gels?

When two 2-D gels are to be compared, simple techniques may not suffice. There are several methods for comparing two gel images: (1) put the images side by side and visually compare them; or (2) slide one gel (autoradiograph or stained gel) over the other while back lighted; or (3) build a 2-D-gel quantitative computer database from both gels after scanning and quantitatively analyzing these gels using a 2-D-gel database system; (4) more recently, dye multiplexing (20) has been used to label different samples in the same gel. A variant of the latter method is to spatially warp two gels to the same geometry and then pseudocolor them differentially. These methods may be impractical for many investigators, since in the first case the physical gel or autoradio-

graph from another lab may not be locally available. The first method may work for very similar gels with only a few differences. The second method will work better for gels that are not so similar but that have local regions that are similar. The third method may be excessive if only a single visual comparison is needed, because of the costs (labor and equipment) of building a multi-gel database solely to answer the question of whether one spot is probably the same spot in the two gels. The fourth method may also not be practical if you want to compare your sample against an existing reference gel.

1.2. The Flicker Program

We describe a computer-based image comparison technique called Flicker that has been used for years in finding differences in star maps in astronomy.

The Flicker program runs on most computers. It is started as one would any program after it is downloaded from the Flicker server and installed (see **Subheading 3., step 1**). One gel image may be read from any Internet 2-D-gel database (e.g., SWISS-2DPAGE), the other may reside on the investigator's Web server where they were scanned or copied, or the two gel images may be from either Web server source.

Figure 1 shows the Flicker application after it has been started with some demo gels. You interact with the program by clicking or dragging the mouse in the left or right images, adjust parameter scrollers (upper right), set interaction modes (checkboxes upper left), keyboard short-cut commands, and primarily pull-down menu commands.

1.3. Notation in This Paper

We use the notation <menu name> | <command> throughout this paper to indicate menu commands. The <menu name> indicates one of the pull-down menus: File, Edit, View, Landmark, Transform, Quantify, Help. The <command> indicates one of the commands in that menu. **Table 1** summarizes the menu commands. Some of the commands have alternative keyboard shortcuts activated by using the Control key with another key, and are indicated as **Control**-<key> or C-<key> (e.g., C-A). The checkbox menu commands are indicated with a prefix. Checkbox commands may be toggled on and off.

The gels in the two lower left and right images are specified by the user with the Flicker Files menu. Gel images may be loaded from: the local computer **File** | **Open image file**, or any Internet site with GIF, JPEG, TIFF, or PPX images (with .gif, .jpg, .tiff or .tif, or .ppx GELLAB-II format (8) file extensions) using the **File** | **Open image URL** command. In addition, the installation provides a few demonstration images **File** | **Open demo images**, which loads pairs of comparable images. You may also specify active gel images from Web servers as described below.

Flicker is also capable of interacting with federated 2-D-gel databases to retrieve data on individual protein spots if one of the gels is a federated gel having an associated clickable gel-map database. After aligning gels in Flicker, you enable federated database access in Flicker and then click on a spot in the gel belonging to the federated database (see **Fig. 2**). This causes a Web page to pop up with information from the federated server describing that protein. We provide menu entries **File** | **Open active map image** | ... to let you load one of the SWISS-2DPAGE gel images.

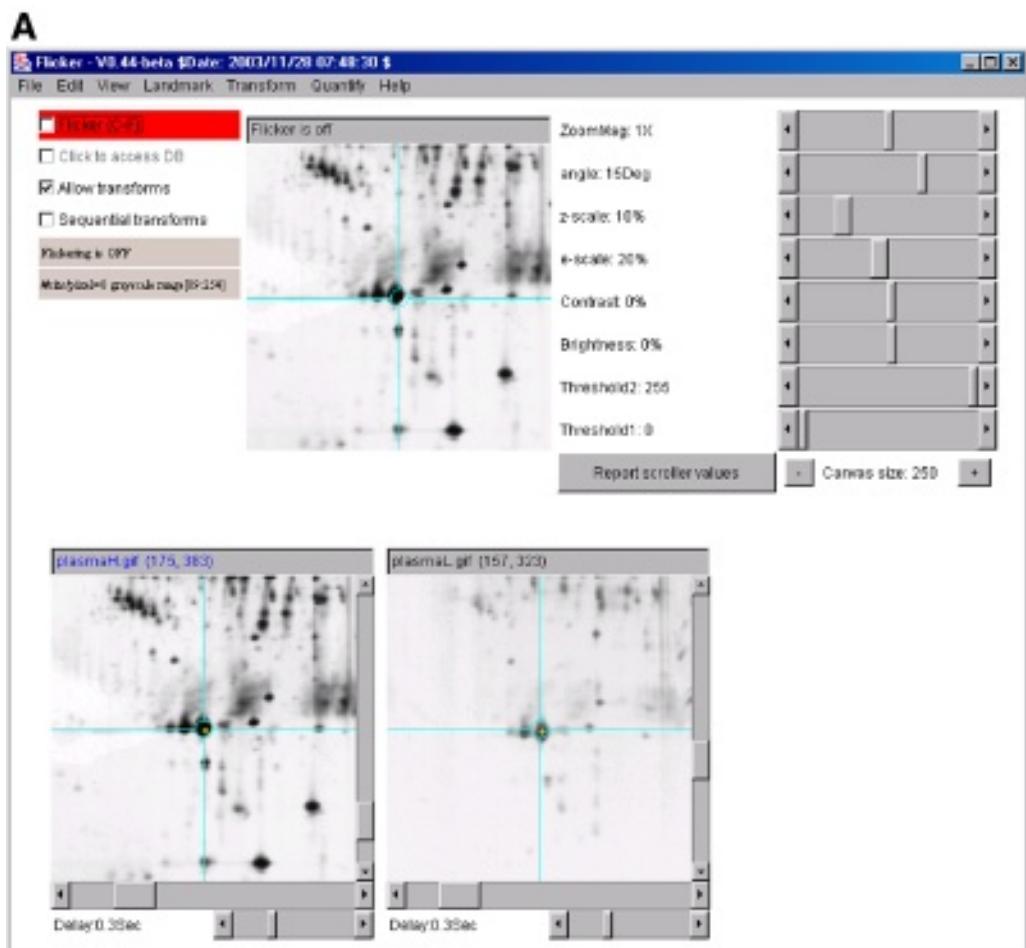


Fig. 1. Screen view of initial Flicker program. (A) shows the pull-down menus at the top used to invoke file operations, editing, view selection, landmarking, image transforms, spot quantification, and help. A set of scroll bars on the right determines various parameters used in some of the transforms. The “File” menu options include opening a new gel image. Checkboxes on the left activate flickering and active gel-map access if the data support it. A set of status lines appears below the checkboxes, indicating the state of operation and various other messages. The flicker image is in the upper middle of the frame when it is enabled. The two labeled human blood plasma gel images are shown in the bottom scrollable windows, which may be positioned to the region of interest. These windows also have associated flicker time-delays used when flickering. Image plasmaH is an immobilized pH gradient nonlinear gradient gel from SWISS-2DPAGE in Geneva, and plasmaL is a carrier-ampholyte linear gradient gel from the Merril Lab at NIMH. Transformed image results are shown in the same scrollable windows. The four checkboxes are: “Flicker (C-F)” to enable/disable flickering; “Click to access DB” enables/disables access to a Web server that is associated with a clickable image DB if it exists for the selected image; “Allow transforms” enables/disables image transforms; “Sequential transforms” enables/disables using the last image transform output as input for the next image transform. The parameters used in various transforms are adjusted directly by first selecting an image and then adjusting its values. You can pop up the scrollable report window using the “Report scroller values” button. These parameters are saved when you save the state. You can



Fig. 1. (continued) change the size of the three image windows by using the “+” and “-” buttons. (B) The pop-up scrollable report window shows a log of all text output that appears in the status lines. It may be saved to a text file on the local disk.

You may load a gel image in the lower left or right image windows. First click on the image you want, to load the new image. Then select the active gel image you want using the entry **File | Open active map image | ...** pull-down menu (e.g., select ... | **SWISS-2DPAGE Human | Plasma**). Next, click on the other image and then use the other gel image you want to compare it with, using either **File | Open image file** or **File | Open image URL** commands.

The Flicker program is written in Java, a general purpose, object-oriented programming language developed by Sun Microsystems (26) <http://java.sun.com/>. Java has become a standard for portable Internet Web applications.

Table 1**Summary of Commands Available in Flicker Menu**

Submenus are indicated by an underline and by the addition of the arrow symbol. Submenu items are indented. For brevity, not all submenus are shown; see Web site for all full documentation. Checkboxes are indicated by the “” prefix to the command. Shortcuts are indicated by the (C-<key>) at the end of the command. Commands are grouped by “----” in the menus (see Notes 5-7 and 9).

File menu - load images, load demo or user images, active map URLs, load/save Flicker .flk state, update program and data (from the Web server).

Open image File - pop up gel image file browser

Open image URL - pop up gel image URL dialog

Open demo images ► - load pairs of demonstration gel images

Open user images ► - load pairs of user's gel images from Images directory

 Pairs of images ► - directories with pairs of images in each user directory

 Single images ► - directories with single images in each user directory

 List user's images by directory - list the user's images

Open active map image ► - load active gel image from the Internet Reference DBs

Open recent images ► - load an image you have used recently

Assign active image URL - to one of the open images to make it active

Open state file - restore the Flicker state of previously saved session

Save state file - save the Flicker state in current .flk state file

SaveAs state file - save the Flicker state in new .flk state file

Update ► - download and update your program and data from server

 Flicker program - to get the most recent release

 Active Web maps image DB - get latest active maps image database

 Demo Images DB - get latest demonstration images database

 Add user's Flicker Demo Images DB by URL - specify user demo data URL site

Save Transformed image - of selected image as .gif file if transformed

SaveAs Overlay image - the current overlay image

Reset images - to the initial state when they were loaded

Abort transform - abort any active image transforms

Quit - exit the program, saving the .flk state of Flicker in the process

Edit menu - change various defaults.

Canvas size ► - change the size of 3 image canvases and overall Flicker window

 Increase size (C-Numpad '+') - increase the canvas size

 Decrease size (C-Numpad '-') - decrease the canvas size

Set colors ► - set default colors for the overlays

 Target colors ► - to change the target color

 Trial object colors ► - to change the trial object color

 Landmarks colors ► - to change the color of landmarks

 Measurement colors ► - to change the color of measurements

Resize Flicker memory limits - set startup memory limit, 30Mb to 1784Mb.

Use linear else log of TIFF files > 8-bits - take log of tiff data if > 8-bits

Enable saving transformed images when do a 'Save(As) state'

Use protein DB browser, else lookup ID and name on active images

Auto measure, protein lookup in active server and Web page popup

Select access to active DB server ► - select active server to use

 Use SWISS-2DPAGE DB access

 Use PIR UniProt DB access

 Use PIR iProClass DB access

 Use PIR iProLink DB access

Reset default view - sets all view options to the defaults

Clear all 'Recent' images entries - clears list of recently accessed images

(continued)

Table 1 (Continued)**Summary of Commands Available in Flicker Menu**

View menu - change the display overlay options.

- Flicker images** (C-F) - toggle flickering on and off
- Set view overlay options ► - enable/disable overlay view options
 - View landmarks - add landmarks to the overlay display in images
 - View target - add target to the overlay display images
 - View trial object - add trial object to the overlay display images
 - View Region Of Interest (ROI) - add ROI to the overlay display images
- Set view measurement options ► - enable/disable measurement view options
 - View measurement circle - add measurement circles to overlay display images
 - Use 'Circle' for measured spot locations - add circles, else '+'
 - Use '+' for measured spot locations - add '+', else circles.
 - Use 'spot number' for spot annotations - add spot number.
 - Use 'spot identifier' for spot annotations - add spot identifier
- Set gang options ► - enable/disable ganged images view options
 - Multiple popups - make multiple popup windows instead of reusing one
 - Gang scroll images - move left and right images scrolling together
 - Gang zoom images - zoom left and right images scrolling together

- Display gray values** (C-G) - show gray values of cursor trial object
- Show report popup** - display the report popup window again if needed

Landmark menu - define landmarks for spatial warping.

- Add landmark** (C-A) - add trial objects (in images) as landmark
- Delete landmark** (C-D) - delete the last landmark defined
- Show landmarks similarity** - compute LSQ error measure of 2 sets of landmarks
- Set 3 pre-defined landmarks for demo images** (C-Y)
- Set 6 pre-defined landmarks for demo images** (C-Z)

Transform menu - perform various image processing transforms.

- Affine Warp** - warp selected image using 3 pairs of landmarks
- Pseudo 3D transform** - do pseudo 3D scaling based on image intensity

- Sharpen Gradient** - gradient + gray scale sharpen selected image
- Sharpen Laplacian** - Laplacian + gray scale sharpen selected image
- Gradient** - gradient of the selected image
- Laplacian** - Laplacian of the selected image
- Average** - average selected image
- Median** - median of selected image
- Max 3x3** - max of 3x3 neighborhood of selected image
- Min 3x3** - min of 3x3 neighborhood of selected image

- Complement** - complement selected image
- Threshold** - threshold the selected image by gray values in [T1:T2]
- Contrast Enhance** - Contrast enhance selected image
- Histogram equalize** - histogram equalize selected image

- Original color** - Restore original data for selected image
- Pseudo color** - compute pseudo color scaling for selected image
- Color to grayscale** - compute NTSC RGB to grayscale transform for image

- Flip Image Horizontally** - flip image horizontally selected image
- Flip Image Vertically** - flip image vertically selected image

- Repeat last transform** (C-T) - repeat last transform, if any
- Use threshold inside [T1:T2] filter** - filter pixels inside (outside) range

(continued)

Table 1 (Continued)**Summary of Commands Available in Flicker Menu**

Quantify menu - contains OD calibration, background and foreground measurements.

Measure circle ► - measure intensity/density within circle

Capture background (C-B) - background measurement at current position

Capture measurement to spot list(C-M) - measure circle at current position

Clear measurement - clear measurement data

Edit selected spot(s) 'id' fields from spot list(s) (C-I)

Edit selected spot(s) from spot list(s) (C-E)

Delete selected spot from spot list (C-K)

List spots in the spot list for selected image - measured spots report

List spots in the spot list (tab-delim) - measured spots report for export

List 'id'-paired annotated mean norm. spots in both spot lists (tab-delim)

List 'id'-paired annotated spots in both spot lists (tab-delim)

Lookup Protein IDs and Names in spot list from active map server

Clear spot list (ask first) for selected image

Print data-window ► - print the data window at the current image location

Set print-window radix (C-V) - print gray-scale window popup report window

Set print-window size ► - set print window size (5x5 to 40x40) pixels

Set print-window radix ► - set data format (decimal, octal, hex, OD) radix

Calibrate ► - calibrate optical density (OD) or other step wedge

Optical density by step wedge - calibrate optical density from ND step wedge.

Use demo leukemia gels ND wedge calibration preloads - to preset OD values

Optical density by spot list wedge - calibrate OD from list of spots

Region of Interest (ROI) ► - region of interest operations

Set ROI ULHC (C-U) - define upper left hand corner of ROI

Set ROI LRHC (C-L) - define lower right hand corner of ROI

Clear (ROI) (C-W) - delete ROI

Show (ROI) grayscale histogram (C-H) - for the current ROI

Capture measurement by ROI (C-R) - measure integrated density less background

Use sum density else mean density - region measurement method to use

List-of-spots else trial-spot measurement-mode (C-J)

Help menu - popup Web browser documentation on Flicker from the Web server.

Flicker Home - pop up Flicker home page open2dprot.sourceforge.net/Flicker

Reference Manual - pop up the reference manual for Flicker application

How-to use controls ► - pop up the references at particular manual sections

Vignettes ► - pop up short vignettes showing how-to-do tutorials

Version on the web site - show current version available on the Web site

About Flicker - show details on Flicker application

Old flicker applet documentation ► old flicker-applet Web home page

2D gel web resources ► - useful 2D gel Web sites

Most often, the original images may be compared directly. However, occasionally, the comparison may be made visually easier by first applying enhancement transforms such as spatial warping, brightness, contrast, or other image transforms. Adjusting image brightness and contrast so the two gels have similar ranges will make the image

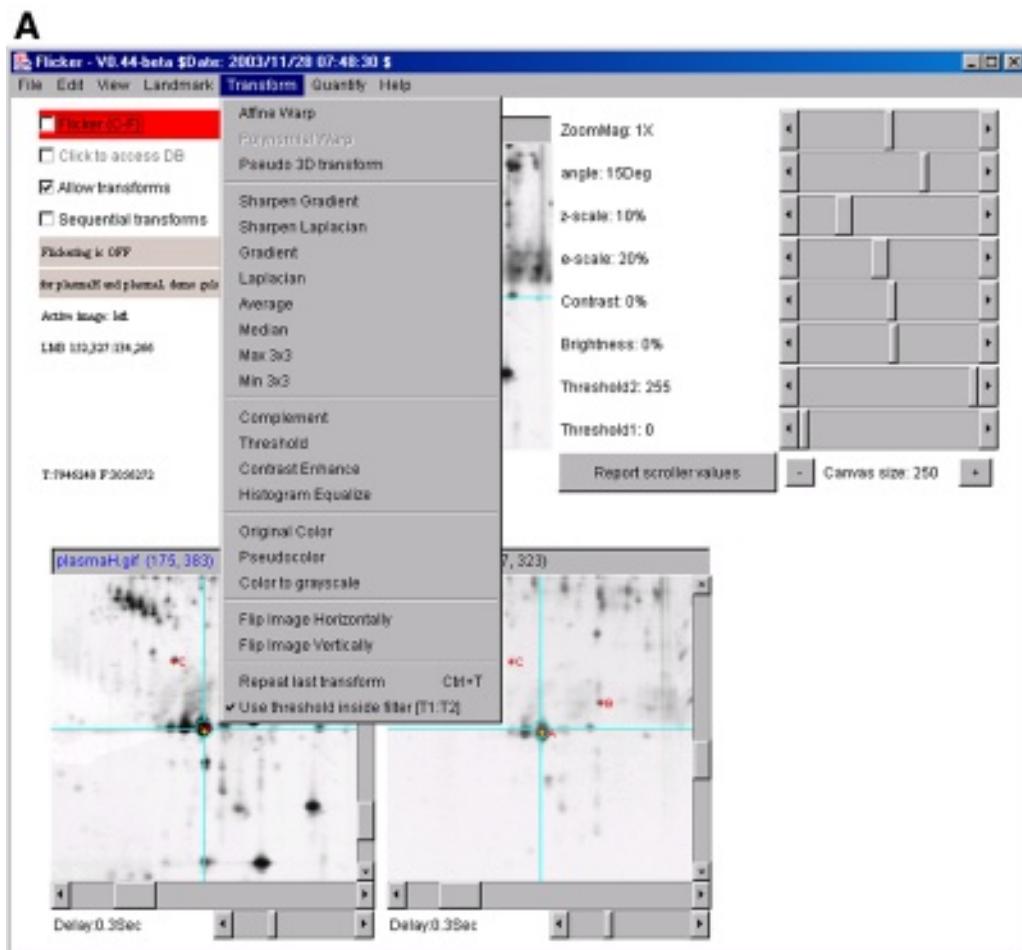


Fig. 2. Screen view of the landmarks used for the affine transform of the human plasma gel images. (A) The transform warps the geometry of a local region defined by the three landmarks so it more closely resembles the geometry of the corresponding local region in the other gel. Scrollable image windows with three “active” landmarks defined in both gel images that were selected interactively in preparation for doing the affine image transform. Corresponding landmark spots are selected so they are defined unambiguously in both gel images. For demonstration purposes, the command **Landmarks | Set 3 pre-defined landmarks for demo gels** will set up the three landmarks shown in this figure. (B) After defining the three landmarks, use the **Transform | Affine warp** command (*Fig. 2. continued on next page*).

fusion easier for the user when flickering. For gels with a lot of geometric distortion, it is useful to adjust the geometry of one gel so that the geometry of the local region being compared approximates that of the other gel. By local geometry, we mean the relative positions, distances, and angles of a set of spots in corresponding regions.

One technique to correct geometry differences is called “spatial warping.” When performing spatial warping, corresponding regions of interest are (1) first marked by the user (we call this “landmarking”) with several corresponding points in each gel

B

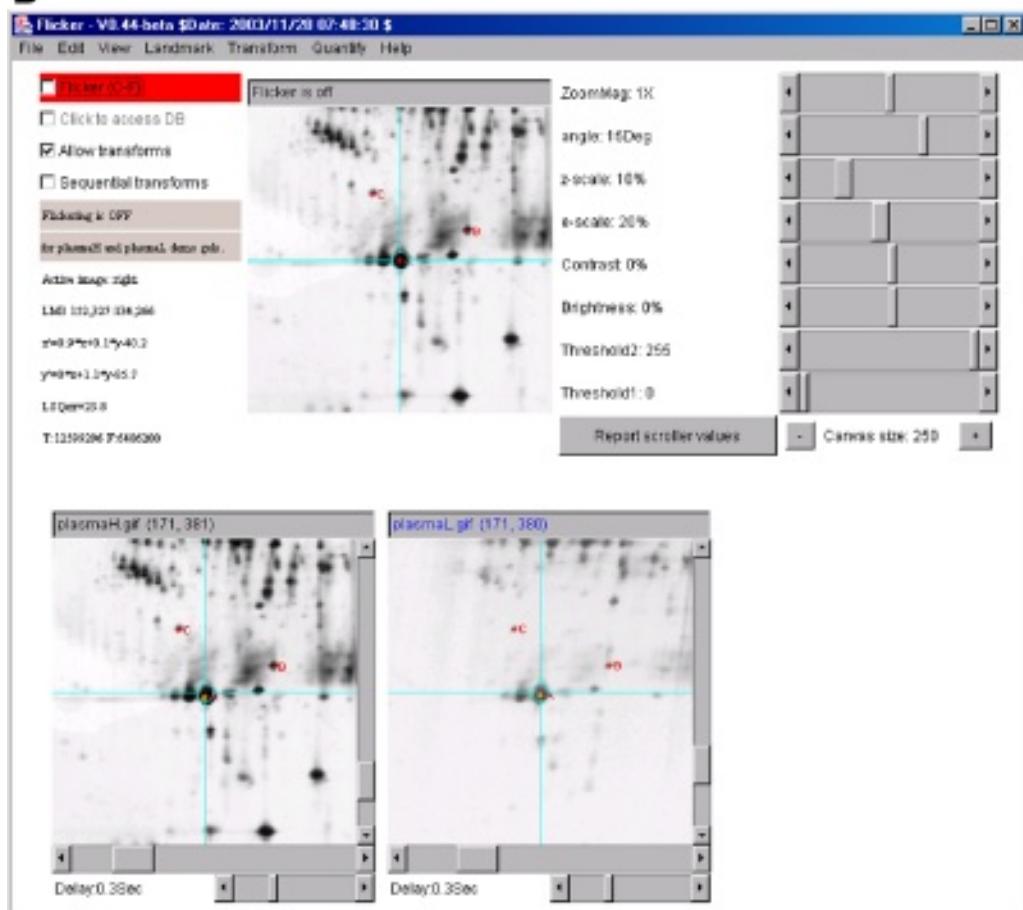


Fig. 2. (continued)

image (three for affine warping and six for poly warping), and (2) then one of the two gel images is warped to the geometry of the other gel (*see Equations 1 and 2*). A landmark is a corresponding spot that is present in both gels. Landmarks are defined by clicking on the spot to mark and selecting the **Landmark | Add Landmark (C-A)** command. The warping is performed by first selecting the image to warp by clicking on it, and then selecting the **Transform | Affine Warp** command. Landmarking and warping are described in more detail below.

Spatial warping doesn't change the underlying grayscale values of the synthesized warped image to the extent that would cause local structural objects to appear and disappear, possibly creating spot artifacts. Instead, it samples pixels from the original image to be transformed, and places them in the output image according to the geometry of the other input image. After warping is finished, gels may then be compared visually by flickering.

1.4. Image Flickering

The basic concept of using flickering as a dynamic visualization technique is simple. If two images may be perfectly aligned, then one could simply align them by laying one over the other and shifting one image until they line up. However, many images such as 2D PAGE gels have rubber-sheet distortion (i.e., local translation, rotation, and magnification). This means there is more distortion in some parts of the image than in others. Although it is often impossible to align the two whole images at one time, they may be locally aligned piece-by-piece by matching the morphology of local regions.

If it appears that a spot and the surrounding region do match, then one has more confidence that the objects are the same. This putative visual identification is our definition of matching when doing a comparison. Full identification of protein spots requires further work, such as cutting spots out of the gels and subjecting them to sequence analysis, amino-acid composition analysis, mass spectrometry, testing them with monoclonal antibodies, or other methods.

1.5. Image Enhancement

It is well known that 2-D gels often suffer from local geometric distortions, making perfect overlay impossible. Therefore, making the images locally morphologically similar while preserving their grayscale data may make them easier to compare. Even when the image subregions are well aligned, it is still sometimes difficult to compare images that are quite different. Enhancing the images using various image transforms before flickering may also help. Some of these transforms involve spatial warping, which maps a local region of one image onto the geometry of the local region of another image while preserving its grayscale values. Another useful operation is contrast enhancement, which helps when comparing light or dark regions by adjusting the dynamic range of image data to the dynamic range of the computer display. Other transforms include image sharpening and contrast enhancement. Image sharpening is performed using edge-enhancement techniques, such as adding a percentage of the gradient or Laplacian edge detection functions to the original grayscale image. The gradient and Laplacian have higher values at the edges of objects. In all cases, the transformed image replaces the image previously displayed. Other functionality is available in Flicker and is described in the Methods and Notes sections of this paper (see Notes 3 and 12), **Table 1**, and on the Web server.

1.6. Image-Processing Transforms

There are a number of different image transforms that may be invoked from the Transform menu. You may display the transformed image, use it as input to another transform, or save it as a .gif file on your local computer. When you save the state, you may also save the transformed images.

1.7. Affine Spatial Warping Transform

The spatial warping transforms require defining several corresponding landmarks in both gels. As we have mentioned, one gel image can be morphologically transformed to the geometry of the other using the affine or other spatial warping transformations. These transforms map the selected image to the geometry of the other image. It does not interpolate the grayscale values of pixels—just their position in the transformed

image. As described in (1–2,4–5), this might be useful for comparing gels that have some minor distortion, comparing local regions, gels of different sizes, or gels run under slightly different conditions. Flicker uses the affine transform as an inverse mapping, as described in (28). Let $(u_{xy}, v_{xy}) = f(x, y)$, where (x, y) are in the output image, and corresponding (u, v) are in the input image. Then, in a raster sweep through the output image, pixels are copied from the input image to the output image. The affine transformation is given in equations 1 and 2:

$$u_{xy} = ax + by + c \quad (1)$$

$$v_{xy} = dx + ey + f \quad (2)$$

When the affine transform is invoked, Flicker solves the system of six linear equations for coefficients a , b , c , d , e , and f using three corresponding landmarks in each gel.

1.8. Pseudo 3D Transform

As described in refs. 1,2,4,5, the Pseudo 3D transform is a forward mapping that generates a pseudo-3D relief image to enhance overlapping spots with smaller spots, seen as side peaks. The gel size is width by height pixels. The gray value determines the amount of y shift scaled by a percentage z_{scale} (in the range of 0 to 50%). Pseudo perspective is created by rotating the image to the right (left) by angle theta (in the range of –45 to +45 degrees). The transform is given in equations 3–5 for image of size width \times height, shift in the horizontal dimension computed as d_x .

$$d_x = \text{width} \sin(\text{theta}) \quad (3)$$

$$x' = (d_x (\text{height} - y)/\text{height}) + x \quad (4)$$

$$y' = y - z_{scale} \times g(x, y) \quad (5)$$

where $g(x, y)$ is in the original input image and (x', y') is the corresponding position in the output mapped image. Pixels outside of the image are clipped to white. The Pseudo 3D transform is applied to both images so that one can flicker the transformed image.

1.9. Edge Sharpening

Edge sharpening may be useful for sharpening the edges of fuzzy spots. The sharpened image function $g'(x, y)$ is computed by adding a percentage of a two-dimensional edge function of the image to original image data $g(x, y)$, as shown in equation 6. The edge function increases at edges of objects in the original image and is computed on a pixel by pixel basis. Typical “edge” functions include the eight-neighbor gradient and Laplacian functions that are described in refs. 1,2,4,5 in more detail. The e_{scale} value (in the range of 0 to 50%) is used to scale the amount of edge detection value added.

$$g'(x, y) = (e_{scale} \times \text{edge}(x, y) + (100 - e_{scale}) \times g(x, y))/100 \quad (6)$$

1.10. Simple Quantitative Spot Measurement

A rudimentary spot measurement facility is available in which variably sized circles can be drawn around isolated spots and which allows for background correction. Lists of spots and paired spots may be created with annotation (see Note 2).

2. Materials

The following lists all items necessary for carrying out the technique. Because it is a computer technique, the materials consists of computer hardware, software, and an Internet connection. We assume the user has some familiarity with computers and the World Wide Web.

1. A Windows PC, MacIntosh with MacOS-X, a Linux computer, or a Sun Solaris computer having a display of at least 1024×768 resolution. At least 30 Mb of memory is required, and more is desirable for comparing large images or performing many transforms. If the computer does not have enough memory, it will be unable to load the images, the transforms may crash the program, or other problems may occur. An Internet connection is required to download the program from the <http://open2dprot.sourceforge.net/Flicker> Web site (see **Note 1**). New versions of the program will become available on this Web site and can be downloaded using the various Update commands described in the Notes section. If you have installer software that someone else downloaded and gave to you, then you do not need the Internet connection to install the program. If you will be using the active gel image maps associated with federated 2-D-gel databases, then you will need the Internet connection for accessing those databases. You do not need the Internet for local image comparisons. We currently distribute a version of Flicker that uses up to 128 Mb. If you want to run it with less or more memory, use the **Edit | Resize memory limits** command to set it to a value in the range of 30 Mb to 1784 Mb.
2. When you install Flicker, it creates several subdirectories (see **Notes 10** and **11**): *Images/*, containing the demonstration images; *DB/*, containing startup database files; and *FlkStartups/*, containing any startup files you create when you do a **File | SaveAs state file**. The *DB/* files are: *FlkDemoDB.txt*, which describes the demo images; *FlkMapDB.txt*, which describes the gel images and their corresponding active image map URLs, and *FlkRecentDB.txt* which lists recently accessed images. An empty database file *FlkRecentDB.txt* contains the file names and active gel map URLs, if any, of recently accessed images.
3. The Internet is a good source of 2-D-gel images. You can find them by searching WORLD-2DPAGE and 2D Hunt on the <http://www.expasy.org/> server or a Google search to find other Web 2-D protein gel image databases. Links to these databases are available in the **Help | 2D gel Web resources** submenu.

3. Methods

We now describe the operation of Flicker from the user's point of view. You first install Flicker. Then run it with either the demonstration images, your own images, or images from the Internet. Then you simply flicker the gel images. If necessary, to improve the image comparability, use image enhancement transforms, before flickering the two images.

3.1. *Installing Flicker From the Web Server*

Click on the Download link on the <http://open2dprot.sourceforge.net/Flicker> Web site. This brings up the Java installer for your computer (we use the commercial InstallAnywhere installer by ZeroG.com). You may either click on the "Download Flicker for <computer type>" button or click on one of the links in the list of available installers. The latter is useful if you want to save the downloaded installer for later installation or for installing it on another computer. You have the option of downloading the "Java Virtual Machine (JVM)," which we recommend. This will not interfere

with any other JVMs you have already installed or may install in the future. Once the installer starts, you may select an installation language (English is the default) and press “OK.” Then press the “Next” button after the Introduction window pops up. It then asks you where to install it, suggesting a reasonable default that you may override; then press the “Next” button. For Windows and some of the other systems, it will ask you where you want to put the startup icon; then press the “Next” button. After it finishes the installation, it will show the “Installation Complete” window. Finally, press the “Done” button to finish the installation. For example, in MS Windows systems, a “Flicker startup” icon will appear on your desktop.

To start Flicker, click on the startup icon. For Unix systems, including MacOS-X, you can start Flicker from the command line by specifying the path to *Flicker.bin*. Normally it comes up with the two demonstration human plasma 2-D-gel images—*plasmaH.gif*, an immobilized pH gradient (IPG) gel from SWISS-2DPAGE, on the left, and *plasmaL.gif*, a carrier ampholyte gel from Dr. Carl Merril/NIMH, on the right.

If you have your own gels (JPEG, GIF, or TIFF formats), you can try loading them. You may want to limit resolution by first decreasing their size using an image editing program like Photoshop or the shareware program ThumbsPlus (www.cerious.org). Large, very high-resolution images that are 20Mb to 40Mb will not work well. We suggest reducing the size to about 1K × 1K for good interactivity if you have any problems with running out of memory or very sluggish response. These image-editing programs can also be used for converting other formats to JPEG, GIF, or TIFF formats that Flicker can read.

3.2. Graphical User Interface for Flickering

Figure 1 shows the initial screen of the Flicker program. Pull-down menus at the top invoke file operations, edit preferences, view overlay options, landmarking, image transforms, and help commands. Scroll bars on the side determine various parameters used in the transforms. The two images to be compared are loaded into the lower scrollable windows. A flicker window appears in the upper middle of the screen. Checkboxes on the left activate flickering and control display options. A group of status lines below the checkboxes indicate the state of operations. **Table 1** shows the summary of the commands in the pull-down menus.

Only part of an image is visible in the scrollable windows. These subregions are determined by setting horizontal and vertical scroll bars. Another, preferred, method of navigating the scrollable images is to click on the point of interest while the Control key is pressed. This will re-center the scrollable image around that point. Note that if you are near the edge of the image when you do this, it will not scroll the image. This lets the user view any sub-region of the image at high resolution. These images may be navigated using either the scroll bars or by moving the mouse with the button pressed in the scrollable image window. Then, each image in the flicker window is centered at the point last indicated in the corresponding scrollable image window.

A flicker window is activated in the upper middle of the screen when the “Flicker” checkbox is selected. Images from the left and right scrollable images are alternatively displayed in the flicker window. The flicker delay for each image is determined by the adjusting the scroll bar below the corresponding scrollable image window. Various

graphic overlays may be turned on and off using the various view overlays selected in the **View | View ...** checkbox menu commands.

Clicking on either the left or right image selects it as the image to use in the next transform. However, clicking on the flicker image window indicates the next transform you might use should be applied to both left and right images. You can change this default by just clicking on any of the images.

You can increase or decrease the size of the three image windows by using the **Edit | Canvas size | Increase size (C-keypad "+")** and **Edit | Canvas size | decrease size (C-keypad "-")** commands. This will resize the main window accordingly.

3.3. Loading Images

As mentioned in the introduction, gel images may be loaded into the left or right selected image from: (a) the local computer using the **File | Open image file** command; (b) any Internet site using the **File | Open image URL** command. You may load pairs of demonstration images that come with Flicker, installing them in the *Images*/directory. Use the **File | Open demo images | ...** command to load them into the left and right images. The demos include a few samples that may be useful for initially learning the system. They include: two human plasma gels—an IPG SWISS-2DPAGE gel vs a carrier ampholyte gel (Merril/NIMH)—and some human leukemias (acute myeloid leukemia [AML], acute lymphoblastic leukemia [ALL], chronic lymphocytic leukemia [CLL], hairy cell leukemia [HCL]) from Lester et al. (9).

You may specify active gel images from the Web using the **File | Open active map image | ...** to let you load one of the Swiss-2DPAGE gel images into the left or right selected image. This list of active images is defined by the tab-delimited *FlkMapDB.txt* file read by Flicker when it is started. “Power users” could edit this file (use Excel and save as tab-delimited) to add active map entries, pointing to other federated 2D-gel Web databases. The *FlkMapDB.txt* file is provided with your download installation in the *DB*/directory.

Gel images are loaded into the lower left or right images. First click on the left or right image you want to replace. Then, select the active gel image you want, using the **File | Open active map image | ...** pull-down menu (e.g., select ... | **SWISS-2DPAGE Human | Human Plasma**). Next, click on the other image and then open the other gel image you want to compare it with, using either the **File | Open image file** or the **File | Open image URL** command.

You may put directories of your own images to be compared in the *Images* subdirectory, and they will appear in **File | Open user images | ...** (see Note 11).

3.4. Flickering

When flickering two images with the computer, one aligns putative corresponding subregions of the two rapidly alternating images. The flicker window overlays the same space on the screen with the two images and is aligned by interactively moving one image relative to the other, using the cursor in either or both of the lower images. Using the mouse, the user initially selects what they suspect is the same prominent spot or object in similar morphologic regions in the two gel images. The images are then centered in the flicker window at these spots. When these two local regions come into

alignment, they appear to pulse and the images fuse together. At this point, differences are more apparent, and it is fairly easy to see which spots or objects correspond, which are different, and how they differ. We have found that the user should be positioned fairly close to the flicker window on the screen to optimize this image-fusion effect (i.e., it does not work as well standing back more than a few feet from the screen).

3.4.1. Selecting the Proper Time Delays When Flickering

The proper flicker delays, or the time each image is displayed on the screen, is critical for the optimal visual integration of image differences. We have also found that optimal flicker rates are dependent on a wide variety of factors, including amount of distortion, similarity of corresponding subregions, complexity and contrast of each image, individual viewer differences, phosphor decay-time of the display, ambient light, distance from the display, and so on. We have found the process of flickering images is easier for some people than for others.

When comparing a light spot in one gel with the putative paired darker spot in the other gel, one may want to linger longer on the lighter spot to make a more positive identification. Because of this, we give the user the ability to set the display times independently for the two images (typically in the range of 0.01 s to 1.0 s, with a default of 0.30 s), using separate Delay scroll bars located under each image. If the regions are complex and have a lot of variation, longer display times may be useful for both images. Differential flicker delays, with a longer delay for the light gel, are useful for comparing light and dark sample gels. This lets you stare at the lighter spots to have more verification that they are actually there.

3.5. Image Processing Methods

As mentioned before, there are a number of different image transforms that may be invoked from the menus. These are useful for changing the geometry, sharpness, or contrast, making it easier to compare potentially corresponding regions. As we go through the transforms, we will indicate how they may be used. Some affect one image, while some affect both. Flickering is deactivated during image transforms to use most computational power for doing the transforms.

The Transform menu has a number of commands, which include warping, grayscale transforms, and contrast functions. The two warp method selections—"Affine Warp" and "Poly Warp"—are performed on only one image (the last one selected by clicking on an image). The "Pseudo 3D" transform makes a 3-D image with the "peaks" created proportional to gray level. Unlike the warp transforms, the grayscale transforms are performed on both images. These include: "SharpenGradient," "SharpenLaplacian," "Gradient," "Laplacian," "Average," "Median," "Max 3 × 3," and "Min 3 × 3." The contrast functions are "Complement" and "ContrastEnhance." You can transform color images to grayscale using the "Color to grayscale" command, and generate a false color image from a grayscale using the "Pseudo color" command. You can flip the image using "Flip image horizontally" or "Flip image vertically" commands.

3.5.1. Landmarks: Trial and Active

The affine transform requires that three active landmarks be defined before it can be invoked. A trial landmark is defined by clicking on an object's center anywhere in a scrollable image window. This landmark would generally be placed on a spot. Clicking

on a spot with or without the Control key pressed still defines it as a trial landmark. After defining the trial landmark in both the left and right windows, selecting the **Landmark | Add Landmark (C-A)** command defines them as the next active landmark pair and identifies them with a red letter label (+A, +B, +C, ...) in the two scrollable image windows. The **Landmark | Delete Landmark (C-D)** command is used for deleting the last landmark you defined.

3.5.2. The Affine Transform for Spatial Warping

The two warping transforms, affine (*see equations 1 and 2*) and polynomial, require three and six landmarks, respectively. Attempting to run the transform with insufficient landmarks will cause Flicker to notify you that additional landmarks are required. The image to be transformed is the one last selected. You must select either the left or right image. **Figure 2A** shows the landmarks the user defined in the two gels before the affine transform. **Figure 2B** shows the transformed image. Then, re-center the transformed image before you flicker. After the transform, the landmarks can be lined up perfectly, and adjacent spots will line up better.

3.5.3. Pseudo 3D Transform

As described in **refs. 1,2,4,5** and as shown in **equations 3–5**, the Pseudo 3D transform generates a pseudo-3-D relief image to enhance overlapping spots, with smaller spots seen as side peaks. The gray value determines the amount of y shift scaled by a percentage (set by scroll bar z_{scale}) in the range of 0 to 50%. Pseudo perspective is created by shifting the image to the right or left by setting by scroll bar “angle” degrees (in the range of -45 to +45 degrees). Negative angles shift it to the right and positive angles to the left. The image to be transformed is the one last selected. If neither was selected (i.e., you clicked on the flicker window), then both images are transformed.

3.5.4. Edge Sharpening

Edge sharpening may be useful for improving the visibility of the edges of fuzzy spots. You can select either a Gradient or Laplacian edge-sharpening function using the “SharpenGradient” or “SharpenLaplacian” operation in the Transform menu where the image to be transformed is the one last selected. The Laplacian filter generates a “softer” edge than the Gradient. You can set the scroll bar e_{scale} value (in the range of 0 to 50%) to scale the amount of edge detection value added. The image to be transformed is the one last selected. If neither was selected (i.e., you clicked on the flicker window), then both images are transformed.

3.5.5. Putative Identification of a Spot in One Gel by Comparison With Federated Database Gel Map (see **Fig. 3**)

Au: Fig. citation ok?

Open an active gel image in the lower left or right window. First click on the window you want to load the new image. Then, select the active gel image to you want, using the entry **File | Open active map image | ...** pull-down menu (e.g., select ... | **SWISS-2DPAGE Human | Plasma**). Next, click on the other image and then use the other gel image you want to compare it with, using one of the other **File | Open ...** commands.

At this point, flicker the two images so that you can make a putative guess as to which spot you are interested in—which spot in the active map gel your gel corre-

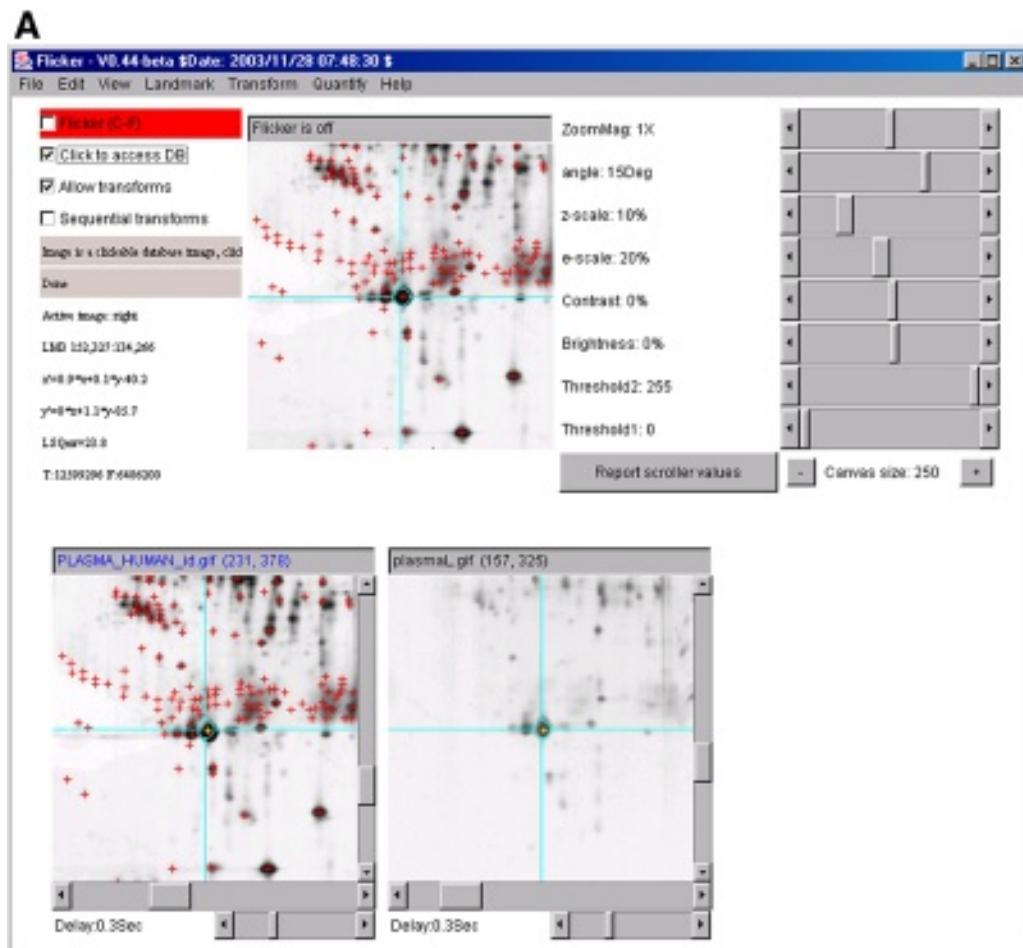


Fig. 3. Screen views of clickable active gel compared with another gel. **(A)** We have loaded the active SWISS-2DPAGE human plasma gel (PLASMA_HUMAN_id) in the left image and the plasmaL gel in the right image. Spots that appear in SWISS-2DPAGE are indicated with red "+" symbols. We then aligned the images using flickering. We then selected the "Click to access DB" checkbox. Finally, we clicked on the indicated spot in the left gel to determine the putative identification of the corresponding spot in the right gel. **(B)** The SWISS-2DPAGE window then popped up as a result of clicking on that spot in the left image, and indicates the putative protein identification of the visually corresponding spot in the right gel. The plasmaH image is the same gel as PLASMA_HUMAN_id but without the graphic overlays. You can load these same gels using the **File | Open demo images | Human Plasma gels | (SWISS-2DPAGE vs Merril)—clickable** command, which should be used when you are connected to the Internet.

sponds to. Then, shut off flickering by turning off the "Flicker" checkbox. Then, turn on the Click to access DB checkbox. Then, click on the spot in the active map image, which will pop up a Web-browser window indicating the SWISS-2DPAGE Web page for that spot, if it is in their database.

B


The screenshot shows a Netscape browser window with the title "Nice2DPage View of SWISS-2DPAGE: P02647 - Netscape". The URL in the address bar is http://www.expasy.org/cgi-bin/imap3/PLASMA_HUMAN1231377. The page content is as follows:

ExPASy Home page **Site Map** **Search ExPASy** **Contact us** **SWISS-2DPAGE**

Search **SWISS-2DPAGE** for

Search

[\[by description\]](#)
[\[by accession number\]](#)
[\[by clicking on a spot\]](#)
[\[by author\]](#)
[\[by serial number\]](#)
[\[by full text search\]](#)
[\[SRS\]](#)

swiss2Dpage : P02647

1 protein has been found in the clicked spot (2D-0005EK):

General information about the entry

[View entry in original SWISS-2DPAGE format](#)

Entry name: APA1_HUMAN
 Primary accession number: P02647
 Entered in: SWISS-2DPAGE in: Release 00, August 1993
 Last modified in: Release 16, May 2003

Name and origin of the protein

Description: Apolipoprotein A-I (Apo-AI).
 Gene name(s): APOA1
 From: Homo sapiens (Human). [TaxID: [9606](#)]
 Taxonomy: Eukaryota, Metazoa, Chordata, Craniata, Vertebrata, Euteleostomi, Mammalia, Eutheria, Primates, Catarrhini, Hominoidea, Homo.

References

[1] MAPPING ON GEL
 MEDLINE=93162045, PubMed=1286669 [[NCBI](#), [ExPASy](#), [EBI](#), [Israel](#), [Japan](#)]
 Hochstrasser D.F., Frutiger S., Paquet N., Baroch A., Rauer F., Pasqua C., Sanchez J.-C., Tisot J.-D., Bjellqvist B., Vargas R., Appel R.D., Hughes G.J.

Fig. 3. (Continued)

4. Notes

4.1. Status of the Flicker Application

As of the time of the submission of this chapter, most of the functionality available in the former Java applet (1,2,4,5) is fully functional in the stand-alone application. The current state of Flicker is documented on the Web server. A future release of Flicker will contain a spot quantification function, with the ability to calibrate the image (either from the calibration information in the image itself, if available, or from a scanned neutral-density step wedge scanned with the image), estimate background density, and

estimate spot intensity with a background subtraction. Documentation is available on the Web site. This documentation may also be invoked from the Help submenus.

The original Flicker program was converted from a Java applet to a Java application by Peter Lemkin and Greg Thornwall, with help from Jai Evans. Code was added from the open source MicroArray Explorer (<http://maexplorer.sourceforge.net/>) program. The new Flicker program uses the Mozilla 1.1 open source license and is available on the open source Web server <http://open2dprot.sourceforge.net/Flicker/>.

You can update your program and data files using the various Update options in the Files menu. The **File | Update | Update Flicker Program** command downloads and installs the latest *Flicker.jar* file. The **File | Update | Update active Web maps DB** command downloads and installs the latest active Web maps database *DB/FlkMapDB.txt* file. The **File | Update | Update Demo images DB** command downloads the latest demo images into the *Images*/ directory.

4.2. *Hints on Measuring Spots*

There are some disadvantages in comparing gels visually. It is useful for doing a rough comparison, and there is currently no simple way available to do adequate quantitative comparison (as can be done with existing 2-D-gel computer database systems) using automatic spot segmentation and global normalization methods. However, you can look at the gray value of the cursor in the left or right image if you enable the **View | Display gray values** menu option. We are working on an extension to the program to allow single-spot quantification with optical-density calibration. This will be announced on the Flicker Web site. These limitations should be kept in mind when using the technique.

In the meantime, we are providing a simple method that can be used for ballpark estimates of density if your gels and scanner are reasonably linear, so that grayscale approximates protein concentration. You can do this in either of two ways: by measuring the area under a circular mask (you can set the radius), or the area inside a rectangular region of interest (ROI). Note that unless the spot fits well inside of the mask or ROI, you will not get a very accurate measurement. Both methods can subtract an optional background value you can capture and so can give intensity corrected for background if defined.

Use the “measure circle” slider to set the measurement circle region size (1 × 1, 3 × 3, 5 × 5... or 51 × 51). Click on a background region near where you want to measure a spot’s density within a circular mask. Then select the **Quantify | Measure by circle | Capture background (C-B)** command. Then click on the center of the spot you want to estimate and select the **Quantify | Measure by circle | Capture measurement (C-M)** command. This will compute and display background-corrected data that appear in the report window as:

```
Setting the circle size to: 9 × 9
plasmaL.gif (83, 278) Bkgrd value: 670 tot mask(9 × 9) gray-value
plasmaL.gif (156, 323) Tot(Meas-Bkgrd): 5452 gray-value
TotMeas: 6122 [143:162], TotBkgrd: 670 at (83, 278) [14:20] gray-value
CircleMask: 9 × 9
```

The **View | View measurement circle** displays the background and foreground measurement circles, if enabled, with “B” and “M” labels. Set the **View | Use sum**

density else mean density menu option to specify that it report either total region density or mean density.

You also can measure intensity inside a rectangular ROI region you set by using both the **Quantify | Region of Interest (ROI) | Set ROI ULHC (C-U)** and the **Quantify | Region of Interest (ROI) | Set ROI LRHC (C-L)** commands. Use the **View | View Region Of Interest (ROI)** to display the ROI as a rectangle from the ULHC to the LRHC (upper left-hand corner and lower right-hand corner). You can measure the integrated density of the **Quantify | Region of Interest (ROI) | Capture measurement by ROI (C-R)** command. If you set it, the (C-R) command will subtract the background computed by the area times the mean background using the **Quantify | Measure by circle | Capture background (C-B)** command. This will compute and display background corrected data that appear in the report window as:

```
Setting ULHC (188,234) of ROI: right image
Setting LRHC (211,256) of ROI: right image
plasmaL.gif (211, 256) Tot(Meas-Bkgrd): 5936.923 gray-value
TotMeas: 15420 [15:127], TotBkgrd: 9483.077 at (83, 278) [14:20] gray-value
ROI: [188:211, 234:256]
```

When grayscale calibration is added in a future version, then the measurements will be in terms of the calibration rather than grayscale.

4.3. Additional Hints on Image Transforms

The intent of applying image transforms is to make it easier to compare regions that have similar local morphologies but some different objects within these regions. Image warping prior to flickering is intended to spatially warp and rescale one image to the geometric “shape” of the other image so that we can compare them at the same scale. This should help make flickering of some local regions on quite different gels somewhat easier. Of the two warping transforms, affine and polynomial, the latter method handles non-linearities better. For those cases where the gels are similar, the user may be able to get away with using the simpler (affine) transform. For demonstration purposes, if you are using the demo *plasmaH* and *plasmaL* gels, the **Landmark | Set 3 pre-defined landmarks for demo gels (C-Y)** and **Landmark | Set 6 pre-defined landmarks for demo gels (C-Z)** define three and six corresponding landmarks for these gels that may be used with the affine and polynomial warping transforms, respectively.

In cases where there is a major difference in the darkness or lightness of gels, or where one gel has a dark spot and the other a very faint corresponding spot, it may be difficult to visualize the light spot. By differentially setting the flicker display-time delays, the user can concentrate on the light spot using the brief flash of the dark spot to indicate where they should look for the light spot. We have found differential flicker to be very helpful for deciding difficult cases. Adjusting one image so that its brightness and contrast are approximately that of the other image also helps when flickering. You change the image brightness and contrast using the Shift/Drag mouse control described in **Note 4.5**.

Other transforms including image sharpening may be useful in cases where spots are very fuzzy, as might be the case when comparing Southern blots. When two corre-

sponding local regions of the two images are radically different so the local morphologies are not even slightly similar (e.g., when high-MW regions of gels are run differently, such as: IPG vs non-IPG, gradient vs nongradient sodium dodecyl sulfate [SDS]), then even using these transforms may not help that much.

4.4. Saving and Restoring the Flicker State

Flicker gives you the option of saving the current state of your session, including the images you are looking at and the parameter values of the sliders, and so on. To save the current state, use the **File | Save (or SaveAs) state file** command. This creates a file with a .flk file extension in the installation *FlkStartups/* folder (default *FlkStartup.flk*). If you have used the Flicker Web site Java installer (ZeroG.com) for installing Flicker, then it lets you click on a specific .flk you have previously saved to restart it where you left off. While running Flicker, you can also use **File | Open state file** command to change it to another state.

4.5. Mouse Control of Images

The following mouse and key-modified mouse operations control various actions.

Pressing the mouse in either the left or right image selects it. If flickering is active, then it will move the center of the flicker image for the selected image to that position. A little yellow “+” indicates the position you have selected. If the “Click to access DB” checkbox is enabled and the image has an associated active map database server

associated with it, then it will request the spot identify when you click on a spot from the map database.

Dragging the mouse is similar to pressing it. However, only pressing it will invoke a clickable database. It also displays the cursor coordinates in the image title.

Control/Press will position the selected image so that the point you have clicked on will be in the center of the crosshairs. If you are near the edge of the image, it will ignore this request.

Shift/Drag activates the brightness/contrast filter with minimum brightness and contrast in the lower left-hand corner.

4.6. Checkbox Control of Flickering and Database Access

There are four checkboxes in the upper left part of the window that control commonly used options.

The “Flicker” checkbox enables/disables flickering.

The “Click to access DB” checkbox enables/disables access to a Web database server that is associated with a clickable image, if it exists for the selected image. Turning on this option will disable flickering.

The “Allow transforms” checkbox enables/disables image transforms.

The “Sequential transforms” checkbox enables/disables using the last image transform output as input for the next image transform (image composition) if “Allow transforms” is enabled.

4.7. Keyboard Shortcut Controls

There are several short-cut key combinations that can be used to perform operations instead of selecting the command from the pull-down menus. The notation **C-<key>** means to hold the Control key (the Apple key on the Macintosh) and then press the following <key>.

- **C-A**—add landmark (you must have selected both left and right image trial objects) (see landmark menu)
- **C-B**—capture background intensity value for current image under circle (see quantify menu)
- **C-D**—delete landmark (the last landmark defined; see landmark menu)
- **C-F**—toggle flickering lower left and right images into the upper flicker window (see view menu)
- **C-G**—toggle displaying gray values in the left and right image titles as move the cursor (see view menu)
- **C-H**—show grayscale ROI histogram. Popup a histogram of the computation region of interest (ROI) (see quantify menu)
- **C-I**—Define or edit selected measured spot(s) annotation “id” field (see quantify menu)
- **C-J**—toggle the spot measurement mode between list-of-spots measurement mode and the single spot trial-spot measurement mode (see quantify menu)
- **C-K**—delete selected measured spot, click on spot to select it (see quantify menu)
- **C-L**—define lower right hand corner (LRHC) of ROI and assign that to computing window (see quantify menu)
- **C-M**—measure & show intensity under circle for current image, report background corrected value defined (see C-B shortcut and quantify menu)
- **C-R**—measure and show intensity under a the computing window defined by the ROI (see C-U and C-L) for current image. Report background-corrected value if circular mask background was defined (see C-B shortcut and quantify menu)
- **C-T**—repeat the last Transform used, if one was previously performed else no-op (see transform menu)
- **C-U**—define upper left hand corner (ULHC) of ROI and assign that to computing window (see quantify menu)
- **C-V**—show data-window of selected pixel in the popup report window (see quantify menu)
- **C-W**—clear the ROI and computing window (see quantify menu)
- **C-Y**—set 3 predefined landmarks for demo gels for Affine transform (see landmark menu)
- **C-Z**—set 6 predefined landmarks for demo gels for Polywarp transform (see landmark menu)
- **C-Keypad “+”**—increase the image canvas size for all three images (see edit menu)
- **C-Keypad “-”**—decrease the image canvas size for all three images (see edit menu)

4.8. Reporting the Status in the Pop-Up Status Window

Information is display in several places in Flicker:

- (a) There are two status lines in the upper left part of the main window. The output into these status lines is also appended to the Report window (c).
- (b) The selected image (clicking on the left or right image) changes its title to blue from black. If neither image is selected, then both titles are black.
- (c) A report pop-up window is created when Flicker is started. It may be temporarily removed by closing it. You can get it back at any time by selecting **View | Show report popup**. All text output is appended to this window. The Clear button clears all text. The SaveAs button lets you save the text in the window into a local text file.

4.9. Sliders for Defining Parameters

The following sliders are in the upper right part of the window and are used for adjusting parameters in the various image transforms:

angle (degrees) used in the Pseudo 3D transform
 brightness (%) set by Shift/Drag to change the image brightness
 contrast (%) set by Shift/Drag to change the image contrast
 eScale(%) used in the sharpening transforms
 threshold1 (grayscale or od) is the minimum grayscale value to show pixels otherwise
 they are shown as whites
 threshold2 (grayscale or od) is the maximum grayscale value to show pixels otherwise
 they are shown as white
 zoomMag (X) to zoom both left and right images from 1/10X to 10X by a transform
 zScale (%) used in the Pseudo 3D transform

4.10. Local Database Files

When Flicker is installed, several tab-delimited (spreadsheet derived) *.txt* files are available in the *DB/* directory (located where the *Flicker.jar* file is installed). These *DB/Flk*DB.txt* files are read on startup and are used to set up the **File | Open ... image | ...** menu trees:

DB/FlkMapDB.txt—contains instances of Web-based active image maps with fields: *MenuName, ClickableURL, ImageURL, baseURL, DatabaseName*

DB/FlkDemoDB.txt—contains instances of pairs of images in the local *Images/* directory and contains fields: *SubMenuName, SubMenuEntry, ClickableURL1, ImageURL1, ClickableURL2, ImageURL2, StartupData*

DB/FlkRecentDB.txt—contains instances of recently accessed non-demo images with fields: *DbMenuName, ClickableURL, ImageURL, DatabaseName,TimeStamp*

4.11. Files Required That Are Included in the Download

The following files are packaged in the distribution and installed when you install Flicker:

Flicker.jar is the Java Archive File for Flicker that is executed when you run Flicker.

jai_core.jar is the core Java runtime from SUN's Java Advanced Imaging (JAI) at sun.com.

jai_codec.jar is the JAI tiff file reader from SUN's Java Advanced Imaging JAI at sun.com.

DB/ is a directory containing the set of tab-delimited DB files *Flk*DB.txt* read at startup.

Images/ is a directory holding demo *.gif, .tif, .jpeg, and .ppx* sample files as well as a user's subdirectories of images.

FlkStartups/ is empty directory into which to put the startup *FlkStartup.flk* files.

4.12. Image Transform and Brightness-Contrast Display Model

There are several display models for combinations of using image transforms, zooming, and brightness/contrast filtering. Zooming is an image transform, and you can de-magnify as well as magnify. These transforms and filtering are applied to the left and right windows and also are shown in the flicker window. Two checkboxes in the upper left of the main window control transforms: "Allow transform" enables/disables transforms, and "Sequential transforms" allows using the previous transform as the input to the next transform—i.e., this lets you implement image composition.

This description applies independently to the left and right images. The original image is denoted *iImg*. If you allow transforms and are also composing image trans-

forms, you may optionally use the previous transformed output image (denoted *oImg*) as input to the next image transform. The output (either *iImg* or *oImg*) is then sent to the *output1*. Then *output1* may be optionally zoomed to *output2* by being sent to the zoom transform (if the magnification is different from 1.0X). Then *output2* may be optionally contrast adjusted by being sent to the brightness-contrast filter (if it is active, as specified by dragging the mouse in the selected window with the Shift key pressed). The *output2* of the brightness-contrast filter is denoted as *bcImg*. If you have not used the zoom or brightness-contrast filtering since loading an image, then *zImg* and *bcImg* are not generated and hence not used in the displayed image. This will speed up display refresh as you navigate the windows.

- (a) If no transforms or brightness-contrast filtering is used on the selected image
 (no transforms)

iImg → *output1*

- (b) The image may be optionally transformed from the original image (*iImg*)
 (transform)

iImg → *oImg* → *output1*

- (c) Image transforms may be optionally composed from the original image or from the sequential composition of image transforms on the selected image

(sequential transforms)
 ↓
iImg → **Transform** → *oImg* → *output1*

- (d) The image may be optionally zoomed if the magnification is not 1.0X
 (zoom)

output1 → *zImg* → *output2*

or

(no zoom)
output1 → *output2*

- (e) The brightness-contrast filter may be optionally applied to the image
 (B-C filter)

output2 → *bcImg* → *display*

or

(no B-C filter)
output2 → *display*

References

1. Lemkin, P. F. (1997) Comparing two-dimensional electrophoretic gels across the Internet. *Electrophoresis* **18**, 461–470.
2. Lemkin, P. F. and Thornwall, G. (1999) Flicker image comparison of 2-D gel images for putative protein identification using the 2DWG meta-database. *Molecular Biotechnology* **12**, 159–172.
3. Lemkin, P. F. (1997) 2DWG meta-database of 2D electrophoretic gel images on the Internet. *Electrophoresis* **18**, 2759–2773.
4. Lemkin, P. F. (1997) Comparing 2D electrophoretic gels across Internet databases. In: *2-D Protocols for Proteome Analysis, Methods in Molecular Biology*, Vol. 112, Andrew Link (ed.), Humana Press, Totowa, NJ: 339–410.

5. Lemkin, P. F. and Thornwall, G (2002) Comparing 2D Electrophoretic gels across databases. In *Protein Protocol Handbook*, J. M. Walker (ed.), Humana Press, Totowa, NJ, pp 197–214.
6. Lemkin, P. F., Merrill, C., Lipkin, L., et al. (1979) Software aids for the analysis of 2D gel electrophoresis images. *Comput Biomed Res* **12**, 517–544.
7. Lipkin, L. E. and Lemkin, P. F. (1980) Data-base techniques for multiple two-dimensional polyacrylamide gel electrophoresis analyses. *Clin Chem* **26**, 1403–1412.
8. Lemkin, P. F. and Lester, E. P. (1989) Database and search techniques for two-dimensional gel protein data: a comparison of paradigms for exploratory data analysis and prospects for biological modeling. *Electrophoresis* **10**, 122–140.
9. Lester, E. P., Lemkin, P. F., Lowery, J. F., and Lipkin, L. E. (1982) Human leukemias: a preliminary 2-D electrophoretic analysis. *Electrophoresis* **3**, 364–375.
10. Lemkin, P. F., Thornwall, G. C., Walton, K. D., and Hennighausen L. (2000) The Microarray Explorer tool for data mining of cDNA microarrays—application for the mammary gland. *Nucleic Acids Research* **28**, 4452–4459.
11. Herbert, B. R., Pederson, S. L., Harry, J. L., et al. (2003) Mastering genome complexity using two-dimensional gel electrophoresis. *PharmaGenomics* **Sept**, 22–36.
12. Hood, L. (2002) A personal view of molecular technology and how it has changed biology. *J Proteome Biology* **1**, 399–409.
13. Herbert, B. R., Pedersen, S. K., Harry, J. L., et al. (2003) Mastering proteome complexity using two-dimensional gel electrophoresis. *PharmaGenomics* **3**, 21–36.
14. Pedersen, S. K., Harry, J. L., Sebastian, L., et al. (2003) Unseen proteome: mining below the tip of the iceberg to find low abundance and membrane proteins. *J Proteome Biology* **2**, 303–311.
15. Pieper, R., Su, Q., Gatlin, C. L., Huang, S.-T., Anderson, N.L., and Steiner, S. (2003) Multi-component immunoaffinity subtraction chromatography: an innovative step towards a comprehensive survey of the human plasma proteome. *Proteomics* **3**, 422–432.
16. Pieper, R., Gatlin, C. L., Makusky, A. J., et al. (2003) The human serum proteome: display of nearly 3700 chromatographically separated protein spots on two-dimensional electrophoresis gels and identification of 325 distinct proteins. *Proteomics* **3**, 1345–1364.
17. Anderson, N. L. and Anderson, N. G. (2002) The human plasma proteome: history, character, and diagnostic prospects. *J. Mol. Cellular Proteomics* **1**, 845–867.
18. Liotta, L. A., Espina, V., Mehta, A. I., et al. (2003) Protein microarrays: meeting analytical challenges for clinical applications. *Cancer Cell* **3**, 317–325.
19. Hoving, S., Gerrits, B., Voshol, H., Muller, D., Roberts, R. C., and van Oostrum, J. (2002) Preparative two-dimensional gel electrophoresis at alkaline pH using narrow range immobilized pH gradients. *Proteomics* **2**, 127–134.
20. Von Eggeling, F., Gawriljuk, A., Fiedler, W., et al. (2001) Fluorescent dual colour 2D-protein gel electrophoresis for rapid detection of differences in protein pattern with standard image analysis software. *Int J Mol Med* **8**, 373–377.
21. Lopez, M. F., Mikulskis, A., Golenko, E., et al. (2003) High-content proteomics: Fluorescence multiplexing using an integrated, high-sensitivity, multi-wavelength charge-coupled device imaging system. *Proteomics* **3**, 1109–1116.
22. Sanchez, J.-C., Appel, R. D., Golaz, O., et al. (1995) Inside SWISS-2DPAGE database. *Electrophoresis* **16**, 1131–1151.
23. Appel, R. D., Bairoch, A., Sanchez, J-C., et al. (1996). Federated two-dimensional electrophoresis database: a simple means of publishing two-dimensional electrophoresis data. *Electrophoresis* **17**, 540–546.
24. Appel, R. D., Sanchez, J-C., Bairoch, A., et al. (1993) SWISS-2DPAGE: A database of two-dimensional gel electrophoresis images. *Electrophoresis* **14**, 1232–1238.

25. Appel, R. D., Sanchez, J.-C., Bairoch, A., et al. (1994) SWISS-2DPAGE database of two-dimensional polyacrylamide gel electrophoresis. *Nucleic Acids Res.* **22(17)**, 3581–3582.
26. Arnold, K., and Gosling, J. (1996) *The Java Programming Language*. Addison Wesley, NY.
27. Wolberg, G. (1990) *Digital Image Warping*. IEEE Computer Press Monograph, Los Alamitos, CA.

Sample Cleanup by Solid-Phase Extraction/Pipet-Tip Chromatography

Alastair Aitken

1. Introduction

Sample cleanup by solid-phase extraction can be achieved with the use of a ZipTip (see Note 1), which is a miniature reverse-phase column packed into a 10 μL pipet tip with a micro volume (approx 0.5 μL) bed of reversed-phase, ion exchange, or affinity chromatography medium fixed at its end without dead volume. It is ideal for concentrating, desalting, fractionating, and enriching 1 to 100 μL of protein, peptide, or oligonucleotide samples prior to analysis. ZipTip μ -C₁₈ contains a micro bed of resin that allows final elution volumes as low as 0.5 μL . It is intended for purifying and concentrating femtomoles to picomoles of protein, peptide, or oligonucleotide samples prior to analysis, providing better data quality. The sample is aspirated and dispensed through ZipTip to bind, wash, and elute. Recovered samples are contaminant free and eluted in 0.5–4 μL for direct transfer to a mass spectrometer matrix-assisted laser desorption/ionization (MALDI) target or vial. Complex peptide and protein mixtures can be partially fractionated to enrich hydrophilic and hydrophobic components by eluting with an acetonitrile step gradient. The separation of different-sized fragments not only simplifies spectra but also minimizes peak suppression of larger fragments.

2. Materials

1. Gloves and lab coat (see Note 2).
2. Polypropylene microcentrifuge tubes (see Note 3): 500 or 1500 μL with snap caps.
3. Pipet such as Gilson or Finn pipet P-10 or compatible automated liquid-handling workstation.
4. ZipTipC₁₈, ZipTip μ -C₁₈, and ZipTipC₄ pipet tips.
5. Wetting solution: 50% acetonitrile in 0.1% trifluoroacetic acid (TFA) in high-purity water (e.g., Milli-Q water).
6. Equilibration and washing solution: 0.1% TFA in high-purity water.
7. Sample preparation solution: 2.5% TFA in Milli-Q water (5X stock solution).

3. Methods

3.1. Sample Concentration and Buffer Removal

1. Place the ZipTip pipet tips on the single- or multi-channel pipet, or automated liquid-handling station. Bind the sample by aspirating through the resin several times (see Note 4).

2. To equilibrate the ZipTip for sample binding, prewet the tip by depressing the pipet plunger to a dead stop using the maximum volume setting of 10 μ L (see **Note 5**).
3. Aspirate the wetting solution into tip. Dispense to discard. Repeat.
4. Equilibrate the tip for binding by washing it twice with the equilibration solution.
5. Bind peptides and proteins to the ZipTip by fully depressing the pipet plunger to a dead stop.
6. Aspirate and dispense sample 3 to 7 cycles for simple mixtures and up to 10 cycles for the maximum binding of complex mixtures. Dilute solutions require increased contact time (see **Note 6**).
7. To bind and wash the peptides or proteins, wash the tip and dispense to discard, using at least two cycles of wash solution. A 5% methanol in 0.1% TFA/water wash can improve desalting efficiency.
8. Wash away contaminants and unwanted biomolecules that do not bind.
9. Elute the concentrated, purified sample in 1 to 4 μ L of compatible solvent such as 50–75% acetonitrile in 0.1% TFA in high-purity water and directly transfer the solution to a MALDI-time-of-flight (TOF) mass spectrometry (MS) target or a polypropylene microcentrifuge tube for electrospray MS. Acetonitrile is volatile, and evaporation can occur rapidly. If this occurs, add more eluant to recover the sample.
10. After use, the tips may be washed and regenerated for reuse up to three times.

3.2. Fractionation of Complex Peptide and Protein Mixtures

1. Prepare varying concentrations of ACN in high-purity water (e.g., 5%, 10%, 20%, 30%, 50%, and 70%) with or without 0.1% TFA.
2. Step-gradient elution of peptides or proteins is achieved by pipetting 1 to 3 μ L of 5% acetonitrile/0.1% TFA into clean vial using a standard pipet tip.
3. Carefully aspirate and dispense this eluant through the ZipTip at least three times without introducing air. If desired, use the final wash cycle to apply the desalted-concentrated peptides or proteins directly onto the MALDI-TOF MS target. The sample can be directly spotted in matrix onto the MALDI plate by eluting with the solution containing the desired matrix (e.g., 0.5 μ L of cyano-4-hydroxy-cinnamic acid).
4. Thoroughly wash the tip immediately, using three cycles of 5% acetonitrile, prior to subsequent elution step, and wash the tip with respective eluant prior to increasing ACN solution, to minimize peptide or protein carry-over.
5. Perform the step gradient (e.g., 10, 20, 30, then 50% ACN) by increasing acetonitrile concentration, and repeat **steps 1–4** until the step-gradient is completed.

4. Notes

1. ZipTip is a trademark of Millipore Corporation. Similar products, such as the Stylus ProTip, can be purchased from other companies, or one can prepare one's own pipet tip columns. ZipTipC₁₈ and μ -C₁₈ are most applicable for low-molecular-weight proteins and peptides, while ZipTipC₄ is most suitable for low- to intermediate-molecular-weight proteins. Higher-molecular-weight proteins tend to adsorb strongly to hydrophobic surfaces; therefore, ZipTipC₄ is recommended for proteins over 100,000 Da. The capacity (when used with saturating amounts of analyte) of the C₁₈ ZipTip is >1.0 μ g (typically 5.0 μ g); capacity of C₁₈ is typically 2.0 μ g, and C₄ is >0.5 μ g (typically up to 3.3 μ g). Resins for a variety of other applications are commercially available. Anion Exchange (AX) ZipTips will remove non-ionic detergents from peptides, proteins, and oligonucleotides. Metal chelate (MC) ZipTips can be used to enrich phosphopeptides (see Chapter 44) and purify His-tagged proteins.

2. Gloves and lab coats must be worn at all times to avoid keratin contamination. Work on a clean surface.
3. Use clean polypropylene microcentrifuge tubes, 500 or 1500 μL , with snap caps. Test first to confirm that they are satisfactory and do not leach out polymers, mold release agents, plasticizers, and so on. Set aside a box for digest use only; handle only with gloves. Use only clean tools, containers, and reagents for anything that will come in contact with the samples. Keep samples capped at all times except while they are being processed.
4. Low or zero organic solvent in the buffer (e.g., 0.1% TFA) will retain peptides and proteins on a ZipTip but remove common buffers and salts, such as: 2 M NaCl, 100 mM phosphate, 8 M urea, 6 M guanidine, and 50% glycerol. If there is a considerable amount of detergent, dilute sample with 0.1% TFA to achieve acceptable binding conditions; for example, reduce SDS to below 0.1%, Triton to below 1%, and Tween to below 0.5%. Optimal binding of protein to ZipTip may also require a chaotropic agent (e.g., guanidine-HCl at a final concentration of 1–4 M). In the case of proteins, 8 M guanidine and 2.5% TFA in high-purity water (5X stock solution) may be used if sample solubility is a problem. Maximum binding to the ZipTip is achieved in the presence of TFA or other ion-pairing agents. The final TFA concentration should be between 0.1% and 1.0% at a pH of <4.0.
5. Since the resin bed produces a slight backpressure, the pipet should not be used as an accurate volumetric dispenser. To achieve optimal sample uptake and delivery, set the pipet to 10 μL and attach tip securely. Depress the plunger to the dead stop and slowly release or dispense plunger throughout operation.
6. Dilute samples can be concentrated by adsorbing analyte from multiple 10- μL aliquots into the ZipTip and eluting into a small volume, effecting a 10- to 50-fold concentration.

Suggested Readings

1. Information on using ZipTips is available at www.millipore.com/ziptip.

Protein Identification by In-Gel Digestion and Mass Spectrometric Analysis

Michele Learmonth and Alastair Aitken

1. Introduction

This chapter describes the analysis of proteins that have been separated by one- (or two-) dimensional gel electrophoresis. In-gel digestion of the protein bands or spots from 1-D gels is detailed. Normally the proteins are digested with trypsin, which cleaves after lysine and arginine. Therefore, due to the specificity of this protease, this results in basic charges at both the amino-terminal and carboxy-terminal ends of most peptides. Not all peptides will have basic groups at both ends, since the N-terminus of an intact protein is commonly blocked by an acetyl or a number of other modifications and the C-terminus may well not be a lysine or arginine. Doubly charged peptide species will, however, predominate. These are of high energy, and fragment more easily in tandem (including ion-trap) mass spectrometry (1). This results in substantial sequence information, leading to more powerful database analysis. There are programs available, such as Sequest (2), that allow the databases to be searched directly with the peptide fragment data (tandem MS-MS data). This results in a large saving of time and effort, since the data need not be interpreted into tentative peptide sequences, which can be extremely tedious. Glu-C may also be used in place of trypsin, although fewer doubly charged species will result and analysis may be limited to the mass fingerprint data. This is perhaps more suitable to mass analysis on the highly sensitive matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometers, where few sequence or mass fragment data are obtained in any case.

2. Materials

1. Sodium dodecyl sulfate (SDS)-polyacrylamide gel electrophoresis (PAGE) apparatus.
2. 0.1% Colloidal Coomassie Blue, GELCODE, from Pierce Warriner, or SYPRO® fluorescent dyes.
3. 0.2 M NH_4HCO_3 .
4. 50% Aqueous acetonitrile.
5. Microcentrifuge.
6. 30 mM Potassium ferricyanide (stock solution, made fresh).
7. 100 mM Sodium thiosulphate (stock solution, made fresh).
8. Dithiothreitol (DTT).
9. Centrifugal evaporator (Savant or Gyrovap).

10. Trypsin (Promega sequencing grade, modified by reductive methylation to remove pseudotrypsin activity and TPCK treated).
11. Digestion buffer: 0.2M NH_4HCO_3 or 50 mM Tris-HCl (pH 7.5).
12. 25- μL Hamilton syringe.
13. Formic acid.
14. Coated nanospray needle (PicoTip from New Objective [Econo 12]).
15. C_8 reverse-phase column (0.8 mm \times 2mm, 300 \AA), L.C. packings.
16. MALDI-TOF plates.
17. Shaking platform.

3. Methods

3.1. SDS Acrylamide Gel

Use of a shaking platform allows more efficient gel washing for these steps.

1. Run 1-D or 2-D SDS acrylamide gel under conditions suitable for proteins of interest (*see Note 1*).
2. Wash gel in 200 mL water for 5 min. Repeat three times.
3. Visualize proteins by staining gel in SYPRO fluorescent dye or 0.1% colloidal Coomassie Blue (GELCODE, *see Note 2*) for approx 1 h.
4. De-stain in water for 1–2 h (*see Note 3*).

3.2. In-Gel Digestion

More efficient removal of buffer and solvent from the gel pieces can be achieved with a P1000 pipet tip with a P10 tip stuck on the end.

1. Excise the stained gel band (identified by SYPRO or colloidal Coomassie Blue) by cutting out center (most concentrated part of band) to minimize the amount of acrylamide.
2. Incubate three times (in approx 200 μL 0.2M NH_4HCO_3 /50% acetonitrile) for 30 min each at 30°C to remove the SDS.
3. Incubate the gel band in DTT (20 mM) in 200–300 μL of 0.2 M NH_4HCO_3 /50% aqueous acetonitrile for 1 h at 30°C to reduce the proteins.
4. Wash three times in approx 200 μL 0.2 M NH_4HCO_3 /50% acetonitrile.
5. Alkylate cysteine residues (*see Note 4*) in fresh iodoacetamide (50 mM) in 100 μL , 0.2 M NH_4HCO_3 /50% acetonitrile for 20 min at room temperature in the dark.
6. Wash three times in 500 μL 20 mM NH_4HCO_3 /50% acetonitrile.
7. Cut the band into 1 \times 2 mm pieces.
8. Centrifuge for 2 min at 10,000g (or top speed) in a microcentrifuge.
9. Cover pieces with acetonitrile (they must turn opaque white).
10. Dry the gel band completely by centrifugal lyophilization in centrifugal evaporator (approx 30 min).
11. Rehydrate gel band with trypsin solution. Use approx 0.5–1.0 μg trypsin freshly made up in 60 μL 50 mM NH_4HCO_3 , for 15–30 min at 4°C (*see Note 5*).
12. Add sufficient digestion buffer to make up to about 100 μL , i.e., enough to cover the gel pieces.
13. Incubate at 32°C, overnight—i.e., approx 16 h.

3.3. Peptide Extraction

1. The following day, centrifuge for 2 min at 10,000g (or top speed) in microcentrifuge.
2. Collect the digest buffer from above the gel pieces.

3. Add 100–200 μ L 50% acetonitrile to the gel pieces, sonicate in a sonic bath for approx 30 min (at 35–40°C), leave 1 h, centrifuge as before, and collect the supernatant.
4. Dry the acetonitrile extract to approx 20–50 μ L in the centrifugal evaporator, then combine with the aqueous extracts.
5. This gives a total volume of approx 100–150 μ L. Store at –20°C if not analyzed by MS immediately.
6. Desalt peptides for mass fingerprinting and/or sequencing by tandem ESMS or MALDI-TOF MS (see Chapters 31 and 33). For nanospray on an ion-trap mass spectrometer, use the protocol described in **Subheading 3.4.**

3.4. Peptide Identification by Nanospray MS Analysis of Smaller Proteins (<60 kDa)

1. For nanospray MS, fit the C₈ reverse-phase column (0.8 mm \times 2 mm) with PTFE tubing (3.5 cm) at the inlet end to allow syringe needle to fit snugly. Fit the outlet end with 5.5 cm of 50- μ fused silica tubing (see **Note 6**) to allow this to insert into nanospray needle.
2. Wash the C₈ reverse-phase column (0.8 mm \times 2 mm) with 200 μ L formic acid (0.01% in 95% acetonitrile).
3. Equilibrate the column with approx 200 μ L aqueous formic acid (0.01%).
4. The sample (dissolved in approx 20 μ L aqueous formic acid, 0.01%) is loaded slowly onto the column using a 25- μ L Hamilton syringe. If the solution is more dilute (i.e., if more sample is required to give good spectra), load up to 3 \times 25 μ L.
5. Wash column with 15 μ L aqueous formic acid (0.01%).
6. Elute directly into the nanospray needle (see **Note 7**) with aqueous formic acid (0.01%) containing 60% methanol to allow 2 μ L to enter needle.
7. After use, store column in 95% aqueous acetonitrile.

3.5. Nanospray MS Analysis of Larger Proteins (>60 kDa)

1. Follow **Subheading 3.4.**, steps 1–5
2. Elute directly into the nanospray needle with 2 μ L aqueous formic acid (0.01%) containing 20% methanol, then in 10% incremental steps of methanol to allow 2 μ L to enter needle at each step (see **Note 8**).
3. If protein is large (i.e., >200 kDa) start with aqueous formic acid (0.01%) containing 10% methanol, then elute in 5% incremental steps of methanol.

4. Notes

1. SDS-PAGE minigels with 0.5-mm spacers are excellent, because this minimizes the amount of acrylamide in the gel piece.
2. The methanol/acetic acid used in conventional Coomassie blue staining procedures will fix the protein in the gel to a varying degree, resulting in much lower recovery.
3. Gel pieces may be stored in water over a weekend. If longer storage is required, keep at –20°C.
4. The advantage of alkylating cysteine residues in gel is to avoid difficulty in removing DTT and iodoacetamide when carrying out the reaction in solution. Some researchers alkylate the cysteines with iodoacetic acid, which is equally effective and is a matter of personal choice.
5. The solution is initially kept at 4°C until the gel is swollen, to minimize autodigestion of trypsin. A rough guide is to use an enzyme substrate ratio of about 1:10. Use 1–3 μ L (0.4 to 1.2 μ g) trypsin solution from stock made by adding 50 μ L NH₄HCO₃ (0.2 M) to one vial (20 μ g). Do not dilute with the Promega resuspension buffer supplied, since it is

50 mM acetic acid. In view of the inadvisability of storing the trypsin solution at this high dilution, better results are obtained if a large number of samples are processed simultaneously.

6. Use 1/32-in ferrules with a reducing ferrule fitting for the fused silica.
7. Preopened needles (PicoTip) from New Objective offer many advantages. With these tips, there is no need to open by pressing against the inlet. Otherwise, there is a risk of loss of precious sample if the tip breaks badly.
8. This reduces the otherwise large number of peptides in each fraction.
9. If necessary, the gel bands may be recovered after identification by a silver-staining protocol that uses thiosulphate sensitization without glutaraldehyde fixation, but this method is less highly recommended. The silver can be removed by incubating with 50 µL/band of 15 mM potassium ferricyanide/50 mM sodium thiosulphate (made fresh from double-strength stock solutions) for 5–10 min until the band pieces go clear (i.e., until all the silver is removed). The supernatant is removed and the procedure continued as in **Subheading 3.2., step 2.**

References

1. Jensen, O. N., Wilm, M., Shevchenko, A., and Mann, M. (1999) Peptide sequencing of 2-DE gel-isolated proteins by nanoelectrospray tandem mass spectrometry. *Methods Mol. Biol.* **112**, 571–588.
2. Yates, J. R., 3rd (1998) Database searching using mass spectrometry data. *Electrophoresis* **19**, 893–900.

Peptide Sequences of 2-D Gel-Separated Protein Spots by Nanoelectrospray Tandem Mass Spectrometry

Alastair Aitken

1. Introduction

This chapter describes the analysis of 2-D gel-separated protein spots by nanospray or nanoelectrospray tandem mass spectrometry (MS). The samples may also have originated from one-dimensional gel electrophoresis separations, or from micro- or nanobore multidimensional chromatography. The principles of the technique described below to obtain nanospray into an ion trap or into hybrid MS instruments are similar (1).

The digestion, desalting, and preparation of peptide samples from an in-gel digestion for “static” nanospray are described in Chapter 30, and 2-D gel electrophoresis is discussed in Chapters 1–20. Static nanospray is particularly useful for MSⁿ to obtain structural information on purified samples or components of a simple mixture. Sample (approx 1 μ L) is loaded into a metal-coated pulled-glass capillary with internal diameter (i.d.) 1 to 4 μ m (see Note 1). Voltage is applied to a metal coating on the tip (there is normally no liquid junction) and an air-filled syringe provides backpressure to initiate and maintain electrospray. Once commenced, flow rate is maintained by the electric field and solvent properties. No external pumping system is employed in static nanospray, and solvent flow is maintained by the electric field itself. Flow rate can be adjusted by modifying the field strength parameters, e.g., the needle or emitter position, typically 1 to 20 nL/min. By contrast, in dynamic mode, a pump is used, and to maintain a stable nanospray, a smaller-i.d. emitter is used. Typical flow rates are 100 to 3000 nL/min (compared to 1 μ L to 1 mL/min in conventional liquid chromatography [LC]/MS). Properties of solvent also affect field strength and optimum flow rate; since solvent must be removed from analyte by evaporation, mixtures of organic/aqueous solvent are used, with volatile acids/bases (ammonium salts may be used). The main difference between electrospray and nanospray is the amount of material loaded. By eluting the peptides in a significantly lower volume, nanospray can reach the same concentrations with much lower amounts of peptide. Since nanospray methods permit direct operation of the electrospray source at low flow rates (nL/min), the inside diameter of the column can be reduced by orders of magnitude with virtually no loss in analysis. A 75- μ m-i.d. nanobore column operating at 250 nL/min has a volume approx 3000 times less than a conventional 4.6-mm-i.d. column. Theoretically, one can obtain an analyte concentration factor of approx 3000-fold, and a much higher signal-to-noise

response with the mass spectrometer. In practice, sensitivity improvements well in excess of 500-fold are commonly obtained. The principal improvement, therefore, is that samples with strict quantity limitations, such as 2-D gel-separated proteins, are now readily analyzed at limits of detection typically in the sub-femtomole to attomole range (see Note 2).

2. Materials

1. Mass spectrometer with NanoES source.
2. Nanospray needles, PicoTip from New Objective (e.g., Econo 12).
3. Gel loading tips, e.g., 350 μ m Eppendorf (cat. no. 22 35 465-6).
4. C₈ and C₁₈ reverse-phase columns (0.8 mm \times 2 mm, 300 \AA), e.g., PepMap, Vydac.
5. Fused-silica capillary.
6. Strong cation exchange columns, e.g., L.C. packings, Poros 10S, or Partisphere SCX, Whatman.
7. 50% Aqueous acetonitrile.
8. Microcentrifuge.
9. 10- or 25- μ L Hamilton syringe.
10. Formic acid.
11. Methanol.

3. Methods

3.1. Filling Nanospray Needles for Static Nanospray

The two main methods for loading samples into nanospray needle tips, using fused-silica needles and gel-loading tips, are described below (see Note 3).

3.1.1. Fused-Silica Needles Connected to a Conventional Syringe

1. Substitute a fused-silica needle for a stainless-steel needle of a conventional syringe (with a maximum volume of 10 or 25 μ L) from Hamilton and other suppliers (see Note 4).
2. Centrifuge the sample in a bench-top centrifuge to remove particulates, and load supernatant into the syringe.
3. Insert the filling needle into the distal end of the nanospray needle (see Fig. 1).
4. Push the needle as far in as possible without damaging either the filling needle or the nanospray needle tip. It is not critical to reach right into the end of the tapered region, since capillary action will fill the tip.
5. Slowly inject the liquid into the PicoTip. Rapid injection can lead to air bubbles, or foaming of the sample. This foaming is especially problematic with concentrated protein and peptide samples. To prevent this, inject sample (1–5 μ L) over a period of approx 5 s.
6. Slowly withdraw the filling needle from the PicoTip. The tip will fill by capillary action. Air bubbles may appear along the shank of the tip, but as the sample sprays from the tip, capillary action will provide a continuous feed and should eliminate the air bubbles in the taper region.

3.1.2. Gel-Loader Tips

Gel-loader disposable pipet tips with an o.d. less than or equal to 0.35 mm are also used to load samples into nanospray needle tips.

1. Centrifuge the sample in bench-top centrifuge to remove particulates and load supernatant into the smallest-o.d. gel-loader tip available for your particular make of pipet (see Note 6).
2. Insert the pipet tip as far as possible into the distal end of the nanospray needle tip (emitter) and deliver 1–5 μ L of the liquid slowly while removing the pipet tip. A typical gel-

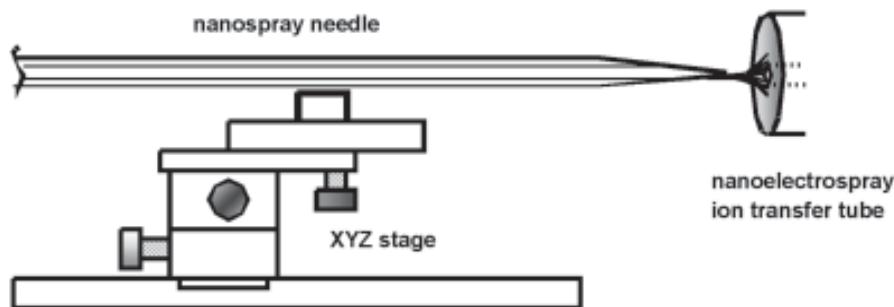


Fig. 1. Nanospray source.

loader tip is not sufficiently long to extend all the way to the tip of the emitter. The sample will fill only the back end of the tip, but capillary action will bring the sample into the tip.

3. Wait a few minutes for this filling action to take place, if necessary.
4. Inspect for proper filling with transmitted light (rather than reflected light) microscope at $\times 50$ –100.

3.1.3. Optimizing Nanospray

Applied voltage is perhaps the most important parameter for stable, efficient operation. To prevent an arc or corona discharge, do not use a turn-on voltage above 500 V unless a stable electrospray ionization (ESI) has been previously established.

1. Position tip with respect to the center of the ion-transfer tube using *x* and *y* adjustments on the *xyz* stage. This should be within 1 mm of center (see Fig. 1).
2. Using the *z* adjustment knob, position emitter approx 2 mm away from ion transfer tube.
3. Start tuning at a low voltage, under 1 kV, and increase the operating potential in 100 V increments until stable operation is achieved. In low-flow electrospray, the flow rate and the applied electric field are interdependent. For a given tip size, stable electrospray ionization can occur over a wide range of flow rates, but only over a narrow range of field strength (50 V or less). Raising the flow rate requires a higher field strength, and vice versa.
4. Tune for an even electrospray ionization plume of droplets, by observing the magnified image of the spray pattern on the video monitor.
5. Flow can be maintained by increasing pressure slightly with the back-pressure syringe.
6. Full optimization will depend on the MS instrument type, and will require different optimization parameters for dynamic nanospray (see Note 7).

4. Notes

1. Preopened needles—e.g., PicoTip from New Objective—offer a number of advantages. With these tips, there is no need to open by pressing against the inlet. Otherwise, there is a risk of loss of precious sample if the tip breaks badly. New Objective markets PicoFrit columns, which couple nanobore-LC directly to ESI/MS. A biocompatible frit adds very little back pressure. Columns can be packed with C₁₈ or strong cation exchange (SCX) material for sample trapping and desalting procedures, and are compatible with flow rates between 20 and 2000 nL/min for direct infusion or capillary LC/MS applications (see Chapter 37).
2. Various “flowing nanospray” sources are commercially available that are compatible with a particular range of ion-trap and electrospray MS instruments. Capillary columns (with

i.d. between 50 and 500 μm) are connected to the emitter tip (with i.d. between 5 and 75 μm) by a T-union (Valco μTee) with a gold wire 0.3-mm electrode in the third port, to provide electrical contact. There are cheaper options that are not enclosed, where the setup is open and the needle is connected to the high voltage by an alligator-type clip. Although much more expensive, choose a nanospray source that is shielded or enclosed to minimize risk of electric shock, as the coated needle is typically held at a 1-3-KV potential in order to facilitate ionization of the spray. These include MicroIonSpray Source for the API 150EX, API 3000, and API QSTAR Pulsar, a very low flow ion spray source. Finnigan supplies a nanospray ionization source (NSI) for their LCQ series ion-trap mass spectrometers. Micromass markets the Z-spray Nanoflow stage for their Q-time-of-flight (TOF). In addition, both New Objectives and Proxeon Biosystems, Denmark (formerly Protana), manufacture a wide range for other manufacturers' instruments. In the nanospray source, the position of the needle is altered to give the optimal spray into MS source by adjusting the position with the x , y , z controls. This is assisted by a $\times 50$ –200 magnification camera, which provides optimal viewing of the position of the needle tip. The camera is also easily positioned and focused with knob-driven controls. A fiber-optic flexible light source is also normally supplied.

3. Filling with a bench-top centrifuge is not recommended, since, unless the speed is kept very low, the sample will be ejected through the tip.
4. Needle-style RNFS from Hamilton, for example, syringe model 1701-RNFS, part number 87404, which has a 10-cm long, 170- μm o.d. flexible fused-silica needle. Fused-silica syringe needles can reach within 0.2 mm of the tip. Remember that the o.d. of the filling needle must be less than the i.d. of the nanospray needle glass tubing.
5. Alternatively, elute the desalted peptides (normally digested with trypsin) from a micro reverse-phase clean-up column fitted with silica capillary tubing (see Chapter 29) directly into the nanospray needle with aqueous formic acid (0.01%) containing 60% methanol to allow 2 μL to enter needle. To reduce complexity of the peptides if a larger protein spot is being analyzed, elute with 2 μL aqueous formic acid (0.01%) containing 20% methanol, then in 10% incremental steps of methanol to allow 2 μL to enter needle at each step.
6. Micro-gel loader tips designed for loading samples onto thin gels are available from Eppendorf (pipet tip number 2235-165-6) and other suppliers.
7. Possible problems that may be encountered, depending on whether some sample is seen to remain in the needle tip, include: bad connection between voltage supply and needle; an air bubble trapped in the needle tip, disrupting flow; or a clogged tip. If analyte spectra are observed for a very short time only, this may be due to a broken or badly opened tip, resulting in high flow rate or ejection of sample due to high back pressure.

References

1. Jensen, O. N., Wilm, M., Shevchenko, A., and Mann, M. (1999) Peptide sequencing of 2-DE gel-isolated proteins by nanoelectrospray tandem mass spectrometry. *Methods Mol. Biol.* **112**, 571–588.
2. www.newobjective.com, Technical Note BG-1. “Filling PicoTips™ for Static Nanospray.”

Identification of Proteins by MALDI-TOF MS

Alastair Aitken

1. Introduction

Matrix-assisted laser desorption/ionization (MALDI) time-of-flight (TOF) mass spectrometry (MS) produces gas-phase protonated ions by excitation of the sample molecules from the energy of a laser transferred via an ultraviolet (UV) light-absorbing matrix. The matrix is a conjugated organic compound (normally a weak organic acid). Different MALDI matrix compounds are optimal for particular biomolecules, and maximally absorb light at the wavelength of the laser, typically a nitrogen laser of 337 nm (*see Note 1*).

Accurate molecular weights of large biomolecules such as proteins of mass greater than 400 kDa (*1*) have been accurately measured by MALDI-TOF (*see Note 2*). The lower mass limit of MALDI-TOF MS is approx 500–800 Da due to interference from fragmentation and adduct ions of the matrix. The particular advantage of MALDI is the ability to produce large-mass ions with high sensitivity, since it is a very “soft” ionization method (*see Fig. 1*); it is also a valuable technique for examining mixtures, since the spectra are mainly composed of singly charged intact protein and peptide ions. The sensitivity is easily sufficient to analyze protein isolated from one-dimensional (1-D) or 2-D polyacrylamide gel electrophoresis (PAGE) gels. Depending on the mass range, internally calibrated MALDI-TOF spectra may be accurate to approx 5 parts per million (ppm), and externally calibrated spectra are accurate to approx 50–100 ppm (*see Note 3*).

Other advantages of MALDI-TOF include its use in high-throughput applications and its relative tolerance of buffer and salt contaminants in the sample, in comparison to electrospray (ES) MS (*see Note 4*). Proteins and peptides are generally analyzed in the positive ion mode, which favors production of protonated ($M+H$)⁺ ions, although in the presence of the appropriate buffer ions, sodium ($M+Na$)⁺, potassium ($M+K$)⁺, and ammonium ($M+NH_4$)⁺ adducts may be formed.

Direct analysis of protein complexes by mass spectrometry is also possible (*2*). Direct analysis by MALDI-TOF mass spectrometry of samples bound to surface plasmon resonance (Biacore) chips is now possible (*3*).

There are a number of commercial developments of hybrid MS instruments, which involve coupling an electrospray, ion trap, or a MALDI ion source with a hybrid quadrupole orthogonal acceleration time-of-flight mass spectrometer. This leads to improved tandem MS performance from MALDI samples. These instruments combine

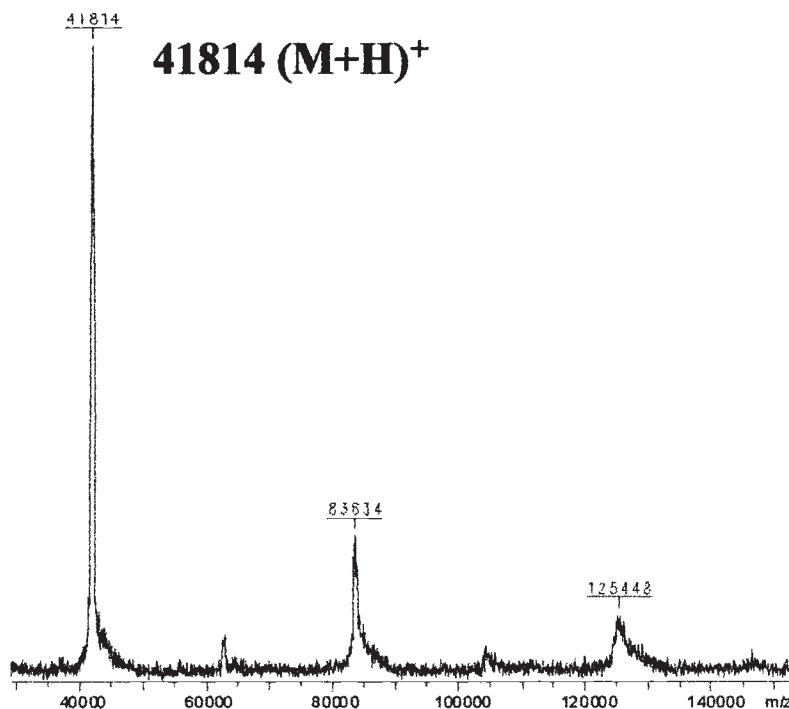


Fig. 1. Matrix-assisted laser desorption/ionization (MALDI)-time-of-flight (TOF) spectrum of actin. The spectrum shows the presence of skeletal muscle actin (41,816.7 Da) and oligomeric forms, determined to high accuracy on a Bruker MALDI-TOF Reflex III instrument by Andrew Illsey and Jackie Jarvis. Approximately 10-pg sample was loaded on the target plate.

the best features of both types of ion source with the best features of all types of analyzer in order to improve MS-MS capability and increase sensitivity (4).

2. Materials

1. MALDI-TOF mass spectrometer (4).
2. MALDI plates (*see Note 5*).
3. 3,5-dimethoxy-4-hydroxycinnamic acid, sinapinic acid (SA, Fluka).
4. α -cyano-4-hydroxycinnamic acid (CHCA/HCCA, Sigma).
5. 2,5-dihydroxybenzoic acid (DHB, Aldrich).
6. Saturated solution of matrix in a 1:1 v/v acetonitrile (ACN)/0.1% aqueous trifluoracetic acid (TFA) (e.g., matrix solution for protein analysis: SA (10 mg/mL) in 30% or 50% acetonitrile/0.1% TFA, v/v). Make up fresh stock weekly.
7. High-purity water.
8. Low-volume pipet such as a 2- μ L Gilson or FinnpipetteTM.
9. Microcentrifuge.
10. 0.5-mL Eppendorf tubes.
11. Vortex mixer.
12. Centrifugal evaporator (Savant or Gyrovap).
13. Propan-2-ol.
14. 2:1 v/v chloroform/methanol solution.

3. Methods

3.1. Sample Preparation for Peptides/Proteins

1. Adjust the concentration of the sample (see Note 6). The optimum concentration for peptides and proteins is 5–50 pmol/μL, giving a final concentration of 0.5–5 pmol/μL after mixing with the matrix (see Note 7).
2. If the concentration is unknown, make a series of dilutions in 0.1% TFA to produce a series of sample spots on the MALDI plate.
3. If sample is very dilute, concentrate by Speed-Vac or pipet-tip chromatography (Chapter 29).
4. Prepare a saturated solution of the matrix in a 1:1 v/v ACN/0.1% TFA solution. Add the aqueous TFA solution, then the ACN. Vortex the mixture for 60 s, then centrifuge for 20 s. Use only the supernatant, since any particulates in the solution can act as nucleation sites and cause inhomogeneity in the crystals.
5. Spot the matrix and the sample on the plate separately by placing 0.5 μL of the analyte solution on the plate, immediately followed by 0.5 μL of the matrix solution. Do not touch the surface of the plate with the pipet tip, otherwise uneven crystallization will occur.
6. Allow to air-dry. The mixture forms a co-crystalline sample/matrix complex.
7. Alternatively, mix the sample in solution (1–10 pmol/μL) with an excess of the matrix (in a ratio between 1000:1 and 10,000:1) in a 0.5-mL Eppendorf tube. Mix with the pipet tip or vortex for 5 s. Transfer 1 μL of this solution to the sample plate and allow to dry completely.
8. Place the target plate containing all the samples in the instrument and evacuate. Pulses of laser light are applied, which causes rapid excitation and vaporization of the crystalline matrix and subsequent ejection into a plume of matrix and analyte ions, which are analyzed by their time of flight (see Note 8).
9. With the aid of the camera that is used to follow the exact position of the laser flash, move the laser beam around the MALDI plate to find so-called sweet spots, where the composition of co-crystallized matrix and sample is optimal, in order to obtain good sensitivity.
10. Once a good quality signal has been obtained by optimizing the parameters, a few hundred (typically) pulses of laser light are used to accumulate sufficient signals to generate a good mass/charge (m/z) spectrum. The laser can be tracked cross the circular spot automatically by a program (a macro) during this time to cover the complete area of the spot, in order to ensure that the sample/matrix sweet spots are targeted. The tracking can be lateral across and down the spot or in a spiral, for example.
11. Calibrate the masses in the spectrum externally or internally using appropriate molecular-weight standards analyzed under similar conditions.

3.2. On-Plate Washing to Remove Buffer and Salts

If there are known (or suspected) high levels of contaminant that may cause problems with the analysis (see Note 4), the dried crystalline sample on the target plate may be washed with a few μL of water or 0.1% aqueous TFA, provided the matrix is not highly water soluble. This simple washing step frequently results in a dramatic improvement in the spectral data. If the sample is sufficiently concentrated, it can simply be diluted to minimize interference from contaminating substances.

1. Dry sample and matrix.
2. Deposit 1–2 μL of cold high-purity water containing 0.1% TFA on the sample spot.
3. Leave on for 5–10 s, then remove.
4. Repeat the wash step as necessary.

5. To reduce or remove detergent contamination, use a similar wash step(s) with 5% aqueous propan-2-ol.
6. To reduce contamination from crude cell extract, use 100% propan-2-ol.

4. Notes

1. An ultraviolet laser is directed at the sample (with a beam diameter of a few microns) for desorption. The laser radiation, a few ns in duration, is absorbed by the matrix molecules, causing rapid heating of the region around the area of laser impact, and electronic excitation of the matrix. The matrix serves as an absorbing medium for the ultraviolet light, converting the incident laser energy into molecular electronic energy, both for desorption and ionization, and as a source of H^+ ions to transfer to, and ionize, the analyte molecule.
2. TOF is coupled to MALDI as this technique has a virtually unlimited mass range. The ions enter the flight tube, where they are accelerated with the same potential at a fixed point and a fixed initial time. Since kinetic energy is purely dependent on mass and velocity, and the kinetic energy of all ions is nominally the same, the ions will separate according to their mass-to-charge ratios (m/z), and the lighter ions will travel faster than the heavier ions to the detector. This time of flight is converted to mass.

Post-source decay (PSD) is the process of fragmentation that may occur after an ion (the “precursor” ion) has been extracted from the source. Many biological molecules, particularly peptides, give rise to ions that dissociate over a time span of microseconds, and most precursor ions will have been extracted before this dissociation is complete. The fragment ions generated will have the same velocity as the precursor and cause peak broadening and loss of resolution in a linear TOF analyzer. The problem is overcome by the use of a reflector (or “reflectron”), a type of “ion mirror” employed to improve resolution. The reflectron is located at the end of the flight tube and compensates for the difference in flight times of the same m/z ions of slightly different kinetic energies. This increases the overall path length for an ion and corrects for minor variation in the energy spread of ions of the same mass. The reflectron has a gradient electric field, and the depth to which ions will penetrate this field, before reversal of direction of travel, depends upon their energy. Higher-kinetic-energy ions arrive at the reflector first but penetrate deeper into the field, thus traveling a longer flight path in the reflector than ions with less kinetic energy (which will travel a shorter distance). Therefore, all the ions arrive at the detector at the same time. This correction leads to an increased mass resolution for all stable ions in the spectrum. This can improve mass accuracy from 20 ppm to 5 ppm (4). Reflectron TOF is limited to the analysis of peptides and small proteins (below approx 10,000 Da) due to the fact that above this mass the peak broadening due to metastable decay is usually greater than the broadening due to kinetic energy distribution.

The reflectron also allows sequence information (in the case of peptides) to be obtained by PSD analysis, but this is less straightforward (and in a large percentage of experiments is unsuccessful) than tandem MS-MS on a quadrupole electrospray or ion trap instrument. At any given setting of the reflector/ion mirror, charged fragments of a particular range of m/z are focused in the reflector. Fragment ions of m/z above and below this narrow range are poorly focused. Therefore, since only fragment ions of a limited mass range are focused for a given “mirror ratio” in the reflector, a number of spectra are run at different settings and “stitched together” to generate a composite spectrum. Some new instruments have “seamless” reflectrons with curved or quadratic field equations to simplify the process (4).

3. Mass accuracy is the m/z of a peak in a spectrum divided by the width of that peak at half height, expressed in parts per million. The detection limit is currently between femtmoles (10^{-15} moles) and picomoles (10^{-12} moles), but attomole (10^{-18} moles) detection can be

- achieved. If 0.5 μ L sample is applied to the MALDI target, the concentration range is between micromolar and picomolar (attomole sensitivity is picomolar concentration).
4. Salts and buffers are in general detrimental, but before attempting to reduce salt and buffer contamination, it may be worthwhile to analyze a small amount of sample rather than risk loss of the sample in an inefficient purification procedure. Salts frequently form adducts that compete with the molecular ion, causing peak broadening, which reduces mass accuracy and signal intensity (especially for protein analysis). High-pH buffers will interfere with ionization. Some buffers also interfere with matrix crystal formation, resulting in signal suppression. It is recommended that salt concentration be less than 10 mM. Maximum concentrations tolerated include (approximately): urea (0.5 M); guanidine-HCl (0.5 M); dithiothreitol (DTT, 0.5 M); glycerol (1%); alkali metal salts (0.5 M); Tris buffer (0.05 M); NH_4HCO_3 (0.05 M); phosphate buffer (0.01 M); detergents other than sodium dodecyl sulfate (SDS) (0.1%); and SDS (0.01%). TFA, formic acid, β -mercaptoethanol, volatile organic solvents, HCl, NH_4OH , and acetic acid do not interfere. HEPES, MOPS, ammonium acetate, and octyl glucoside may be tolerated below 50 mM. Minimizing buffer concentrations will undoubtedly improve performance; therefore, use the minimum needed to control pH, and avoid, if possible, glycerol, sodium azide, DMSO, SDS, Triton X-100, phosphate, NaCl, 2 M urea, and 2 M guanidine. Sample clean-up can be done by simply diluting a sufficiently concentrated sample, by washing, and by ion exchange or reverse-phase pipet tip column chromatography using ZipTips (see Chapter 29). Although detergents and organic solvents such as hexafluoro-2-propanol are used to extract hydrophobic proteins (including membrane proteins), once solubilized the normal sample preparation techniques work well using CHCA, SA, and DHB matrices. If analyte is dissolved in chloroform, the matrix can be prepared in 2:1 v/v chloroform/methanol. Formic acid (70% aqueous) has also been used for preparation of matrix for hydrophobic samples, but formylation of polypeptides can occur; therefore, TFA is recommended. For example, a receptor protein was dissolved in acetone/trifluoroacetic acid (5%) saturated with sinapinic acid. Hexafluoroisopropanol was added to a final concentration of 12.5%, and after thorough mixing, 1 μ L of the sample was placed on the sample plates for MALDI-TOF MS (5).
5. Types of MALDI sample plates include 100-well stainless steel flat plates, which are good for multiple sample analysis with close external calibration employing compound(s) of known molecular weight on an adjacent spot. It is also easier to see crystallization of the matrix on this type of surface. 400-spot Teflon-coated plates have particular application for concentrating sample for increased sensitivity and better location of sample. The surface of the plate is exposed only in the center of each spot; therefore, the sample does not wet over the whole surface, but concentrates itself into the center of each spot as it dries. Other high-density formats include 384-well plates, which are good for automated sample application, although in this case a robot is required for accurate spotting. It is, however, unlikely that a sufficiently large number of protein samples need to be analyzed at one time. Gold-coated plates allow on-plate reactions within the well, with thiol-containing reagents, for example.
6. Maximum sensitivity is achieved if the samples are diluted to a particular concentration range. If the concentration is unknown, a series of dilutions may be needed to produce a satisfactory sample/matrix spot of suitable concentration on the MALDI plate. Some proteins, particularly glycoproteins, may yield better results at concentrations up to 10–50 pmol/ μ L. Oligonucleotides give better spectra at around 10 to 100 pmol/ μ L, while polymers require a concentration around 100 pmol/ μ L.
7. If the sample mass is completely unknown, start with SA, to which all peptides and proteins usually respond. The conditions can then be optimized depending on the spectrum.

Common MALDI matrices at 337 nm include 3,5-dimethoxy-4-hydroxycinnamic acid (sinapinic acid, SA) for peptides, larger proteins (>10,000 Da) and glycoproteins. 2,5-DHB is used for peptides, proteins, lipids and oligosaccharides. α -cyano-4-hydroxycinnamic acid (CHCA/HCCA) is used for analysis of polypeptides with a mass below 10,000 Da, proteins, lipids, and oligonucleotides. The cinnamic acid derivative matrices (e.g., SA) are useful if a wash cycle is necessary, because these matrices are only slightly water soluble, whereas DHB is readily soluble in water. Not all samples run well in every matrix, since each polypeptide/protein has a unique structure to be incorporated into a specific matrix crystal lattice. Each matrix also has unique physical properties and interacts with the analyte molecules in a different way.

Placing the matrix on the plate first will cause it to crystallize before the addition of the analyte, and will prevent homogeneous matrix/analyte crystallization. It is important that neither the matrix nor the analyte precipitate when the two solutions are mixed. Once the sample is applied to the sample support, the sample must be allowed to air evaporate. Do not heat the sample to increase the evaporation rate. Changing the temperature alters the crystal growth and protein incorporation, which may give a poor result. The dried samples are stable and can be stored at room temperature in the dark or in a vacuum for several days or weeks on the plate. It is important to use fresh matrix solutions (made up weekly) whenever possible. Nonvolatile solvents should be avoided, since they can interfere with the crystal growth.

8. In MALDI, the laser pulse generates a plume of material from the sample where energy transfer occurs between matrix and analyte. This occurs tens of micrometers from the probe surface. These ions are accelerated from this region into the TOF analyzer. This usually broadens the peak corresponding to any particular ion, which, combined with fragmentation occurring during this initial extraction period, leads to lower mass accuracy. However, if extraction is delayed until all the ions have formed, this spread is minimized. The procedure is known as delayed extraction (DE), whereby the ions are formed in either a weak field or no field during a predetermined time delay, and then extracted by the application of a high-voltage pulse. The degree of fragmentation of ions can also be controlled, to some extent, by the length of the time delay.

References

1. Smith, R. D. (2002) Trends in mass spectrometry instrumentation for proteomics. *Trends Biotechnol.* **20(12 Suppl)**, S3–7.
2. Sobott, F. and Robinson, C. V. (2002) Protein complexes gain momentum. *Curr. Opin. Struct. Biol.* **12**, 729–734.
3. Nelson, R. W., Nedelkov, D., and Tubbs, K. A. (2000) Biosensor chip mass spectrometry: a chip-based proteomics approach. *Electrophoresis* **21**, 1155–1163.
4. A mass spectrometry product overview of MALDI-TOF instruments, which gives a comprehensive list of suppliers, mass accuracy with and without relectron, target plate and laser type, and features such as DE and PSD, can be found at <http://www.hyperionms-demon.nl/index.html>. This site also lists the features, including mass/charge range, of electrospray, ion trap, and hybrid instruments.
5. Pawate, S., Schey, K. L., Meier, G. P., Ullian, M. E., Maisi, D. E., and Halushka, P.V. (1998) Expression, characterization, and purification of C-terminally hexahistidine-tagged thromboxane A2 receptors. *J. Biol. Chem.* **273**, 22,753–22,760.

Sequencing of Tryptic Peptides Using Chemically Assisted Fragmentation and MALDI-PSD

John Flensburg and Maria Liminga

1. Introduction

Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) has become the preferred method for high-throughput identification of proteins using peptide mass fingerprinting (PMF), due to its ease of automation, short analysis time, relatively high tolerance towards contaminants, high sensitivity, and mass accuracy (1). In this technique, proteins are typically separated by two-dimensional (2-D) gel electrophoresis, enzymatically in-gel digested with trypsin, extracted from the gel, and analyzed by MALDI-TOF MS. The resulting peptide mass fingerprints are compared to theoretical fingerprints from a protein- or DNA-sequence database for identification. In comparison to other ionization techniques such as electrospray ionization (ESI), the soft ionization induced by MALDI predominantly generates singly charged ions, which allows for a relatively easy interpretation of acquired spectra. Unfortunately, identification is not always unambiguous for a substantial fraction of the peptides analyzed, and it is not unusual that only a few peptides are recovered from an in-gel digest, especially when the protein is poorly expressed. To further improve the protein identification rate, amino acid sequence information from tryptic peptides is necessary. However, it is a well known fact that direct sequencing using MALDI post-source decay (PSD) often results in poor and unpredictable fragmentation patterns, which are mostly impossible to interpret (2). Singly charged tryptic peptides, formed during MALDI ionization, do not fragment readily because there is not enough internal energy available to move the ionizing proton from the basic C-terminal to the peptide backbone to induce fragmentation. In ESI, this problem is easily avoided by selecting doubly protonated peptides, which fragment readily. Keough et al. (3) introduced a derivatization chemistry based on sulfonation of the N-terminus of tryptic peptides, which greatly facilitated fragmentation during PSD. The original chemistry required a non-aqueous environment during derivatization. Recently, a novel reagent that is stable in aqueous buffers was introduced, and the reactions were performed on a solid support (4,5). The new reagent is now commercially available as Ettan™ CAF MALDI Sequencing Kit, where chemically assisted fragmentation (CAF) is used to eliminate the most common problems related to PSD analysis. Several investigators have used the kit for *de novo* and confirmative sequencing of peptides (6–8). CAF chemistry is applicable to peptides generated by trypsin

From: *The Proteomics Protocols Handbook*
Edited by: J. M. Walker © Humana Press Inc., Totowa, NJ

cleavage that contain a C-terminal lysine or arginine. In order to get fragmentation, lysine-terminated peptides have to be guanidinated prior to CAF derivatization, resulting in the conversion of lysines to homoarginines and increased peptide masses of 42 u (9) (Fig. 1). Direct sulfonation of lysine terminated peptides often result in disulfonate derivatives, with one sulfonate group attached to the N-terminus of the peptide and the other one attached to the lysine. These derivatives are undesirable because they reduce sensitivity and fragment less well during PSD analysis. The guanidination step is followed by derivatization using the CAF reagent, which adds a sulfonic acid group to the N-terminus of the peptides, thereby increasing their molecular mass by 136 u (Fig. 1). The sulfonated N-terminus will be negatively charged due to its low pKa value (-2), which is counterbalanced by the positive charge on the basic C-terminus (arginine or homoarginine), resulting in a net charge of zero. After MALDI ionization, one additional proton is added, which, according to the mobile proton model (10), can move randomly along the backbone of the CAF-derivatized peptide. The result of this will be a singly charged and doubly protonated peptide that requires less energy for breakage of peptide bonds, which enhances fragmentation towards b- and y-fragments during PSD (3). Due to the fixed negative charge at the N-terminus, the b-fragments will be neutral and not detected, resulting in a PSD spectrum consisting exclusively of y-ions, thus simplifying the interpretation of acquired spectra. The amino acid sequence of a CAF-derivatized tryptic peptide can thus be easily calculated manually or with software, from the mass differences between adjacent y-ions.

In this chapter, the Ettan CAF MALDI Sequencing Kit was evaluated in terms of sensitivity and functionality by PSD analyses of several peptides extracted from 2-D gels and from partial sequencing of a CAF-derivatized trypsin digest of haptoglobin.

2. Materials

1. Ettan CAF MALDI Sequencing Kit, Ettan MALDI-TOF Pro, and Ettan Spot Handling Workstation (AB, GE Healthcare, Uppsala, Sweden).
2. Human adrenocorticotrophic hormone fragment 18-39 (hACTH 18-39) and ile⁷-Angiotensin III (AngIII) (AB, GE Healthcare, Uppsala, Sweden).
3. Coomassie™ Brilliant Blue and Deep Purple (AB, GE Healthcare, Uppsala, Sweden).
4. Sequencing-grade trypsin (Promega Madison, WI).
5. Recrystallized α -cyano-4-hydroxy cinnamic acid (CHCA) (LaserBio Labs, Cedex, France).
6. Acetonitrile (HPLC grade), trifluoroacetic acid (TFA), and human haptoglobin (Sigma, St. Louis, MO).
7. ZipTip™μ-C₁₈ (Millipore, Bedford, MA).
8. The ultrapure water used was of Milli-Q quality (Millipore, Bedford, MA).

3. Methods

3.1. Procedures for Peptide Labeling

1. The derivatization of the tryptic peptides was performed on a solid support (ZipTipμ-C₁₈) as described in the detailed instructions included in Ettan CAF MALDI Sequencing Kit. The kit contains everything needed for derivatization with the exception of ZipTips, acetonitrile, TFA, ultrapure water, and tubes. In those cases where only Arg-terminated peptides were to be sequenced, the lysine protection step was omitted.

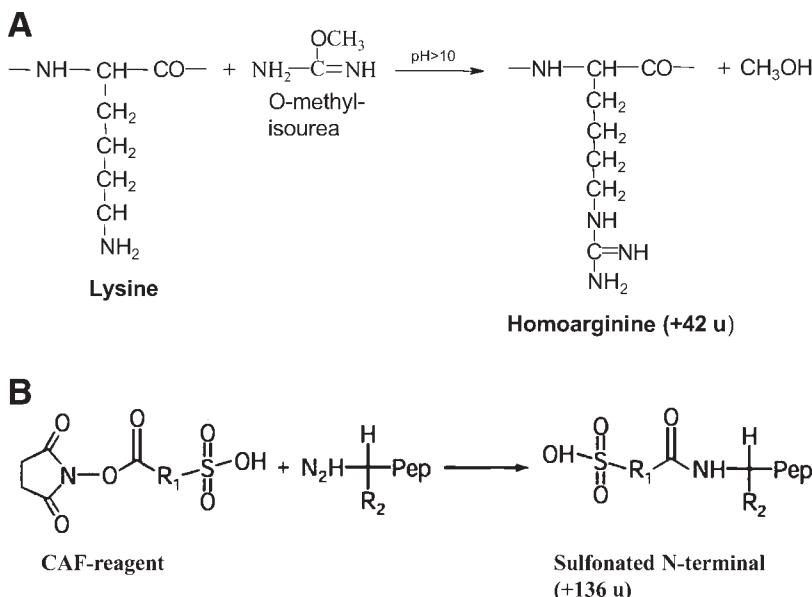


Fig. 1. Reaction scheme for the chemically assisted fragmentation derivatization step of Ettan™ CAF™ MALDI Sequencing Kit. (A) Protection of the ϵ -amino group of the lysine side chain. (B) Sulfonation of the α -amino N-terminus of the peptide.

2. The ZipTip μ -C₁₈ was attached to a 10- μ L pipet, and the RPC matrix was activated by drawing and dispensing 10 μ L of 50% acetonitrile and 0.1% TFA five times through the tip, which was then washed five times with 0.1% TFA, again by drawing and dispensing the liquid through the ZipTip. The peptides, in 5–10 μ L 0.1% TFA, were adsorbed to the ZipTip μ -C₁₈ by pipetting the sample up and down at least 10 times (see **Notes 1** and **2**).
3. The guanidination reaction was next performed by drawing and dispensing 10 μ L of the lysine modifier solution through the ZipTip. After the last drawing, with the solution above the C₁₈ RPC matrix, the tip was removed and put in a 1.5-mL Eppendorf tube, which was closed and incubated at 37°C for 2 h. After the reaction, the tip was washed five times with ultrapure water.
4. The sulfonation agent was prepared by adding 60 μ L CAF buffer to a vial of CAF reagent. Ten μ L of the CAF reagent solution was dispensed into 0.5-mL Eppendorf tubes, which were vortexed for a few seconds and briefly centrifuged to remove air bubbles. The dissolved CAF reagent is stable for only approx 20 min, which means that it is extremely important to continue directly with the sulfonation reactions (see **Note 4**). The CAF reagent solution was added to the ZipTip by slowly pipetting up and down approx 10 times. During this process it is of utmost importance to avoid air bubbles. If any bubbles are observed in the solution, they can easily be removed by lightly tapping the tube against the bench. The sulfonation reagent was left on the ZipTip, which was put in the Eppendorf tube and left to react for at least 3 min. Meanwhile, other samples can be processed for sulfonation.
5. After the sulfonation reaction, 1 μ L of stop solution (hydroxylamine) was added to the tubes, mixed with the sulfonation solution, and drawn up and down through the ZipTip about 10 times. This step will reverse unwanted sulfonation of side-chain hydroxyl groups (serine, threonine, and tyrosine). The CAF derivatization is completed by washing the sample on the tip five times with 0.1% TFA, and the sample is now ready for elution.

6. If more than 200 fmoles of material was loaded on the tip, the derivatized material was eluted in 5 μ L 80% acetonitrile, 0.1% TFA, which was dispensed into a suitable tube or microtiter plate. The liquid was drawn and dispensed about 10 times through the tip and finally collected and dried *in vacuo*. The sample was reconstituted in a suitable volume (1–2 μ L) of recrystallized CHCA (5 mg/mL in 50% acetonitrile, 0.1% TFA), and 0.3 μ L was dispensed onto the MALDI sample slide.
7. If the tip contained smaller amounts than 200 fmoles, the peptides were directly eluted from the ZipTip using 1 mg/mL of recrystallized CHCA in 50% acetonitrile, 0.1% TFA. During the last washing step with 0.1% TFA, the plunger of the pipet was pushed carefully all the way to the bottom, to get rid of all TFA, and the tip was removed and put on a pipet handling volumes from 0.2–3 μ L. 0.8 μ L of 1 mg/mL recrystallized CHCA was carefully drawn through the ZipTip, dispensed until a small liquid drop could be seen on the edge of the tip, and drawn again, which was repeated six or seven times. Finally, the contents of the tip was dispensed onto a MALDI sample slide and analyzed.

3.2. MALDI-TOF MS

1. All mass spectrometric analyses were performed using Ettan MALDI-ToF Pro (11) equipped with a quadratic field reflectron (12) and a timed ion gate. Protein identification (using PMF of trypsin-digested proteins) was conducted in reflectron mode with extraction of positive ions at 18 kV.
2. MALDI-PSD was performed by first acquiring a reflectron-mode spectrum with external calibration. Based on this analysis, peaks were chosen for timed ion gating, and the instrument was switched to PSD mode at an acceleration voltage of 18 kV. Due to the quadratic field reflectron of the instrument, PSD spectra were obtained over the entire m/z (mass-to-charge ratio) range without the need of data stitching.
3. The software of Ettan MALDI-ToF Pro contains functions for fully automated PMF analyses, including data acquisition, spectrum processing, and database searches using the ProFoundTM search engine (13). PSD data can be used for automatic protein identification and/or automatic sequencing by SonarTM (14).
4. Proteins from mouse liver and *Arabidopsis thaliana*, respectively, were separated by 2-D gel electrophoresis and visualized by staining with Coomassie Brilliant Blue or Deep Purple, a novel fluorescent stain (15,16). Selected protein spots were automatically picked and processed (in-gel trypsin digested, extracted from the gel, mixed with matrix, and loaded onto MALDI sample slides) using Ettan Spot Handling Workstation.

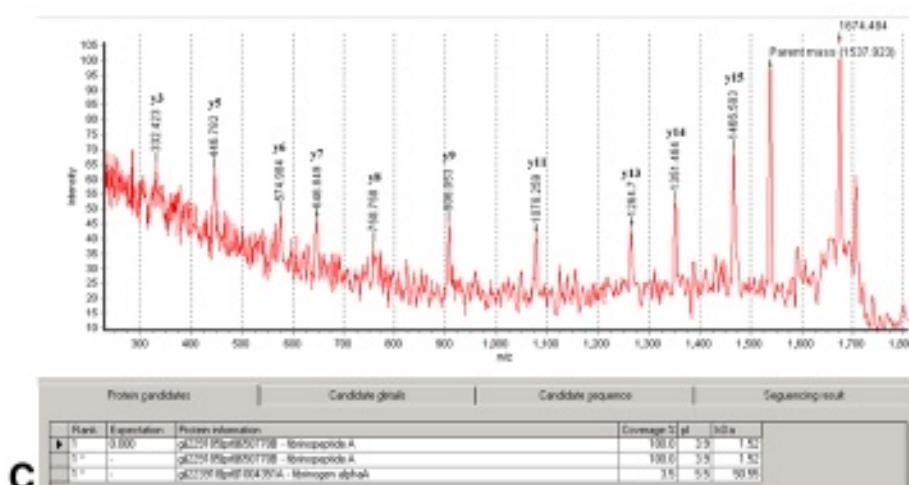
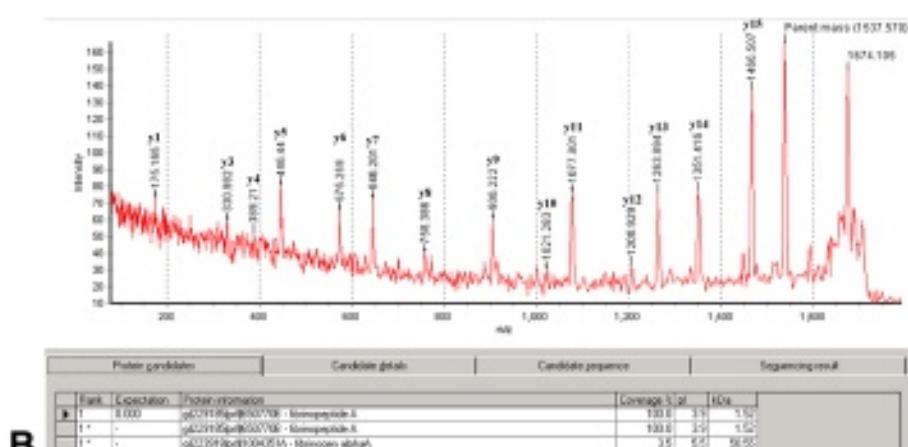
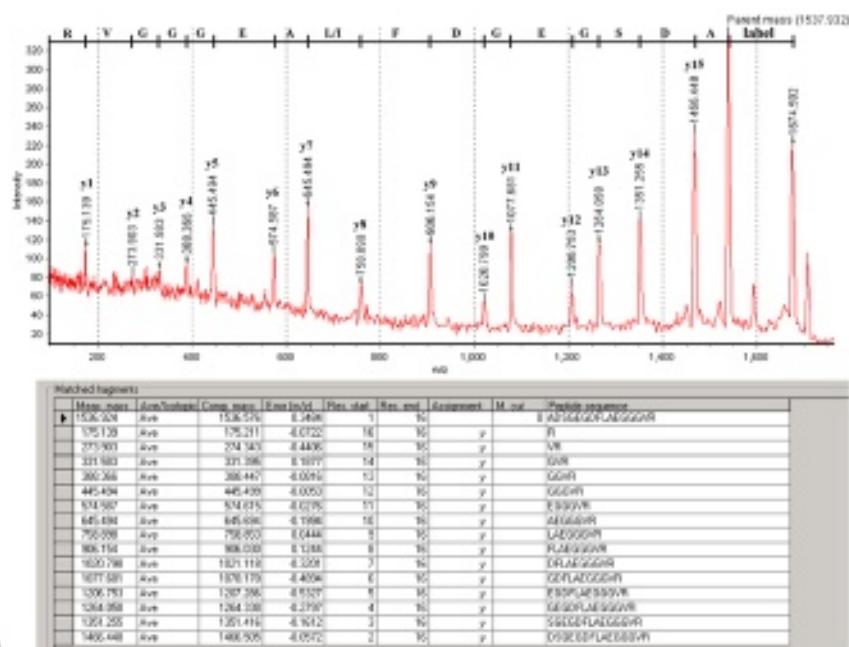
3.3. Minimal Detection Levels

1. To test the sensitivity of Ettan CAF MALDI Sequencing Kit in conjunction with Ettan MALDI-ToF Pro, increasingly smaller amounts of fibrinopeptide A (Fib A), which is included in the kit as a test substance, were sulfonated and analyzed by MALDI-PSD. The amounts of Fib A loaded to the ZipTip μ -C₁₈ were 150, 100, and 50 fmoles, respectively, and the results are shown in **Fig. 2**. In this case, only the sulfonation reaction was performed, since the sequence of the peptide (ADSGEGDFLAEGGGR) is terminated with an arginine. From the spectra, the derivatized Fib A can be seen at an m/z around 1674. The peak at an m/z around 1538 corresponds to the underderivatized parent ion resulting from the loss of the CAF label (136 u), which is followed by the series of y-ions. The parent ion and the derivatized parent ion are usually the most signal-intensive peaks in a PSD spectrum acquired by a MALDI-TOF MS using quadratic field reflectron technology.
2. The PSD analysis of 150 fmoles of CAF-derivatized Fib A (**Fig. 2A**) resulted in the detection of a complete series of y-ions and an unambiguous identification (expectation value 0.000) by the software. The software-generated table under the spectrum shows that the largest mass error for the y-fragments was 0.53 u, with an average error of 0.20 u.

3. The PSD spectrum originating from 100 fmoles of CAF-derivatized Fib A (**Fig. 2B**) also showed an unambiguous identification with 14 matched y-ions. Only one fragment was missing at an m/z value of 274.3 (y2). It is not surprising that this ion was missing, since the fragmentation occurs at a Gly residue, which usually results in lower signal intensity. This is also obvious for the other Gly residues of Fib A at m/z values of 1021.1 (y10) and 1207.3 (y12). In these cases, the signal intensities recovered at the adjacent fragments. The missing y2-ion does not cause any practical problem, since it can be calculated by a gap of 57 u to the next fragment, resulting in a difference of 155.8 u between the y1 and y3 ions ($330.992 - 175.165 = 155.8$ u). This difference fits only the combination Gly-Val.
4. The PSD analysis of 50 fmoles of CAF-derivatized FibA (**Fig. 2C**) revealed that 10 y-ions were matching with most of the missing fragments at Gly residues. Anyway, the number of detected y-ions was still high enough to get an unambiguous identification from the software. In many cases, a complete series of y-ions is not necessary for identification; often, a consecutive series of four of five fragments is sufficient (*see Note 7*). When CAF derivatization was performed on 25 fmoles of fibrinopeptide A, only three or four fragments could be detected upon PSD analysis, resulting in a less confident identification result (data not shown).
5. These results indicate that the practical lowest detection limit resulting in unambiguous identification for the MALDI-TOF MS used in this study for a CAF-derivatized peptide eluted directly from a ZipTip μ -C₁₈ is approx 50 fmoles.

3.4. Confirmative Sequencing of Haptoglobin

1. An important application for Ettan CAF MALDI Sequencing Kit is confirmative sequencing of purified proteins. This have recently been reported for the recombinant forms of human serum albumin (7) and transposase (8), where tryptic peptides were sequenced by MALDI-PSD after chemical derivatization using the kit. These studies indicated that the recombinant serum albumin and the recombinant transposase were identical to their natural counterparts.
2. To illustrate this type of application, haptoglobin of human origin was reduced, alkylated, and digested with trypsin. Five hundred fmoles of the digest was mixed with an equal volume of 50% acetonitrile, 0.1% TFA containing 5 mg/mL CHCA, and 0.3 μ L was dispensed on the MALDI slide for reflectron analysis. The resulting PMF spectrum and identification result with a coverage exceeding 45% is shown in **Fig. 3A**.
3. One picomole of the digest was guanidinated and sulfonated as described in **Subheading 3.1**. The derivatized material was eluted from the ZipTip with 5 μ L 80% acetonitrile and 0.1% TFA, dried, and dissolved in 1 μ L 5mg/mL CHCA, of which one-third was loaded onto the MALDI slide and analyzed in reflectron mode (**Fig. 3B**). Some of the peptides differing in molecular mass, with 42 + 136 u (K-terminated peptides) or 136 u (R-terminated peptides) compared to the underivatized peptide masses (**Fig. 3A**), were chosen for PSD analysis.
4. In **Fig. 4A-F**, the resulting PSD spectra of the CAF-derivatized peptides are shown, together with identification and sequencing results calculated by the software. By looking at the m/z value for the y-1 ion it is easy to discriminate between arginine- and lysine-terminated peptides. For CAF-only derivatized peptides, this value should be 175.3 (Arg), while the corresponding value for a lysine-terminated peptide (guanidinated and sulfonated) should be 189.11 (guanidinated Lys). As can be seen from the figure, complete series of y-ions (complete sequences) were obtained for all analyzed peptides with the exception of the one in **Fig. 4F**.
5. It can be noted (from **Fig. 4D,E**) that in vitro-induced modifications were easily detected, as seen from those peptides containing half-cystine residues. The mass of carboxy-methylated cys (CM-C), which was produced as a result of treatment with iodoacetamide



- prior to digestion with trypsin, was 161 u (an increase in mass of 58 relative to half-cystine).
6. The spectrum in **Fig. 4F** constitutes a schoolbook example of a typical incomplete fragmentation pattern. The sequence of the fragmented peptide is SPVGVQPILNETH-FCAGMSK, and the first 15 amino acid residues could be deduced from the fragment masses by the software, resulting in an unambiguous identification. As mentioned earlier, fragmentation at the C-terminal side of Gly residues can result in impaired fragmentation, and thus reduced signal intensity. This is even more pronounced for fragmentation at the C-terminal side of Pro residues, which often result in non-detectable signal intensities. In **Fig. 4F**, this can be seen between the fragments with m/z 2128.80 and 1932.34 and the fragments with m/z 1647.98 and 1437.48. In both cases, no fragment corresponding to a Pro residue could be detected. However, the presence of a Pro residue is evident, since the difference in m/z 196.46 ($2128.80 - 1932.34$) matches only the combination Pro-Val, while the difference 210.50 ($1647.98 - 1437.48$) matches only Pro-Leu/Ile. 7. The average mass error for detected amino acid residues (spectra A-C) was ± 0.1 u, with a largest deviation of 0.25 u, clearly demonstrating that the mass accuracy in PSD mode of the MALDI-TOF MS used was sufficient for correctly assigning all amino acid residues with the exception of the isobaric Leu/Ile.
 8. These PSD spectra, including identification and sequencing results, were obtained in less than 10 min and from the same sample spot, which contained less than 300 fmoles of the derivatized digest. Due to the quadratic field reflectron of Ettan MALDI-ToF Pro, all fragments, independent of size, could be focused in a single run. In contrast to a MALDI-TOF MS instrument utilizing a linear reflectron, no time-consuming stitching procedure of several spectra covering different m/z ranges was necessary in order to acquire a complete PSD spectrum.
 9. The sequencing information obtained from these six derivatized tryptic peptides covered almost 20% of the haptoglobin sequence. The total time for achieving these results was less than 2 d, which included reduction, alkylation, desalting, trypsin digestion, PMF analysis, derivatization, and MALDI-PSD. It would have taken considerably more time to obtain the same information using traditional non-MS methods.

3.5. Proteomics Applications

3.5.1. Identification of In-Gel Digested Mouse-Liver Proteins

1. In order to test the derivatization kit for proteome investigations, 150 μ g of an extract prepared from mouse-liver cells was separated by 2-D gel electrophoresis. The separated proteins were visualized by staining with Deep Purple, a novel fluorescent stain (**15,16**). Low- to medium-intensity protein spots were selected and automatically picked from the gel and prepared for MS analysis using Ettan Spot Handling Workstation, a robot for spot picking, controlled in-gel trypsin digestion, extraction, and preparation for MALDI-TOF MS analysis.

Fig. 2. (opposite page) Matrix-assisted laser desorption/ionization (MALDI)-post-source decay (PSD) analyses of varying amounts of sulfonated FibA, together with identification results.

(A) Analysis of 150 fmoles of derivatized FibA. The amino acid sequence deduced by the software is inserted on top of the spectrum, and the mass errors of acquired fragments are tabulated at the bottom of the figure.

(B) PSD spectrum resulting from the analysis of 100 fmoles of derivatized FibA.

(C) PSD spectrum resulting from the analysis of 50 fmoles of derivatized FibA.

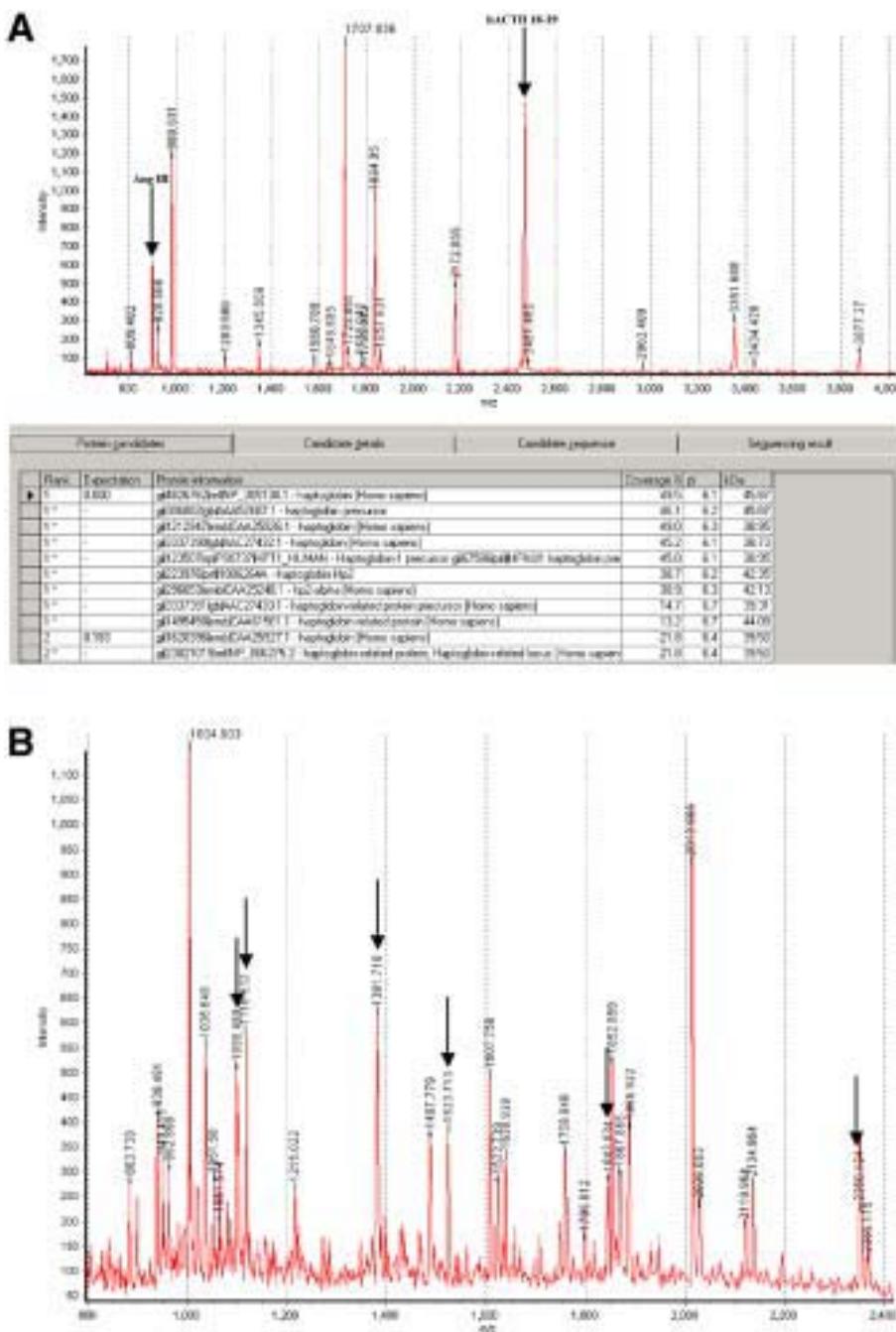


Fig. 3. (A) Matrix-assisted laser desorption/ionization (MALDI)-time-of-flight (TOF) mass spectrometry analysis in positive reflectron mode of a trypsin digest of haptoglobin. The tabulated results at the bottom of the figure show the protein identification results using Ettan MALDI-TOF Pro software. The mono-isotopic m/z values are shown at the top of each peak. Ang III (m/z 897.531) and hACTH 18-39 (m/z 2465.199) were used for internal calibration.

(B) Reflectron spectrum of trypsin-digested, guanidinated, and CAF-derivatized haptoglobin. Peaks denoted with arrows were chosen for sequencing by MALDI post-source decay.

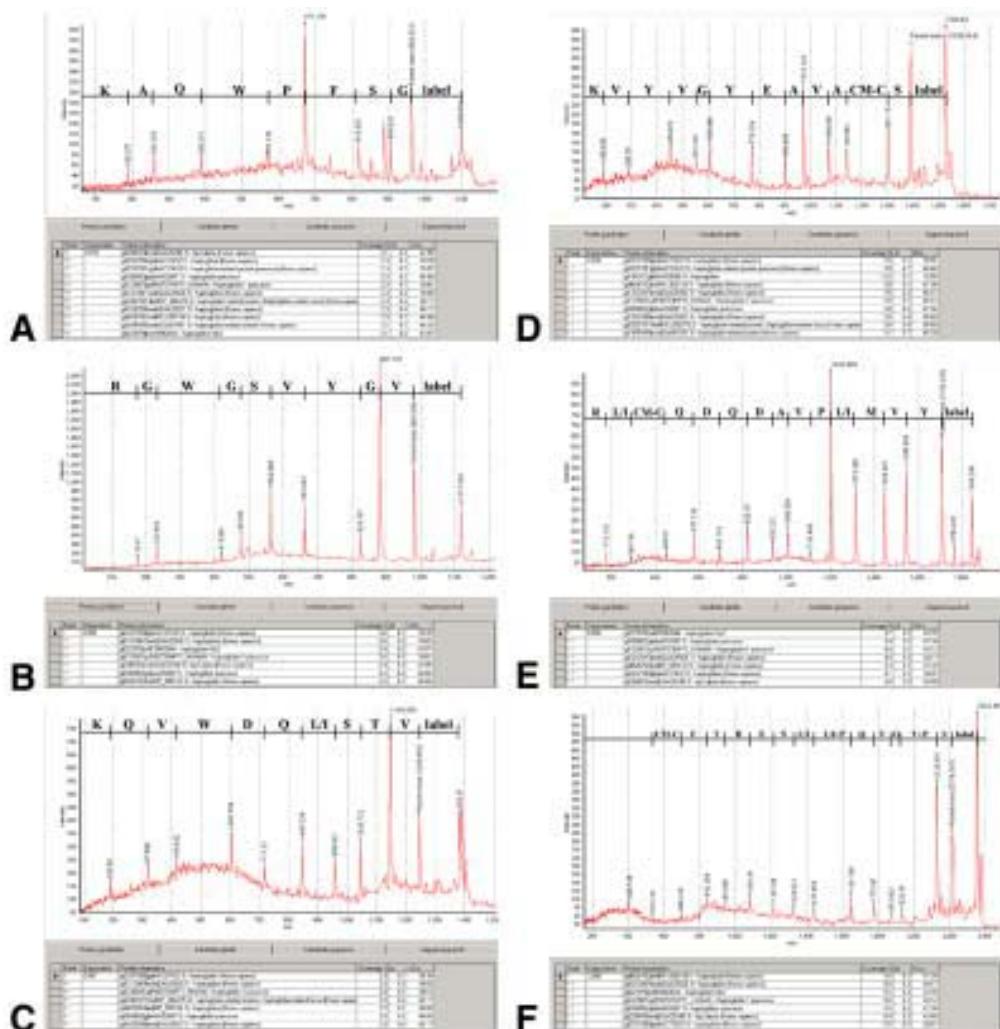


Fig. 4. Chemically assisted fragmentation (CAF)-enhanced matrix-assisted laser desorption/ionization (MALDI)-post-source decay (PSD) analyses of six tryptic peptides from haptoglobin, and identification results from the software. The average m/z values are shown at the top of each peak. The resulting amino acid sequence is inserted on top of each spectrum.

- (A) PSD spectrum of the guanidinated and CAF-labeled 1099 peptide.
 - (B) PSD spectrum of the CAF-labeled 1117 peptide.
 - (C) PSD spectrum of the guanidinated and CAF-labeled 1382 peptide.
 - (D) PSD spectrum of the guanidinated and CAF-labeled 1525 peptide.
 - (E) PSD spectrum of the CAF-labeled 1846 peptide.
 - (F) PSD spectrum of the guanidinated and CAF-labeled 2353 peptide.

2. **Figure 5A** shows a PMF spectrum in reflectron mode after analysis of approx 1/10 of the material from one of the 2-D gel spots (see Note 8). The acquired m/z values were processed by the software, which resulted in a coverage of about 12% for the highest-ranked candidate, methionine adenosyltransferase. Due to the low coverage, a less confident identification result was obtained, which is reflected in an expectation value of 0.632 for the

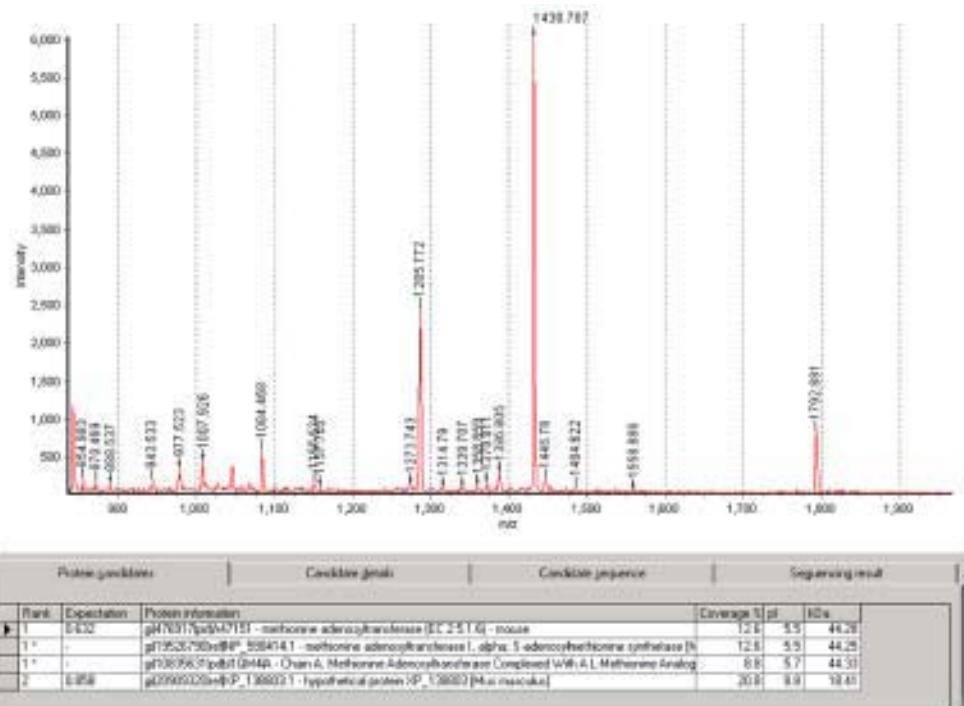
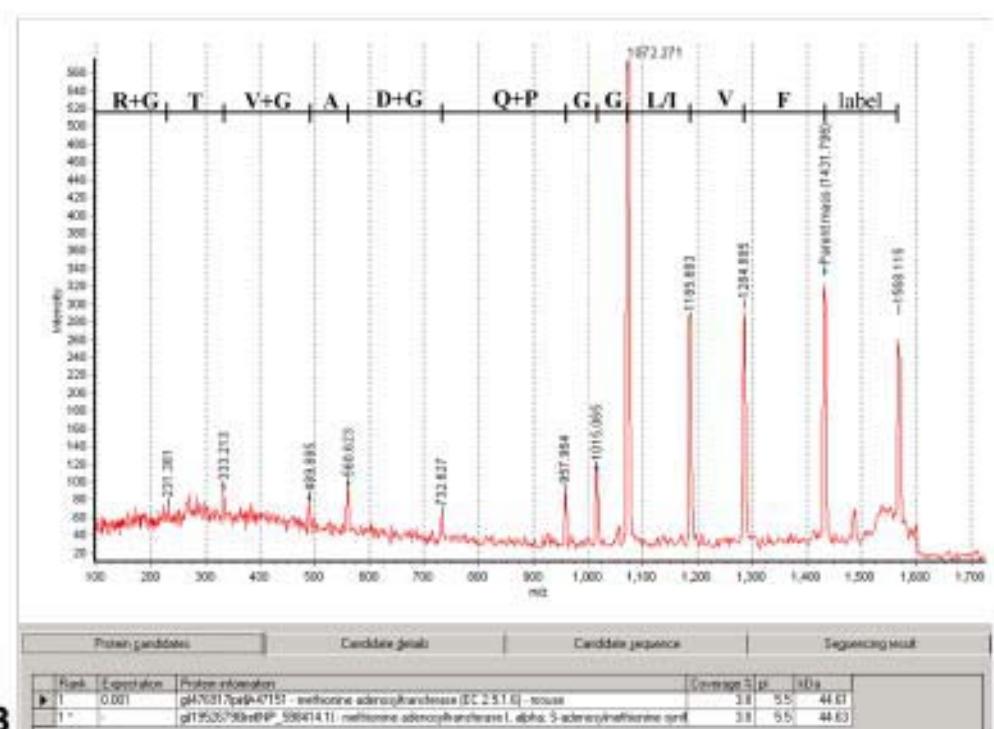
- highest-ranked protein candidates. For an unambiguous identification, the expectation value must be 0.05 or less.
- The remaining parts of the material from the gel plug were subjected to sulfonation, and after reflectron analysis of the derivatized digest, a peptide with a monoisotopic m/z value of 1566.8 was selected by the timed ion gate and fragmented using MALDI-PSD (Fig. 5B). As can be seen from the spectrum, an almost complete series of y-ions was obtained, resulting in a highly confident sequencing and identification result. Missing fragments could again be detected at Pro and Gly residues, but did not negatively affect sequencing and identification results.
 - Another example from a gel plug where the protein could not be identified by PMF but was readily identified through sulfonation and MALDI-PSD, is shown in Fig. 6. In this case, no useful information was obtained after the initial PMF analysis (data not shown). After sulfonation of the remaining material from the gel plug, a derivatized peptide with an average m/z value of 1352.7 was subjected to MALDI-PSD. From the complete y-ion series, the protein was identified as sterol carrier protein with an expectation value of 0.000 (Fig. 6). It can also be noted that the N-terminal amino acid consisted of sulfoxidized methionine, which could be the result of a posttranslational modification.
 - In this survey, proteins from 38 different gel plugs were analyzed, of which 23 proteins were identified directly using automated PMF, a success rate of 60.5%. The unidentified samples were analyzed by MALDI-PSD after CAF derivatization, which resulted in identification of 5 additional proteins. In this case, the success rate for CAF-MALDI PSD was 33.3% (5/15), thereby increasing the total protein identification rate to 73.7%.

3.5.2. Identification of In-Gel Digested Proteins from *Arabidopsis thaliana*

- Another analysis was performed on *Arabidopsis thaliana*, a small flowering weed often used as a model system for plant development, physiology, and genetics. The sequencing of the whole genome of *A. thaliana* was completed during 2000 by the international Arabidopsis Genome Initiative (17).
- A. thaliana* mitochondria were isolated and the cell extracts separated by 2-D gel electrophoresis. The gel was stained with colloidal Coomassie Brilliant Blue, and fifty spots of low staining intensities were selected for automated spot picking, in-gel trypsin digestion, extraction, and preparation for automated MALDI-TOF reflectron analysis. Also, in this case 1/10 of the material from the gel plugs was used for PMF analysis, while the remaining parts were saved for CAF derivatization, in case no unambiguous identification should be achieved through PMF analysis (see Note 8). Using this approach, 37 of the 50 spots were directly identified (74%) by automatic PMF.
- The remaining material from the 13 unidentified samples was sulfonated and analyzed by MALDI-PSD, resulting in unambiguous identification of 6 of the 13 samples (success rate of 46%). Two of the resulting PSD spectra are shown in Fig. 7. In this study, CAF derivatization further improved protein identification rates from 74% to 86%.

Fig. 5. (A) (opposite page) Matrix-assisted laser desorption/ionization (MALDI)-time-of-flight (TOF) mass spectrometry spectrum in reflectron mode of a trypsin-digested mouse-liver protein eluted from a two-dimensional gel, and the ambiguous software identification results. The monoisotopic m/z values are annotated at the top of each peak. Trypsin autolysis peaks at m/z 842.51 and 2211.10 were used for internal calibration.

(B) Chemically assisted fragmentation-enhanced MALDI-post-source decay spectrum of the 1430.8 peak, and the resulting nonambiguous software identification result. The average m/z values are shown at the top of each peak, and resulting amino acid sequence is inserted on top of the spectrum.

**A****B**

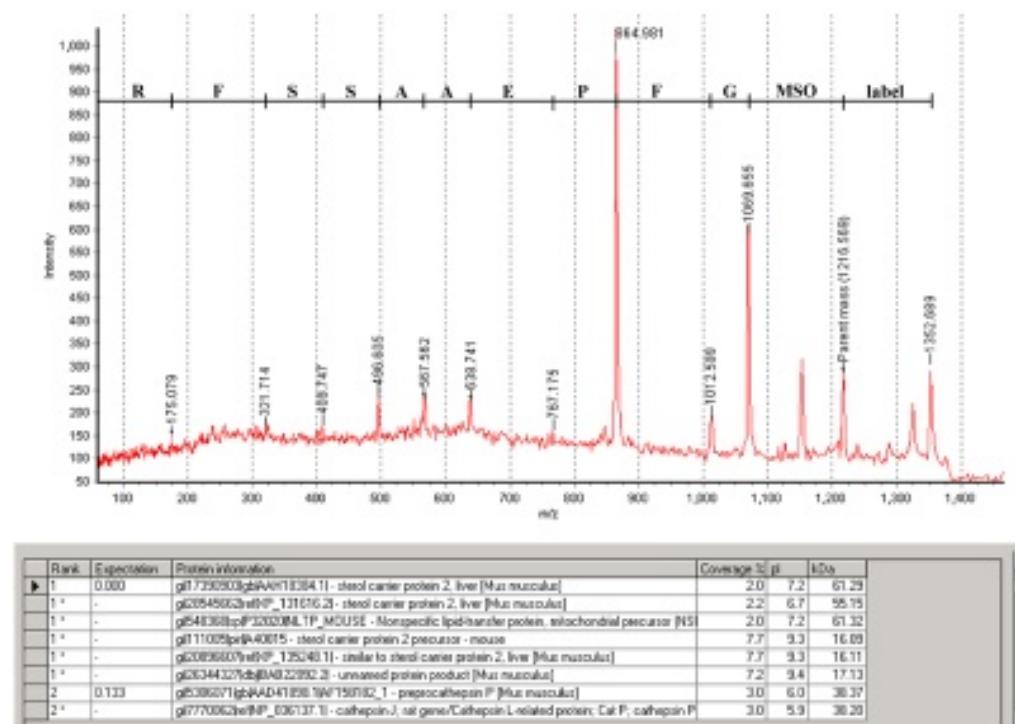


Fig. 6. Matrix-assisted laser desorption/ionization-post-source decay analysis of a sulfonated peptide originating from a trypsin-digested mouse-liver protein eluted from a two-dimensional gel, together with the software identification result (nonambiguous). No identification result was obtained using automated peptide mass fingerprint. The average m/z values are shown at the top of each peak, and the sequence of amino acids deduced by the software is inserted on top of the figure.

3.6. Sequencing of Phosphopeptides

1. Phosphorylation is one of the important posttranslational modification processes in cell regulation and metabolism. A phosphorylated protein can be identified by reflectron analysis of the trypsin-digested protein. An increased peptide mass of 80 u compared to the theoretical one, could indicate the presence of a phosphorylated amino acid residue (18). One way to ascertain that the observed mass difference is the result of a phosphorylation is through MS/MS or PSD analysis of the underderivatized peptide, which often generate spectra with characteristic patterns (19–21). This pattern consists of the formation of two fragment ions from the losses of HPO_3 ($[MH-HPO_3]^+$) and H_3PO_4 ($[MH-H_3PO_4]^+$), corresponding to mass losses of 80 and 98 u, respectively.
2. Recently, CAF derivatization has been used for the analysis of peptides containing phosphorylated Tyr, Ser, or Thr (6,22). Those results showed that for peptides containing phosphoserine or phosphothreonine residues, beta-elimination occurred and a fragment that was 98 u (H_3PO_4) less than the intact CAF-derivatized phosphopeptides could be detected, indicating removal of phosphoric acid from the peptide. The fragmentation data also revealed a mass difference of 69 u for dehydroalanine and 83 u for dehydroamino-2-butyric acid, formed by beta-elimination of phosphoserine and phosphothreonine, respectively. For the peptide containing phosphotyrosine, a mass difference of 243 u between

two adjacent fragments indicated a phosphotyrosine residue (Tyr 163 u, HPO₃ 80 u).

3.7. CAF Sequencing Using ESI and other MS Platforms

1. Sequencing of sulfonated peptides has been performed using MALDI-TOF MS with quadratic, curved, and linear reflectrons (6–8,23). Quadratic and curved reflectrons enable the acquisition of complete PSD mass spectra without the need of stitching, which saves time and sample.
2. CAF-enhanced sequencing has also been carried out on MALDI TOF-TOF instruments, on an atmospheric pressure MALDI ion trap mass spectrometer, and on an ion trap-TOF hybrid instrument (23). CAF chemistry has also been applied to MS/MS studies using ESI (triple quadrupole, ion traps, and Q-TOF instruments). In this study by Keough et al. (23), very similar fragmentation patterns were observed with both MALDI and ESI techniques, showing that CAF derivatization is applicable to a wide range of commercially available MS instruments.

4. Notes

“Ettan, CAF and Deep Purple are trademarks of Amersham Biosciences Limited. Amersham and Amersham Biosciences are trademarks of Amersham plc.

Ettan™ CAF™ MALDI Sequencing Kits are protected by patents owned by Procter & Gamble company and exclusively licensed to AB, GE Healthcare, and by joint patents issued to both companies.

The purchase of Ettan CAF MALDI Sequencing Kits includes a limited license to use the technology for internal research and development, but not for any commercial purposes”.

Copyright Amersham Biosciences AB, GE Healthcare, 2003—All rights reserved.

1. When derivatization is performed with low sample amounts, it is especially important to perform all steps on the ZipTip carefully, to push and release the plunger of the pipet slowly. Also, make sure that all solutions, especially the acetonitrile-containing ones, are freshly prepared.
2. Incomplete binding of the peptides to the ZipTip can occur if the sample is not sufficiently acidic. The pH should be below 4. Sometimes, binding can be improved by spiking the sample with a few microliters of 1% TFA.
3. If the peptides are not completely guanidinated or sulfonated, nonreacted peptides will also be observed upon MS analysis, and the signal for derivatized peptides will be reduced. If this happens constantly, check the pH of the reaction buffers. The pH of the lysine modifier should be 10.0 ± 0.2 for optimal lysine modification. The pH of the sulfonation buffer should be 9.4 ± 0.2 for optimal labeling.
4. The CAF reagent is unstable in aqueous environments, and should be used within 20 min from the time it was dissolved.
5. Sometimes, fragment masses from several peptides are present in the PSD spectrum. If multiple parent ions are seen, it is advisable to narrow the gate in the PSD acquisition method.
6. If no protein identification is obtained from good-quality PSD spectra, the system calibration in PSD mode should be checked. The width of the error window for the identification of fragment masses could be increased to ± 2 u.
7. In most cases, four or five fragment peaks are needed for protein identification using CAF derivatization and MALDI-PSD. However, in many cases, two or three fragment peaks are enough to improve a bad expectation value achieved by PMF. Tools and search engines

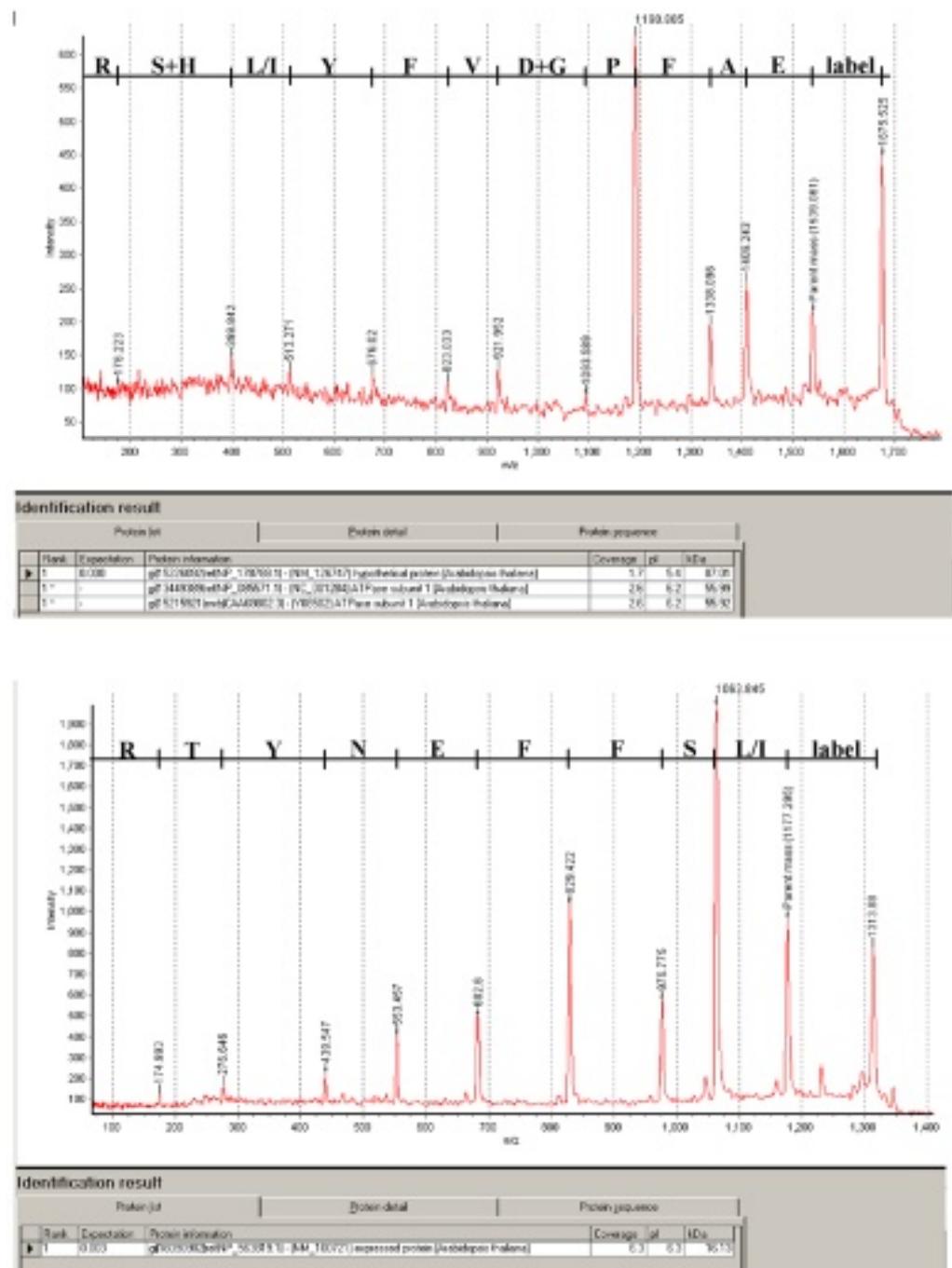


Fig. 7. Matrix-assisted laser desorption/ionization-post-source decay analyses of two sulfonated peptides from *A. thaliana* mitochondrial proteins eluted from different two-dimensional gel spots, together with identification and sequencing results. No useful identification was obtained using automated reflectron analysis.

for combined fragment and peptide mass fingerprinting can be found at www.matrix-science.com.

- When tryptic peptides are eluted from 2-D gel plugs, it is advisable to use 10% of the sample for PMF studies. In those cases where ambiguous or no identification is achieved, the remaining 90% can be CAF derivatized and sequenced by MALDI-PSD. In these cases, many users of the kit might find it convenient to just analyze the Arg-terminated peptides, since it will save a lot of time.

References

- Shevchenko, A., Jensen, O. N., Podtelejnikov, A., et al. (1996) Linking genome and proteome by mass spectrometry: large-scale identification of yeast proteins from two dimensional gels. *Proc. Natl. Acad. Sci. USA* **93**, 1440–1445.
- Spengler, B. (1997) Post-source decay analysis in matrix-assisted laser desorption/ionization mass spectrometry of biomolecules. *J. Mass Spectrom.* **32**, 1019–1036.
- Keough, T., Lacey, M. P., and Youngquist, R. S. (1999) A method for high-sensitivity peptide sequencing using postsource decay matrix-assisted laser desorption ionization mass spectrometry. *Proc. Natl. Acad. Sci. USA* **96**, 7131–7136.
- Liminga, M., Carlsson, U., Larsson, C., et al. (2001) New water stable chemistry for improved amino acid sequencing by derivatization postsource-decay (dPSD) using Ettan MALDI-TOF with a quadratic field reflection. Proc. 49th ASMS Conf. Mass Spectrometry and Allied Topics, Chicago, IL.
- Keough, T., Lacey, M. P., and Youngquist, R. S. (2002) Solid-phase derivatization of tryptic peptides for rapid protein identification by matrix-assisted laser desorption/ionization mass spectrometry. *Rapid Commun. Mass Spectrom.* **16**, 1003–1015.
- Hellman, U. and Bhikhambhai, R. (2002) Easy amino acid sequencing of sulfonated peptides using post-source decay on a matrix-assisted laser desorption/ionization time-of-flight spectrometer equipped with a variable voltage reflector. *Rapid Commun. Mass Spectrom.* **16**, 1851–1859.
- Flensburg, J. and Belew, M. (2003) Characterization of recombinant human serum albumin using matrix-assisted laser desorption ionization time-of-flight mass spectrometry. *J. Chromatogr. A* **1009**, 111–117.
- Eklund, P., Andersson, H. O., Kamali-Moghaddam, M., Sundström, L., and Flensburg, J. (2003) Purification and partial characterization by matrix-assisted laser desorption ionization time-of-flight mass spectrometry of the recombinant transposase, Tn1A. *J. Chromatogr. A* **1009**, 179–188.
- Keough, T., Lacey, M. P., and Youngquist, R. S. (2000) Derivatization procedures to facilitate *de novo* sequencing of lysine-terminated tryptic peptides using postsource decay matrix-assisted laser desorption/ionization mass spectrometry. *Rapid Commun. Mass Spectrom.* **14**, 2348–2356.
- Wysocki, V. H., Tsaprailis, G., Smith, L. L., and Breci, L. A. (2000) Mobile and localized protons: a framework for understanding peptide dissociation. *J. Mass Spectrom.* **35**, 1399–1406.
- Flensburg, J., Haid, D., Blomberg, J., Bielawski, J., and Ivansson, D. (2004) Applications and performance of a MALDI-TOF mass spectrometer with quadratic field reflectron technology. *J. Biochemical and Biophysical Methods* **60**, 319–334.
- Anderson, U. N., Colburn, A. W., Makarov, A. A., et al. (1998) In-series combination of a magnetic-sector mass spectrometer with a time-of-flight quadratic-field ion mirror. *Rev. Sci. Instrum.* **69**, 1650–1660.
- Zhang, W. and Chait, B. T. (2000) ProFound: An expert system for protein identification using mass spectrometric peptide mapping information. *Anal. Chem.* **72**, 2482–2489.

14. Field, H. I., Fenyö, D., and Beavis, R. C. (2002) RADARS, a bioinformatics solution that automates proteome mass spectral analysis, optimises protein identification, and archives data in a relational database. *Proteomics* **2**, 36–47.
15. Bell, P. J. L. and Karuso, P. (2003) Epicocconone, a novel fluorescent compound from the fungus *Epicoccum nigrum*. *J. Am. Chem. Soc.* **125**, 9304–9305.
16. Mackintosh, J. A., Choi, H., Bae, S., et al. (2003) A fluorescent natural product for ultra sensitive detection of proteins in 1-D and 2-D gel electrophoresis. *Proteomics* **3**, 2273–2288.
17. The Arabidopsis Genome Initiative. (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815.
18. Aebersold, R. and Goodlett, D. R. (2001) Mass spectrometry in proteomics. *Chem. Rev.* **101**, 269–295.
19. Hoffmann, R., Metzger, S., Spengler, B., and Otvos, L. (1999) Sequencing of peptides phosphorylated on serines and threonines by post-source decay in matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *J. Mass Spectrom.* **34**, 1195–1204.
20. Metzger, S. and Hoffmann, R. (2000) Studies on the dephosphorylation of phosphotyrosine-containing peptides during post-source decay in matrix-assisted laser desorption/ionization. *J. Mass Spectrom.* **35**, 1165–1177.
21. Qin, J. and Chait, B. T. (1997) Identification and characterization of post translational modification of proteins by MALDI ion trap mass spectroscopy. *Anal. Chem.* **69**, 4002–4009.
22. Bhikhabhai, R., Algotsson, M., Carlsson, U., et al. (2004) Amino acid sequencing of sulfonic acid-labeled tryptic peptides using post-source decay and quadratic field MALDI-ToF mass spectrometry. In Kamp, R. M., Calvete, J. J., and Choli-Papadopoulou, T. (eds.), *Principles and Practice, Methods in Proteome and Protein Analysis* (Springer-Verlag, Berlin, Heidelberg, pp. 279–296.
23. Keough, T., Youngquist, R. S., and Lacey, M. P. (2003) Sulfonic acid derivatives for peptide sequencing by MALDI MS. *Anal. Chem.* **75**(7), 156A–165A.

The *In Situ* Characterization of Membrane-Immobilized 2-D PAGE-Separated Proteins Using Ink-Jet Technology

Patrick W. Cooley, Janice L. Joss, Femia G. Hopwood, Nichole L. Wilson, and Andrew A. Gooley

1. Introduction

A progression in technology development for proteomics research is occurring at an ever-increasing rate (1). The monitoring of the physiological changes of healthy and diseased tissues with linkage to the expression of the proteome is fast becoming a method to identify molecular disease targets for creating novel drugs, as well as providing data for basic research. Improvements in the preparation of protein samples and mass spectrometry (MS) equipment are leading to better identification and characterization of proteins (2,3). Advances in protocols for protein sample prefractionation, solubilization strategies, and two-dimensional polyacrylamide gel electrophoresis (2-D PAGE) are oriented with the development of automated high-throughput proteomic analysis platforms (4–6).

Within this process, protein arrays or chips are being developed for high-throughput proteomic applications (7,8). These protein chips can be comprised of polystyrene film, glass slides, micro-wells, or membranes printed with protein arrays. Applications for protein chips include diagnostics, disease monitoring, screening proteins in affinity binding and protein–protein interaction studies, expression profiling, and disease marker development and drug discovery (9–11). The protein chips have the advantage of high-throughput multi-screening capability and are created from arrays of known proteins, such as antibodies, or proteins derived by recombinant expression methods. However, these protein chips do not address protein isoforms, such as co- and posttranslationally modified forms of the identical translated gene product, members of gene families, and variable spliced variants of mRNA and proteins.

Improvements in the methods of protein sample preparation and 2-D PAGE analysis have enhanced the number of protein isoforms that can be revealed by 2-D PAGE arrays (12–15). These 2-D PAGE arrays can be considered protein chip macroarrays when they are blotted onto a polyvinylidene fluoride (PVDF) or a nitrocellulose membrane (16). Within this macroarray, the coordinates of each protein are determined by the intrinsic properties of the individual protein/protein isoform—the protein’s isoelectric point and apparent molecular weight. An image acquisition device can capture and assign an *x* and *y* location for the proteins within the macroarray for the purpose of manipulation on a motion platform of a robotic system.

From: *The Proteomics Protocols Handbook*
Edited by: J. M. Walker © Humana Press Inc., Totowa, NJ

The methods described in this chapter are for the application of inkjet technology to perform *in situ* processing of membrane-immobilized protein macroarrays with subsequent matrix-assisted laser desorption/ionization (MALDI)-time-of-flight (TOF) MS analysis directly on the membrane or indirectly on a MALDI-TOF target surface (16). This method utilizes a piezoelectric drop-on-demand microdispensing device to dispense minimal quantities of fluids to selected protein spots on the membrane for the parallel processing of numerous tests. Thus, the proteolysis of a protein spot on the PVDF membrane requires 50-nL reagent volumes rather than the 50- μ L volumes necessary for in-gel digests. This technique also eliminates the gel excision steps, multiple liquid handling, and clean-up steps required when performing in-gel digests. A great advantage of the methods described is the ability to archive the protein samples on the membrane prior to and after performing the analysis, and the ability to perform multiple chemical reactions at different locations on an a selected protein spot within the membrane. The ability to apply multiple endoproteinases to a single protein spot can significantly increase the sequence coverage of a protein and increase the confidence of a successful identification (16). This is an important advantage when considering the characterization of low-abundance proteins, or proteins having minimal cleavage sites for a single enzyme.

2. Materials

2.1. Piezoelectric Microdispenser Set-Up

1. Glass capillary piezoelectric microjet devices (MicroFab Technologies, Inc., Plano, TX).
2. 0.25% (v/w) Polyvinylpyrrolidone (PVP360) in 50% (v/v) methanol.
3. PNGase F (Roche, Mannheim, Germany).
4. Porcine trypsin (Sigma-Aldrich, St. Louis, MO).
5. Matrix solution B: 10 mg/mL α -cyano-4-hydroxycinnamic acid in 30% (v/v) methanol, 20% (v/v) 2-propanol, 20% (v/v) 2-butanol, 0.1% (v/v) trifluoroacetic acid.
6. Adrenocorticotrophic hormone (ACTH) (Sigma-Aldrich, St. Louis, MO).

2.2. Electrophoretic Separation and Blotting of Human Plasma Proteins

1. Human plasma samples.
2. ProteoPrepTM Sample Extraction Kit (Sigma-Aldrich, St. Louis, MO). Human plasma electrophoresis reagent: 7 M urea, 2 M thiourea, 2% (w/v) CHAPS, 5 mM Tris (pH 10.2) (Proteome Systems, Inc., Woburn, MA).
3. Reducing reagent: 5 mM tributylphosphine.
4. Alkylation reagent: 10 mM iodoacetamide.
5. Dithiothreitol.
6. 11.0-cm immobilized pH gradient (IPG) strips (Amersham-Pharmacia Biotech, Uppsala, Sweden).
7. Multi Compartment Electctrolyser, IsoelectrIQ^{2TM} (Proteome Systems, Inc., Woburn, MA).
8. Immobilon PSQTM PVDF membrane (Millipore, Bedford, MA).
9. ElectrophoretIQ^{3TM} isoelectric focusing, second-dimension focusing, and electroblotting apparatus (Proteome Systems, Inc., Woburn, MA).
10. 0.008% (w/v) Direct Blue 71, (Sigma-Aldrich, St. Louis, MO), 40% (v/v) ethanol, 10% (v/v) acetic acid in DI H₂O.
11. Equilibration solution: 6 M urea, 2% (w/v) sodium dodecyl sulfate (SDS), 50 mM Tris-HCl (pH 7.0).
12. 6–15% (w/v) tris-acetate SDS-PAGE precast 10 cm \times 15 cm GelChipsTM (Proteome Systems, Inc., Woburn, MA).

2.3. Release and Analysis of N-Linked Oligosaccharides by On-Membrane Digestion Using PNGase F

1. AXIMA-CFR MALDI-TOF target plate (Kratos, Manchester, UK).
2. 3MTM electrically conductive tape 9713 (3M, St. Paul, MN).
3. Prototype ChIP Chemical Inkjet Printer (Proteome Systems, Ltd., Sydney, Australia).
4. PNGase F (Roche, Mannheim, Germany).
5. 0.25% (w/v) Polyvinylpyrrolidone (PVP360) in 50% (v/v) methanol.
6. Humidity chamber (container).
7. Incubator (37°C).
8. 0.1% (v/v) Trifluroacetic acid solution.
9. 96-Well microtiter plate.
10. Micropipettor.
11. Elution buffer: acetonitrile in 10 mM NH₄HCO₃.
12. ThermoHypersil 5 μm Hypercarb column (Thermo Hypersil-Keystone, Bellefonte, PA).
13. ThermoFinnigan LCQ Deca mass spectrometer (Thermo Finnigan, San Jose, CA).

2.4. On-Membrane Protein Digestion and Offline MALDI-TOF MS Analysis

1. Porcine trypsin (Sigma-Aldrich, St. Louis, MO).
2. C18 ZipTipTM (Millipore, Bedford, MA).
3. Humidity chamber (container).
4. Incubator (37°C).
5. Micropipettor.
6. 0.1% (v/v) Trifluroacetic acid solution.
7. Matrix solution A: 2.0 mg/mL α-cyano-4-hydroxycinnamic acid in 90% (v/v) acetonitrile, 0.1% (v/v) trifluroacetic acid.
8. AXIMA-CFR MALDI-TOF target plate (Kratos, Manchester, UK).
9. 3MTM electrically conductive tape 9713 (3M, St. Paul, MN).
10. Prototype ChIP Chemical Inkjet Printer (Proteome Systems, Ltd., Sydney, Australia).
11. Adrenocorticotrophic hormone (ACTH) (Sigma-Aldrich, St. Louis, MO).
12. AXIMA-CFR MALDI-TOF mass spectrometer (Kratos, Manchester, UK).

2.5. On-Membrane Protein Digestion and On-Membrane MALDI-TOF MS Analysis

1. Porcine trypsin (Sigma-Aldrich, St. Louis, MO).
2. Gel piece wash solution: 25 mM NH₄HCO₃, pH 8.5.
3. Humidity chamber (container).
4. Incubator (37°C).
5. Matrix solution B: 10 mg/mL α-cyano-4-hydroxycinnamic acid in 30% methanol, 20% 2-propanol, 20% 2-butanol, 0.1% (v/v) trifluroacetic acid.
6. AXIMA-CFR matrix-assisted laser desorption/ionization time of flight (MALDI-TOF) target plate (Kratos, Manchester, UK).
7. 3MTM electrically conductive tape 9713 (3M, St. Paul, MN).
8. Prototype ChIP Chemical Inkjet Printer (Proteome Systems, Ltd., Sydney, Australia).
9. Adrenocorticotrophic hormone (ACTH) (Sigma-Aldrich, St. Louis, MO).
10. AXIMA-CFR matrix-assisted laser desorption/ionization time of flight (MALDI-TOF) mass spectrometer (Kratos, Manchester, UK).

3. Methods

Peptide mass fingerprinting (PMF) analysis is performed conventionally by isolating individual protein spots of interest from a 2-D PAGE gel for tryptic digestion and

matrix-assisted laser desorption/ionization time of flight mass spectrometry (MALDI-TOF MS) analysis. In this process, the spots are extracted for in-gel tryptic digestion followed by peptide extraction, sample clean-up, and finally, the loading of the digested protein extracts onto a MALDI-TOF target for MS analysis to generate a mass peak list for a protein database query (17). The following methods describe an alternative rapid high-throughput approach with the dispensing of enzymes onto a macroarray of electroblotted human plasma proteins on a PVDF membrane adhered to a MALDI-TOF target plate.

3.1. Piezoelectric Microdispenser Set-up

Described below are the processes required to prepare the piezoelectric microdispensers on the chemical inkjet printer prototype for dispensing of solutions for the *in situ* endoglycosidase and endoproteinase digests of the membrane-bound proteins.

1. Load the piezoelectric microdispenser fluid reservoirs on the chemical inkjet printer prototype with: Reservoir 1: 0.25% (w/v) polyvinylpyrrolidone (PVP360). Reservoir 2: 1 U/ μ L PNGase F. Reservoir 3: 200 μ g/mL trypsin. Reservoir 4: 10 mg/mL α -cyano-4-hydroxycinnamic acid containing 25 fmoles of adrenocorticotrophic hormone (ACTH) (*see Note 1*).
2. Prime each microdispenser by applying pressure to fill the glass capillary to the orifice with fluid. Adjust the vacuum control to maintain the fluid within the orifice.
3. Input the electronic drive parameters for operation of the microdispenser. Initially, set the rise and fall time to 3 μ s, dwell time to 30 μ s, dwell voltage to 25 V, and frequency to 240 Hz (*see Note 2*).
4. Increase the dwell voltage in 3- to 5-V increments if no dispensing is observed (*see Note 3*).
5. An unstable drop or drop with satellites can often be corrected by adjustment of the dwell time in 1- μ s to 3- μ s increments (*see Note 4*).
6. It is very important to clean the microdispenser after use (*see Note 5*).

3.2. Electrophoretic Separation and Blotting of Human Plasma Proteins

The serum albumin-depleted human plasma sample is separated based on molecular weight and pI, using two-dimensional polyacrylamide gel electrophoresis (2-D PAGE) followed by electroblotting onto Immobilon P^{SQ} PVDF membrane (**Fig. 1**). The membrane array is adhered to a MALDI-TOF MS target plate, and selected proteins are digested using chemical inkjet printer noncontact chemical microdeposition.

1. Deplete the human plasma sample of serum albumin using ethanol precipitation of plasma proteins (serum albumin will stay in solution).
2. Resuspend the pellet following the ethanol precipitation in the ProteoPrep resuspension buffer.
3. Reduce and alkylate proteins using 5 mM tributylphosphine (TBP) and 10 mM iodoacetamide according to the method described in the ProteoPrep kit.
4. Quench the reaction with 100 μ L of dithiothreitol (DTT) solution to a final concentration of 10 mM DTT.
5. Pipet 200 μ L of the supernatant for passive rehydration of an 11-cm immobilized pH gradient (IPG) (pI 4–7) strip. Rehydrate the IPG strip for 6 h at room temperature.
6. Separate the proteins by electrophoresis for a total of 36,000 V·h.
7. Equilibrate IPG strips for 10 min in IPG equilibration buffer.
8. Place the IPG strip into the top of the 6–15% Tris acetate ProteomIQTM GelChips and separate proteins by electrophoresis.

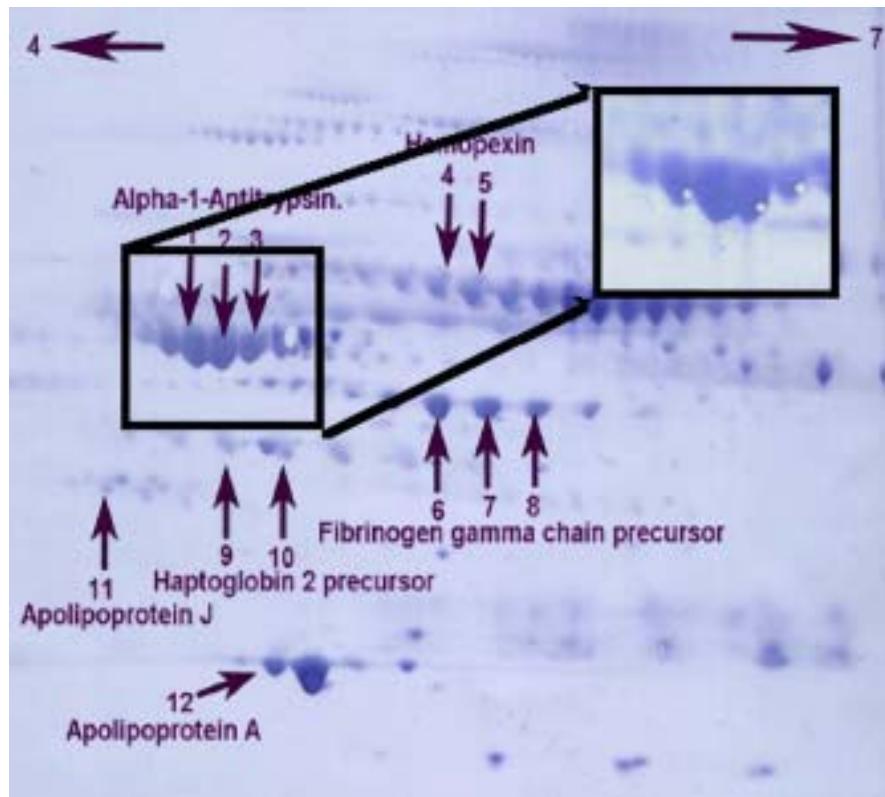


Fig. 1. Human plasma separated by two-dimensional polyacrylamide gel electrophoresis and electroblotted onto polyvinylidene fluoride membrane. Labeled protein spots are digested *in situ* using the Chemical Inkjet Printer. Inset of α -1-antitrypsin protein isoforms shows the Direct Blue 71 stain removed after the microdeposition of reagents using the Chemical Inkjet Printer.

9. Following second-dimension electrophoresis, electroblot proteins onto Immobilon P^{SQ}™ membrane using semi-dry electroblotting technique at a constant current of 300 mA for 80 min.
10. Stain the membrane with Direct Blue 71 for 15 min.
11. Dry membrane and store in a plastic bag to reduce keratin contamination (see Note 6).

3.3. Release and Analysis of N-Linked Oligosaccharides by On-Membrane Digestion Using PNGase F

The following describes the use of the chemical inkjet printer to extract oligosaccharides from proteins electroblotted onto a PVDF membrane (Fig. 1) using PNGase F digestion. The N-linked glycosylation sites are identified using MALDI-TOF MS (Fig. 4) and structural analysis by liquid chromatography (LC)-electrospray ionization (ESI) MS. (Fig. 2 and Table 1).

1. Adhere the membrane to the MALDI-TOF MS target plate with double-sided conductive tape from 3M™.
2. Insert the membrane adhered to the MALDI-TOF MS target plate onto the ChIP motion stage.
3. Acquire a scanned image of the membrane attached to the MALDI-TOF MS target plate.

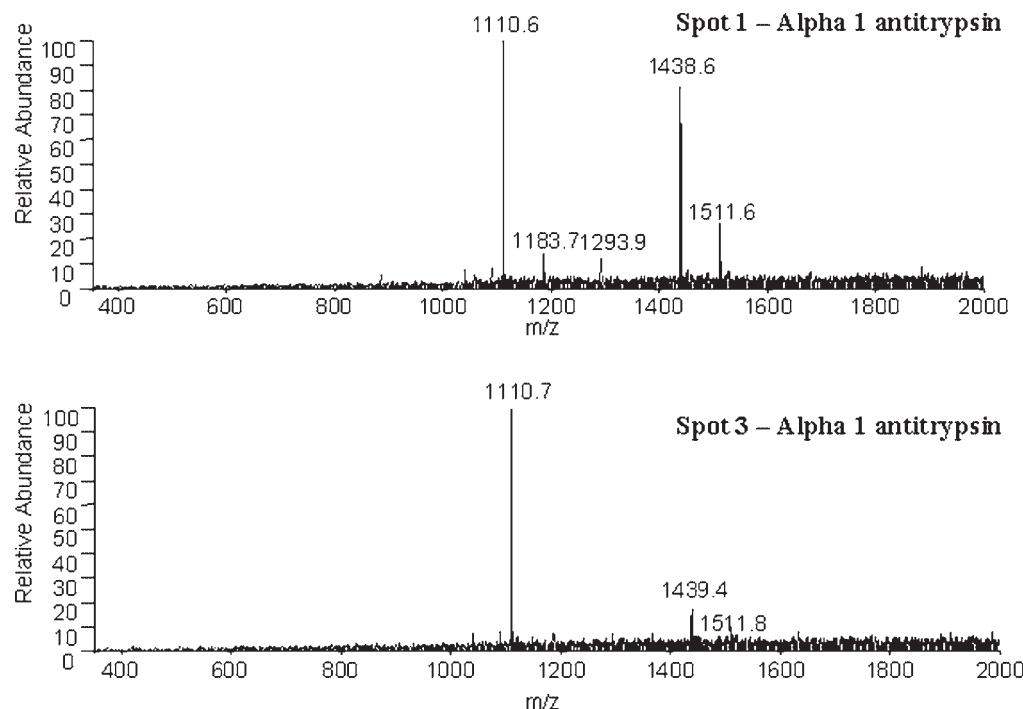


Fig 2. Liquid chromatography (LC)-electrospray ionization (ESI) mass spectrometry analysis showing α -1-antitrypsin chain oligosaccharides cleaved by PNGase F digestion. The different isoforms from the chain have different ratios of sugars.

4. Manually or automatically select protein spots using the ChIP software, and adjust registration points to calibrate the selected *x* and *y* coordinates to the membrane target plate and motion stage. Create two print (microdeposition) positions on the selected protein spot for peptide mass fingerprinting (PMF) and PNGaseF/PMF digestion.
 5. Select appropriate jetting parameters to dispense 10 nL (100 microdrops) of 0.25% w/v PVP onto selected protein spots (see Note 7).
 6. Dispense 10 iterations of 20 microdrops (20 nL or 0.02 U total) (see Note 8) of 1 μ L of PNGase F onto PNGase F print positions (see Note 9).
 7. Place membrane into a humidified container and incubate at 37°C for 3 h (see Note 10).
 8. Remove the oligosaccharides from the membrane by applying 2 μ L of 0.1% trifluoroacetic acid solution to the selected protein spots, followed by aspiration using a micropipettor. Transfer the aspirated oligosaccharides into a microtiter plate and dilute with 8 μ L 0.1% trifluoroacetic acid solution for liquid chromatography electrospray ionization mass spectrometric (LC-ESI MS) analysis (see Note 11).
 9. Load the oligosaccharide samples using the ThermoHypersil 5 μ m Hypercarb column into the ThermoFinnigan LCQ Deca mass spectrometer. Elute over a period of 30 min using 0–25% gradient of acetonitrile in 10 mM NH_4HCO_3 (18).
 10. The resulting masses are used to search GlycoSuiteDB (Proteome Systems, Ltd. <http://www.glycosuite.com/>) to predict possible oligosaccharide structures.
- The deamidase PNGase F is used to remove *N*-linked oligosaccharides from the glycoproteins in the methods described above. The released oligosaccharides are then analyzed using LC-ESI MS. **Figure 2** displays an example of the five different masses of the carbohydrate moieties released from two of the protein isoforms in the alpha-1-antitrypsin

Table 1

Predicted oligosaccharide structures released by PNGase F treatment of the protein a-1-antitrypsin isoforms using the methods described in Subheading 3.3. Structures were obtained from GlycoSuiteDB database searches (error \pm 1 Da) of the calculated monoisotopic mass of the oligosaccharide determined by liquid chromatography (LC)-electrospray ionization (ESI) mass spectrometry.

Predicted Structure (GlycoSuiteDB)	
$[M-2H]^2$ Mass: 1110.6	Monoisotopic Mass: 2222.78
<chem>Neu5c[2]--6Gal[1]--4GlcNAc[1]--2Man[1]</chem>	
<chem>Neu5c[2]--6Gal[1]--4GlcNAc[1]--2Man[1]</chem>	
$[M-2H]^2$ Mass: 1183.7	Monoisotopic Mass: 2368.84
<chem>Neu5c[2]--6Gal[1]--4GlcNAc[1]--2Man[1]</chem>	
<chem>Neu5c[2]--6Gal[1]--4GlcNAc[1]--2Man[1]</chem>	
$[M-2H]^2$ Mass: 1293.9	Monoisotopic Mass: 2587.92
<chem>Gal[1]--4GlcNAc[1]--2Man[1]</chem>	
<chem>Neu5c[2]--3Gal[1]--4GlcNAc[1]</chem>	
<chem>Neu5c[2]--3Gal[1]--4GlcNAc[1]</chem>	
$[M-2H]^2$ Mass: 1438.6	Monoisotopic Mass: 2879.01
<chem>Neu5c[2]--3Gal[1]--4GlcNAc[1]</chem>	
<chem>Neu5c[2]--6Gal[1]--4GlcNAc[1]</chem>	
<chem>Neu5c[2]--6Gal[1]--4GlcNAc[1]--2Man[1]</chem>	
$[M-2H]^2$ Mass: 1511.6	Monoisotopic Mass: 3025.07
<chem>Neu5c[2]--6Gal[1]--4GlcNAc[1]--2Man[1]</chem>	
<chem>Neu5c[2]--3Gal[1]--4GlcNAc[1]</chem>	
<chem>Neu5c[2]--6Gal[1]--4GlcNAc[1]</chem>	

chain. Predicted structures, shown in **Table 1**, are obtained by searching the monoisotopic masses in GlycoSuiteDB.

3.4. On-Membrane Protein Digestion and Offline MALDI-TOF MS Analysis

The following describes the methods to perform tryptic digestion of selected proteins spots on a 2-D PAGE PVDF protein blot, such as on the PNGase F treated spots and non-PNGase F treated spots from the methods described in **Subheading 3.3.** using the chemical inkjet printer. The MALDI-TOF MS analysis is performed offline by removing the digested peptides from the membrane by aspiration using a C18 ZipTip attached to a micropipet. The digested peptides are then transferred to a MALDI-TOF MS target plate. The C18 ZipTip is useful to extract and concentrate the digested peptides when they are removed from the PVDF membrane.

1. Insert the membrane from **Subheading 3.3** adhered to the MALDI-TOF MS target plate onto the chemical inkjet printer (ChIP) motion stage.
2. Acquire a scanned image of the membrane attached to the MALDI-TOF MS target plate.
3. Manually or automatically select protein spots using the ChIP software, and adjust registration points to calibrate the selected *x* and *y* coordinates to the membrane target plate and motion stage.
4. Dispense 25 iterations of 20 micropipet drops (50 nL or 10 ng total) of 200 μ g/mL trypsin in 25 mM NH_4HCO_3 (see **Note 8**) onto PNGase F treated spots and non-PNGase F treated spots on the PVDF protein membrane from **Subheading 3.3.** (see **Note 9**).
5. Return the membrane to the humidified container and incubate at 37°C for 3 h (see **Note 10**).
6. Aspirate the peptides from the membrane using a C18 ZipTip attached to a micropipet using 2 μ L of 0.1% trifluoroacetic acid solution, and dispense onto a hydrophobic MALDI-TOF MS target plate using 1 μ L of 2.0 mg/mL of α -cyano-4-hydroxycinnamic acid, 90% acetonitrile, 0.1% trifluoroacetic acid (matrix solution A) containing 25 fmoles of ACTH.
7. Place the MALDI-TOF MS target plate containing the digested peptides, matrix solution A, and ACTH into the AXIMA-CFR MALDI-TOF mass spectrometer to collect mass spectra for the digested protein spots.
8. Calibrate all spectra using an internal two-point calibration. The two peaks to use for calibration are the ACTH peak at 2466.20 Da (second isotope, most abundant) and the trypsin autolysis peak at 842.51 Da (monoisotopic, most abundant). Protein identifications are made using PMF search engines, such as MASCOT (<http://www.matrixscience.com/>), ProFound (<http://prowl.rockefeller.edu/>), and IonIQ (Proteome Systems, Ltd., in-house tool), searching protein databases.

3.5. On-Membrane Protein Digestion and On-Membrane MALDI-TOF MS Analysis

The following describes the methods to perform tryptic digestion of selected protein spots on a 2-D PAGE PVDF protein blot, such as on the PNGase F treated spots and non-PNGase F treated spots from the methods described in **Subheading 3.3.** using the chemical inkjet printer. The MALDI-TOF MS target plate containing the PVDF membrane with the digested peptides is placed directly into the MALDI-TOF mass spectrometer to collect mass spectra.

1. Insert the membrane from **Subheading 3.3.** adhered to the MALDI-TOF MS target plate onto the ChIP motion stage.
2. Acquire a scanned image of the membrane attached to the MALDI-TOF MS target plate.
3. Manually or automatically select protein spots using the ChIP software, and adjust reg-

- istration points to calibrate the selected *x*-*y* coordinates to the membrane target plate and motion stage.
4. Dispense 25 iterations of 20 microdrops (50 nL or 10 ng total) of 200 µg/mL trypsin in 25 mM NH₄HCO₃ (see Note 8) onto PNGase F treated spots and non-PNGase F treated spots on the PVDF protein membrane from Subheading 3.3. (see Note 9).
 5. Return the membrane to the humidified container and incubate at 37°C for 3 h (see Note 10).
 6. Dispense 25 iterations of 40 microdrops of 10 mg/mL of α-cyano-4-hydroxycinnamic acid and 0.1 % (v/v) trifluoroacetic acid in 30% (v/v) methanol, 20% (v/v) 2-propanol, 20% (v/v) butanol (matrix solution B) containing 25 fmoles of ACTH onto trypsin-digested protein spots.
 7. Create a plate file to transfer the *x*-*y* coordinate map from ChIP to the AXIMA-CFR MALDI-TOF mass spectrometer (see Note 12).
 8. Place the MALDI-TOF MS target plate containing the PVDF membrane with peptides (see Note 13), matrix solution B, and ACTH into the AXIMA-CFR MALDI-TOF mass spectrometer to collect mass spectra for the digested protein spots (see Note 14).
 9. Calibrate all spectra using an internal two-point calibration. The two peaks to use for calibration are the ACTH peak at 2466.20 Da (second isotope, most abundant) and the trypsin autolysis peak at 842.51 Da (monoisotopic, most abundant).
 10. Protein identifications are made using PMF search engines, such as MASCOT (<http://www.matrixscience.com/>), ProFound (<http://prowl.rockefeller.edu/>), and IonIQ (Proteome Systems, Ltd., in-house tool), searching protein databases.

Proteins identified by peptide mass fingerprinting using the methods described in Subheadings 3.4. and 3.5. are shown in Table 2. Peptides derived from the tryptic digestion of protein spots 1 to 8 (Fig. 1) were desorbed directly off the membrane surface in the MALDI-TOF MS. Peptides derived from the tryptic digestion of protein spots 9 to 12 (Fig. 1) were aspirated off the membrane and deposited onto a metal target plate for analysis with MALDI-TOF MS. Figure 3 displays two examples of spectra collected directly from the membrane surface. Minimal loss of resolution from the porous membrane surface is achieved by collecting spectra from the same position on the protein spot (see inset ACTH calibrant at mass 2465 Da).

Precise nanoliter dispensing, using piezoelectric microdispensing, enables dual analyses on the same protein spot. Figure 4 shows two MALDI-TOF MS mass spectra from two print positions on the same protein spot. At one print position, a tryptic digestion is performed on the glycoprotein (Fig. 4A). At the other print position, the oligosaccharides are removed from the protein with a PNGase F digestion (Fig. 4B), and subsequently trypsin is dispensed directly onto the PNGase F-digested area. Digested samples are aspirated off the membrane using a C18 ZipTip, and the peptides are eluted with matrix onto a hydrophobic MALDI-TOF target plate for analysis using the AXIMA-CFR MALDI-TOF MS. Deglycosylation deamidates asparagines to aspartic acid, resulting in the gain of 1 Da to the parent peptide. A deglycosylated peptide of mass 1405.74 Da is observed in the PNGase F-treated protein (Fig. 4B). This peptide is absent at the print position that was treated with only the tryptic digestion, which confirms the presence of the oligosaccharide on the parent peptide (Fig. 4A).

4. Notes

1. The solutions to be microdispensed should be filtered using a 5.0-µm or smaller mesh filter prior to use to remove any particulate material that may block the orifice.
2. The surface tension and viscosity properties of each solution determine the electronic drive parameters required to obtain dispensing. Thus, each solution will require different

Table 2

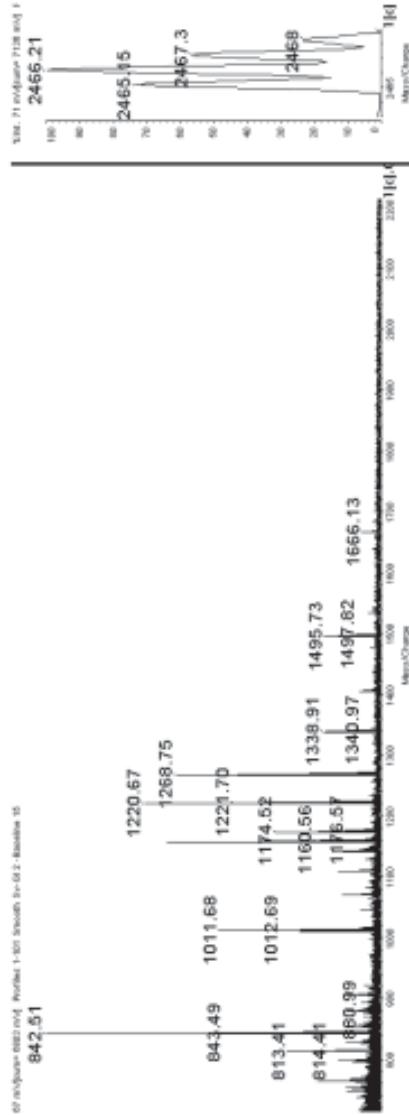
Protein identifications and percent coverage of human plasma proteins from the polyvinylidene fluoride protein blot shown in Fig. 1 after performing tryptic digestions described in methods Subheadings 3.4 and 3.5.

PVDF Protein Spot	Accession No.	Protein	MW (kDa)	pI	Number of Peptide Hits	Percentage of Amino Acid Coverage
1	P01009	Alpha-1-antitrypsin precursor	46.7	5.37	12	30.14
2	P01009	Alpha-1-antitrypsin precursor	46.7	5.37	8	18.9
3	P01009	Alpha-1-antitrypsin precursor	46.7	5.37	7	16.99
4	P02790	Hemopexin precursor	51.6	6.57	8	15.58
5	P02790	Hemopexin precursor	51.6	6.57	5	9.8
6	P02679	Fibrinogen gamma chain precursor	51.5	5.37	12	26.71
7	P02679	Fibrinogen gamma chain precursor	51.5	5.37	9	18.98
8	P02679-2	Fibrinogen gamma chain precursor	49.5	5.7	11	24.72
9	P00738	Haptoglobin - 2 precursor	45.2	6.1	8	19.6
10	P00739	Haptoglobin - 2 precursor	45.2	6.1	10	27.95
11	P10909	Clusterin (Apolipoprotein J) alpha chain	25.9	5.66	9	45
12	P02647	Apolipoprotein - A	30.8	5.56	11	38.2

electronic drive parameters. It is useful to record the dwell time and dwell voltage parameters required for each solution for future reference. However, these may vary depending on the microdispenser used.

- Particulate contamination can block the orifice, resulting in drops that dispense intermittently or at an angle. Carefully wipe the microdispenser tip to remove such contamination. If this fails, the particulate material can be removed by immersing the tip of the microdispenser into methanol and applying vacuum to back flush the device.
- Check the following if the microdispenser is not dispensing:
 - Does the fluid reach the orifice of the microdispenser? Adjust the pneumatics for positive pressure if the fluid does not reach the orifice.
 - Are there air bubbles present in the glass tip of the microdispenser? Remove air bubbles by manipulating the pneumatics.
 - Is there sufficient fluid in the reservoir? Check to confirm that there is sufficient volume in the reservoir.
 - Are the wires of the microdispenser securely connected? Check to confirm that the wires are connected.
- The microdispenser should be back flushed immediately after use by applying vacuum to the back inlet of the device, while applying fluid to the tip. The fluid used for flushing should be the solvent for the fluid that was dispensed. This will aid in the removal of any material adhering to the glass capillary. Then, immerse the tip of the device and back flush with a 2% solution (80°C) of Micro-90 Cleaning Solution (Cole-Parmer Cat. no. U-18100-

Sample 4 - Hemopexin



Sample 6 - Fibrinogen

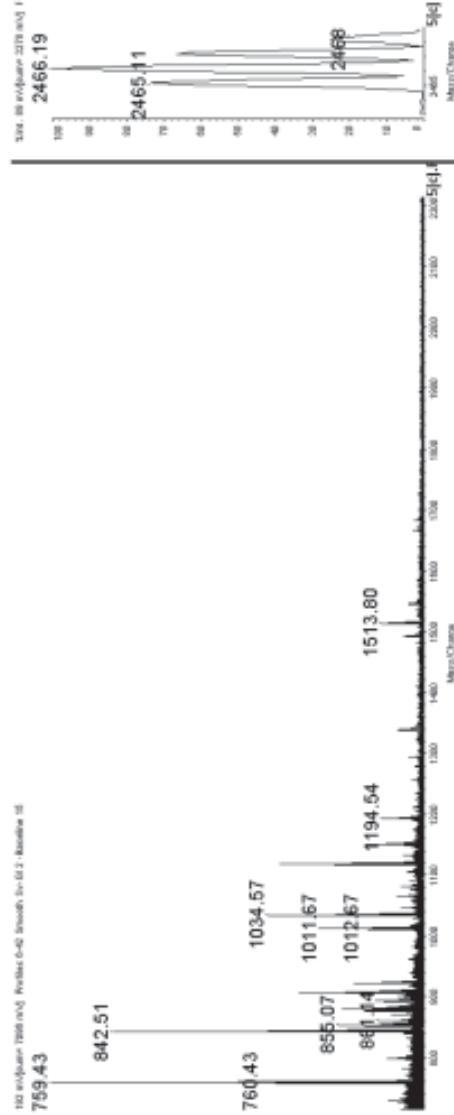


Fig. 3. Matrix-assisted laser desorption/ionization-time-of-flight mass spectrometry analysis spectra for hemopexin and fibrinogen collected directly from the membrane surface.

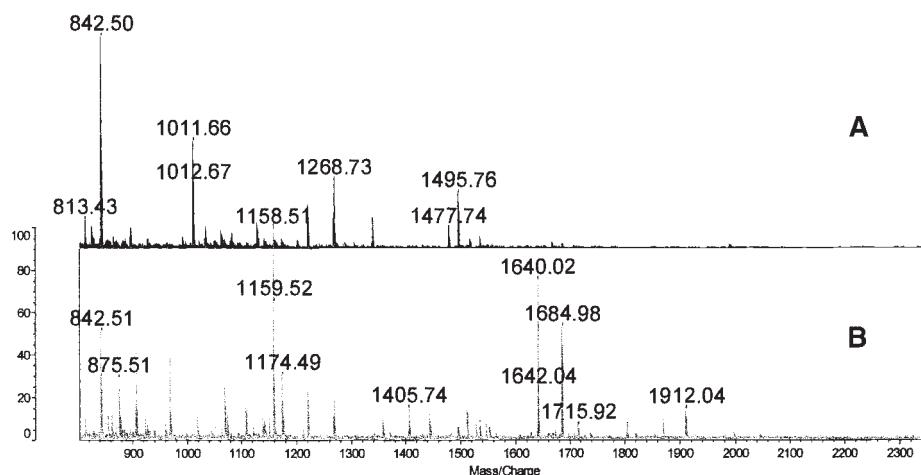


Fig. 4. Detecting the deglycosylated peptide detected by matrix-assisted laser desorption/ionization-time-of-flight mass spectrometry analysis results in an increased coverage and confidence in identification of hemopexin (**Fig. 4B**). Hemopexin is co-migrated with serum albumin protein in the PNGase F treated protein. The identification of a hemopexin glycopeptide is confirmed by the observation of the deglycosylated-deamidated peptide SWPAVGDCSSALR (1405.74, NCS → DCS).

00). Observe the tip of the microdispensing device after 1 min of back flushing to verify that all residues, particles, or condensations have been removed. Repeat the procedure until the tip is clear. Next, back flush the tip using MilliQ water for 1 min, followed by a 1-min back flush using acetone. Dry the microdispenser at 50°C before storing.

6. Keratin contamination is detrimental to any proteomics experiment. To reduce keratin contamination, use only methanol-cleaned tweezers to handle the PVDF membrane. Place the electroblotted membrane on filter paper in a covered container to allow membrane to dry. Store the membranes in plastic Ziplock bags to reduce exposure to keratin.
7. Polyvinylpyrrolidone blocks the membrane to ensure that endoprotease, peptides, and oligonucleotides do not stick to the membrane.
8. Dispensing iterations of a small number of drops ensures that digested areas remain between 300 and 400 µm in diameter.
9. If there is a difference between the spots selected and the areas where solution is being deposited: (a) Confirm the alignment of the microdispenser tip to the MALDI-TOF target plate; (b) confirm the alignment of the MALDI-TOF target to the motion stage; (c) confirm the alignment of the motion stage to the scanned image of the protein blot; (d) confirm that the MALDI-TOF target plate is inserted properly onto the motion stage; (e) confirm that the microdispenser is dispensing properly.
10. Use an airtight container for humidified incubation of the membrane. Fill the bottom of the container with MQ water and place the MALDI-TOF target plate onto an elevated platform to prevent contact of water with the target plate or the membrane.
11. A large volume of 0.1% trifluoroacetic acid solution (2 µL) can be used to extract the released oligosaccharides from the membrane, due to the difference in solubility between the polar oligosaccharides and the immobilized protein.

12. A plate file is the coordinate map required to drive the MALDI-TOF MS stage to the specific digested protein spots.
13. Due to the porous nature of the membrane surface, the MALDI-TOF MS analysis will produce broad peaks and lower resolution. Ensure that the resolution of high peptides is sufficient to calibrate and analyze. If the software version allows for summing of spectra, calibrate spectra and sum together.
14. Collecting spectra from the AXIMA-CFR. (a) Collect from one position on the membrane or collect only a small number of profiles from each spectrum and calibrate each spectrum. Sum the spectra together to increase the signal-to-noise ratio. (b) Most peptides are found in the outer ring of the digested protein region. This is due to the “corona effect,” where peptides are washed to the outer areas of the digested spot. The best results are obtained by analyzing peptides located around the outer area of digested spot.

References

1. Lee, K. H. (2001) Proteomics: a technology-driven and technology-limited discovery science. *Trends Biotechnol.* **19**, 217–222.
2. Hamdan, M. and Righetti, P. G. (2002) Modern strategies for protein quantification in proteome analysis: Advantages and limitations. *Mass Spectrom. Rev.* **21**, 287–302.
3. Smith, R. D. (2002) Trends in mass spectrometry instrumentation for proteomics. *Trends Biotechnol.* **20**, S3–S7.
4. Righetti, P. G., Castagna, A., and Herbert, B. (2001) Prefractionation techniques in proteome analysis. *Anal. Chem.* **73**, 320A–326A.
5. Govorun, V. M., and Archakov, A. I. (2002) Proteomic technologies in modern biomedical science. *Biochemistry* **67**, 1109–1123.
6. Gygi, S. P., Corthals, G. L., Zhang, Y., Rochon, Y., and Aebersold, R. (2000) Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proc. Natl. Acad. Sci. USA* **97**, 9390–9395.
7. Wilson, D. S. and Nock, S. (2003) Recent developments in protein microarray technology. *Angew. Chem. Int. Ed. Engl.* **42**, 494–500.
8. Lee, Y. S. and Mrksich, M. (2002) Protein chips: from concept to practice. *Trends Biotechnol.* **20**, S14–18.
9. Zhu, H. and Snyder, M. (2003) Protein chip technology. *Curr. Opin. Chem. Biol.* **7**, 55–63.
10. Sanders, G. H. and Manz, A. (2000) Chip-based Microsystems for genomic and proteomic analysis. *Trends Anal. Chem.* **19**, 364–378.
11. Figeys, D. and Pinto, D. (2001) Proteomics on a chip: promising developments. *Electrophoresis* **22**, 208–216.
12. Candiano, G., Musante, L., Bruschi, M., et al. (2002) Two-dimensional maps in soft immobilized pH gradient gels: a new approach to the proteome of the Third Millennium. *Electrophoresis* **23**, 292–297.
13. Gorg, A., Boguth, G., Kopf, A., Reil, G., Parlar, H., and Weiss, W. (2002) Sample prefractionation with Sephadex isoelectric focusing prior to narrow pH range two-dimensional gels. *Proteomics* **2**, 1652–1657.
14. Westbrook, J. A., Yan, J. X., Wait, R., Nelson, S. Y., and Dunn, M. J. (2001) Zooming-in on the proteome: very narrow-range immobilized pH gradients reveal more protein species and isoforms. *Electrophoresis* **22**, 2865–2871.
15. Pedersen, S. K., Harry, J. L., Sebastian, L., et al. (2003) Unseen proteome: mining below the tip of the iceberg to find low abundance and membrane proteins. *J. Proteome Res.* **2**, 303–311.

16. Sloane, A. J., Duff, J. L., Wilson, N. L., et al. (2002) High throughput peptide mass finger-printing and protein macroarray analysis using chemical printing strategies. *Mol. Cell. Proteomics* **1**, 490–499.
17. Henzel, W. J., Billeci, T. M., Stults, J. T., Wong, S. C., Grimley, C., and Watanabe, C. (1993) Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proc. Natl. Acad. Sci. USA* **90**, 5011–5015.
18. Wilson, N. L., Schulz, B. L., Karlsson, N. G., and Packer, N. H. (2002) Sequential analysis of N- and O-linked glycosylation of 2-D PAGE separated glycoproteins. *J. Proteome Res.* **1**, 521–529.

Protein Identification by Peptide Mass Fingerprinting

Alastair Aitken

1. Introduction

This chapter describes the database identification of proteins that have been separated by one- (or two-) dimensional gel electrophoresis followed by in-gel digestion of the protein bands or spots from the gels. Database searches using mass spectrometry data are particularly important to identify a specific protein and to determine whether post-translational modifications are present. Use of some of the many websites for “proteomic” identification of proteins is described here. Various software, including commercial software packages such as SEQUEST (*see Note 1*), is available to use the information on the fragment ions obtained from a tandem MS experiment to search protein (and DNA translation) databases to identify the sequence and the protein from which it is derived. The increasing number and public availability of complete genome sequences has greatly increased the possibility of protein identification using intact peptide mass data (fingerprints). Fingerprints of digests of complex protein mixtures are unsuitable, since it would not be possible to ascertain which peptide(s) originated from which protein component in such a mixture. Matrix-assisted laser desorption/ionization (MALDI)-time-of-flight (TOF) mass spectrometry (MS) has a particular application here, due to the possibility of high throughput with the assistance of robotics to excise large numbers of spots from 2-D gels, digest, and spot onto a MALDI plate. Modern MALDI-TOF MS instruments now have the sufficiently high mass accuracy necessary for unambiguous identification based on fingerprint data (*1*). Expressed sequence tag (EST) databases do not usually contain sufficiently long sequences to be generally useful for identifying proteins using fingerprint data from intact peptide masses alone, since the part of the coding sequence that is present may be too short to include a sufficient number of peptides that are represented in the experimental data. Some sequence data (even if these are very short and fragmentary) obtained by tandem MS is required. The isotope-coded affinity tagging (ICAT) strategy for quantifying differential protein expression and sequence tag methodology that now enable the identification of a protein from only one peptide from that particular protein, are described in Chapter 38.

The databases contain hyperlinks to enable one to view the full protein summary and link to protein structure, Swiss 2-D polyacrylamide gel electrophoresis (PAGE), nucleic acid databases containing the available published literature on the protein (e.g., Medline), and so on. Other chapters in this volume describe these applications in detail (e.g., Chapters 26 [“2-DE Databases on the World Wide Web”]; 51 [“Computational

Identification of Related Proteins: *BLAST and PSI-BLAST and Other Tools*”]; 52 [“Protein Identification and Analysis Tools in the ExPASy Server”]; and 53 [“Protein Sequence Databases”]).

In addition to the more conventional combination of 2-D gel separation with mass spectrometry, it is worth noting the recent development of the molecular scanner technique that processes thousands of proteins simultaneously (2). Proteins separated in a 2-D system are digested in parallel in the gel and transferred onto a polyvinylidene fluoride (PVDF) membrane which retains their relative positions. The membrane is then sprayed with matrix and introduced into a MALDI-TOF mass spectrometer. A peptide mass fingerprint is obtained at each site on the scanned grid. Since all fingerprints are measured at the same time, information on the presence of chemical noise, a potential source for erroneous identification, is obtained. This can then be removed from the mass fingerprints, allowing identification of many weakly expressed proteins in the 2-D gel. The membrane can be reused after chemical derivatization to improve identification of particular proteins. Post-translational modifications can also be detected by addition of phosphatases and glycosidases, for example. The construction of this detailed 2-D map of the identified proteins is possible with the great advances in personal computers of even relatively modest processing power, speed, and memory.

2. Materials

1. Networked personal computer or terminal with good minimum configuration.
2. Mass spectrometer data-handling software.

3. Methods

Methods for protein digestion and extraction, and methods for obtaining peptides for mass fingerprinting and/or sequencing by tandem electrospray or MALDI-TOF MS are described in Chapters 31 and 33. The range of parameters to be entered (“input”) into the search form are roughly similar whichever program is used.

3.1. Case Study of Identification of a Protein by Interpretation of Mass Fingerprint Data

Proteins in a sheep brain cytosolic protein extract that interact with centaurin- α_1 were identified by affinity chromatography (4). Proteins that bound to the column were eluted with high salt, concentrated, and analyzed by sodium dodecyl sulfate (SDS)-PAGE. The proteins were visualized by GelCODE staining. The five major proteins that eluted specifically from the centaurin- α_1 affinity column were analyzed by MALDI-TOF mass spectrometry. The identification of one of the proteins, isoforms of leucine- and alanine-rich nuclear protein (LANP) by database searching of the mass fingerprint is illustrated (A). The home page, search form, and top “hits” of the results obtained from both Mascot (B) and Protein Prospector (C) search engines are shown. The numbers indicated in **Fig. 1B** relate directly to the numbered steps below.

1. Open the form page for the particular website search engine being used (see Note 2 and **Fig. 1**).
2. Input the list of peptide masses to four decimal places if possible. In the initial search, use masses from the higher signal intensity peaks and set the “Minimum Number of Peptides” low compared to number of masses in the peptide list. To increase the specificity of the

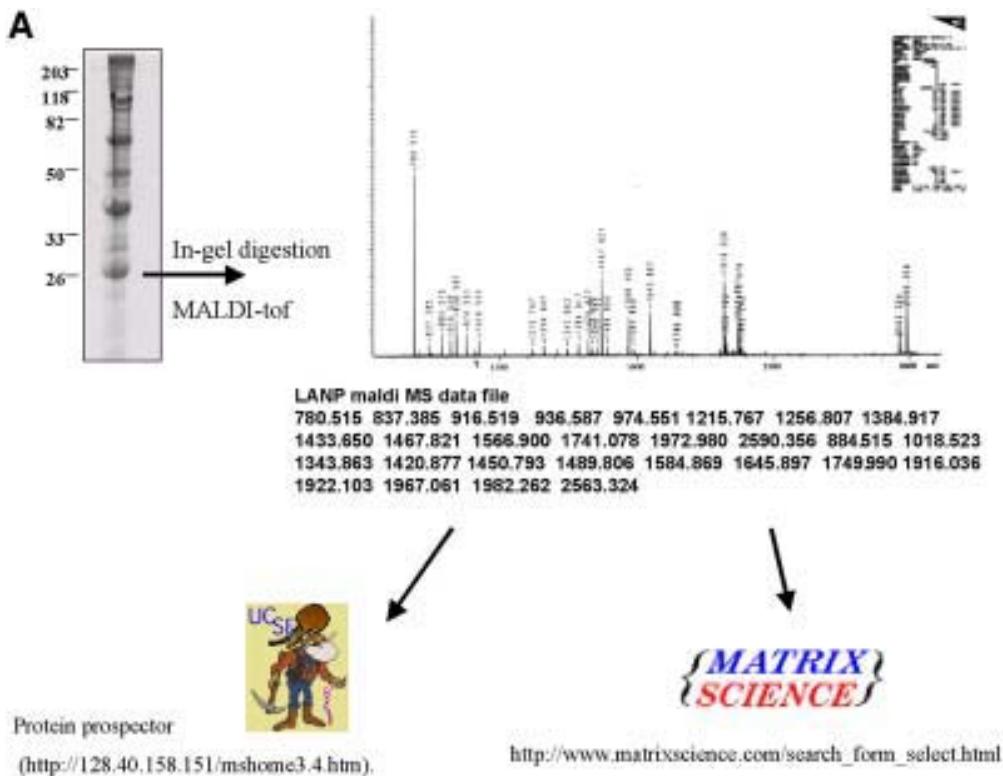


Fig. 1. Peptide mass fingerprint analysis of a sodium dodecyl sulfate–polyacrylamide gel electrophoresis band.

search, increase this number. If no hits are found, then decrease this number in subsequent passes.

3. Limit the search to particular species or genera, e.g., “mammalia” only, thus increasing the speed of the search. However, when looking for a homologous sequence in another eukaryote, for example, the species should not be defined.
4. Set the “Missed Cleavages” parameter to 1 or 2. Setting this to the latter value may be particularly important if a number of larger-size peptides are seen in the mass file (*see Note 3*).
5. The particular modification of cysteine residues, if any, should be included; otherwise, the number of peptides matched to the theoretical list will be decreased, producing a worse “hit” (*see Note 4*). Select the top three options if the possibility of post-translational or other modification is uncertain, i.e., acetylation of the N-terminus, oxidation of Met, and conversion of Glu to pyro-Glu. If phosphorylation of Ser, Thr, or Tyr is selected when not suspected, this may lead to false hits. More than one amino acid can be listed in this option (using the “Control” key).
6. If an internal calibration has been performed, the “mass accuracy” parameter can be set to 20 ppm. For a close external calibration, this should be set to 50 ppm. This is equivalent to ± 0.1 Da for a 2000 Da peptide (setting the tolerance to \pm fractions of a Dalton is an option). The choice is to a large extent dependent on the mass spectrometer from which the data have been acquired. If a hit is not found with the first search, this parameter can be increased to around 100 ppm or ± 1 Da without in most cases producing a major probabil-

B-1

Mascot Search

- Peptide Mass Fingerprint: The experimental data are a list of peptide mass values from an enzymatic digest of a protein.
 - Example of results report
 - More information
- Sequence Query: One or more peptide mass values associated with information such as partial or ambiguous sequence strings, amino acid composition information, MS/MS fragment ion masses, etc. A subset of a sequence tag query.
 - Example of results report
 - More information
- MS/MS Ion Search: Identification based on raw MS/MS data from one or more peptides.
 - Example of results report
 - More information

Search Form Defaults: Follow this link to save your preferred search form defaults as a browser cookie.

MASCOT Peptide Mass Fingerprint

8 Your name: Aitken Adrien Email: Aitken.Aitkend@ucl.ac.uk

9 Search Site: SwissProt

10 Database: Mammalia (mammals)

2 Protein mass: 500

3 Taxonomy: Mammalia (mammals)

4 Enzyme: Trypsin

5 Allow up to 1 missed cleavage

6 Variable modifications: Carbamidomethyl (C) - Carbamid (N)

7 Monoisotopic: Average

8 Data file: Query

950.525
837.385
884.529
916.529
926.587
974.581

100.525
120.525
140.525
160.525
180.525
200.525
220.525
240.525
260.525
280.525
300.525
320.525
340.525
360.525
380.525
400.525
420.525
440.525
460.525
480.525
500.525
520.525
540.525
560.525
580.525
600.525
620.525
640.525
660.525
680.525
700.525
720.525
740.525
760.525
780.525
800.525
820.525
840.525
860.525
880.525
900.525
920.525
940.525
960.525
980.525
1000.525

Fig. 1B1.

ity of chance “hits” with protein(s) that contain some peptides of quite different composition but somewhat similar masses (*see also step 13*, this section).

7. Select the option of whether the peptide or fragment and precursor ions have been calculated from monoisotopic or average masses (*see Note 5*).
8. Search different databases. NCBInr is the largest database, while Swiss-Prot is smaller. However Swiss-Prot provides the most information with the protein hits.
9. Define the enzyme used for the digestion (most frequently trypsin) (*see Note 3*).
10. If it is known that the protein is definitely not larger than 50 kDa, for example, then limit the “Mass Range” parameter to prevent false hits (*see Note 6*). Similar limits may be placed on the range of pI in MS-fit.
11. Examine thoroughly the ranking and scoring of your results to verify that the “MOWSE” score is significant (*see Note 7*).

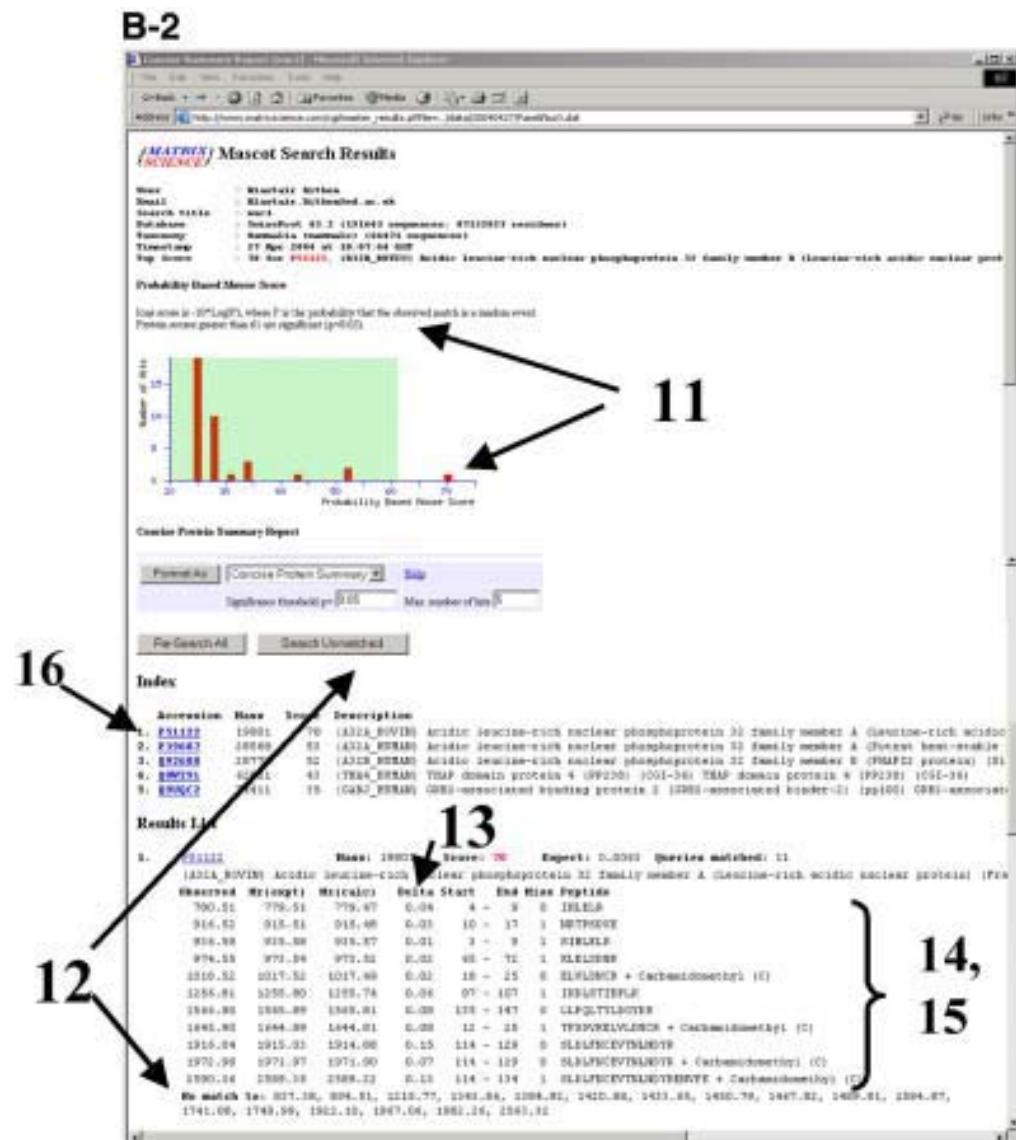


Fig. 1B2.

12. Re-search with the “Unmatched masses” option, since sometimes two or more proteins run together even on 2-D electrophoresis.
13. Note the “delta ppm” (the difference between the theoretical and the experimental mass of a particular peptide), which should be low and consistent. This gives an indication of whether the result is genuine. Errors or drift in the mass spectrometer calibration are inevitable, and this may be less accurate over a particular part of the spectrum; therefore, experimentally derived masses of the peptides from a genuine protein hit will differ from the theoretical masses, although they should all lie on the same side. For example, in the MS-fit search shown in **Fig. 1**, the delta ppm for the top hit peptides are around 10 to 40 ppm on the positive side (mean is +30 ppm) (see 13b). For the Mascot search, the

C-1

Fig. 1C1.

differences between experimentally determined and theoretical masses for all the peptides in the top protein hit are between 0.1 and 0.15 Da on the positive side (which is also approx +40 to 50 ppm). This indicates that the instrument calibration was tens of ppm to the high side when these data were collected. Another protein might by chance contain peptides with somewhat similar masses (but quite different composition), some of which in all probability would be higher than the mean and some, much lower.

14. Make some guesses depending on what is known about the mass fingerprint data file. For example, if the enzyme used for digestion is trypsin, then look for evidence of R or K at the C-terminus (*see Note 3*).
15. Make some assumptions on the amino acids that must be present or must be absent, if information is available on peptide composition.
16. There are hyperlinks to the database entry (Swiss-Prot or NCBI). In Mascot, this will first

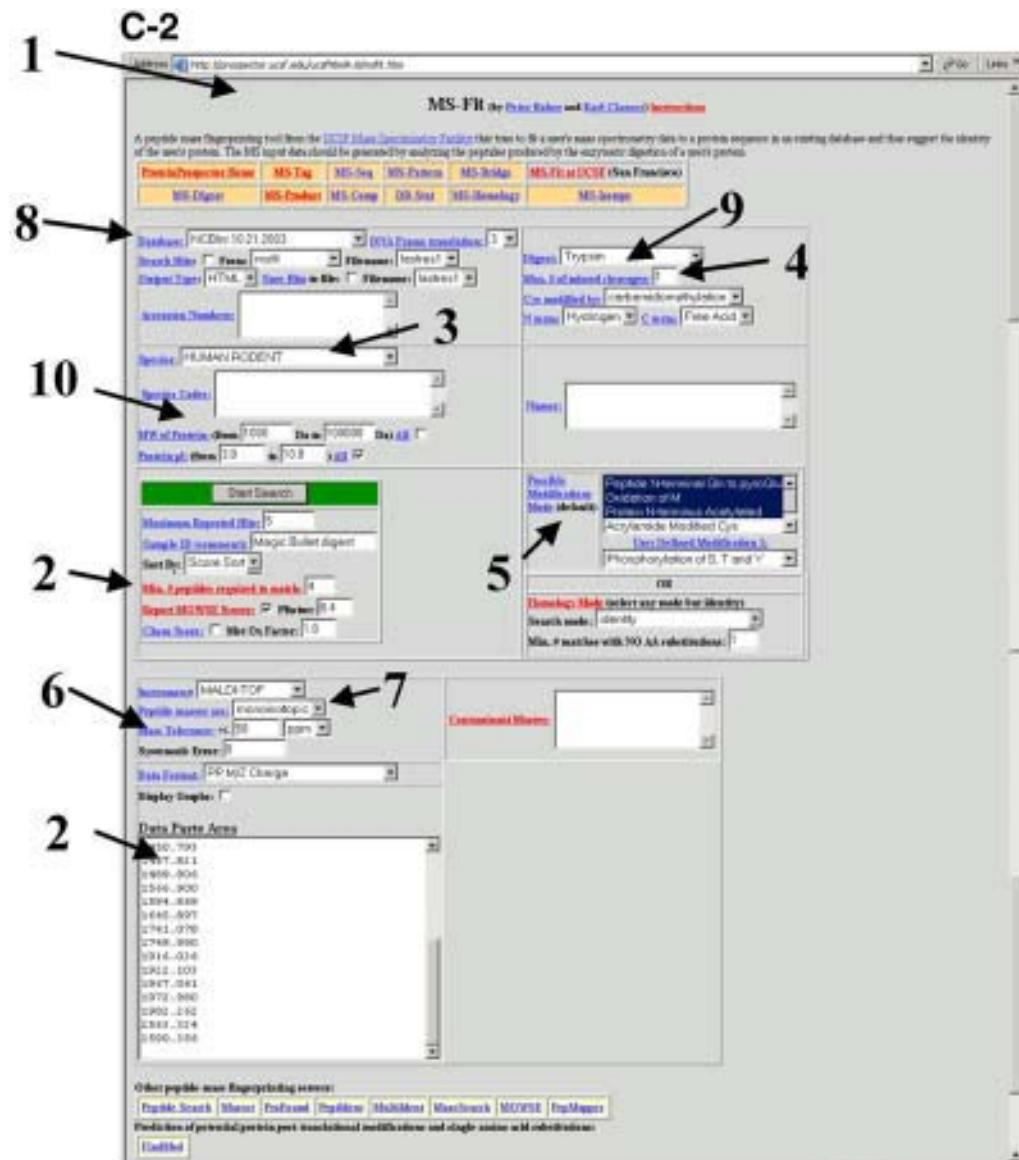


Fig. 1C2.

give a detailed "Protein View" of the matches, followed by the link through the Accession number to the database entry.

4. Notes

1. The SEQUEST commercial software uses the fragmentation data from a tandem mass spectrum to search complete protein databases to identify the sequence that best fits the fragmentation pattern. SEQUEST matches unknown MS/MS spectra to sequences in a database. It finds all the peptides that match the input masses and calculates preliminary

C-3

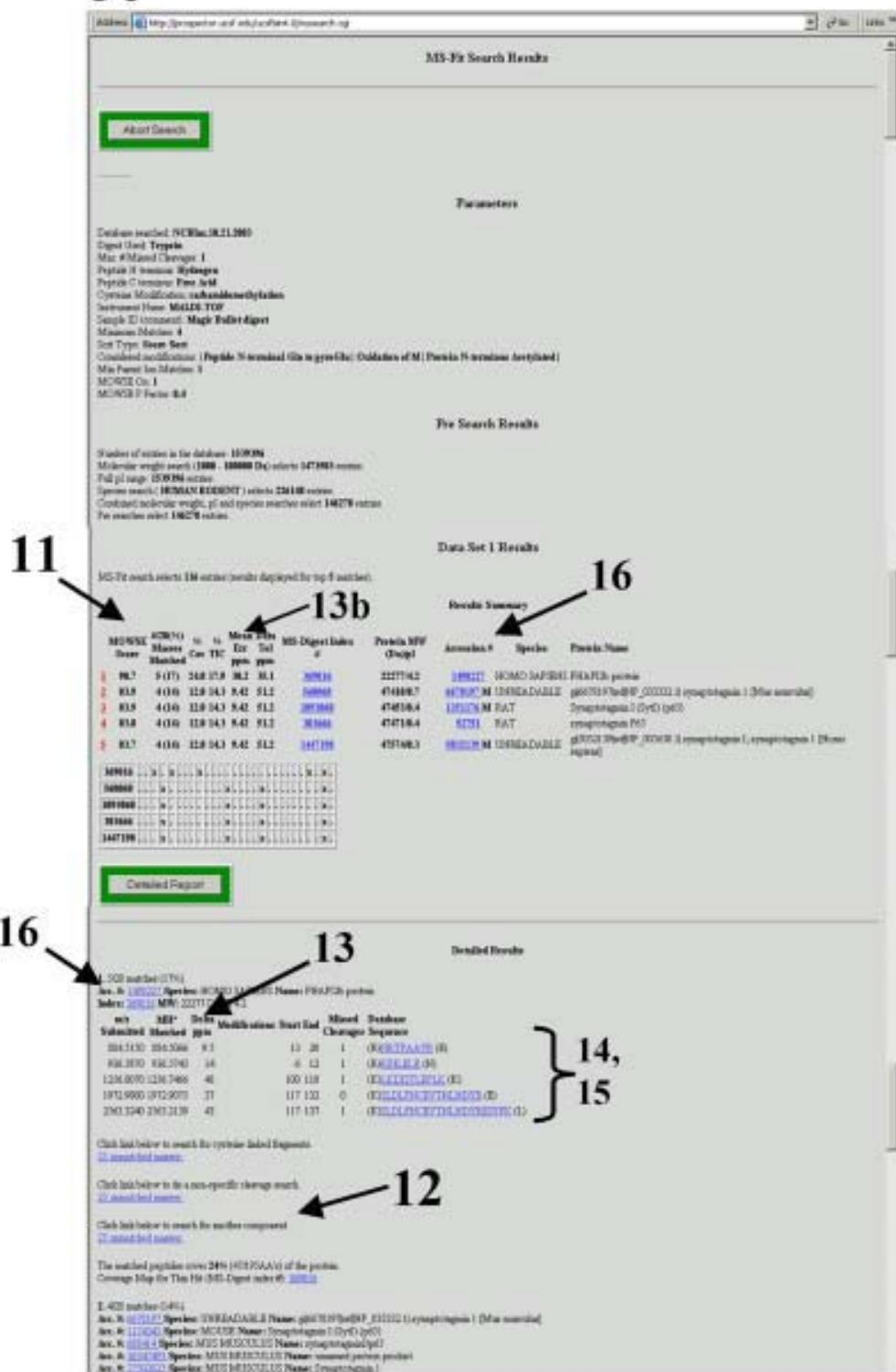


Fig. 1C3.

scores based on matching ion intensities of predicted fragment ions to peaks in the experimental spectrum, then performs cross-correlations of theoretical spectra of the top N scoring peptides against the input spectrum. Expected fragmentation patterns can be predicted from sequence and then compared to the spectrum. Other commercial software includes proteinLynx (<http://www.waters.com/WatersDivision/>). Publicly available search engines are listed at the end.

2. The search form pages, listed under References to Search Engines, include a similar range of options; therefore, the recommended parameters that are inputted are largely independent of which search engines are used.
3. The specificity of trypsin is cleavage at the C-terminus of lysine and arginine peptide bonds (3). This protease does not normally cut Lys- or Arg-Pro bonds, which is allowed for in the algorithm, but cleavage at valine and isoleucine bonds may also be poor, and since it is an endopeptidase, trypsin will not further cleave consecutive basic residues from the amino- or the carboxy- amino terminus once the initial cut has been made; e.g., in the example shown, the sequence –LKKLSDNR– has been cleaved to produce the peptide KLSDNR, and trypsin did not further cleave the lysine in the new amino-terminal KL peptide bond. Despite the fact that the protein will have been denatured during SDS-PAGE, since it is being digested while held in a gel matrix, this usually means that digestion may not go to completion (unlike that of a protein in solution). Clearly, while setting the “missed cleavages” parameter to 0 will lessen the chance of a false hit, there are almost always some, or many, incompletely cleaved peptides in any such digest, which may mean that there may very well be insufficient numbers of peptides to produce a genuine match.
4. This is dictated by the procedure used for reduction and alkylation, if any. Modification of cysteine residues is strongly recommended. However, if no cysteine modification has been carried out and if the protein originates from a gel sample, then much of this residue will have been converted to acrylamide-modified Cys.
5. “Deisotoping” software is available on mass spectrometers to artificially remove the ^{13}C peaks arising from identical peptides. This simplifies the spectrum, but more importantly this will ensure that the search algorithm will not be confused and look for two or more distinct peptides that each differ by 1Da. This is particularly valuable when analyzing peptide mixtures, since overlapping isotope clusters are thus identified correctly and only the genuine ^{12}C peaks are reported. If the resolution of the mass spectrum is not sufficient to resolve individual isotope peaks, then the average mass is often reported. This is still the case with larger polypeptides and proteins, but in modern instruments, depending on the resolution obtained, the peptide forms containing all ^{12}C , one ^{13}C , two ^{13}C , and so on, may be resolved up to approx 20,000 Da.
6. Although the search will be refined by limiting to a particular mass range of the intact protein, the possibility of subunits or fragments must be considered. Some information on the isoelectric point of a protein will also be known for a 2-D gel sample, but this should also be treated cautiously.
7. The MOWSE scoring algorithm for Mascot is described in (5), and details can be found by clicking the “HELP” button and “scoring schemes.” The final score is the absolute probability that the observed match is a random event. The scores cover a very wide range of magnitude, and a “high” score is a “low” probability, which can be confusing. Therefore, scores are reported as $-10\log_{10}(P)$, where P is the absolute probability. Thus, a probability of 10^{-20} becomes a score of 200. The results page for a typical peptide mass fingerprint search includes a histogram with the “probability based Mowse score” (see **Fig. 1B**). This example reports that “Scores greater than 61 are significant ($p < 0.05$).” The score reported by MS-Fit is also based on this MOWSE scoring system.

References

1. Gevaert, K., and Vandekerckhove, J. (2000) Protein identification methods in proteomics. *Electrophoresis*. **21**, 1145–1154.
2. Muller, M., Gras, R., Appel, R. D., Bienvenut, W. V., and Hochstrasser, D. F. (2002) Visualization and analysis of molecular scanner peptide mass spectra. *J. Am. Soc. Mass Spectrom.* **13**, 221–31.
3. Aitken, A., Geisow, M. J., Findlay, J. B. C., Holmes, C., and Yarwood, A. (1989). Peptide preparation and characterization. In Geisow, M. J., and Findlay, J. B. C. (eds), *Protein Sequencing: A Practical Approach*, IRL Press, Oxford, pp. 43–68.
4. Dubois, T., Zemlickova, E., Howell, S., and Aitken, A. (2003) Centaurin- α_1 associates in vitro and in vivo with nucleolin. *Biochem. Biophys. Res. Commun.* **301**, 502–508.
5. Perkins, D. N., Pappin, D. J., Creasy, D. M., and Cottrell, J. S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551–3567.

Suggested Reading: Databases

EMBL—EMBL Nucleotide sequence db (EBI), at <http://www.ebi.ac.uk/embl/>

Genbank—GenBank Nucleotide Sequence db (NCBI), at <http://www.ncbi.nlm.nih.gov/Genbank/GenbankSearch.html>

DDBJ—DNA Data Bank of Japan, at <http://www.ddbj.nig.ac.jp/>

dbEST—dbEST (Expressed Sequence Tags) db (NCBI), at <http://www.ncbi.nlm.nih.gov/dbEST/>

Suggested Reading: Search Engines

The ExPASy (Expert Protein Analysis System) server of the Swiss Institute of Bioinformatics (SIB) contains a large suite of programs for the analysis of protein sequences, structures, and proteomics, as well as 2-D PAGE analysis (2-D gel documentation and 2-D gel image analysis programs). The ExPASy suite of programs, including Swiss-Prot, is at <http://www.expasy.ch/> and <http://us.expasy.org/tools/>

These include:

AACompIdent, which identifies a protein by its amino acid composition.

AACompSim, which compares the amino acid composition of a Swiss-Prot entry with all other entries.

MultiIdent, which identifies proteins with pI, Mw, amino acid composition, sequence tag, and peptide mass fingerprinting data. One or more species and a Swiss-Prot keyword can also be specified for the search. MultiIdent is at <http://ca.expasy.org/tools/multiident/>.

PeptIdent, which identifies proteins with peptide mass fingerprinting data, pI, and Mw. Experimentally measured, user-specified peptide masses are compared with the theoretical peptides calculated for all proteins in Swiss-Prot, making extensive use of database annotations.

TagIdent, which identifies proteins with pI, Mw, and sequence tag, or generates a list of proteins close to a given pI and molecular weight.

PepMAPPER is a peptide mass fingerprinting tool from UMIST, UK.

Mascot (see **Fig. 1**) is a peptide mass fingerprint, sequence query, and MS/MS ion search from Matrix Science, Ltd., London. This is an extension of the original “MOWSE” search engine. Mascot is available from Matrix Science (<http://www.matrixscience.com/home.html>). http://www.matrixscience.com/search_form_select.html.

ProteinProspector (see **Fig. 1**) includes a variety of tools from UCSF (MS-Fit, MS-Tag, MS-Digest, and so on) for mining sequence databases in conjunction with mass spectrometry experiments. There are mirror sites at UCL-Ludwig, UK, and Ludwig Institute, Melbourne (Australia). Protein prospector is at (<http://128.40.158.151/mshome3.4.htm>). PepSea—Protein identification by peptide mapping or peptide sequencing, from Protana, Denmark.

PeptideSearch—Peptide mass fingerprint tool, from EMBL, Heidelberg.

PROWL—Protein chemistry and mass spectrometry resource, from Rockefeller and New York Universities.

Analysis of the Proteomes in Human Tissues by In-Gel Isoelectric Focusing and Mass Spectrometry

Francesco Giorgianni and Sarka Beranova-Giorgianni

1. Introduction

The mammalian proteomes are inherently complex, and therefore a global analysis of these proteomes presents a great technical challenge. It is recognized that there is no single method today that is able to probe an entire human proteome, and that a combination of several methodological approaches provides the most flexible strategy (1). To date, the most widely used methodology for proteome analysis involves a combination of two-dimensional gel electrophoresis (2-DE), mass spectrometry, and bioinformatics tools. Although this “classical” approach has been successfully applied in many studies, it suffers from a number of limitations, such as incomplete proteome coverage, low throughput, and so on (2). In order to overcome these limitations, a number of modifications of the 2-DE-based approach have been introduced, such as new strategies for sample pre-fractionation, introduction of medium- and narrow-range immobilized pH gradient (IPG) strips, development of fluorescent protein stains and differential gel electrophoresis, and design of new robotics systems for protein processing. In addition to modifications in the 2-DE-based proteomics, new methodologies were developed that utilize only one of the dimensions of 2-DE, i.e., isoelectric focusing (IEF) or sodium dodecyl sulfate (SDS)-polyacrylamide gel electrophoresis (PAGE). For example, several strategies that combine in-gel IEF with mass spectrometry (MS) were described, such as direct mass spectrometric characterization of intact proteins in IEF gels (3,4), or identification of proteins in selected regions of IEF gels via proteolytic digestion and mass spectrometry (5–7). Furthermore, in-solution IEF has been also been used for protein fractionation, and new devices for liquid IEF have been designed (8–12).

In this chapter, an in-gel IEF-liquid chromatography (LC)-MS/MS strategy suitable for the mapping of human tissue proteomes is described (7,13). This method combines isoelectric focusing in IPG strips with mass spectrometry and bioinformatics. The principle of the in-gel IEF-LC-MS/MS methodology is outlined in **Fig. 1**. The proteins are extracted from the tissue under study, and the protein mixture is separated by in-gel IEF in a conventional IPG strip. The entire strip is cut into a set of gel sections, and the proteins in each section are digested with trypsin. The tryptic digests are subjected to LC-MS/MS to obtain MS/MS data that are diagnostic of the peptides’ sequences. The

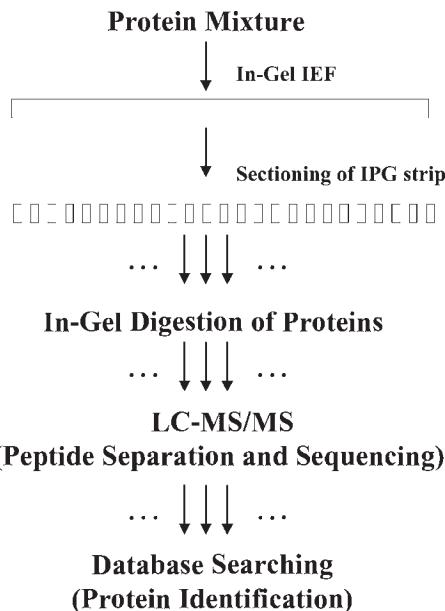


Fig. 1. General strategy for proteome mapping by in-gel IEF-LC-MS/MS. In this method, proteins extracted from the tissue of interest are resolved by IEF, using an immobilized pH gradient (IPG) strip. After in-gel IEF, the whole IPG strip is sectioned and the proteins in each gel section are digested. The digests are analyzed by LC-MS/MS, and the MS/MS data are used in database searches to identify the proteins in each gel section of the IPG strip.

MS/MS data are used to identify the proteins through searches of a protein-sequence database. Here, the method is demonstrated for the analysis of human pituitary tissue.

2. Materials

2.1. Tissue Processing and Protein Extraction

1. A postmortem human pituitary tissue specimen was analyzed. The tissue was obtained from the National Disease Research Interchange (NDRI) (14).
2. Polytron homogenizer (KINEMATICA AG, Littau-Lucerne, Switzerland).
3. Probe sonicator (Heat Systems-Ultrasonics, Farmingdale, NY).
4. Tissue homogenization buffer: acetic acid (2 M)/0.1% mercaptoethanol or TRIZOL reagent kit (Invitrogen, Carlsbad, CA).
5. EttanTM 2-D Quant protein concentration assay kit (Amersham Biosciences, Piscataway, NJ).

2.2. In-Gel IEF

1. In-gel IEF separations were carried out with a Multiphor II 2-DE system (Amersham Biosciences, Piscataway, NJ), which consisted of a Multiphor II flatbed electrophoresis unit, a MultiTemp II thermostatic circulator, and an EPS 3501 XL power supply. Additional accessories were an Immobiline DryStrip reswelling tray for the rehydration of IPG strips, and a Immobiline DryStrip kit (containing a running tray, IPG strip aligner, electrodes, and electrode paper strips).
2. IEF rehydration buffer: urea (7 M), thiourea (2 M), CHAPS detergent (2% w/v), IPG buffer (carrier ampholyte mixture; 2% v/v), dithiothreitol (0.3% w/v), and a trace of bromophenol blue dye (see Note 1). All reagents from Amersham Biosciences except thio-

- urea (Sigma, St. Louis, MO) and bromophenol blue (Sigma).
3. IPG gels: 3–10NL, 18-cm IPG strips (Amersham Biosciences, Piscataway, NJ).
 4. DryStrip cover fluid (oil) from Amersham Biosciences.

2.3. Processing of the IPG Strip and Protein Digestion

1. Sequencing-grade trypsin (Promega, Madison WI).
2. ZipTip C18 pipet tips (Millipore, Billerica, MA).
3. Gel-loading pipet tips, 100 μ L (Fisher Scientific, Pittsburgh, PA).
4. Other reagents: ammonium bicarbonate (Sigma, St. Louis, MO); acetonitrile (J. T. Baker, Phillipsburg, NJ); trifluoroacetic acid (TFA; Pierce, Rockford, IL); heptafluorobutyric acid (Sigma).

2.4. LC-MS/MS

1. LC-MS/MS measurements were performed with an LCQ Deca nanoelectrospray-quadrupole ion trap mass spectrometer (ThermoFinnigan, San Jose, CA).
2. LC column: PicoFritTM column (360 μ m o.d., 75 μ m i.d., 15 μ m tip i.d.) from New Objective (Woburn, MA) packed with 9–10 cm of C18 silica-based reversed-phase packing material (5 μ m, 200 \AA MAGIC C18) from Michrom Bioresources (Auburn, CA). Columns were packed under nitrogen gas pressure using a homemade pressure-vessel (see Note 2).
3. LC solvents: water (J. T. Baker, Phillipsburg, NJ); acetonitrile (J. T. Baker, Phillipsburg, NJ); formic acid (EM Science–EMD Chemicals, Gibbstown, NJ).
4. Mobile phases for LC: A = water/0.1% formic acid; B = 90% acetonitrile/10% water/0.1% formic acid (see Note 3).

2.5. Database Searches

1. The database searches were carried out with the program SEQUEST, part of the LCQ Deca software package.

3. Methods

3.1. Tissue Processing and Protein Extraction

3.1.1. Extraction With Acetic Acid/Mercaptoethanol

A number of methods for protein extraction from tissues for proteomics applications have been developed. The AcOH/mercaptoethanol method is well suited for larger tissue specimens where a sufficient amount of protein for multiple experiments can be obtained. The lyophilizate produced with this method can be stored for long periods of time, and appropriate amounts can be weighed out on a microbalance. The lyophilizate can be re-dissolved in a rehydration buffer of a chosen composition, which provides flexibility to carry out different types of experiments with the same tissue extract. It should be kept in mind that with this method, no separation of the proteins from other cellular components is performed.

1. Quickly weigh the frozen tissue specimen (typically 0.4–0.5 g of pituitary tissue) and cut it in several pieces; place the tissue pieces in a 50-mL plastic tube (see Note 4).
2. Rinse the tissue three times with 10 mL of ice-cold saline solution (the tissue will have thawed by this time).
3. Add to the tissue 10 mL of the homogenization buffer (chill buffer on ice before use); homogenize with the Polytron homogenizer power set to 6 (keep the mixture on ice during homogenization, alternate 10 s homogenization and 30 s of rest; repeat until complete disruption of the tissue is achieved).

4. After homogenization, sonicate the homogenate for 10 s (keep the homogenate on ice during sonication); repeat this step.
5. Aliquot the homogenate into 1-mL fractions; lyophilize these fractions.
6. Determine protein concentration in the lyophilizate (see Note 5).
7. Store the lyophilizate at -80°C until further use.

3.1.2. Extraction With TRIZOL Reagent

The TRIZOL-based method is well suited for large tissue specimens. With this method, cellular proteins are separated from RNA and DNA (15). Prolonged storage of the protein extract may result in problems with subsequent solubilization.

1. Quickly weigh the frozen tissue specimen and cut it in several pieces; place the pieces in a 14-mL plastic tube.
2. Rinse the tissue three times with 5 mL of ice-cold saline solution.
3. Add 4 mL of the TRIZOL reagent per gram of tissue.
4. Homogenize the mixture with the homogenizer power set to 6; (keep the mixture on ice during homogenization, alternate 10 s homogenization and 30 s of rest; repeat this step until the tissue is fully disrupted).
5. After homogenization; sonicate the sample (keep suspension on ice, alternate 10 s of sonication and 30 s of rest; repeat this step three times).
6. Vortex the sample for 4 h at 4°C.
7. Add 0.2 mL of chloroform per 1 mL of TRIZOL reagent used.
8. Centrifuge at 12,000g for 10 min. The mixture will separate into three phases: upper (aqueous) phase, interphase, and lower (organic) phase (see Note 6).
9. Remove the upper (aqueous) phase.
10. To the lower phases (interphase and organic), add 0.3 mL of ethanol per 1 mL of TRIZOL reagent used; mix by inversion (see Note 7).
11. Centrifuge at 2500g for 5 min.
12. Draw off the supernatant, and add to this supernatant a volume of acetone five times the volume of the supernatant solution; mix by inversion (see Note 8).
13. Centrifuge at 12,000g for 10 min. After centrifugation, remove and discard the supernatant.
14. Wash the protein pellet with 95% ethanol containing 0.3 M guanidine-HCl (2 mL per 1 mL of TRIZOL reagent used); repeat three times (see Note 9).
15. Vortex the protein pellet in 2 mL of ethanol, store the protein pellet in ethanol for 20 min, centrifuge at 7500g for 5 min.
16. Remove ethanol and dissolve the pellet in 2–3 mL of IEF rehydration buffer containing no dye (see Note 10).
17. Determine protein concentration (see Note 5).
18. Aliquot the sample and store at -80°C until further use.

3.2. In-Gel IEF

3.2.1. Sample Preparation for IEF

1. Weigh out a required portion of the pituitary lyophilizate (Subheading 3.1.1.) (see Note 11); or measure out the TRIZOL extract (Subheading 3.1.2.) (see Note 12).
2. Add 360 µL of IEF rehydration buffer (see Notes 12, 13).
3. Vortex the sample mixture for 1 h; at 20, 40, and 60 min, sonicate the sample for 20 s with the probe sonicator (see Note 14).
4. Centrifuge the sample at 21,000g for 30 min.
5. Take out the IPG strip(s); with a permanent marker, outline on the plastic backing the sections to be cut later (see Notes 15–17).
6. Apply 350 µL of the sample solution into a slot in the DryStrip reswelling tray.

7. Remove the protective cover from the IPG strip and place the IPG strip gel-side down on top of the sample solution (*see Note 18*).
8. Overlay the strip with 3 mL of oil to prevent evaporation.
9. Rehydrate the IPG strip overnight.

3.2.2. Isoelectric Focusing (see Note 19)

1. Set the MultiTemp circulator to 20°C.
2. Pour 4 mL of oil (cover fluid) onto the cooling plate of the Multiphor II unit; position the DryStrip running tray onto the cooling plate; connect electrode wires to appropriate positions.
3. Pour 10 mL of oil into the DryStrip tray; place IPG strip aligner on top of oil.
4. Cut two electrode paper strips to a length of 110 mm and moisten them with 0.5 mL of water (they should be just damp).
5. Lift the rehydrated IPG strip out of the reswelling tray; rinse the gel surface with water; gently sweep the strip backing over a sheet of paper to remove excess oil.
6. Place the IPG strip into a groove in the strip aligner, gel side up and in the correct orientation (*see Notes 20 and 21*).
7. Place the moist electrode paper strips across the cathodic and anodic ends of the IPG strip; the electrode paper strips must be in contact with the gel ends of the IPG strip. Position the electrodes over the electrode paper strips and press them down to come in contact with the electrode paper strips.
8. Pour 70–80 mL of oil into the tray to completely cover the IPG strip(s).
9. Perform IEF according to the following protocol: 0–100 V (gradient over 1 min); 100 V (fixed for 120 min); 100–500 V (gradient over 1 min); 500–3500 V (gradient over 90 min); 3500 V (fixed for 8 h) (*see Note 22*).
10. After completion of IEF, remove the IPG strip, wipe off excess oil, and loosely cover the IPG strip with plastic wrap; store the wrapped IPG strip at –80°C until further processing.

3.3. Processing of the IPG Strip and Protein Digestion

1. Rinse the IPG strip with 200 mM ammonium bicarbonate for 10 s.
2. Incubate the strip in 10 mL of 200 mM ammonium bicarbonate for 10 min.
3. Use a clean scalpel to separate the strip into gel sections; place each gel section into a siliconized 0.5-mL Eppendorf tube (*see Note 23*).
4. Dehydrate each section with 50 µL of acetonitrile; repeat twice (*see Note 24*).
5. Dry the gel sections in a vacuum centrifuge for 30 min.
6. Rehydrate each gel section with 100 µL of ammonium bicarbonate (50 mM) containing trypsin (20 µg/mL).
7. Incubate the samples in a water bath at 37°C overnight.
8. After digestion, centrifuge the samples at 10,000g for 1 min and transfer each supernatant into a clean 0.5-mL tube.
9. To extract the remaining peptides from each gel section, add 70 µL of acetonitrile/water/trifluoroacetic acid (50:45:5; v:v:v) and sonicate the sample in a sonicator bath for 20 min; collect the supernatant and repeat the extraction one more time; combine all extracts from each gel section.
10. Dry the samples in a vacuum centrifuge.
11. Reconstitute the samples in 15 µL of water/0.1% trifluoroacetic acid.
12. Centrifuge the samples at 10,000g for 1 min.
13. Purify the digests with ZipTip tips, following manufacturer's instructions; elute the peptides from the ZipTip with 3 µL of acetonitrile/water (50:50; v:v) (*see Note 25*).
14. Add to each eluate 3 µL of water/1% acetic acid/0.02 % heptafluorobutyric acid (*see Note 26*).
15. Store the peptide samples at –20°C until analysis.

3.4. LC-MS/MS

To obtain peptide sequence data, the digest from each section of the IPG strip is analyzed by LC-MS/MS. In this experiment, the peptide mixture is separated by reversed-phase high-performance liquid chromatography (HPLC), using a gradient of water/acetonitrile (*see Note 27*). The peptides eluting from the LC column are introduced online into the mass spectrometer and analyzed. The mass spectrometric analysis is performed in data-dependent mode, in which the instrument switches between MS and MS/MS: usually, a full-scan MS spectrum is acquired to determine peptide molecular masses, followed by acquisition of several peptide sequence-diagnostic MS/MS (product-ion) spectra; this cycle is repeated throughout the run, thus producing extensive MS/MS data for peptides in a given digest. In our experiments, the mass spectrometer cycles between acquisitions of a full-scan MS spectrum, and five MS/MS scans of the most abundant ions from the MS scan.

3.5. Database Searches

The MS/MS data are used to search a protein sequence database (Swiss-Prot or NCBInr) to identify the proteins in each section of the IPG strip. The database search outputs are examined manually to pinpoint proteins retrieved with significant scores and to remove false-positive hits (*see Note 28*).

4. Notes

1. The rehydration buffer (without bromophenol blue dye and dithiothreitol [DTT]) can be prepared in advance and frozen in aliquots. The dye (which would interfere with protein assay) and DTT should be added just before use.
2. Pre-packed columns for nanoflow LC are commercially available from New Objective (Woburn, MA).
3. High purity solvents must be used (see recommendations in the main text), and extreme care must be taken to avoid contamination. Dedicated glassware should be used for mobile phase preparation.
4. A face shield, gloves, and protective clothing should be worn during tissue processing. Rules for handling of biohazard materials and for disposal of biohazard waste must be followed.
5. Protein concentration should be determined in the IEF rehydration buffer. The protein assay used must be compatible with the components of the rehydration buffer.
6. The aqueous phase contains RNA; the interphase and the organic phase contain DNA and proteins.
7. This will precipitate the DNA.
8. This will precipitate the proteins.
9. During each wash cycle, the protein pellet should be stored in the wash solution for 20 min, then centrifuged at 7500g for 5 min.
10. Prolonged storage of the protein pellet can make subsequent solubilization difficult.
11. A standard protein load for the in-gel IEF-LC-MS/MS method is 100–250 µg.
12. The TRIZOL-extracted proteins are already dissolved in the rehydration buffer in the last step of the extraction procedure; therefore, the extract should just be diluted with additional rehydration buffer to the required protein concentration.
13. The final sample volume applied to the IPG strip (*see step 6*) is 350 µL. This volume is appropriate for 18-cm IPG strips from Amersham Biosciences. Strips of different length

- and/or from a different manufacturer may require a different volume for proper strip rehydration.
14. Care must be taken not to heat up the sample.
 15. Marking the IPG strip prior to IEF greatly facilitates subsequent sectioning of the gel (**Subheading 3.3.**)
 16. In our experiments to date, 25 gel sections were used (each section was 0.72 cm).
 17. If possible, another IEF separation of the sample (or another suitable protein mixture) should be performed in parallel, and SDS-PAGE (i.e., the second dimension of 2-DE) should be carried out with this second IPG strip. This will confirm that the IEF separation was adequate.
 18. There should be no bubbles in the sample solution that is in contact with the gel.
 19. Steps 1–4 should be completed before removal of the IPG strip from the reswelling tray.
 20. The IPG strips have markings indicating the anodic and cathodic ends.
 21. The IPG strip should not be left exposed for more than 5 min.
 22. This IEF protocol is used in our laboratory for separations of proteins from human pituitary or human prostate. Note that optimum IEF conditions are sample dependent and that the protocol may need to be adjusted to achieve optimum focusing.
 23. The dull edge of the scalpel should be placed almost flat against the strip at the appropriate mark (see **Note 15**), and the gel piece should be scraped off the plastic backing and transferred into the Eppendorf tube.
 24. Gel-loading pipet tips must be used; regular tips would clog.
 25. The purification step is needed even when LC-MS/MS is used for digest analysis because of the presence of gel residue. It is necessary to centrifuge the sample before ZipTip purification. When aspirating sample solution during ZipTip purification, the tube should be tilted and the ZipTip should not reach all the way to the bottom.
 26. We prefer to dilute the samples over drying them down (the binding of the peptides to the LC column is good despite the relatively high acetonitrile percentage). Note that the sample volume may need to be adjusted depending on the particular injection setup for LC-MS/MS.
 27. The LC conditions used in our laboratory are: (a) sample injection volume: 4 μ L; (b) gradient of solvents A and B (A = water/0.1% formic acid, B = 90 % acetonitrile/10% water/0.1% formic acid): 5 min initial isocratic elution with 0% B, followed by 50 min in a linear gradient 0–70% B, 5 min isocratic elution with 70% B, and 15 min in a linear gradient 70–0% B; (c) flow rate: 400 nL/min; (d) to minimize carryover, the injector port is washed between injections with 200 μ L of methanol/water (50:50), followed by 200 μ L of water/0.1% formic acid.
 28. The SEQUEST searches retrieve proteins that were identified based on correlations of the MS/MS data for multiple unique peptides to the protein sequence; these matches are accepted as valid identifications. In addition, the database searches identify proteins for which only a single peptide was matched. These single-peptide matches should be examined manually to confirm or to rule out the identification. In our laboratory, the following criteria are used to evaluate single-peptide matches (**16**): (a) a SEQUEST Xcorr (cross-correlation score) \geq 2.0 (doubly or triply charged peptides); (b) a good-quality MS/MS spectrum with most of the abundant product ions assigned; (c) a continuous stretch of the peptide sequence covered by either the y- or b-ion series; (d) intense y-ions corresponding to a proline residue (if Pro is predicted in the sequence); (e) approximately similar values of pI estimated from IEF and the theoretical pI (although the estimated pI of a protein can be affected by experimental error and/or by the presence of posttranslational modifications, a large discrepancy between experimental and theoretical values should be viewed with caution).

References

1. Hancock, W. S., Wu, S. L., and Shieh, P. (2002) The challenges of developing a sound proteomics strategy. *Proteomics* **2**, 352–359.
2. Beranova-Giorgianni, S. (2003) Proteome analysis by two-dimensional gel electrophoresis and mass spectrometry: strengths and limitations. *Trends Anal. Chem.* **22**, 273–281.
3. Ogorzalek Loo, R. R., Stevenson, T. I., Mitchell, C., Loo, J. A., and Andrews, P. C. (1996) Mass spectrometry of proteins directly from polyacrylamide gels. *Anal. Chem.* **68**, 1910–1917.
4. Loo, J. A., Brown, J., Critchley, G., Mitchell, C., Andrews, P. C., and Ogorzalek Loo, R. R. (1999) High sensitivity mass spectrometric methods for obtaining intact molecular weights from gel-separated proteins. *Electrophoresis* **20**, 743–748.
5. Steinberg, T. H., Chernokalskaya, E., Berggren, K., et al. (2000) Ultrasensitive fluorescence protein detection in isoelectric focusing gels using a ruthenium metal chelate stain. *Electrophoresis* **21**, 486–496.
6. Castellanos-Serra, L., Vallin, A., Proenza, W., Le Caer, J. P., and Rossier, J. (2001) An optimized procedure for detection of proteins on carrier ampholyte isoelectric focusing and immobilized pH gradient gels with imidazole and zinc salts: its application to the identification of isoelectric focusing separated isoforms by in-gel proteolysis and mass spectrometry analysis. *Electrophoresis* **22**, 1677–1685.
7. Le Bihan, T., Pinto, D., and Figgeys, D. (2001) Nanoflow gradient generator coupled with μ -LC-ESI-MS/MS for protein identification. *Anal. Chem.* **73**, 1307–1315.
8. Zuo, X., and Speicher, D. W. (2000) A method for global analysis of complex proteomes using sample prefractionation by solution isoelectrofocusing prior to two-dimensional electrophoresis. *Anal. Biochem.* **284**, 266–278.
9. Herbert, B. and Righetti, P. G. (2000) A turning point in proteome analysis: sample prefractionation via multicompartment electrolyzers with isoelectric membranes. *Electrophoresis* **21**, 3639–3648.
10. Hoffman, P., Ji, H., Moritz, R. L., et al. (2001) Continuous free-flow electrophoresis separation of cytosolic proteins from the human colon carcinoma cell line LIM 1215: a non two-dimensional gel electrophoresis-based proteome analysis strategy. *Proteomics* **1**, 807–818.
11. Ros, A., Faupel, M., Mees, H., et al. (2002) Protein purification by off-gel electrophoresis. *Proteomics* **2**, 151–156.
12. Righetti, P. G., Castagna, A., Herbert, B., Reymond, F., and Rossier, J. S. (2003) Prefractionation techniques in proteome analysis. *Proteomics* **3**, 1397–1407.
13. Giorgianni, F., Desiderio, D. M., and Beranova-Giorgianni, S. (2003) Proteome analysis using isoelectric focusing in immobilized pH gradient gels followed by mass spectrometry. *Electrophoresis* **24**, 253–259.
14. <http://www.ndriresource.org>
15. <http://www.invitrogen.com/content/sfs/manuals/15596026.pdf>
16. Link, A. J., Eng, J., Schieltz, D. M., et al. (1999) Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* **17**, 676–682.

Liquid Chromatography Coupled to MS for Proteome Analysis

Alastair Aitken

1. Introduction

The advantages of nanoLC include direct on-line injection into a mass spectrometer without the attendant sample loss associated with gel electrophoresis. Very acidic or basic proteins and membrane proteins still pose problems for two-dimensional (2-D) gels. Two-dimensional nano-LC as a complementary separation technique to 2-D gels has recently been developed. More than 1000 proteins have been identified with nano-LC in one run, and more than 10^4 peaks can be mass analyzed (1,2). Multidimensional liquid chromatography (LC)–mass spectrometry (MS) methods also have a much greater sample dynamic range than 2-D gels (3).

On-line coupling with MS, e.g., electrospray ionization (ESI)/MS, matrix-assisted laser desorption/ionization (MALDI)-time-of-flight (TOF), and so on can analyze minute samples (low fmol range) with no sample loss. The systems are designed to provide the smallest possible dead volume and high reproducibility (with micro- and nanoliter gradients). In order to achieve this, connecting tubing, including low dead-volume connections, should be kept as short as possible. Such systems consist of a micro pump for accurate and reproducible micro- and nanoflow delivery, with autosampler for automated injections. Automated sample preparation and multidimensional (e.g., 2-D, 3-D) LC can be carried out with column-switching modules. High-precision reciprocating pumps with microflow processing and stream splitting generate highly reproducible flow rates ranging from 50 nL/min to 200 μ L/min. Online fraction collection onto MALDI targets can be combined with direct LC-MS. The materials in contact with solvents and sample should be inert—i.e., PTFE, PEEK, ceramic, titanium, and fused silica. Therefore, the sample does not come in contact with stainless steel. Helium sparging devices are normally incorporated for optimal solvent degassing. This improves check valve reliability (i.e., avoids gas bubbles in the lines). This is especially important at low flow rates, and this diminishes baseline disturbances at low-wavelength ultraviolet (UV) detection. Each solvent bottle should be equipped with a shut-off valve for minimal helium consumption.

An autosampler allows fully automated sample injection from sample vials or micro-well plates for unattended routine high-throughput analytical applications controlled by the MS software (see **Note 1**). A column oven controlled by the system software can

allow for thermostatting of the microcolumn up to approx 70°C. A scanning UV-visible detector is an integral part of most systems (*see Note 2*).

A wide variety of analytical microcolumns are available, packed with a variety of reverse-phase materials (*see Note 3*). Many of the new reverse-phase packings have a C₁₈ stationary phase with virtually no silanophilic activity, resulting in superior separation of peptides with minimal band broadening. Trifluoroacetic acid (TFA) is therefore no longer required, allowing for the use of solvent additives such as formic acid with substantially lower signal suppression and higher sensitivity in MS. Guard columns are important to prolong the lifetime of the analytical microcolumns (*see Note 4*).

A wide variety of applications related to advanced microcolumn switching (such as sample preconcentration, sample clean-up, multidimensional separations, desalting, selective extraction, and so on) can be performed with the instrument configurations described below. Typical applications include sample desalting and detergent removal (*see Note 5*); concentration of minute and/or dilute samples; isolation of phosphorylated peptides from a complex protein digest (*see Note 6*); and extraction of drugs from biological fluids and affinity extractions, such as separations of isotope-coded affinity tags (ICAT) (*see Chapter 38*). A typical example of a method for separation of a complex mixture of peptides by 2-D ion-exchange and reverse-phase chromatography is given below. The separated peptides can be analyzed directly online by electrospray MS and/or collected for MALDI-TOF MS. Binding assays or immune assay by enzyme-linked immunosorbent assay (ELISA) and radioactivity measurements, and so on, can also be performed on the remaining sample collected. With a simpler setup (*see Note 7*) or if there are budget constraints, one can carry out any part of the techniques described, as the system is very flexible. Complex peptide mixtures can also be loaded onto one biphasic microcapillary column packed with strong cation exchange (SCX) and reverse-phase (RP) material for direct elution into a tandem MS.

2. Materials

1. Mass spectrometer with NanoES source (*see Note 8*).
2. Microbore or capillary high-performance liquid chromatography (HPLC) system with helium sparging device and solvent organizer; micropump; column oven; UV/VIS detector (e.g., the Ultimate nano-HPLC system from Dionex or NanoLC from Presearch) (*see Note 9*).
3. Reverse-phase columns (e.g., PepMap, Vydac) and guard columns.
4. Strong cation exchange columns, e.g., L.C. packings, Poros 10S, or Partisphere SCX, Whatman.
5. Fused-silica capillary.
6. Mobile-phase ion exchange buffers: (a) 5 mM KH₂PO₄, 5% acetonitrile (ACN) (pH 3.0), and (b) 5 mM KH₂PO₄, 5% ACN (pH 3.0), with 500 mM KCl.
7. Reverse-phase solvents: Buffer A, 0.1% aqueous formic acid (FA). Buffer B, 0.08% formic acid in water/acetonitrile (ACN, 20:80, v/v).
8. 250 and 500 mM ammonium acetate.
9. Formic acid (FA).
10. Heptafluorobutyric acid (HFBA).
11. Precolumn stream splitter, from Upchurch or made in-house from “zero-dead volume” T-piece that may add only 20 nL to the flow path. A wide variety of stream-splitter devices, T-pieces, and the capillary PEEK tubing are available from Upchurch.

12. Rheodyne valves, 10 port.
13. Loading pump for precolumn(s).
14. Robotic microfraction collector (e.g., the Probot from Dionex).
15. An automated micro-autosampler (e.g., the FAMOS micro-autosampler from Dionex).

3. Methods

3.1. Generating Micro- and Nanoliter Flow Rates With a Precolumn Stream Splitter

This allows HPLC pumps, which function more reproducibly at flow rates above, say, 10 $\mu\text{L}/\text{min}$, to deliver a typical flow of solvent in the range of 0.1–100 $\mu\text{L}/\text{min}$. These are configured with a static micromixer and a variable microsplitter to generate specific flow rates required for a particular microcolumn. Highly reproducible gradients with virtually no time delay are thus achieved. Split ratios of approx 1:1000 would be preferable to 1:100, for example. With the latter, pump adjustment is less effective in fine tuning of flow rate.

1. To obtain a flow rate in the range of 0.1–100 $\mu\text{L}/\text{min}$, connect the T-pieces with capillary tubing of different lengths and internal diameters to give the required percent of the total flow into each line.
2. Test the actual flow rate obtained over a fixed period of time by connecting the outlet of the HPLC column to a disposable micropipet (1 to 5 μL , e.g., Fisher Scientific) and with a stopwatch note time to fill the micropipet.
3. Alternatively, inject a known volume of a test solution, e.g., MRFA that is not chromatographically retained in the column, or, using high organic phase, inject a compound, e.g., peptide that is not retained under those conditions, and start data acquisition. Measure the time over which the signal due to the test compound is obtained on the TIC graph. Divide volume of sample by total time observed.
4. If necessary, adjust the flow rate by cutting an appropriate length from tubing. Typical flow rates are 4 $\mu\text{L}/\text{min}$ to 180 nL/min for 1.0 mm, 300 μm , and 75 μm i.d. columns, respectively.
5. A postcolumn stream splitter can be set up in a similar manner to collect part of the eluate from the column and to analyze the rest directly in an electrospray MS.

3.2. Micro-Column Switching for 2-D Chromatography (4,5)

A dual-gradient pumping system consists of a capillary LC pump for the first-dimension separation, and a nanoLC pump for the second-dimension separation (see **Fig. 1**). A column-switching module, equipped with micro-10-port valves and a loading pump, is employed. As well as gaining an extra dimension of chromatographic separation, this enables faster loading of sample from relatively large volumes, resulting in faster analysis (see **Note 9**). An ion-exchange or similar chromatographic step always precedes the reverse-phase step for reasons of compatibility with nanoelectrospray MS of the solvent that elutes the peptide from the latter column.

1. Reduce and alkylate the sample (1 to 400 mg of total protein) with iodoacetamide (see Chapter 30).
2. Digest the sample with trypsin for 24 h at 37°C (see Chapter 30).
3. Dilute digested sample with 0.05% TFA to stop the reaction and store at –20°C until required for analysis.

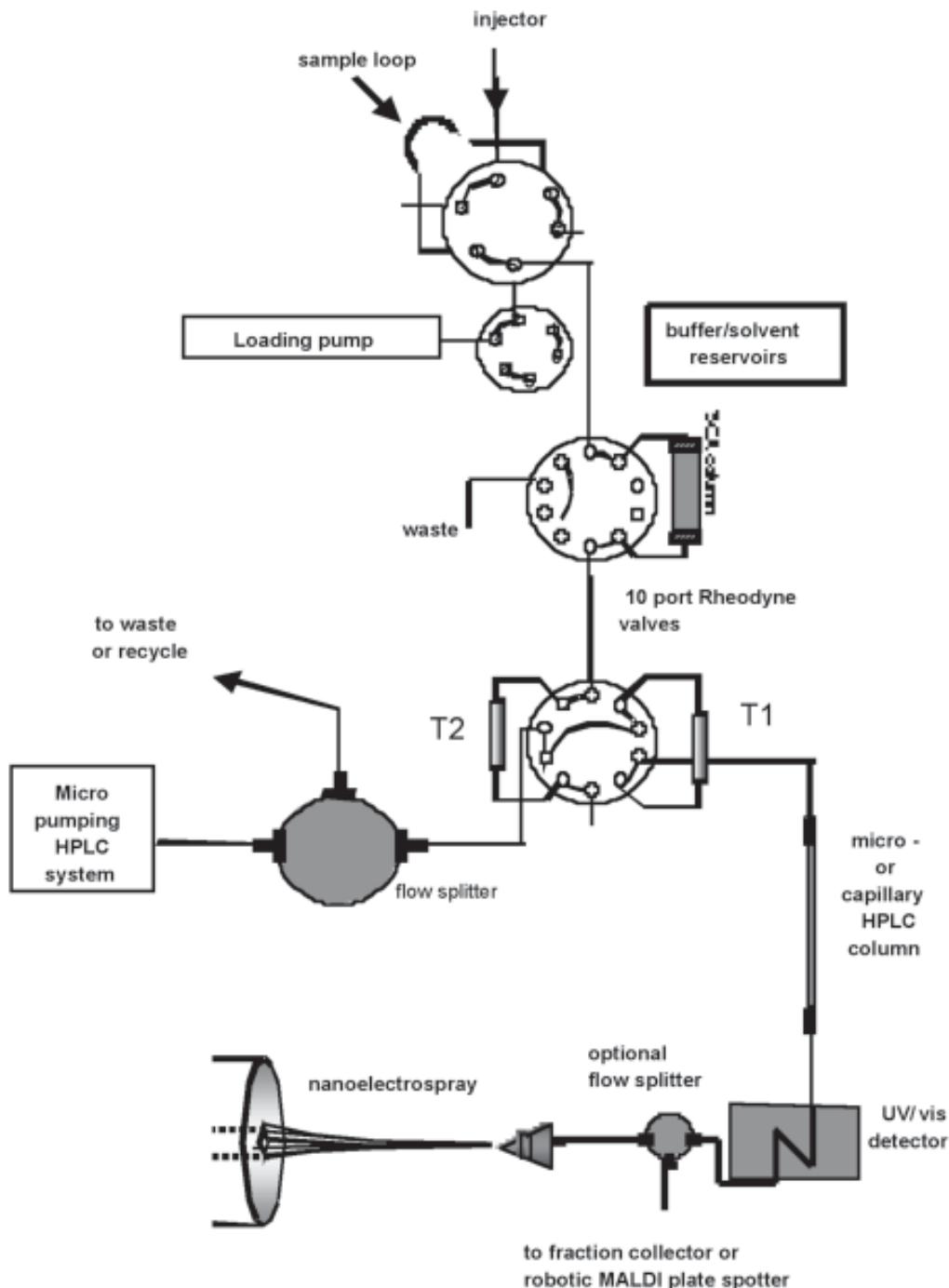


Fig. 1. Schematic of a dual gradient system, consisting of column switching module and an UltiMate dual-gradient pumping module. The drawing is schematic and the position of the lines to ports depends on the particular valve configuration. (1) The sample is loaded onto the strong cation exchange (SCX) column. (2) After washing, a step increase in salt concentration is applied (see Subheading 3.2.) and the eluted peptides are pumped onto trapping column 1 (T1). (3) After valve switching, the peptides are eluted onto the micro-analytical column, where

4. Pump sample onto a strong cation-exchange (SCX) column as a first-dimension separation.
5. Apply a gradient at 5–10 μ L/min to the SCX column.
6. Elute fractions containing peptides from the column by applying a linear salt gradient (the first dimension). Typical ion-exchange gradient conditions are: start with 0% B (100% buffer A) followed by a 5% increase in buffer B for each fraction over 65 min.
7. Using column switching, subsequently separate the component peptides in each fraction by an acetonitrile/water/formic acid gradient on the reversed-phase column (the second orthogonal separation dimension).
8. Typical reverse-phase gradient conditions are as follows. Mobile phase A comprises 0.1% FA in water. Mobile phase B is 0.08% FA in water/ACN (20:80, v/v%). The gradient is 0% B to 40% B in 80 min, then up to 90% B over 10 min and hold at that percentage for 5 min to wash the column before re-equilibration. Flow rate is 100–300 nL/min.
9. Record the peptide peaks in the UV detector at 214 nm.
10. Analyze the eluant online through a nanospray source (see Chapter 31).
11. All of each peptide peak may be analyzed or the eluant from the micro-reverse-phase column may be stream split to allow part to be analyzed by electrospray MS (typically 20%) and the rest to be fraction collected or spotted directly onto a MALDI plate using a suitable robot (see **Fig. 1** and **Subheading 3.5.**).

3.3. Peptide Identification by Continuous Nanospray MS Analysis; Use of Self-Pack and Prepacked Fused Silica Columns; Multidimensional Protein Identification Technology (MudPIT) (2)

In these techniques, a microcapillary column can be packed with one or two independent chromatography phases. Once the complex peptide mixture is loaded, no additional sample handling is required because the capillary column connects directly into the mass spectrometer (see **Note 10**). The peptides are eluted directly off the column into the tandem mass spectrometer due to the kV potential that is directly interfaced with the microcapillary column. Instead of two individual columns and multiple switching valves, a biphasic microcapillary column with sequential strong cation exchange and RP particles is used.

3.3.1. Prepacked Fused Silica Columns

These are usually prepacked with reverse-phase media for compatibility of eluting solvent with electrospray MS—e.g., PicoFrit nanobore column, 75 mm i.d., with a 5-cm packed bed of C₁₈ 5- μ m particle size material with a 15-mm integral PicoTip emitter.

1. Identify the nontapered (distal) end of the column.
2. Connect the distal end to a nanoflow pump and condition column with 50:50 organic: aqueous solvent, per manufacturers' instructions. This also "repacks" the column bed if necessary.
3. Cut column to required length by making a light score on surface to "nick" the tubing with a diamond-edged scribe.

Fig. 1. (*continued from opposite page*) the peptides are separated by a gradient of acetonitrile and analyzed by nanospray mass spectrometry (MS)MS. This is followed by re-equilibration of both this column and trap column T1. (4) Meanwhile another step increase in salt concentration elutes another batch of peptides from the SCX column onto trap column 2 (T2). (5) The peptides are eluted from T2 onto the micro-analytical column, then separated by the gradient of acetonitrile and analyzed by nanospray MSMS as before. (6) T1 is now ready to receive another batch of peptides eluted by the next increment in salt applied to the SCX column.

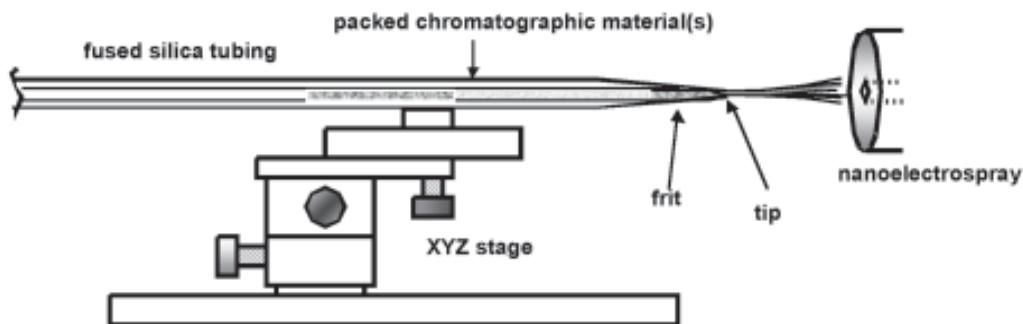


Fig. 2. Nanospray source with packed fused-silica columns.

4. Snap end off and examine at a magnification of $\times 10$ –30 to ensure cut is square and smooth. Practice first with fused silica tubing.
5. Sample is loaded onto this micro-column off-line by a micro-flow LC pump.
6. Mount column in flowing/dynamic nanospray head (see Fig. 2) and set up parameters to provide stable electrospray at flow rates of 100–500 nL/min.
7. Occasionally, trim the head of the column to remove accumulated insoluble material if necessary, for prolonged column use.

3.3.2. Use of Self-Packed Fused Silica Columns

1. Pack a polyimide-coated fused-silica microcapillary column (e.g., 360 o.d. and 75–100 μm i.d., tapered at one end) with 10 cm of 5 μm C₁₈ RP material followed by 4 cm of 5 μm particle strong cation exchange material (SCX, see Note 2).
2. Load up to 400 μg of digested protein sample into the microcapillary column, using a microbore pump.
3. Couple the column to the ion-trap mass spectrometer equipped with a nanoLC electrospray ionization source.
4. Set up an automated 15-step chromatography run as follows, to first displace the peptides from the SCX to the RP by the salt gradient, then elute these off the RP column into the MS/MS:

Buffer solutions: 5% ACN/0.02% HFBA (buffer A), 80% ACN/0.02% HFBA (buffer B), 250 mM ammonium acetate/5% ACN/0.02% HFBA (buffer C), 500 mM ammonium acetate/5% ACN/0.02% HFBA (buffer D).

The first step of 80 min is a 70-min gradient from 0 to 80% buffer B and a 10-min hold at 80% buffer B.

The next 12 steps are 110 min each with the following profile: 5 min of 100% buffer A, 2 min of x% buffer C (where “x” represents increments in C, i.e., 10, 20, 30, 40, 50, 60, 70, 80, 90, 90, and 100%, in steps 2–13); 3 min of 100% buffer A; 10 min gradient from 0 to 10% buffer B; 90 min gradient from 10 to 45% buffer B.

Step 14 is: 5 min 100% buffer A wash, 20 min 100% buffer C wash, 5 min 100% buffer A wash, 10 min gradient from 0 to 10% buffer B, and a 90-min gradient from 10 to 45% buffer B.

Step 15 is identical to step 14 except the 20-min salt wash is with 100% buffer D.

5. Re-equilibrate the microcolumn and apply an additional salt step of higher concentration to displace more peptides from the SCX to the RP.
6. Elute these peptides with the RP gradient into the mass spectrometer and repeat the process (steps 2–15).

3.4. Peak Parking for NanoLC/Nanospray/MS/MS

Peak parking is the ability to lower the gradient flow rate rapidly. This provides more time for the mass spectrometer to analyze a particular peak. For example, it can take 10 to 15 s to complete MS/MS of a peptide. If a peak is 30 s wide, the mass spectrometer will be able to analyze only the three most abundant peptides, and the less abundant peptides will be lost. Using peak parking, the flow rate can be slowed by a factor of 10 or more, allowing all of the peptides in the peak to be analyzed. This leads to a large increase in nanospray MS acquisition time to enable interrogation of more precursor ions with no effect on chromatographic separation. This is particularly important for extending the MS/MS interrogation time over peaks of interest eluting from the nanoLC column into the MS. The system requires the MS to send a signal (contact closure) at the beginning and end of peak recognition, or when the instrument switches from MS to MS/MS mode. After the initial portion of the fraction (signaled by the peak) enters the MS, the flow rate is immediately reduced (for example, by a factor of eight, from 200 nL to 25 nL/min), which results in a significant increase in the interrogation time for the remaining portion of the peak. Consequently, enhanced precursor ion selection as well as increased MS/MS acquisition times are possible with no loss in separation performance or MS sensitivity.

3.5. Robotic System for Fractionation, Spotting, and Preparation

Robots are available for online collection of fractions from capillary/nanoLC or CZE systems in volumes as small as a few nanoliters. Collection is possible on many different types of substrates—i.e., MALDI-TOF/MS targets; polyvinylidene fluoride (PVDF) membranes for subsequent protein sequencing; 96, 384, or 1536-well plates; or any other collection vessel. The most precise robots move the target table and not the needle. The needle remains in a fixed position, allowing accurate spotting of the nanoliter volumes to a precision better than 20 μ m. These robotic instruments can pipet or dispense sample and reagents. Robots can add MALDI matrices or make-up solutions with a dosage pump while spotting fractions onto MALDI targets. The matrix is added co-axially at the needle tip to generate optimal spot sizes.

4. Notes

1. A micro-autosampler allows the automated injection of volumes ranging from nL to μ L and uses a separate needle to pierce the septum of the sample vial. This allows for the protected insertion of a fused-silica capillary through the septum to the bottom of a tapered microvial. Consequently, extremely small sample volumes can be picked up from open, capped, or sealed wells, vials, or tubes, even, it is claimed, 1 μ L out of 1 μ L without sample loss. The needle is washed between injections, which results in very low carryover (less than 0.02%). The sample tray is equipped with Peltier cooling to avoid degradation of thermo-labile samples.
2. The most sensitive detectors consist of longitudinal, U- or Z-shaped capillary flow cells to achieve highest sensitivity with virtually zero dispersion. These have a long path length to enable detection of very small quantities of material. Interchangeable flow cells with volumes down to 3 nL and 10-mm path length are available for highest sensitivity. Simultaneous detection at multiple wavelengths is normally included. The enhanced optics of the “UZ-View Flow Cells” in the Dionex system allow for low noise levels comparable to those of conventional UV flow cells (20 μ AU at 254 nm). Because the entire flow cell

- consists of a single piece of fused-silica capillary, the extremely small cell volume results in negligible dead volume, which maintains excellent chromatographic resolution. In contrast to on-column detection transversely through a fused-silica capillary “window” with its very short path length determined by the i.d. of the fused silica capillary, the bending of the capillary into a U or Z shape in a longitudinal alignment results in path lengths of 8–10 mm for micro- or capillary LC, and thus a large increase in sensitivity.
3. For TFA-free LC/MS, PepMap is an excellent choice for LC/MS. Novel silica-based PepMap (LC Packings) is available in microcolumns with inner diameters from 75 μm up to 1.0 mm in different lengths (5, 15, and 25 cm) and configurations, to permit easy connection with MS. The outlet fused-silica capillary is 280 μm o.d. or 1/16" PEEK tubing for 1-mm i.d. columns. “Low TFA,” Vydac reverse-phase columns are also highly suitable. Disposable-cartridge precolumns consist of a cartridge holder and a set of disposable cartridges for precolumn concentration of diluted samples, or are used in various column-switching techniques. The resultant substantially shorter loading times for large-volume injections leads to greatly reduced overall analysis times.
 4. Guard columns prolong the life of micro- and capillary columns and can be configured to ensure virtually no dispersion. The guard column can be installed directly on top of the analytical micro-column to trap particles and absorb impurities that are present in the mobile phase or originate from the sample. The guard column should be replaced periodically for optimal protection. Guard columns packed with the same stationary phase as in the microcolumns may be used to provide maximum compatibility.
 5. Sodium dodecyl sulfate (SDS) is sometimes used to solubilize proteins and is often present in proteins and peptides electroeluted from SDS-polyacrylamide gel electrophoresis (PAGE) gels. The SDS in the sample often ruins subsequent runs by reverse-phase HPLC. A major application of multidimensional LCMS is automated online SDS removal, which can be achieved in a number of ways. Specific trapping columns for SDS and other detergents are available. These include strong anion exchange (SAX) SDS-trapping columns. Alternatively, SDS can also be removed by hydrophilic-interaction chromatography (HILIC) using polyhydroxyethyl aspartamide columns. In the HILIC mode, SDS and Coomassie blue elute immediately after the void volume. Peptides and proteins are retained and can be eluted with a decreasing gradient of acetonitrile for peptide separation, or propanol for protein separation. This method also eliminates neutral surfactants such as Triton X-100 and Nonidet P-40. SDS guard cartridges, which selectively remove SDS from peptide mixtures, may also be used. Accumulated SDS is washed off at concentrations of acetonitrile above 70%.
 6. Phosphorylated peptides in a complex digest of proteins from tissue can be enriched by employing IMAC packing material (*see* Chapter 44) in the first column, followed by elution into the reverse-phase microcolumn as described in **Subheading 3.2**.
 7. One-dimensional nanoLC systems with peak parking capability have recently been launched by Presearch. These are stated to produce stable and reproducible flow rates from 20 to 1000 nL/min without stream splitting.
 8. NanoES sources are manufactured by a number of companies for various electrospray mass spectrometers (*see* Chapter 31).
 9. Dual gradient capillary/nanoLC set-ups have been developed that allow for parallel capillary/nanoLC separations. In this method, the throughput of the MS can be increased substantially by performing overlapped injections (*see* **Fig. 1**). A typical nanoLC separation consists of a 30-min gradient, with 10 to 15 min for equilibration and 0 to 5 min for sample loading. The peptides are eluted only during the course of the gradient, usually 10–50% acetonitrile. Typically, therefore, of the total analysis time, only 40–50% is relevant for the identification of peptides by MS. Using a dual gradient system, the loading

and equilibration time can be used for running a second sample, and the MS acquisition is used only during the relevant part of the sample separation. Consequently, the MS throughput can increase by 30–40%. In a parallel capillary/nanoLC set-up, the solvent gradient on pump 1 runs simultaneously with the column wash and equilibration step on pump 2. Samples are injected alternately on trap column 1 then trap column 2. To optimize data management and avoid redundant MS data, the MS data acquisition can be started, for example, 20 min after the sample injection.

10. The nanospray ionization of eluate directly from the column outlet minimizes post-column losses and increases sensitivity. Improved chromatographic performance may also be achieved by eliminating post-column band broadening. Clearly, no UV detection is possible, but this need not be a great disadvantage.

References

1. Shen, Y. and Smith, R. D. (2002) Proteomics based on high-efficiency capillary separations. *Electrophoresis* **23**, 3106–3124.
2. Wagner, Y., Sickmann, A., Meyer, H. E., and Daum, G. (2003) Multidimensional nano-HPLC for analysis of protein complexes. *J. Am. Soc. Mass Spectrom.* **14**, 1003–1011.
3. Aebersold, R., and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature* **422**, 198–207.
4. Washburn, M. P., Wolters, D., and Yates III, J. R. (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19**, 242–247.
5. Bodnar, W. M., Blackburn, R. K., Krise, J. M., and Moseley, M. A. (2003) Exploiting the complementary nature of LC/MALDI/MS/MS and LC/ESI/MS/MS for increased proteome coverage. *J. Am. Soc. Mass Spectrom.* **14**, 971–979.

Quantitative Analysis of Proteomes and Subproteomes by Isotope-Coded Affinity Tag and Solid-Phase Glycoprotein Capture

Eugene Yi, Hui Zhang, Kelly Cooke, Ruedi Aebersold, and David R. Goodlett

1. Introduction

Chemical probes for isolating specific subsets of a proteome in conjunction with mass spectrometry have had a profound influence on quantitative analysis of complex protein mixtures. Because of the dynamic range of protein abundance, comprehensive profiling of complex proteomes has been an exceedingly challenging analytical problem. However, selective isolation of a subset of proteins (i.e., a protein class) from a proteome via chemistries selective for moieties such as phosphates or sulfhydryls substantially reduces the sample complexity by one or two orders of magnitude and enriches a subclass of the proteome prior to mass spectrometric analysis (1). In this chapter, two commonly used chemical probes for selective isolation of cysteine-containing and *N*-linked carbohydrate-containing peptides for the quantitative analysis of a proteome are described (2–5). The first is based on stable isotope affinity tagging of the cysteine residues in a protein; i.e., the original isotope-coded affinity tag (ICAT) method. The second method uses specific chemical probes that selectively isolate *N*-glycosylated proteins (i.e., the glycopeptide capture method) and subsequently labels the amino groups with light (d0, contains no deuteriums) or heavy (d4, contains four deuteriums) forms of succinic anhydride for quantitative measurement.

The ICAT approach uses a reagent with a biotin affinity group, an acid-cleavable linker, a sulfhydryl-specific iodoacetic moiety, and a mass tag linker that carries nine ¹²C (light reagent) or ¹³C (heavy reagent) atoms (Fig. 1). Proteins isolated from two different cell states are labeled with light and heavy reagents in the presence of denaturing and reducing agents. The two separate heavy and light alkylated protein mixtures are combined and digested with trypsin. Peptide mixtures are then simplified by fractionation using strong cation exchange chromatography followed by selective isolation of cysteine-containing peptides by avidin affinity chromatography. Subsequently, the biotin affinity tag connected to the rest of the ICAT reagent via an acid-cleavable linkage is removed by an acid treatment. Peptide mixtures are then subjected to analysis by microcapillary reversed-phase high-performance liquid chromatography (HPLC) (RP-microLC). Tandem mass spectrometry (MS/MS) data are searched against protein-sequence databases to identify proteins. Furthermore, the relative abundance of

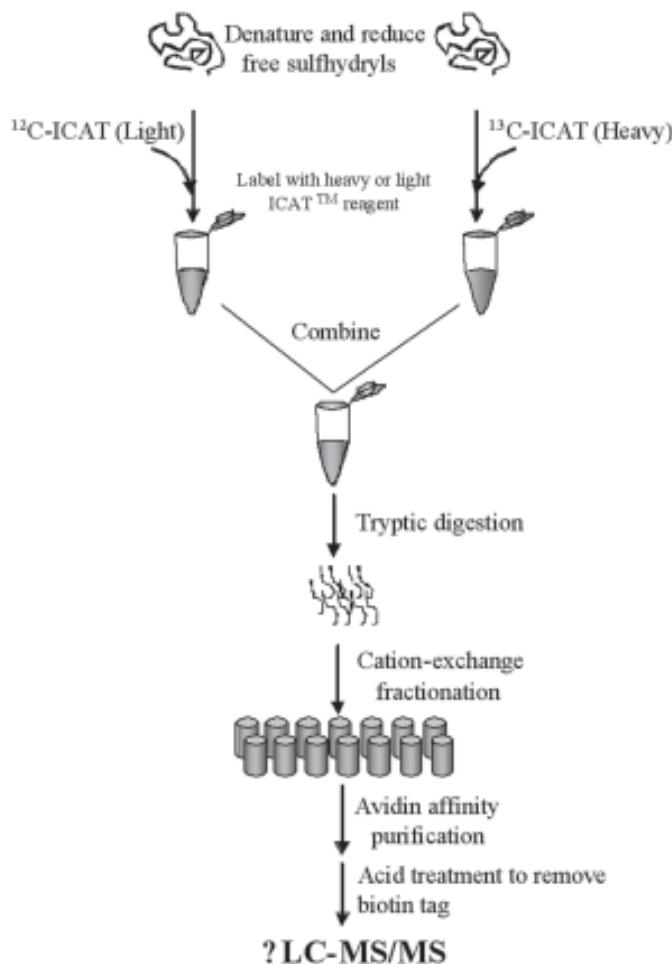


Fig. 1. Flow chart of the quantitative analysis of protein profiles with cleavable isotope-coded affinity tag (ICAT) reagent.

identified proteins is measured by comparing the mass-spectrometric signal intensities of the ICAT-labeled peptide pairs, which contain mass differences of 9 m/z (singly charged peptides), 4.5 m/z (doubly charged peptides), or 3 m/z (triply charged peptides), respectively.

The second method reduces proteome complexity by segregation of the *N*-linked glycoproteins. Typically, carbohydrates are linked to serine or threonine residues (*O*-linked glycosylation) or to asparagine residues (*N*-linked glycosylation) (1). *N*-linked glycosylation sites generally fall into the N-X-S/T sequence motif, in which X denotes any amino acid except proline (2). Protein glycosylation, and in particular *N*-linked glycosylation, is prevalent in proteins destined for extracellular environments (3). These include proteins on the extracellular side of the plasma membrane, secreted proteins, and proteins contained in body fluids. The method for quantitative glycoprotein profiling described here is based on the conjugation of glycoproteins to a solid support using hydrazide chemistry, stable isotope labeling of glycopeptides, and the specific release of formerly *N*-linked glycosylated peptides via peptide-*N*-glycosidase

F (PNGase F) (4). The recovered peptides are then identified by MS/MS and quantified in a manner identical to that of the ICAT method.

2. Materials

2.1. ICAT

1. Cleavable ICAT reagents (heavy and light) (Applied Biosystems, Foster City, CA).
2. Labeling buffer: 0.05% sodium dodecyl sulfate (SDS), 200 mM Tris-HCl (pH 8.3), 5 mM ethylenediamine tetraacetic acid (EDTA), 6 M urea.
3. Bovine IgG.
4. Tris(2-carboxyethyl) phosphine (TCEP).
5. Bio-Rad Protein Assay (Bio-Rad Laboratories).
6. TPCK-treated sequencing grade trypsin (Promega).
7. SDS-polyacrylamide gel electrophoresis (PAGE).
8. Milli-Q water or equivalent.
9. Sonicating water bath.
10. 1.5-mL Eppendorf tubes.
11. Tube Rocker (cat. no. R4185-10, Baxter Scientific Products).
12. HPLC-grade methanol.
13. PolysulfoethylATM column (PN 202SE0503, PolyLC, Inc) or ICAT Cation Exchange Cartridge (4326695, 5 pack, Applied Biosystems).
14. ICAT Avidin cartridge (4326694, Applied Biosystems).
15. Cartridge holder (4326688, Applied Biosystems).
16. Ring stand for mounting the cartridge.
17. SpeedVac.
18. Cation-exchange buffer A: 5 mM potassium phosphate, 25% acetonitrile (pH 3.0), store at room temperature (RT).
19. Cation-exchange buffer B: 5 mM potassium phosphate, 350–600 mM KCl, 25% acetonitrile (pH 3.0), store at RT.
20. 1000- μ L gas-tight syringe (81316, Hamilton).
21. Loading buffer: 2X phosphate buffered saline (PBS), pH 7.2.
22. Wash 1 buffer: 1X PBS, pH 7.2.
23. Wash 2 buffer: 50 mM ammonium bicarbonate (pH 8.3), 20% methanol.
24. Elution buffer: 0.4% trifluoroacetic acid (TFA), 30% acetonitrile.
25. Storage buffer: 1X PBS, 0.1% sodium azide.
26. 1-mL glass vial with polyethylene snap cap (WAT025054, Waters).
27. 2-mL Micro Tube (72.694.006, Sarstedt).
28. Cleaving reagent A and B (provided by Applied Biosystems).

2.2. Glycopeptide Capturing

1. Coupling buffer: 100 mM NaOAc, 150 mM NaCl (pH 5.5).
2. Sodium periodate (153-6055, Bio-Rad Laboratories).
3. Tube rocker (Cat. No. R4185-10, Baxter Scientific Products).
4. Affi-Gel Hz Hydrazide Gel (153-6047, Bio-Rad Laboratories).
5. Denaturing buffer: 8 M urea, 0.4 M NH₄HCO₃.
6. 80% Acetonitrile.
7. Methanol.
8. 100 mM Ammonium bicarbonate.
9. Milli Q water or equivalent.
6. Peptide-*N*-glycosidase F (P0705S, New England Biolabs).
7. Desalting column (732-2010, Bio-Rad Laboratories).

3. Methods

3.1. Labeling Proteins With ICAT Reagent

Typically, a protein solution will need to be cleaned up (to remove salts and detergents) by a trichloroacetic acid (TCA) precipitation or other compatible clean-up methods. After the salts and detergents have been removed, the sample is dissolved in a minimal volume of labeling buffer. The volume of the labeling buffer and the stoichiometric ratio of ICAT reagent can be adjusted proportionally to the amounts of sample to achieve approx 3 mM ICAT concentration. The following protocol is based on labeling 500 µg of protein with each of the two ICAT reagents. If an amount other than 500 µg is used, then the volume of the labeling buffer and the stoichiometric ratio of ICAT reagent can be adjusted proportionally to the amounts used here.

1. Dissolve the dried or precipitated protein from control and test samples in a minimal volume of freshly prepared labeling buffer (*see Note 1*).
2. Measure the amount of protein in the control and test samples separately. Save an aliquot (approx 1 µg) from each sample to monitor the labeling efficiency by SDS-PAGE at the end of labeling (*see Note 2*).
3. Add Tris(2-carboxyethyl) phosphine (TCEP) to a 5-mM final concentration in the protein sample solutions.
4. Add fivefold molar excess of ICAT reagent (*see Note 3*) to achieve a final concentration of 2.5–3.0 mM in the protein sample solution. Mix well and gently shake the suspension using a tube rocker for 2 h at 37°C in dark.
5. Add fivefold molar excess of dithiothreitol (DTT) to quench the reactions and incubate for 5 min at room temperature. Save an aliquot (approx 1 µg) of proteins from each sample to monitor labeling efficiency (*see Note 2*).
6. Combine the control and the test labeled samples, and dilute combined sample with 50 mM Tris-HCl (pH 8.3) buffer until the urea is at final concentration of 1 M.
7. Add trypsin 1:100 w/w (trypsin/protein) and incubate the mixture overnight at 37°C.
8. Analyze ICAT labeling efficiency and tryptic digestion by SDS-PAGE followed by silver staining with the saved aliquots from **steps 2 and 5**.

3.2. Sample Clean-Up and Fractionation by Strong Cation-Exchange Chromatography

Purification of the combined sample using the strong cation-exchange chromatography method is sample amount and complexity dependent. For less complex samples or where limited protein is available (e.g., approx 100 µg of total protein), use of the cation-exchange cartridge found in the ICAT Reagent Kit (Applied Biosystems) to prepare a single fraction for avidin affinity purification should be considered. For highly complex samples (i.e., 0.5–1 mg of total protein), however, fractionation by strong cation exchange (SCX) into 25–50 fractions is recommended.

This section describes the sample fractionation by the SCX chromatography system.

1. Acidify the sample with diluted phosphoric acid to pH 3.0 and load onto the SCX column (*see Note 4*).
2. Elute ICAT-labeled peptides off of the SCX column using a linear gradient of KCl from 0 to 100% B in 60 min and collect fractions into a glass vial (placed in a 2-mL Micro Tube) at a suitable interval.

3.3. Affinity Purification of Peptides With Avidin Affinity Cartridge

1. Neutralize each collected cation exchange fraction to pH 7.2 with 500 μ L loading buffer (see Note 5).
2. Slowly load (approx 1 drop/s) the neutralized fraction onto the avidin cartridge and collect and save the flow-through into a glass vial.
3. Add an additional 500 μ L loading buffer and collect the flow-through into the same tube (see Note 6).
4. Inject 1 mL of wash 1 buffer. Discard the output.
5. Inject 1 mL of wash 2 buffer. Discard the output.
6. Elute the labeled peptides (approx 1 drop/second) with 800 μ L of the elution buffer and collect into a glass vial.
7. Repeat steps 1–6 for the rest of the cation-exchange fractions.

3.4. Cleavage of the Biotin Affinity Tag From the ICAT-Labeled Peptides

The cleavage procedure described in this section is based on the protocol provided by Applied Biosystems, Foster City, CA.

1. Dry each fraction completely with a SpeedVac (see Note 7).
2. Prepare the cleaving reagent in a glass vial by combining 95 μ L of cleaving reagent A and 5 μ L cleaving reagent B (see Note 8).
3. Vortex to mix, and transfer the cleaving reagent to the dried fraction in the glass vials.
4. Cap each glass vial and incubate for 2 h at 37°C.
5. Dry the samples completely with the SpeedVac.
6. Add the appropriate solvent to the dried pellet for MS analysis.
7. Vortex, centrifuge, carefully transfer to the appropriate tube, and store at –80°C until ready for microLC-MS/MS.

3.5. Isolation and Analysis of Glycopeptides

The glycopeptide capture method is schematically illustrated in [Fig. 2](#). The method includes the following steps:

1. Glycoprotein oxidation: oxidation with sodium periodate converts the cis-diol groups of carbohydrates to aldehydes ([Fig. 2](#)).
2. Coupling: the aldehydes react with hydrazide groups immobilized on a solid support to form covalent hydrazone bonds ([Fig. 3](#)). Nonglycosylated proteins are removed.
3. Proteolysis: the immobilized glycoproteins are proteolyzed with trypsin on the solid support. The nonglycosylated peptides are removed by washing, and the glycosylated peptides remain covalently bound to the solid support.
4. Isotope labeling: the α -amino groups of the immobilized glycopeptides are labeled with isotopically light (d0) or heavy (d4) forms of succinic anhydride after the ϵ -amino groups of lysine are converted to homoarginines.
5. Release: formerly *N*-linked glycopeptides are released from the solid-phase by PNGase F treatment.
6. Analysis: the isolated peptides are identified and quantified using microcapillary high-performance liquid chromatography electrospray ionization tandem mass spectrometry (microLC-ESI-MS/MS) or microLC separation followed by matrix-assisted laser desorption/ionization (MALDI) MS/MS.

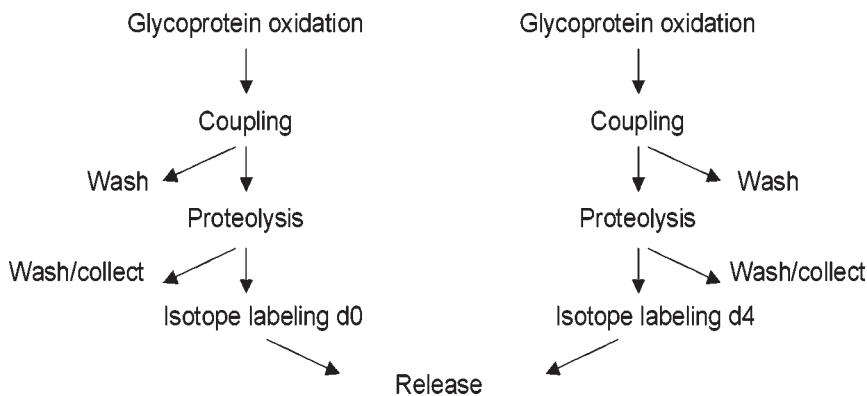


Fig. 2. Schematic diagram of an exemplary method of identifying and quantifying glycopolypeptides/glycoproteins and for determining quantitative changes in the glycosylation state of proteins.

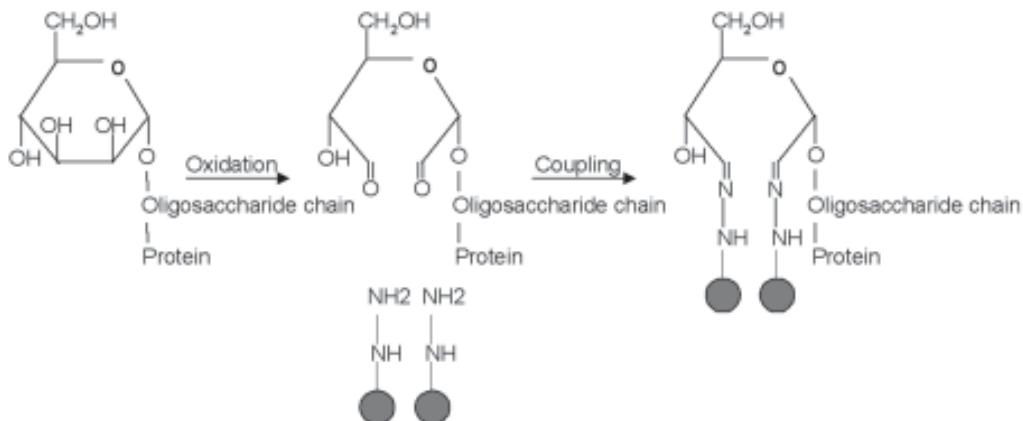


Fig. 3. Oxidation of a carbohydrate to an aldehyde followed by covalent coupling to hydrazide beads.

3.5.1. Isolation of Glycopeptides

1. Suspend 1 mg of protein in 100 μ L of coupling buffer containing 100 mM NaOAc, 150 mM NaCl (pH 5.5) (see Note 9).
2. Add sodium periodate solution to a 15-mM final concentration in the sample solutions. Mix well and gently shake the suspension using a tube rocker for 60 min at room temperature in the dark.
2. Remove excess sodium periodate from the sample using a desalting column.
3. Add hydrazide resin (Bio-Rad; Hercules, CA) equilibrated in coupling buffer to the sample (1 mL gel/5 mg protein). Mix well and gently shake the suspension using a tube rocker for 10–24 h at room temperature in dark.
4. Spin down the resin at 1000g for 10 min, and wash the resin to remove nonglycoproteins three times with 1 mL of 8 M urea/0.4 M NH_4HCO_3 .
5. Suspend the proteins on the resin in 2 M urea/0.1 M NH_4HCO_3 , add trypsin 1:100 w/w (trypsin/protein), and incubate the mixture overnight at 37°C.

6. The peptides can be reduced by adding 8 mM TCEP (Pierce, Rockford IL) at room temperature for 30 min, and alkylated by adding 10 mM iodoacetamide at room temperature for 30 min after reduction.
7. Remove unbound trypsinized peptides (see **Note 10**).
8. Wash the peptides on the resin with 1 mL of 1.5 M NaCl three times, 80% acetonitrile three times, 100% methanol three times, and 0.1 M NH₄HCO₃ six times (see **Note 11**).
9. Suspend resin with 0.5 mL of 0.1 M NH₄HCO₃, and release *N*-linked glycopeptides from the resin by incubating with 0.5 μ L of peptide-*N*-glycosidase F at 37°C overnight.
10. Collect the solution from the beads to a clean glass tube, wash the resin twice with 200 μ L of 80% CH₃CN/0.1% TFA, and combine the supernatant in the glass tube.
11. Dry the combined supernatant completely with a SpeedVac and suspend in 0.4% acetic acid for LC-MS/MS analysis.

3.5.2. Isotope-Labeling Glycopeptides

1. For isotopic labeling of glycopeptides with succinic anhydride following methanol wash at **step 8**, wash the glycopeptides attached to the resin twice with 15% NH₄OH in water (pH > 11.0).
2. Add methylisourea at 1 M in 15% NH₄OH (NH₄OH/H₂O = 15/85 v/v) in 100-fold molar excess over amine groups and incubated at 55°C for 10 min.
3. Wash the beads twice with water, twice with dimethylformamide (DMF)/pyridine/H₂O = 50/10/40 (v/v/v), and resuspend in DMF/pyridine/H₂O = 50/10/40 (v/v/v).
4. Add succinic anhydride solution to a final concentration of 2 mg/mL. Incubate the mixture at room temperature for 1 h.
5. Wash the mixture three times with DMF, three times with water, and six times with 0.1 M NH₄HCO₃.
6. Release peptides from the beads using peptide-*N*-glycosidase F as described.

4. Notes

1. In order to maintain the effective concentration of ICAT reagent (approx 2.5–3.0 mM), an appropriate sample volume should be kept.
2. ICAT-labeling process can be monitored by SDS-PAGE with aliquots saved before and after labeling. A gel mobility shift differentiating labeled and unlabeled proteins should be observed if the protein samples are properly labeled.
3. To estimate molar amount of cysteines in samples, we assume the average molecular weight of proteins is 50 kDa and the average number of cysteins per protein is 6.
4. Acidification may precipitate salts and trypsin. Remove the precipitation by centrifugation before loading onto the SCX column. Deactivate and equilibrate the strong cation-exchange column as recommended by the manufacturer. It is good practice to run standard peptide mixtures to ensure the column performance prior to running sample. Typical gradient profile is 0 to 100% B over 1 h. However, the length of the gradient and the sample collection size can be optimized to suit samples of various sizes and complexities.
5. Use a 1000- μ L gas-tight syringe with stainless steel needle for transferring buffers. Wash the syringe thoroughly with water between the buffers. The avidin cartridge can be used up to 50 times.
6. Store the combined flow-through at –80°C for troubleshooting the avidin purification efficiency. Optionally, this flow-through can be used to corroborate protein identifications based on collision-induced dissociation (CID) of the ICAT-labeled peptides or to locate peptides with posttranslational modifications not present on the ICAT-labeled peptides.
7. Avoid drying any sample containing base together with a sample containing acid in the same SpeedVac, because salts will form in the sample and inside the SpeedVac.

8. Be sure to use a glass vial and a metal syringe for transferring the cleaving reagent, because the reagent contains a strong acid.
9. Protein in other solutions can be exchanged to coupling buffer using a desalting column.
10. The trypsin-released peptides were removed and collected, and can be saved to later corroborate identifications from single-glycosylated peptides. Optionally, the bounded peptides can be isotopically labeled at this step.
11. After methanol wash, the glycopeptides conjugated to the resin can be labeled with stable isotope as described previously.

References

1. Adam, G. C., Sorensen, E. J., and Cravatt, B. F. (2002) Chemical strategies for functional proteomics. *Mol. Cell. Proteomics*. **1**, 781–790.
2. Gygi, S. P., Rist, B., Gerber, S. A., Frantisek, T., Gelb, M. H., and Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* **17**, 994–999.
3. Aebersold, R. and Goodlett, D. R. (2001) Mass spectrometry in proteomics. *Chem. Rev.* **101**, 269–295.
4. Goodlett, D. R. and Yi, E. C. (2002) Proteomics without polyacrylamide: qualitative and quantitative uses of tandem mass spectrometry in proteome analysis. *Funct. Integr. Genomics*. 138–153.
5. Zhang, H., Li, X. J., Martin, D. B., and Aebersold, R. (2003) Identification and quantification of N-linked glycoproteins using hydrazide chemistry, stable isotope labeling and mass spectrometry. *Nat. Biotechnol.* **21**, 660–666.

Amino Acid-Coded Mass Tagging for Quantitative Profiling of Differentially Expressed Proteins and Modifications in Cells

Xian Chen

1. Introduction

To date, large-scale protein quantitation mainly relies on high-resolution separation of complex mixtures using two-dimensional gel electrophoresis (2-DE). In comparison with 2-DE-based protein intensity displays for quantitating differentially expressed proteins, mass spectrometry (MS)-based quantitation of protein expression remains a challenging task because a mass spectrometer itself is a poor quantitative analyzer, as a result of the uneven ionization efficiency of different peptides. Recently, several techniques coupling stable isotope labeling (SIL) to MS have emerged as the primary approach for rapid, large-scale protein quantitation (1–4). There are two major strategies to quantitate proteins through SIL—chemically introducing SIL tags either after cell lysis, or *in vivo/in vitro* during cell growth (5–7). Chemical SIL approaches usually target at a particular residue of tryptic peptides through chemical reactions after cell lysis, thus reducing the complexity of a sample. The most representative methods are isotope-coded affinity tags (ICAT) (1) and mass-coded abundance tagging (MCAT) (4). Disadvantages include the relatively low efficiency of these chemical modification reactions and the limited abundance of target residues. On the other hand, during cell growth, for example, Oda et al. used ^{15}N uniformly labeled medium to label all nitrogen atoms in the whole proteome, and applied this strategy to quantitate protein expression and to identify modifications (5). Our laboratory originally developed a different isotope-tagging strategy, using SIL amino acids as tag precursors that can be incorporated into cellular proteins in a residue-specific manner. These amino acid-coded mass tags (AACTs) then provide a signature for each individual protein or modification for quantitation and concurrent identification (7–13). In our experimental design, *in vivo/in vitro* cell culturing is first performed for one cell line/condition, which then is compared with another cell line/condition in the presence of a particular type of AACT. After cell mixing, protein extraction, and separation, MS is used to recognize and measure the relative population ratios of different isotope-forms of individual peptide sequences. Identification of the proteins that these peptides are derived from is carried out using peptide mass fingerprinting or tandem MS/MS.

2. Materials

2.1. Amino Acids, Stable Isotope-Labeled Amino Acids, Strains, Cell Lines, Chemicals, and Buffers

1. SIL amino acid precursors: Cambridge Isotope Laboratories, Inc. (Andover, MA). The frequently used SIL for AACT include [5,5,5-d₃]leucine (leu-d₃), [4,4,5,5-d₄]lysine (lys-d₄), [methyl-d₃]methionine (met-d₃), [3,3-d₂]tyrosine (tyr-d₂), [2,3,3-d₃]serine (Ser-d₃), and [2,2-d₂]glycine (gly-d₂). Twenty naturally occurring or unlabeled types of amino acids can be obtained from Sigma (St. Louis, MO).
2. Twenty *Escherichia coli* strains of bacteria, each containing a different genetic defect closely linked to a selectable transposon marker, are generous gifts from Dr. David Waugh of NCI (7). These strains of *E. coli* with ideal genotypes were constructed for residue-specific, selective labeling of proteins with almost any stable isotope-labeled amino acid. Using those strains that have been modified to contain the appropriate genetic lesions to control amino acid biosynthesis, dilution of the isotope label by endogenous amino acid biosynthesis and scrambling of the label to other types of residues can be avoided. For example, *E. coli* strain CT2 is constructed by transduction of the BL21(DE3) strain to tetR with a P1 lysate from MF14, and then screening for the gly phenotype. This derivative of BL21(DE3) is used for the selective labeling of proteins with the stable isotope-labeled glycine. Similarly, CT13 is constructed by transducing BL21(DE3) to tetR with a P1 lysate from MF 21, and then screening for the met phenotype (metA⁻). This metA⁻ derivative of BL21(DE3) has the ideal genotype for selective isotope labeling with methionine.
3. *Schizosaccharomyces pombe* strain 972h⁻: American Type Culture Collection (Manassas, VA).
4. A human skin fibroblast (HSF) cell line: American Type Culture Collection (Manassas, VA).
5. Essential vitamins, salts, and trace elements for cell growth, ammonium sulfate, dextrose, and so on: GIBCO-BRL (Grand Island, NY). For *E. coli* bacteria cultures, M9 minimal medium: GIBCO-BRL. For yeast cultures, yeast nitrogen base (YNB): Fisher Scientific (Pittsburgh, PA). For mammalian or human cell cultures, all medium components except for hygromycin B (Calbiochem, San Diego, CA) and doxycycline (Sigma, St. Louis, MO) of the cell culture medium can be obtained from GIBCO-BRL (Grand Island, NY); these include α-minimum essential medium (MEM), McCoy's 5 medium, the regular and dialyzed fetal bovine serum (FBS), genetin (G418), penicillin, and streptomycin.
6. The human colorectal cancer DLD-1.p53 cell line is a "Tet-Off" tetracycline-inducible and stably transfected cell line (14,15). DLD-1.p53 cells are maintained in noninduction medium (Non-Ind): McCoy's 5A base (GIBCO-BRL) supplemented with 10% FBS, 100 units/mL of penicillin, 100 µg/mL of streptomycin, 400 µg/mL G418, 0.02 ng/mL doxycycline, and 250 µg/mL hygromycin B. The induction medium (Ind) has exactly the same composition as NonInd except doxycycline is omitted.
7. Protease inhibitors and other chemicals for gel electrophoresis, peptide extraction, and sample preparation for mass spectrometric analysis: Sigma (St. Louis, MO). Dithiothreitol (DTT) and sequencing-grade trypsin: Roche Diagnostics Corporation (Indianapolis, IN). C₁₈ ZipTips: Millipore Corp. (Billerica, MA). All the chemicals are sequence or high-performance liquid chromatography (HPLC)-grade unless specifically mentioned.
8. Cell re-suspending buffer: 50 mM Tris-HCl (pH 8.0), 10 mM ethylenediamine tetraacetic acid (EDTA), 100 mM NaCl, 10 mM DTT or 5% β-mercaptoethanol, 0.1% sodium dodecyl sulfate (SDS), and protease inhibitor cocktail (1 tablet/50 mL).
9. Rehydration buffer: 7 M urea, 2 M thiourea, 4% CHAPS, 2% DTT, 2% immobilized pH gradient (IPG) buffer (pH 4.0–7.0), 1 mM benzamidine, 1.5 mM EDTA/ethyleneglycol tetraacetic acid (EGTA), 1 mM sodium vanadate, 1 µM microcytin-LR, 2 µg/mL pepstatin-A, 10 µg/mL aprotinin, 20 µg/mL leupeptin.

10. Equilibration buffer: 6 M urea, 2% SDS, 0.375 M Tris-HCl (pH 8.8), 20% glycerol.
11. IPGphor II isoelectric focusing system: (Amersham Biosciences, Piscataway, NJ).

2.2. Cell-Culture Media for AACT Precursor Incorporations

1. For residue-specific incorporation of a selected type of SIL amino acids in bacterial cells, M-9 minimal media is supplemented with 200 mg/L of each of the 19 naturally occurring or unlabeled amino acids except for cysteine; the latter is at a concentration of 20 mg/L. The 20th amino acid, containing stable isotope labels, is added for a final concentration of 200 mg/L (7).
2. For residue-specific tagging of yeast cells, yeast strains can be grown in a synthetic defined (SD) medium. The SD medium consists of 20 g dextrose, 1.7 g yeast nitrogen base without amino acids, 5.0 g ammonium sulfate, and 100 mg each of adenine, histidine, tryptophan, and leucine in a liter of H₂O. SIL amino acids can be supplied at 100 mg/L. The medium components are sterilized by autoclaving (13).
3. α -MEM medium: 20 amino acids plus cystine, inorganic salts, trace elements, and so on. The concentration of each component is indicated in the manufacturer's manual (GIBCO-BRL, Grand Island, NY). For AACT of human cells, the α -MEM medium depleted with particular essential amino acids can be obtained from the UCSF cell-culture facility. A particular AACT- α -MEM medium is supplemented with 10% dialyzed fetal bovine serum (FBS), appropriate antibiotics, and the selected isotope-enriched amino acid precursor—leu-*d*₃, ser-*d*₃, or tyr-*d*₂, which completely substitutes for its unlabeled counterpart, depleted in the α -MEM medium (9).

3. Methods

This quantitative technique of AACT allows for large-scale analyses of simultaneous changes of multiple proteins in their expressions at a proteome level. Through the natural incorporation of AACT in cellular proteins, this MS-based approach maintains the same ionization efficiency for both reference and sample signals in quantitative measurements of differential protein expression/modification profiles. Meanwhile, the high-throughput capability of mass spectrometry is fully utilized. This AACT method allows for an instant comparison of protein expression/modification changes in a whole proteome without performing tedious 2-DE gels.

3.1. In Vivo/In Vitro Cell Culturing With AACT for Comparative Proteomic Studies

The general design is illustrated in **Fig. 1**. First, two pools of cells are grown under different isotope environments—one population is cultured in normal or “light” medium, the other in “heavy” medium containing amino acid residues labeled with heavier isotopes. Note that both light and heavy media are essentially identical except for the difference in isotope form of particular types of amino acids. In practice, two biologically relevant cell lines for which protein expression comparison is desired (e.g., diseased vs normal, or untreated vs to-be-treated cells) are grown pairwise in the light and heavy cell pool, respectively, whereas one cell pool will be exposed to or stimulated by a stress. After cell growth in the absence or presence of stress treatment, equal amounts of cells from both pools are mixed to minimize artifacts in sample handling, thus allowing an accurate quantitation of the relative protein expression levels in the two cell populations. Through AACT-assisted MS analyses, the level of differential protein expressions correlates to the relative intensities of paired isotope peaks in MS spectra.

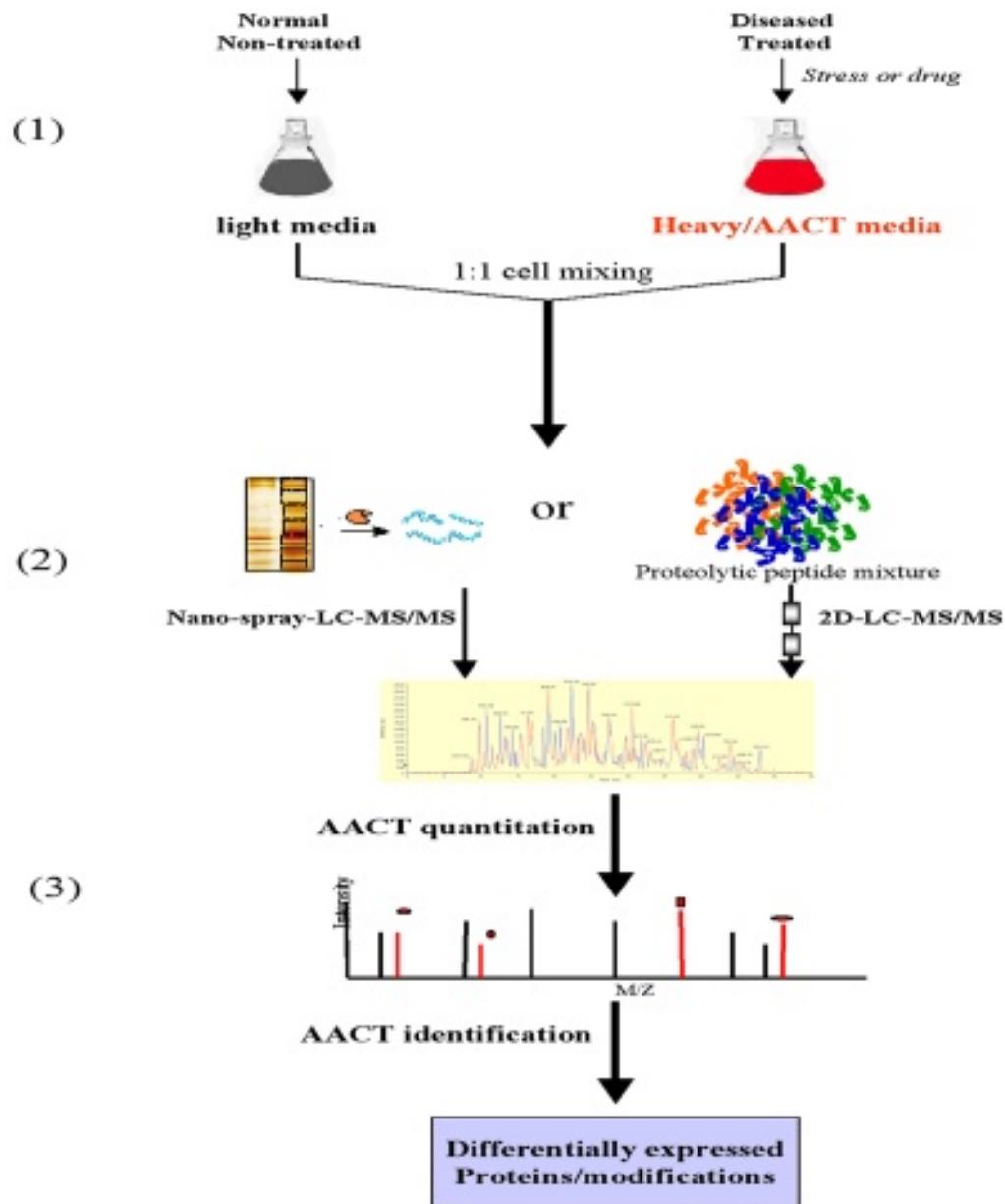


Fig. 1. The overall experimental design of amino acid-coded mass tagging (AACT)-assisted quantitative proteomics. (A) In vivo/in vitro cell culturing; (B) sample preparation; (C) mass spectrometry (MS) measurements and data analyses.

These isotope pairwise signals originate from two different cell states, allowing the immediate comparison of any pair of proteomic states.

1. Without isotope scrambling, residue-specific incorporation of each selected type of SIL-amino acids, essential amino acids in particular, were observed for a variety of cell species, including *E. coli*, yeast, and human cell lines (7-9,13). For the quantitation of posttranslational modifications, those amino acid residues that have a potential to be modi-

- fied, such as tyr-*d*₂, and ser-*d*₃ in protein phosphrolation, can be used as SIL precursors to signature the proteome of interest during in vivo/in vitro cell culturing (11).
2. Yeast cells are inoculated into 10 mL SD medium and grown overnight at 30°C. The overnight yeast culture can be diluted to a starting optical density (OD) of 0.02 in 100 mL SD medium containing either naturally occurring amino acids or AACT. No growth difference is observed between cells in regular and labeled media—i.e., isotope-labeled amino acids did not affect cell growth as a source of nutrition. As an example, in the study of cellular proteome responses to heavy-metal exposure (16), cadmium ions in the form of CdSO₄ are added to the mid-exponentially growing *S. pombe* culture (A_{600} approx 0.35) in the heavy medium at a defined concentration. The control *S. pombe* cells are separately grown in the light medium under the same conditions without Cd²⁺ treatment. The ODs of cell cultures grown in the normal and the labeled medium are measured at different time points. After mixing the same OD number of cells from both pools at each designated time after Cd²⁺ addition, the cell mixture is harvested and washed twice with milli-Q H₂O to remove excess medium. The cell pellet is then resuspended in cell resuspending buffer. The cell lysate is prepared by vortexing with glass beads for 10 min at 4°C, followed by centrifugation for 10 min at 4°C to clarify the soluble proteins.
 3. As a model mammalian or human cell line to study the cellular responses to different stresses such as heat shock and radiation, human skin fibroblast (HSF) cells are grown in α-MEM medium at 37°C with 5% CO₂ and 90% relative humidity (13,17). The heavy medium is made of lysine- and leucine-depleted α-MEM medium supplemented with 10% dialyzed fetal bovine serum, 100 units/mL penicillin, 100 µg/mL streptomycin sulfate, and 88 mg/L of 100% lys-*d*₄ or leu-*d*₃ precursors. The cells grown in the heavy medium are subjected to an incubation at 43°C for 8 h (9) or irradiated with 500 Rads of γ-rays before recovering for 4 h under normal condition (10), while the cells without exposure to a stress are grown in light medium. At about 80% confluence, equal numbers of cells from the treated or nontreated cell pool are mixed and harvested with trypsin-EDTA treatment. The cell mixture is washed with phosphate-buffered saline (PBS) to remove any trypsin and FBS residues.
 4. As a cancer model cell line to profile biomarkers responsible for p53-mediated apoptosis (18), normal DLD-1 cells are grown in α-MEM medium containing naturally occurring amino acids, while cells from the DLD-1.p53 cell line that harbors the gene for tetracycline-inducible p53 expression are cultured in the same medium except for the substitution of a selected type of amino acid with its heavy form, both at 37°C with 5% CO₂ and 90% relative humidity in the presence of 20 ng/mL of doxycycline (Dox). To achieve 100% AACT for the induced DLD-1.p53 cells, at approx 80% confluence, the adherent cells are rinsed in PBS and re-fed with the induction (heavy) medium, i.e., AACT-α-MEM, in the absence of Dox, for the induction of p53 and downstream apoptosis.

3.2. Sample Separation and Preparation

A variety of methods, such as one-dimensional sodium dodecyl sulfate (1-D SDS) polyacrylamide gel electrophoresis (PAGE), 2-DE, or high-throughput two-dimensional liquid chromatograph (2-D LC) can be used for protein separation prior to mass spectrometric analyses.

1. After cell mixing and ultra-centrifugation, a cell pellet containing 0.5–3 million cells can be suspended in a corresponding buffer containing 5 mg/mL protease inhibitors (leupeptin, pepstatin A, and chemostatin A). Cells are lysed according to the species—*E. coli* (7), yeast (10,13), or human (9). Each cell lysate is then centrifuged or fractionated to different subcellular compartments (19), collected, and subjected to subsequent PAGE or LC separations.

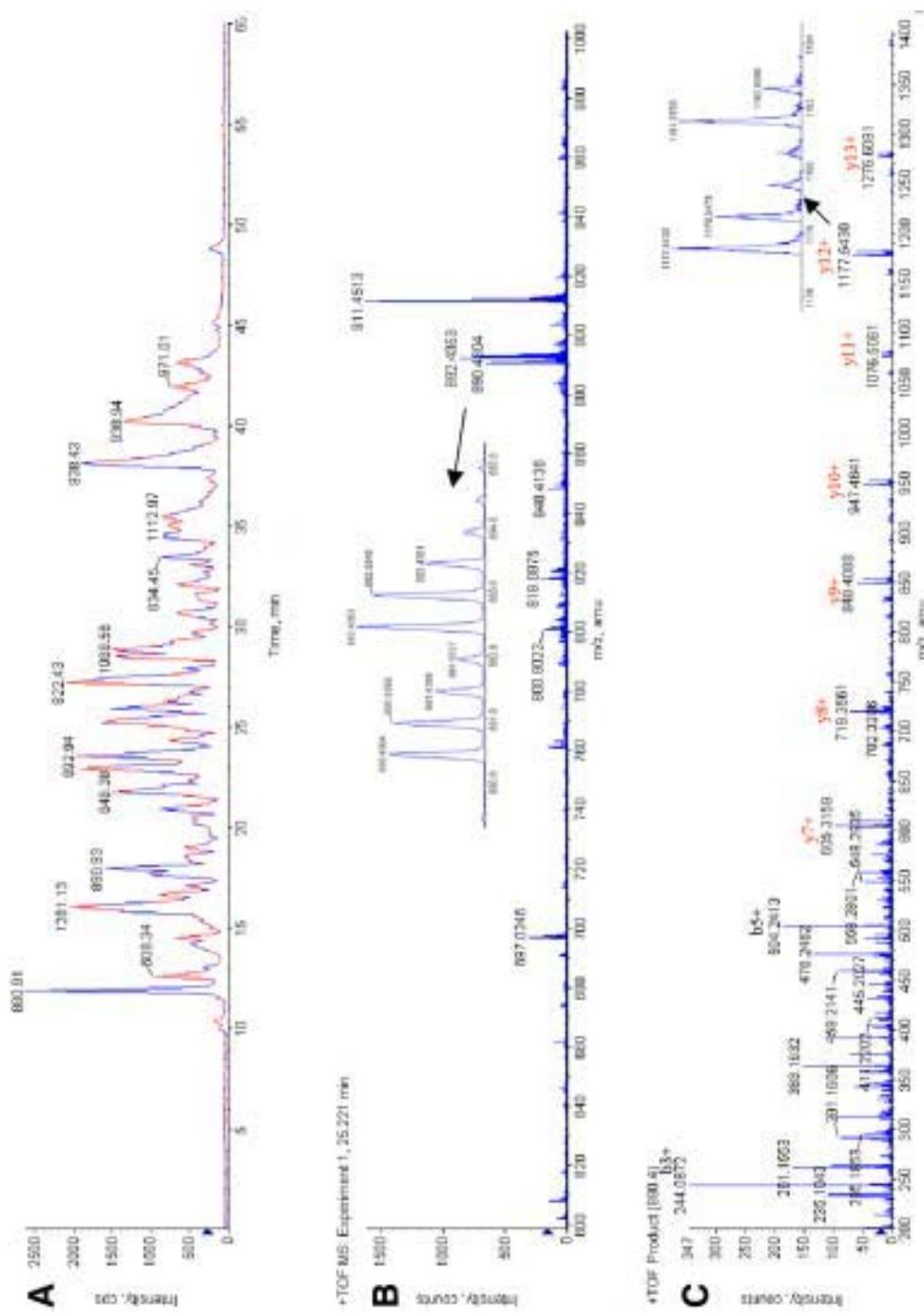
2. For high-resolution separation, the lysate of a cell mixture can be resolved using one 2-DE gel. For example, 50–200 µg protein mixture can be added to 350 µL of rehydration buffer and rehydrated at 20 V for 12 h. Isoelectric focusing is performed at ambient temperature on an 18-cm immobiline pH dry strip with a defined pH range. The voltage is held at 500 V for 30 min, stepped to 1000 V and held for 2 h; then a 3-h gradient from 1000 to 7000 V is applied to the strips. The voltage is held at 7000 V for another 4 h. Prior to the second molecular-weight dimension, the strips are incubated for 10 min in equilibration buffer, first with 130 mM DTT and second with 135 mM iodoacetamide. Each equilibrated strip is then put onto a 20 × 20 cm 10% duracryl gel, and the second dimension is run at 500 V for 5 h in a 2-D gel system. The gel can be silver- or SYPRO® Ruby-stained as previously described (20). The protein spots are excised from the gel for digestion.
3. One-dimensional SDS gels can maximize the solubility of hydrophobic or membrane proteins in a cell lysate. The coupling between 1-D SDS separation and microLC-electrospray ionization (ESI)-MS/MS (see **Subheading 3.3.**) is particularly useful for the quantitative analyses of subcellular fractions of membrane proteins. For example, a cell fraction is fully solubilized in the SDS-PAGE loading buffer in boiling water for 10 min. Fifty to 200 µg of total protein can be loaded onto a 4–20% gradient SDS-PAGE gel (10 cm in length, 1 mm in thickness). Protein bands are visualized by Coomassie blue (G-250) staining and are excised continuously with a 1- to 2-mm step. The resulting gel slices are placed into 1 mm³ cubes for destaining and digestion.
4. As previously described (13), gel spots or bands are destained with 50% (v/v) acetonitrile in 50 mM NH₄HCO₃ and dried with 100% acetonitrile for 20 min followed by speed-vacuum centrifuge for 10 min, and then subjected to overnight in-gel trypsin digestion (11). Briefly, 15 µL sequencing-grade modified trypsin (10 µg/mL in 50 mM NH₄HCO₃) is added to the lyophilized gel with 20 µL 50 mM NH₄HCO₃ to cover the rehydrated gel, and the reaction is incubated overnight at 37°C. The tryptic peptides are extracted twice from the gel slices by 45 min sonication for each, first in 200 µL 5% acetic acid solution, and then in 200 µL 5% acetic acid/50% acetonitrile solution. The supernatants are combined, lyophilized, and resuspended in a 0.1% trifluoracetic acid (TFA) solution. In the integrated 1-D SDS-microLC MS/MS analytical mode, gel elutes of proteolytic digests are directly subjected to further microLC separation before nanospray-MS/MS experiments (see **Subheading 3.3.**). Meanwhile, this process can be done in a high-throughput mode by linking the autosampler to the LC apparatus that loads the digests of 1-D bands automatically; e.g., the digest elute from 24 gel bands can be analyzed within 12 h. For matrix-assisted laser desorption/ionization (MALDI)-time-of-flight (TOF) MS analysis, additional desalting steps are performed using C₁₈ ZipTips, and the final elution solution is 50% acetonitrile/0.1% TFA.

3.3. Mass-Spectrometric Measurements and Data Analyses

The ABI Voyager DE-STR MALDI-TOF mass spectrometer can be employed to examine the peptide mass map of each in-gel digest. These measurements can be accomplished automatically on a 96-spot format. An ABI MDS SCIEX Qstar XL quadrupole TOF MS with both nanospray atmospheric pressure ionization (API) and oMALDI sources is fully capable of MS/MS fragmentation of proteolytic peptides. This instrument can perform nanospray ESI LC-MS/MS (LC Packings Ultimate microcapillary LC system) experiments for 1-D-gel band extracts. An upstream auto-sampler handles multiple sample (currently with 24 samples) injections automatically. A Finnigan LCQ ESI MS is online with a Microtech capillary LC system, which can be used to generate tandem MS/MS spectra of the peptides separated by 2-D LC.

1. As shown in **Fig. 2**, the incorporated AACT serve as markers for the quantitation of differentially expressed proteins or modifications. Meanwhile, AACT represents a parameter in addition to m/z values of fragment ions, to enhance MS signal specificity. For example, in MS spectra, a mono-isotopic distribution pattern of proteolytic peptides can be resolved for each individual isotope species— M^+ , $(M + 1)^+$, $(M + 2)^+$, and so on. (M refers to the mass of the most abundant isotopes). At natural abundance, biomolecules are composed of over 99% of the isotopes ^{12}C , ^{14}N , and ^1H , whereas the heavier isotopes— ^{13}C , ^{15}N , and ^2H (D)—are close to or less than 1%. Stable isotope enrichment or labeling is a process whereby the amount of the least abundant isotope(s) is enriched to a higher level. The incorporation of a selective amino acid enriched by heavy isotopes, such as Leu- d_3 or lys- d_4 , into any given peptide increases its mass. This mass shift leads to a pair of mono-isotopic peaks—the lower-mass component corresponding to the naturally occurring isotopes and the higher-mass component containing heavier isotopes. The intensity ratios between the lower- and upper-mass components of these peak pairs will correlate to the relative abundance of the natural and heavy isotopes.
2. As previously described (8,12), MALDI-TOF mass spectra of proteolytic digests can be acquired with oMALDI-TOF of Qstar XL or a PE Voyager-DE STR biospectrometry workstation equipped with a N_2 laser (337 nm, 3-ns pulse width, 20-Hz repetition rate) using the reflector mode with delayed extraction. The matrix, α -cyano-4-hydroxycinnamic acid, was prepared as a saturated solution in 50% acetonitrile/0.1% TFA solvent. For MALDI-TOF analysis, 0.5 μL of the matrix solution was mixed with 0.5 μL of sample on the sample plate, and the mixture was air-dried to form the crystal analyte.
3. A 1-D SDS-microLC-MS/MS platform can be used to perform protein separation, protein quantitation, and concurrent identification. In a high-throughput mode, in-gel digest of each 1-D SDS gel slice can be analyzed by a microLC-nanospray-MS/MS using a Qstar XL mass spectrometer coupled with LC Packings Ultimate microcapillary LC system. The PepMap C18 column (3 μm , 100 \AA , 75 μm i.d., 15 cm length) employed for peptides separation is also obtained from Dionex. An autosampler is configured using the partial loop injection mode involving a 10- μL sampling loop; a preconcentration C_{18} cartridge is connected with the analytical column through a 10-port switch valve. The partial loop injection method loads a 3- μL sample into the 10- μL loop, which is then pumped onto the preconcentration column at a flow rate of 30 $\mu\text{L}/\text{min}$ by a sample-loading pump. Three minutes after the start of the sample loading, the 10-port valve can be switched to the preconcentration cartridge in line with the nanoflow solvent delivery system, thus enabling the trapped peptides to be eluted onto the analytical column. Mobile phase A is 0.1% formic acid and 5% acetonitrile. Mobile phase B is 0.1% formic acid and 95% acetonitrile. The gradient is kept at 5% B for 5 min, then ramped linearly from 5 to 50% B in 50 min, then jumped to 75% B and kept for 10 min. Then the gradient is returned to the starting point and the column is equilibrated for 10 min. The flow rate can be 200 nL/min . The end of the analytical column is connected to a 10 μm i.d. PicoTip nanospray emitter (New Objective, Woburn, MA) by a stainless steel union mounted on the nanospray source. The spray voltage (usually set between 1800 and 2100 V) is applied to the emitter through the stainless steel union and tuned to get the best signal intensity using standard peptides. The two most intense ions, with charge states between 2 and 4 in each survey scan, are selected for the MS/MS experiment provided they passed the switching criteria of the MS/MS scan. The rolling collision energy feature is employed to fragment the peptide ions according to their charge states and m/z values (18).
4. ProICAT software with AACT constraint can be used to perform a data-dependent search to locate those isotope pairs and their derivative MS/MS spectra (**Fig. 3**). The light peptide isotope peak in the MS spectrum derived from the protein expressed in the normal/light-

Base Peak Chem.



medium pool will form a pair with the peptide from the same protein expressed in the heavy-medium culture. Isotope pairs of equal intensity show that the parent protein expression is the same in both cell pools, whereas isotope pairs of different intensities corresponding to protein expression triggered by a particular biological process or environmental stimulus (18).

5. Proteolytic peptide masses are typically in the range of 500 to 4000 amu, where both ESI and MALDI-TOF mass spectrometers have sufficient resolution and sensitivity to distinguish AACT peptides and their isotopic distribution patterns. To facilitate large-scale protein identification, AACT is highly specific and characteristic so as to be recognized by MS. As shown in **Fig. 4**, the amino acid-coded mass tag increases signal specificity in database searches. Thereby, both parameters, the *m/z* value of a fragment ion and its content of a particular amino acid(s), can be used to characterize the corresponding ions. For example, in a peptide mass map (PMM), those peptides containing AACTs can be distinguished from other peptides by their characteristic mass-split patterns. Proteolytic peptides derived from the theoretical digestion of various proteins translated from the genomic sequence or expressed sequence tag (EST) databases are filtered with the content of the labeled amino acid residue(s), resulting in an AACT-constrained proteolytic peptide library. The database search is then constrained by two parameters, allowing for far more selective and confident protein sequence searches than those by peptide *m/z* values alone. Therefore, the multiple AACT approach can reveal the partial amino acid composition of particular AACT peptides (12).
6. For all peptides with complete AACT patterns showing induced protein expressions on MS spectra, the observed *m/z* and the peptide composition of the four labeled amino acids can be submitted to the National Center for Biotechnology Information (NCBI) nonredundant database to determine the protein identity with 500 ppm mass tolerance, using the MS-Seq program (12). Both MASCOT and ProID loaded on the Qstar instrument can be used to interpret the LC-MS/MS data by searching against the NCBI protein database. Peptides matches with a score larger than 39 in MASCOT or larger than 43 in ProID were considered as a significant match or homology (*p* < 0.05) (21).

4. Notes

1. During microLC separation runs, although the most of the AACT peptides co-elute from the LC column with their unlabeled counterparts, some of the labeled peptides can appear slightly (usually from 1 to 5 s) ahead of the unlabeled peptides. Therefore, we can average the chromatographic profiles of both unlabeled and labeled peptides rather than using single time spectra for quantitative analysis (18).
2. Using those AACT with mass tags less than 3 Da, for those peptides containing only one precursor residue, the first mono-isotopic ion peak, M^+ , of the labeled peptide could overlap with the fourth isotope peak, $(M + 3)^+$, of the unlabeled peptide. To assure accuracy of quantitation, the theoretical intensity of the $(M + 3)^+$ isotope peak of each unlabeled peptide can be calculated from the atomic compositions and then subtracted from the overall apparent peak intensity to give the accurate intensity of the labeled peak (18).

Fig. 2. (*opposite page*) On-line high throughput amino acid-coded mass tagging (AACT)-assisted liquid chromatography (LC)-tandem mass spectrometry (MS/MS) for quantitation and identification of differentially expressed proteins. **(A)** Base peak chromatograph of in-gel digests of a one-dimensional sodium dodecyl sulfate band. The change of color indicates the change of base peak during time-of-flight (TOF) MS scan; **(B)** TOF MS scan spectrum of the LC retention time at 25.2 min and the peptide at 890.43 $lys-d_4$ split pattern indicates the protein is differentially expressed; **(C)** MS/MS spectrum of the lysine-containing peptide in **B**.

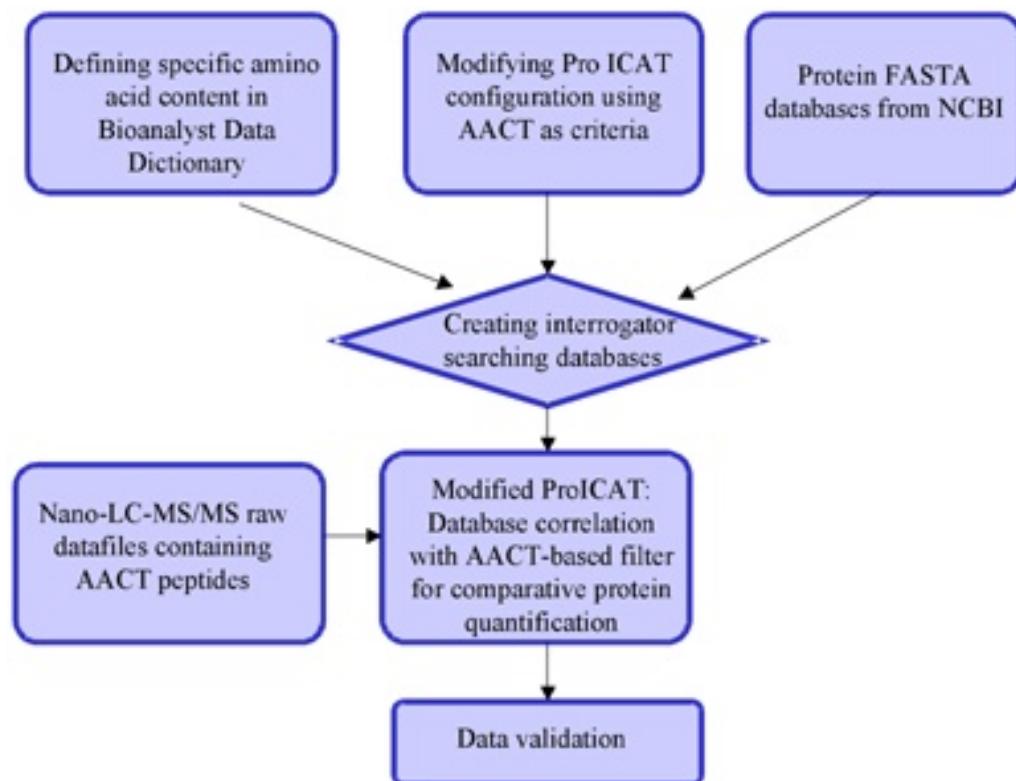


Fig. 3. Amino acid-coded mass tagging (AACT)-assisted modifications of ProICAT program in Qstar XL for simultaneous protein quantitation and identification.

3. In general, among various chemical or enzymatic tag methods that provide MS-recognizable markers, this AACT strategy improves the efficiency and accuracy of the isotope-based quantitation for a proteome-scale quantitative analysis. Unlike ICAT, AACT does not require any further chemical modifications, and the problems caused by side reactions and the relatively low efficiency of the desired reaction can be avoided. Except for the mass differences, these AAC tags maintain the same chemical and physical properties as their naturally occurring counterparts during quantitative MS measurements in a good dynamic range. More importantly, the presence of AACT in peptides has no effect on their MS/MS fragmentation, signal ionization, and fragment assignment; thus, high accuracy is intrinsic for protein quantitation.
4. For many technical reasons, our previous results suggest that this AACT strategy is superior to the popular ICAT-based approaches where the cysteine-containing peptides are sole reference signals for protein quantitation. AACT is amino acid-specific; essentially, most of the 20 types of amino acid residues can be chosen as the labeled precursors, while essential amino acids are ideal. Consequently, in comparison with the 1.8% natural distribution of cysteine in cellular proteins, the availability of a variety of amino acid signatures in cellular proteins gives much larger sequence coverage than ICAT, which frequently quantifies and identifies proteins relying on single peptide sequences. Therefore, AACT opens the window for observing multiple AACT-containing peptide signals for more accurate protein/modification quantitation and unambiguous identification. As a direct result, different isomers in the same protein family can be quantitated and identi-

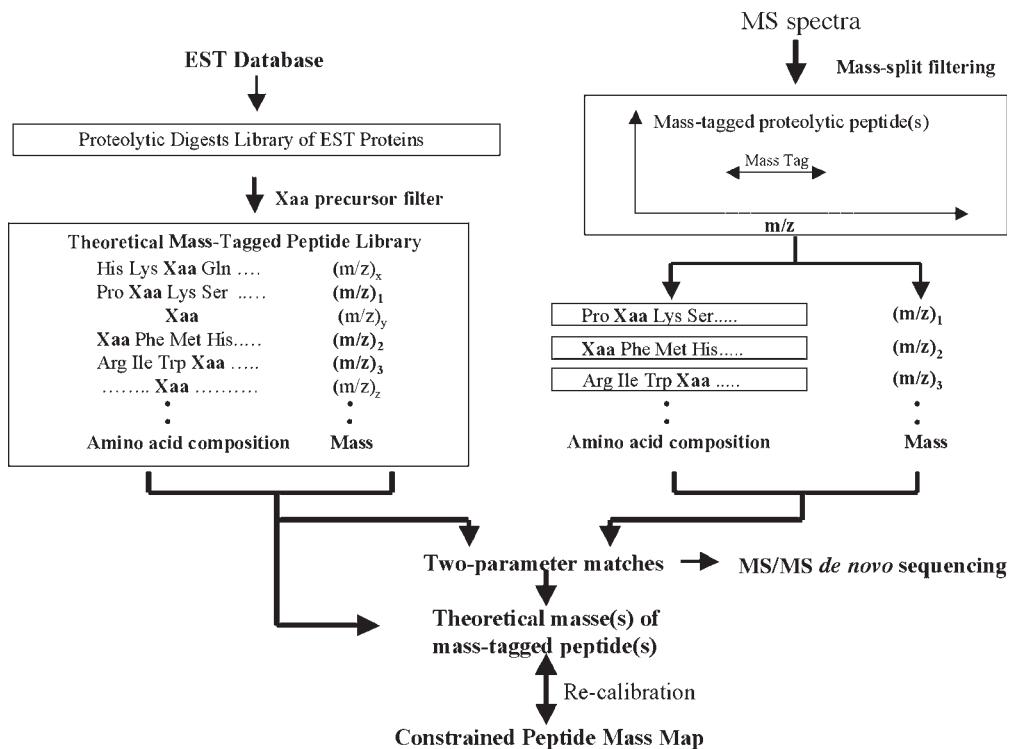


Fig. 4. Two-parameter constrained data search, the m/z value, and the content of particular amino acid(s) of a fragment ion.

fied, because the signal specificity provided by these residue-specific tags helps to reduce spectral complexity in MS spectra of complex proteomes. These unique features facilitate comprehensive analysis of all possible regulated proteins for a whole proteome (18).

5. The high specificity and efficiency of incorporation of these mass tags have been demonstrated for various cell lines with low cost. Because of the high sensitivity of MS, the cost of the labeled amino acids is minimum for each small-scale cell culture. The sample consumption using the AACT approach, usually 100–300 μ g for each run, is far less than that required by the ICAT method, which usually requires milligram quantities. Also, the AACT strategy can be coupled to the 2-D LC-based methods for higher throughput and less sample consumption.
6. Through AACT, individual MS strengths such as throughput, specificity, accuracy, and dynamic range are integrated. The use of AACT as internal markers will enhance the signal specificity of mass spectra, allowing us to overcome technical difficulties in identifying post-translational modifications, proteins of low abundance, cell membrane-spanning proteins, and proteins of high molecular charge, extreme pH, or very low molecular mass.

Acknowledgments

This work was supported by DOE grants ERW9923, ERW9840, and Los Alamos National Laboratory LDRD 20030508ER. X. C. is a recipient of a Presidential Early Career Award for Scientists and Engineers (PECASE) (2000–2005). The author thanks Dr. Sheng Gu for technical supports.

References

1. Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., and Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol* **17**, 994–999.
2. Geng, M., Ji, J., and Regnier, F. E. (2000) Signature-peptide approach to detecting proteins in complex mixtures. *J Chromatogr. A* **870**, 295–313.
3. Shevchenko, A., Chemushevich, I., Ens, W., et al. (1997) Rapid “de novo” peptide sequencing by a combination of nanoelectrospray, isotopic labeling and a quadrupole/time-of-flight mass spectrometer. *Rapid Commun. Mass Spectrom.* **11**, 1015–1024.
4. Cagney, G. and Emili, A. (2002) De novo peptide sequencing and quantitative profiling of complex protein mixtures using mass-coded abundance tagging. *Nat. Biotechnol.* **20**, 163–170.
5. Oda, Y., Huang, K., Cross, F. R., Cowburn, D., and Chait, B. T. (1999) Accurate quantitation of protein expression and site-specific phosphorylation. *Proc. Natl. Acad. Sci. USA* **96**, 6591–6596.
6. Veenstra, T. D., Martinovic, S., Anderson, G. A., Pasa-Tolic, L., and Smith, R. D. (2000) Proteome analysis using selective incorporation of isotopically labeled amino acids. *J. Am. Soc. Mass Spectrom.* **11**, 78–82.
7. Chen, X., Smith, L. M., and Bradbury, E. M. (2000) Site-specific mass tagging with stable isotopes in proteins for accurate and efficient protein identification. *Anal. Chem.* **72**, 1134–1143.
8. Gu, S., Pan, S., Bradbury, E. M., and Chen, X. (2002) Use of deuterium-labeled lysine for efficient protein identification and peptide de novo sequencing. *Anal. Chem.* **74**, 5774–5785.
9. Gu, S., Pan, S., Bradbury, E. M., and Chen, X. (2003) Precise peptide sequencing and protein quantification in the human proteome through in vivo lysine-specific mass tagging. *J. Am. Soc. Mass Spectrom.* **14**, 1–7.
10. Zhu, H., Pan, S., Gu, S., Bradbury, E. M., and Chen, X. (2002) Amino acid residue specific stable isotope labeling for quantitative proteomics. *Rapid Commun. Mass Spectrom.* **16**, 2115–2123.
11. Zhu, H., Hunter, T. C., Pan, S., Yau, P. M., Bradbury, E. M., and Chen, X. (2002) Residue-specific mass signatures for the efficient detection of protein modifications by mass spectrometry. *Anal. Chem.* **74**, 1687–1694.
12. Pan, S., Gu, S., Bradbury, E. M., and Chen, X. (2003) Single peptide-based protein identification in human proteome through MALDI-TOF MS coupled with amino acids coded mass tagging. *Anal. Chem.* **75**, 1316–1324.
13. Hunter, T. C., Yang, L., Zhu, H., Majidi, V., Bradbury, E. M., and Chen, X. (2001) Peptide mass mapping constrained with stable isotope-tagged peptides for identification of protein mixtures. *Anal. Chem.* **73**, 4891–4902.
14. Donald, S. P., Sun, X. Y., Hu, C. A., et al. (2001) Proline oxidase, encoded by p53-induced gene-6, catalyzes the generation of proline-dependent reactive oxygen species. *Cancer Res.* **61**, 1810–1815.
15. Yu, J., Zhang, L., Hwang, P. M., Kinzler, K. W., and Vogelstein, B. (2001) PUMA induces the rapid apoptosis of colorectal cancer cells. *Mol. Cell* **7**, 673–682.
16. Bae, W. and Chen, X. (2004) Proteomic study for the cellular responses to cadmium in *Schizosaccharomyces pombe* through amino acid-coded mass tagging and LC-MS/MS. *Mol. Cell. Proteomics* **3(6)**, 596–607.
17. Zhu, H., Hunter, T. C., Pan, S., Yau, P. M., Bradbury, E. M., and Chen, X. (2002) Residue-specific mass signatures for the efficient detection of protein modifications by mass spectrometry. *Anal. Chem.* **74**, 1687–1694.

18. Gu, S. et al. (2004) Global investigation of p53-induced apoptosis through quantitative profiling regulatory proteins using comparative amino acid-coded tagging proteomics. *Mol. Cell Proteomics* **3**, 998–1008.
19. Gu, S., Chen J., Dobos K. M., Bradbury E. M., Belisle J. T., and Chen X. (2003) Comprehensive proteomic profiling of the membrane constituents of a *Mycobacterium tuberculosis* strain. *Mol. Cell. Proteomics* **1**, 1284–1296.
20. Mortz, E., Krogh, T. N., Vorum, H., and Gorg, A. (2001) Improved silver staining protocols for high sensitivity protein identification using matrix-assisted laser desorption/ionization-time of flight analysis. *Proteomics* **1**, 1359–1363.
21. Perkins, D. N., Pappin, D. J., Creasy, D. M., and Cottrell, J. S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551–3567.

Mass-Coded Abundance Tagging for Protein Identification and Relative Abundance Determination in Proteomic Experiments

Gerard Cagney and Andrew Emili

1. Introduction

Advances in mass spectrometry have led to the emergence of the distinct field of proteomics. One aim of proteomics, the identification of the protein components of complex biological mixtures, is now routinely realized, typically by peptide mass mapping following matrix-assisted laser desorption/ionization (MALDI) mass spectrometry (MS) or by peptide sequence determination from tandem mass spectra obtained by electrospray ionization followed by collision-induced dissociation (CID) (1). Both approaches rely on the identified proteins being present in DNA or protein sequence databases. This is because the behavior of ionized peptides in MS experiments is somewhat unpredictable and the resulting spectra are searched against “idealized” spectra generated from the sequence databases to find the nearest match. Nevertheless, both approaches have been highly successful, with thousands of proteins identified in a single large-scale analysis (reviewed in ref. 2). A method that is independent of databases would be useful in certain cases, however, especially for protein samples deriving from organisms whose genomes remain unsequenced, proteins with erroneous sequences deposited in the databases, or proteins whose splicing patterns or modification states are unknown. Another partially fulfilled goal of proteomics is to determine the quantities of each protein present in a mixture, or at least the relative abundance of proteins present in two different samples, such as a test sample and a reference control. Several approaches for determining relative abundance in proteomic experiments have involved differential incorporation of stable isotopes into one of the samples, using either labeled growth media (3) or postexperimental chemical labeling (4,5). At least two methods using nonisotopic reagents for purposes of proteomic quantification have recently been reported (6,7).

Here we describe the use of one of these methods, mass-coded abundance tagging (MCAT), to identify and compare the relative abundances of proteins present in two proteomics samples (6). The samples may be complex protein mixtures—for example, immunoprecipitates or cell lysates—and differ in any respect—for example, before and after an experimental perturbation such as the treatment of cells with a drug. The samples are first extensively digested with trypsin, to produce peptides with C-terminal lysine or arginine residues. One sample is then treated with the reagent *O*-methyl-

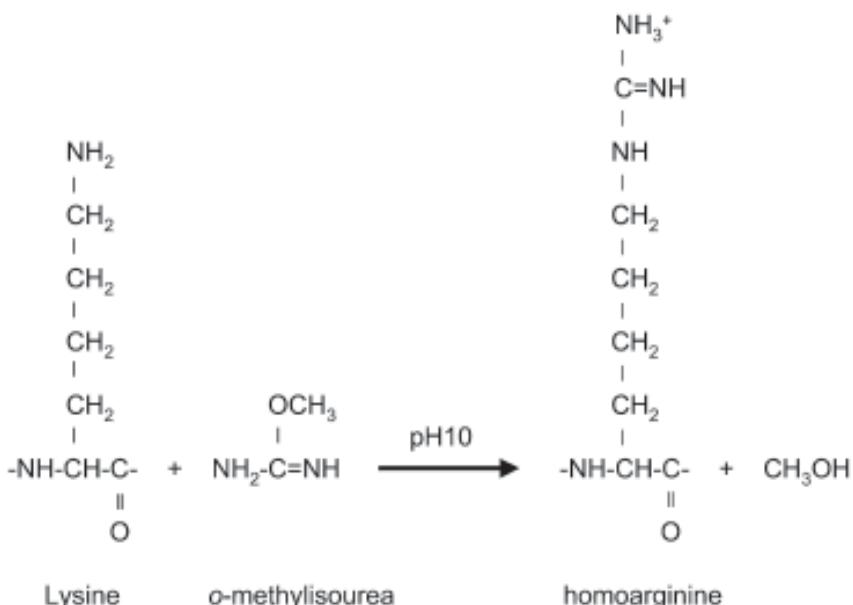


Fig. 1. Selective conversion of peptide lysine residues to homoarginine at elevated (basic) pH using the reagent *o*-methylisourea.

sourea, which modifies the ϵ -amino group of lysine side-chains, resulting in quantitative conversion to homoarginine (Fig. 1). This increases the mass of the modified peptides by 42 amu per lysine residue, without significantly affecting the ionization or fragmentation efficiencies using standard liquid chromatography (LC)-MS conditions. (The modification does increase the signal produced by lysine-containing peptides in MALDI-MS experiments, presumably because the more basic homoarginine residues have better ionization properties than lysine [8–11].) The reaction is quenched, and the samples pooled and analyzed using capillary-scale LC-MS using an electrospray ionization source. The modified and nonmodified peptide pairs typically elute within 1 min of each other, and hence their relative abundance can be determined by comparing the peak intensities of both peptides over approximately this period (Fig. 2).

As a result of low-energy CID, peptides primarily fragment along their amide bonds, generating two main types of fragment ion: *b*-type and *y*-type (where the charge is retained at the N- or C-terminus of the ion, respectively). If present, the peaks of these series are separated by the unit masses of the amino acid residues between them, thus allowing the sequence to be determined (except in cases where the residue masses are identical or nearly so). However, *de novo* sequencing—that is, the interpretation of peptide mass spectra without the use of protein sequence databases to facilitate interpretation—is difficult to automate, primarily because the researcher does not know which peaks on a mass spectrum belong to which ion series. Using MCAT, the *y*-ion series can be readily determined, because modification with *o*-methylisourea adds 42 Da to the expected value of the *y*-series fragment ions (Figs. 2 and 3). This greatly simplifies the process of peptide sequencing, and facilitates automation by computer. This has been exploited to enable *de novo* sequencing of peptides (6,8–11).

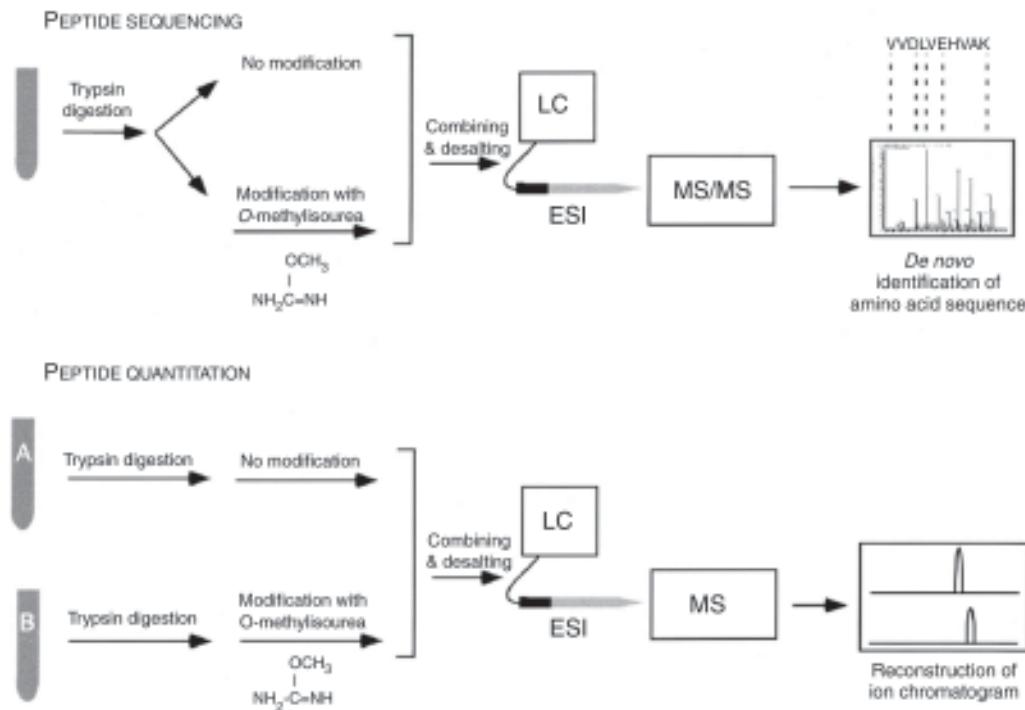


Fig. 2. Determination of peptide sequence and measurement of relative protein abundance using MCAT. (Top) Peptide sequence is determined by treating half of a tryptic protein digest with *O*-methylisourea, combining the fractions, followed by reverse-phase capillary liquid chromatography (LC) and electrospray tandem mass spectrometry (MS). Comparison of the fragmentation patterns of modified and unmodified pairs of sister peptides reveals the C-terminal y-ion series, allowing the order of amino acid residues to be deduced. (Bottom) Protein quantification based on differential labeling of tryptic digests of one of two protein mixtures of interest using *O*-methylisourea. The modified and unmodified peptides are then pooled and analyzed by LC-MS. Relative protein levels are determined by comparing the ion currents produced by respective pairs of labeled/unlabeled sister peptides, which closely co-elute during reverse-phase chromatography.

2. Materials

2.1. Trypsin Digestion

1. μ C18 ZipTips (Millipore or NEST group).
2. 100 mM Ammonium bicarbonate, pH 8.5 (stable 3 mo at room temperature).
3. 8 M Urea, pH 8.5 (stable 6 mo at room temperature).
4. 100 mM Dithiothreitol (DTT) (stable 3 mo at -20°C).
5. 100 mM Iodoacetamide (freshly made in the dark).
6. 100 mM Calcium chloride (stable 6 mo at room temperature).
7. Sequencing-grade trypsin (Promega or Roche) or immobilized trypsin beads (Porozyme) (both stable 6 mo at 4°C).

2.2. MCAT Modification

1. Crystalline ultra-pure [3]-*O*-methylisourea (Sigma).

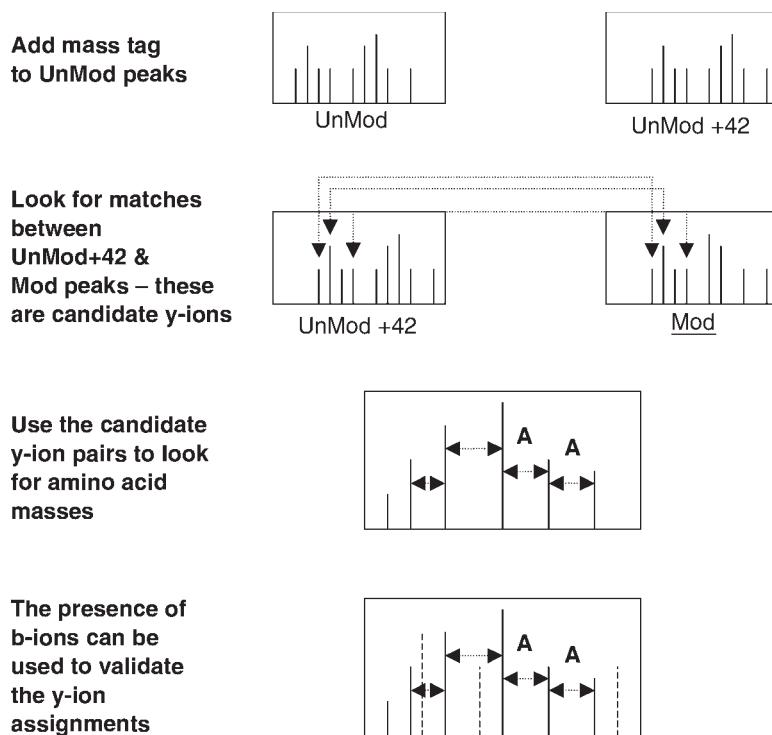


Fig. 3. MCAT-assisted *de novo* peptide sequencing. Lysine containing C-terminal y-ions become apparent in comparisons of the tandem mass spectrometry fragmentation patterns of unlabeled and labeled tryptic peptides. Modified y-ions correspond to peaks exhibiting a shift in apparent mass of +42 amu per lysine residue. Adjacent pairs of putative y-ion peaks define candidate amino acid residues, allowing the peptide sequence to be inferred. The presence of complementary b-ion peaks (which do not change in mass) provides additional support for an amino acid assignment.

2.3. Peptide Fractionation and LC-MS Analysis

1. Buffer A: 5% acetonitrile/0.2% heptafluorobutyric acid (HFBA).
2. Buffer B: 80% acetonitrile/0.2% HFBA.
3. Buffer C: 250 mM ammonium acetate/5% acetonitrile/0.2% HFBA.
4. Buffer D: 500 mM ammonium acetate/5% acetonitrile/0.2% HFBA.
5. Formic acid.
6. Fused-silica microcapillary tubing and microcolumn: 100 μ m i.d. \times 365 μ m o.d.
7. C18 reverse-phase material: 5 μ m (XDB-C18, Hewlett-Packard).
8. Strong cation exchange material: 5 μ m (Partisphere SCX, Whatman, Clifton, NJ).
9. High-performance liquid chromatography (HPLC) quaternary gradient pump, with split.
10. Ion trap or qTOF tandem mass spectrometer.

3. Methods

Trypsin digestion and modification with *O*-methylisourea is described in **Subheading 3.1**. Following modification, the peptides are pooled, fractionated by HPLC, and analyzed by online electrospray tandem mass spectrometry in a procedure common to many proteomic experiments (**Subheading 3.2**). Full scan spectra and tandem mass

spectra are collected in an automated, data-dependent manner for MCAT-modified and unmodified peptide pairs. These fragmentation spectra can be used to facilitate partial peptide sequence determination without the use of a reference protein sequence database (**Subheading 3.3.**). The last method (**Subheading 3.4.**) explains how peptide (and protein) relative abundance is determined through comparison of the total ion signal of the MCAT-modified and unmodified peptide pairs.

3.1. Trypsin Digestion and Peptide Modification

This method is used to digest biological samples to enable analysis by LC-MS. The MCAT method allows quantitative and qualitative comparison of two separate protein mixtures. Following digestion with trypsin, one mixture is treated with *O*-methylisourea, while the other is mock treated.

1. Adjust the protein solution (*see Note 1*) to pH 8.5, using 1 *M* ammonium bicarbonate if necessary.
2. Add urea to 2 *M*. This facilitates trypsin digestion by partially denaturing the proteins.
3. Add DTT to 1 mM to reduce the proteins. Incubate at room temperature for 10 min.
4. Add iodoacetamide to 10 mM to carboxyamidomethylate the free thiols. Incubate at 37°C for 20 min in the dark. Note that static modification on cysteine of +57.02 should be used in database searches and sequencing algorithms when cysteines are alkylated using this reagent.
5. Add calcium chloride to 1 mM. Calcium is essential for trypsin activity. If a calcium chelating agent (e.g., ethyleneglycol tetraacetic acid [EGTA]) is present in the protein solution, add more calcium chloride as needed.
6. Add soluble trypsin (e.g., Roche) or immobilized trypsin beads (Porozyme) to approx 1/50 stoichiometry, or according to the manufacturer's recommendations.
7. Incubate the mixture 4–24 h at 30°C with rotation or agitation.
8. Recover the peptides by solid-phase extraction using immobilized C18, followed by desolvation of the eluate to near dryness in a SpeedVac.
9. Resuspend the peptides in 500 μL aqueous solution, and adjust to pH 10 using sodium hydroxide (*see Note 2*).
10. Gently add solid *O*-methylisourea to a final concentration of 0.5 *M*, allow to dissolve, and adjust to pH 10 using a 1 *N* NaOH solution (*see Note 2*).
11. Incubate at 37°C for at least 1 h to overnight (*see Note 3*).
12. Desalt and recover the peptides again by solid-phase extraction, followed by drying in a SpeedVac.
13. Resuspend the peptides in 20 μL of buffer A, followed by the addition of 5 μL of formic acid.
14. Pool and store the modified and unmodified samples at –20°C prior to analysis.

3.2. Peptide Fractionation and Analysis by Mass Spectrometry

The MudPIT (multidimensional protein identification technology) method developed by the Yates group is used to optimally separate the modified and unmodified peptides using cation exchange and reverse-phase capillary-scale chromatography (**6,12,13**). For relatively simple proteomics mixtures (<20 proteins), a single reverse-phase column may be used.

1. Pack a microcolumn with 5–10 cm of C18 reverse-phase material followed by 5 cm of strong cation exchange (SCX) material (Partisphere SCX, Whatman, Clifton, NJ).
2. Load the desalted peptide mixture onto the column using a pressure cell.
3. Place the loaded column inline with the LC-MS system.

4. A typical 15-step chromatography run is described here (adapted from **ref. 13**). The number of steps may be increased or decreased depending on sample complexity. Step 1 is 60 min, comprising a 60-min gradient from 0 to 80% buffer B and a 10-min hold at 80% buffer B. The next 12 steps are 60 min each, with the following profile: 5 min of 100% buffer A, 2 min of x% buffer C, 3 min of 100% buffer A, a 10-min gradient from 0 to 10% buffer B, and a 400-min gradient from 10 to 45% buffer B. The 2-min buffer C percentages (x) in steps 2–13 were as follows: 10, 20, 30, 40, 50, 60, 70, 80, 90, 90, 100, and 100%. Step 14 comprises a 5-min 100% buffer A wash, a 10-min 100% buffer C wash, a 5-min 100% buffer A wash, a 5-min gradient from 0 to 10% buffer B, and a 40-min gradient from 10 to 45% buffer B. Step 15 is identical to step 14, except that the 10-min salt wash is with 100% buffer D.
5. The mass spectrometer is operated in dual mode, cycling from full scans to data-dependent MS/MS mode, wherein a full-scan mass spectrum is followed by up to three CID experiments in a row (where the parent ion target list is obtained from the previous full scan). The parent ions subject to MS/MS analysis are dynamically excluded for 1 min to prevent repeated fragmentation. Proteins from the mixture are later identified using the SEQUEST algorithm (**14**) and validated using the STATQUEST probabilistic scoring program (**15**).

3.3. De Novo Peptide Sequencing

1. The MCAT de novo sequencing approach is based on two principles: first, a short sequence of contiguous amino acid sequence from a peptide (minimally, 5–8 residues) usually contains sufficient information to identify the corresponding parental protein; second, the C-terminal daughter ions produced during MS/MS fragmentation of alternatively unmodified and *O*-methylisourea-modified lysine-containing peptides exhibit a characteristic mass differential of 42 amu. Therefore, when high-quality CID fragmentation spectra are obtained, the MCAT approach allows the y-ion series to be easily discerned. This greatly simplifies the spectral interpretation process, allowing for determination of a stretch of the amino acid sequence of a peptide to be readily deduced by assigning amino acid residue identities corresponding to the observed y-ion series by referring to a reference table of amino acid masses (see **Note 4**). The MCAT sequencing process can be formalized computationally to facilitate computer-based automation (**Fig. 3**). We describe the process here. (Computer software to automate the process of sequence prediction has been developed and is freely available for downloading and academic use at www.utoronto.ca/emililab/emililab_software.htm.)
2. The mass of the tag (or a factor of it resulting from multiple charges—i.e., 42 mass-to-charge units (Th) for +1 ions, 21 Th for +2 ions, or 14 Th for +3 ions) is added to each peak observed in the unmodified MS/MS spectrum (above some signal peak intensity threshold).
3. The spectrum of the modified peptide is then searched for peaks corresponding to these mass-tagged peaks, with all such peaks being candidate y-ions.
4. Peaks common to both spectra likely represent b-ions, or other ion products, and are excluded from the initial analysis.
5. The mass differentials observed between all candidate y-ions are calculated. Delta masses matching known masses of single or double amino acids are noted, and attempts are made to extend the sequence from this starting point in both directions (i.e., higher and lower m/z) in an iterative, stepwise manner.
6. Putative sequences can be ranked using a score function incorporating factors such as unbroken peak series, isotopic shoulders, and sites of preferred fragmentation.

7. For each predicted y-ion series, the remaining peaks (i.e., those conserved in the unmodified and modified spectra) represent candidate b-ions and therefore can be used to impose further statistical limits on the y-ion designations. The presence or absence of the complementary ion peaks can be factored into an overall score.

3.4. Determination of Peptide Relative Abundance

Pairs of peptides alternatively unmodified and modified with *O*-methylisourea can be readily told apart based on their mass differential, thereby serving as mutual internal references during MS analysis (see Note 5). In MS mode, the signal ratio between the recorded signal intensities of the lower and upper mass components of a peptide ion pair provides a direct measure of the relative abundance of the two forms of a peptide and, by inference, the corresponding proteins present in the original samples (see Note 6).

1. The peptides present in the proteome mixture are identified, using either the *de novo* method described above or database searching software (e.g., SEQUEST).
2. Parent ion masses for the identified peptides are searched for pairs that are offset by 42 (+1 ions), 21 (+2 ions), or 14 (+3 ions) Th. We have written a program called TWINPEAKS to facilitate this for peptides identified using SEQUEST. TWINPEAKS can be freely downloaded at http://www.utoronto.ca/emililab/emililab_software.htm.
3. An ion intensity profile is reconstructed for each sequenced peptide using the MS dataset and the relative abundance of modified and unmodified peptides calculated by integrating the peak area under the curve. Automated software for this can be obtained from the Aebersold group (Xpress; www.proteomecenter.org/software.php).

4. Notes

1. Careful preparation of protein and peptide solutions for electrospray MS analysis is critical to the success of a proteomic experiment. Factors to bear in mind include the need to (a) protect proteins from degradation (working at low temperature and using suitable protease inhibitors), (b) avoid excessive detergent use (detergents can suppress peptide ionization and interfere with liquid chromatography), (c) avoid protein losses resulting from excessive contact with container surfaces, and (d) keep the number of steps in the purification to a minimum.
2. The modification is pH and temperature dependent. The MCAT reaction takes place efficiently at pH 10 or greater, but is quenched at <pH 10. Reaction tubes should be carefully sealed with parafilm, and incubated at >30°C.
3. Some workers have reported guanidination of N-terminal amines upon prolonged reaction when the N-terminal residue is glycine (8).
4. Limitations to the MCAT sequencing method include the need for good-quality spectra exhibiting a rich fragmentation pattern and a near-continuous y-ion series. Furthermore, as with all *de novo* sequence efforts, some ambiguity remains, due to the isobaric or near-isobaric nature of certain amino acids (e.g., leucine and isoleucine). Of necessity, the MCAT approach is limited to peptides that terminate with a lysine residue. Tryptic fragments ending with arginine residues are not modified and, therefore, cannot be sequenced by this approach. If necessary, endoproteinase LysC can be used instead of trypsin to generate peptides ending exclusively in lysine residues (apart from peptides derived from the C-terminus). It should be noted that the presence of N-terminal or internal lysines due to incomplete protease digestion can complicate the MCAT sequencing process by causing a mass shift in at least a subset of b-ions. However, the presence of multiple modified

residues can be detected *a priori* by searching for parent ion mass shifts of multiples of 42 amu (adjusted for the charge on the ion).

5. The MCAT method compares the relative abundance of the proteins present in two mixtures. In this regard, it is analogous to a typical gene-expression profiling experiment. One sample, A, is labeled with the *O*-methylisourea tag, while the other, B, is mock-labeled with buffer containing no tagging reagent. The reciprocal experiment (i.e., A is unlabeled while B is labeled) is also carried out in order to normalize for any experimental artifact (preferential labeling of one sample, nonequal protein recovery, and so on). The relative abundances determined by both orientations of the experiment are then averaged.
6. When a peptide is observed in a labeled sample but not in the unlabeled one, or vice versa, the cognate protein may be completely absent in one sample. To confirm this, two criteria should be met: (a) the peptide should be present in the labeled fraction in one experiment orientation and the unlabeled fraction in the opposite orientation; (b) if possible, an independent peptide from the same protein should confirm this result.

Acknowledgments

The authors thank Jimmy Eng, David Schieltz, and John Yates for technical guidance, and Duy Mai, Mike Lindo, Thanuja Primariwala, Pete St. Onge, Alex Ignatchenko, and Shiva Amiri for expert assistance with software programming.

References

1. Aebersold, R. and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature* **422**(6928), 198–207.
2. Kislinger, T. and Emili, A. (2003) Going global: protein expression profiling using shotgun mass spectrometry. *Curr. Opin. Mol. Ther.* **5**(3), 285–293.
3. Ong, S-E., Blagoev, B., Kratchmarova, I., et al. (2002) Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* **1**(5), 376–386.
4. Munchback, M., Quadroni, M., Miotto, G., and James, P. (2000) Quantitation and facilitated de novo sequencing of proteins by isotopic N-terminal labeling of peptides with a fragmentation-directing moiety. *Anal. Chem.* **72**, 4047–4057.
5. Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelf, M. H., and Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* **17**, 994–999.
6. Cagney, G. and Emili, A. E. (2002) De novo peptide sequencing and quantitative profiling of complex protein mixtures using mass-coded abundance tagging. *Nat. Biotechnol.* **20**(2), 163–170.
7. Beardsley, R. L. and Reilly, J. P. (2003) Quantitation using enhanced signal tags: a technique for comparative proteomics. *J. Proteome Res.* **2**, 15–21.
8. Beardsley, R. L., Karty, J. A., and Reilly, J. P. (2000) Enhancing the intensities of lysine-terminated tryptic peptide ions in matrix-assisted laser desorption/ionization mass spectrometry. *Rapid Commun. Mass Spectrom.* **14**, 2147–2153.
9. Brancia, F. L., Oliver, S. G., and Gaskell, S. J. (2000) Improved matrix-assisted laser desorption/ionization mass spectrometric analysis of tryptic hydrolysates of proteins following guanidination of lysine-containing peptides. *Rapid Commun. Mass Spectrom.* **14**, 2070–2073.
10. Hale, J. E., Butler, J. P., Knierman, M. D., and Becker, G. W. (2000) Increased sensitivity of tryptic peptide detection by MALDI-TOF mass spectrometry is achieved by conversion of lysine to homoarginine. *Anal. Biochem.* **287**, 110–117.

11. Keough, T., Lacey, M. P., and Youngquist, R. S. (2000) Derivatization procedures to facilitate de novo sequencing of lysine-terminated tryptic peptides using postsource decay matrix-assisted laser desorption/ionization mass spectrometry. *Rapid Commun. Mass Spectrom.* **14**, 2348–2356.
12. Link, A., Eng, J., Schieltz, D. M., et al. (1999) Direct analysis of proteins complexes using mass spectrometry. *Nat. Biotechnol.* **17**, 676–682.
13. Washburn, M. P., Wolters, D., and Yates, J. R. III. (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19**, 242–247.
14. Eng, J. K., McCormack, A. L., and Yates, J. R. I. (1994) An approach to correlate tandem mass-spectral data of peptides with amino-acid-sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **11**, 976–989.
15. Kislinger, T., Rahman, K., Radulovic, D., Cox, B., Rossant, J., and Emili, A. (2003) PRISM, a generic large scale proteomic investigation strategy for mammals. *Mol. Cell. Proteomics* **2**(2), 96–106.

Virtual 2-D Gel Electrophoresis by MALDI Mass Spectrometry

Angela K. Walker, Gary Rymar, and Philip C. Andrews

1. Introduction

Two-dimensional (2-D) gel electrophoresis (2-DE) is capable of separating several hundred to thousands of proteins on a single gel, depending on the format (1,2). The most commonly used methods for visualization of proteins in gels include Coomassie Brilliant Blue (CBB), silver stains, or fluorescent stains (1–3). Each staining method exhibits different levels of overall sensitivity, and they vary widely in their relative sensitivities for specific proteins. Two-dimensional gels also allow only the approximate sizes of proteins to be determined. It is desirable to develop methods that combine the high resolution of electrophoretic separations with the accurate mass measurements that can be provided by mass spectrometry (MS).

In our laboratory, we have developed a methodology to produce “virtual” 2-D gels (V2DGs) by utilizing conventional first-dimension separation of isoelectric focusing (IEF) using thin-layer immobilized pH gradient (IPG) gels and replacing the conventional second-dimension electrophoretic separation (sodium dodecyl sulfate [SDS]-polyacrylamide gel electrophoresis [PAGE]) with matrix-assisted laser desorption/ionization (MALDI)-MS, which provides more accurate molecular masses (4–8), independent of changes in the abilities of proteins to bind SDS. The methods described in this chapter detail the steps required to prepare an IPG for MALDI-MS analysis, to acquire mass spectra from the gel, and to assemble the spectra into a “virtual” 2-D gel. This approach also provides an alternative method for visualization of proteins from gels that can be complementary to classical 2-D gels.

2. Materials

2.1. Isoelectric Focusing in IPG Gels

1. Cell extract (or other complex protein sample).
2. Rehydration buffer: 7 M urea, 2 M thiourea, 0.8% CHAPS, 1.0% Triton X-100, trace bromophenol blue.
3. Dithiothreitol (DTT).
4. 18-cm, pI 4–7 IPG gels.
5. IPG buffer.
6. IEF gel apparatus.

2.2. Processing of IPG Gels for MALDI-MS Analysis

Chemicals should be of high-performance liquid chromatography (HPLC) grade or better.

1. Sinapinic acid (*see Note 1*).
2. Acetonitrile.
3. Isopropanol.
4. *n*-Propanol.
5. Formic acid.
6. Trifluoroacetic acid (TFA).
7. Deionized water.
8. Conductive adhesive (Adhesives Research, Glen Rock, PA) or glue stick.

3. Methods

The following subheadings describe (1) isoelectric focusing of the proteins in IPG gels; (2) processing of IPG gels for MALDI-MS analysis; (3) acquisition of mass spectra from IPG gels; and (4) assembly and visualization of the spectra in a “virtual” 2-D gel. The first two subheadings are generic, but **Subheadings 3. and 4.** are specific to the Applied Biosystems MALDI-time-of-flight (TOF) mass spectrometers used in our laboratory.

3.1. Isoelectric Focusing of Proteins in IPG gels

Typically, IPG gels are supplied by the manufacturer as washed and dried gels cast on a plastic support. Gels are rehydrated with an appropriate solution to maintain proteins in a denatured state. Reduction of disulfide bonds prior to IEF results in separation of protein polypeptides. In the following protocol, proteins (polypeptides) are loaded into the IPG gel during the rehydration process. Volumes provided are for preparing 12 IPG gels with an 18-cm separation distance.

1. To 4.875 mL rehydration buffer, add 325 μ L protein solution containing approx 700 μ g of protein. Each gel will contain approx 50 μ g protein.
2. Add DTT to 100 mM (*see Note 2*).
3. Add pH 3.0–10.0 IPG buffer (0.5%, v/v, final).
4. In a gel rehydration tray, rehydrate each 18-cm IPG gel with 400 μ L of rehydration buffer, gel side down, for 18 h.
5. IEF is for 50,000–100,000 Volt-hours, per manufacturer’s directions, under a layer of mineral oil.
6. Following IEF, store IPG gels at –80°C.

3.2. Processing of IPG Gels for MALDI-MS Analysis

Co-crystallization of matrix and analyte is necessary for a MALDI-MS analysis to be successful. In a typical MALDI-MS experiment, the protein and matrix solutions are co-deposited onto the surface of the MALDI target plate, and incorporation of the proteins into the matrix occurs during the crystallization process. In IPG gels, the proteins are located within the gel, presenting a twofold problem for MALDI-MS analysis: the proteins must co-crystallize with the matrix, and the proteins must be at or near the surface of the gel (the ultraviolet [UV] laser is unable to penetrate very far into the gel) to successfully acquire MALDI mass spectra. Commercial IPG strips are rela-

tively thin layers and have a low percentage of acrylamide, both of which improve diffusion of proteins to the surface of the gel, but these features alone are not sufficient to allow efficient MALDI directly from gels. Additionally, a number of chemicals (detergents or ampholines) used in running IPG gels can interfere with the MALDI ionization process, and oil may interfere with matrix crystallization. For these reasons, the majority of the procedures we have evaluated focus on optimizing the co-crystallization of the protein and the matrix on the surface of the gel.

1. Prepare matrix: saturated sinapinic acid in 2:3:3:2 formic acid:H₂O:acetonitrile:isopropanol (or *n*-propanol) (FWAI or FWAN) (*see Notes 3 and 4*).
2. Remove IPG gels from -80°C freezer and thaw 5–10 min at room temperature.
3. Place in a tray with approx 20–30 mL of DI H₂O and wash twice for 1 min to remove surface oils (*see Note 5*).
4. Transfer gel to a clean tray.
5. Add matrix solution and soak 10 min (*see Note 6*).
6. Pour off matrix solution and allow IPG gel to dry approx 10 min. Fine crystals should begin to form.
7. Layer approx 200 μL of matrix onto the surface of gel (*see Note 7*).
8. Allow to dry, loosely covered, for 3–5 d at ambient temperature (*see Notes 8–10*).
9. Cut into five pieces and mount onto MALDI plate using glue stick or conductive adhesive tape (*see Notes 11 and 12*).

These steps result in a dry, stable IPG gel with a homogenous layer of fine matrix crystals coating the surface.

3.3. Acquisition of Mass Spectra From IPG Gels

Spectra have been acquired from IPG gels using an Applied Biosystems (ABI) Voyager-DE STR and an ABI Voyager-DE Pro. Each instrument is controlled by the Voyager Control Panel software, which utilizes plate files (*.plt) to describe the size, shape, and location of each well on a MALDI plate. A custom *.plt file must be created to represent the layout of the gels on the MALDI plate. The software uses this file to create a graphic representation of the gel strips and the “wells” (**Fig. 1**) and to provide coordinates from which to acquire spectra in an automated run.

1. The coordinates of the left, right, top, and bottom edges of each strip of gel are determined by firing the laser and visually observing where the boundary of the gel is located.
2. An ABI *.plt file is created based on these coordinates and the desired width of the wells. Typically our wells are defined to be 0.25 mm wide, for a total of approx 720 spectra across an 18-cm gel (*see Note 13*).
3. Spectra can be collected manually by using the joystick to maneuver around the gel within each of the wells. However, this method is extremely time-consuming, and automatic acquisition is preferred (*see Note 14*).
4. The pI for each spectrum is calculated based on the well from which it was acquired.

3.4. Processing and Assembly of Mass Spectra into V2DGs

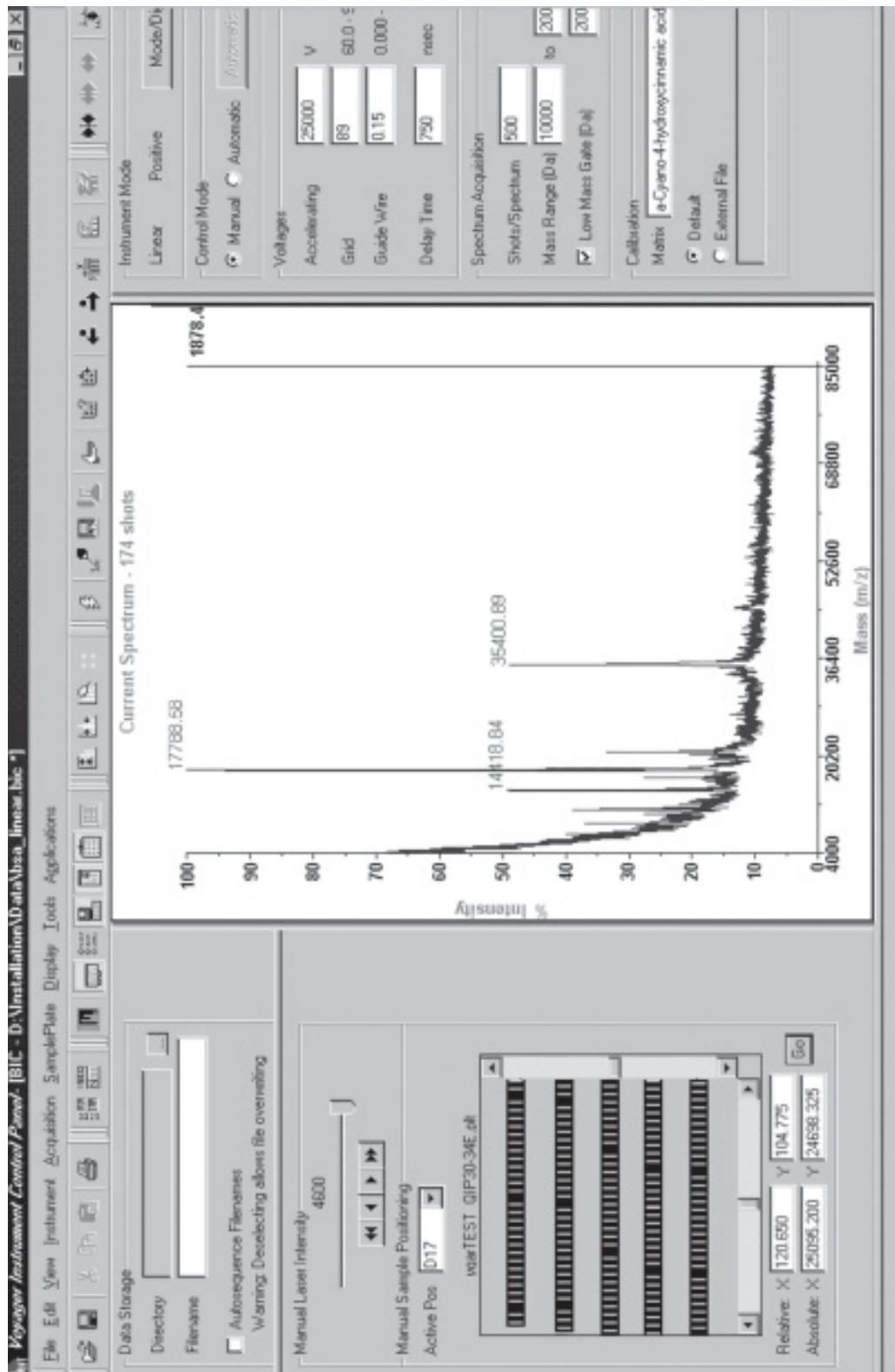
Prior to visualization, spectra should be smoothed and baseline corrected. This process can be performed manually, but due to the large number of spectra acquired, it has been automated using software written in-house. For visualization of the spectra, the software Transform (Research Systems, Inc.) can be used to create a virtual gel image.

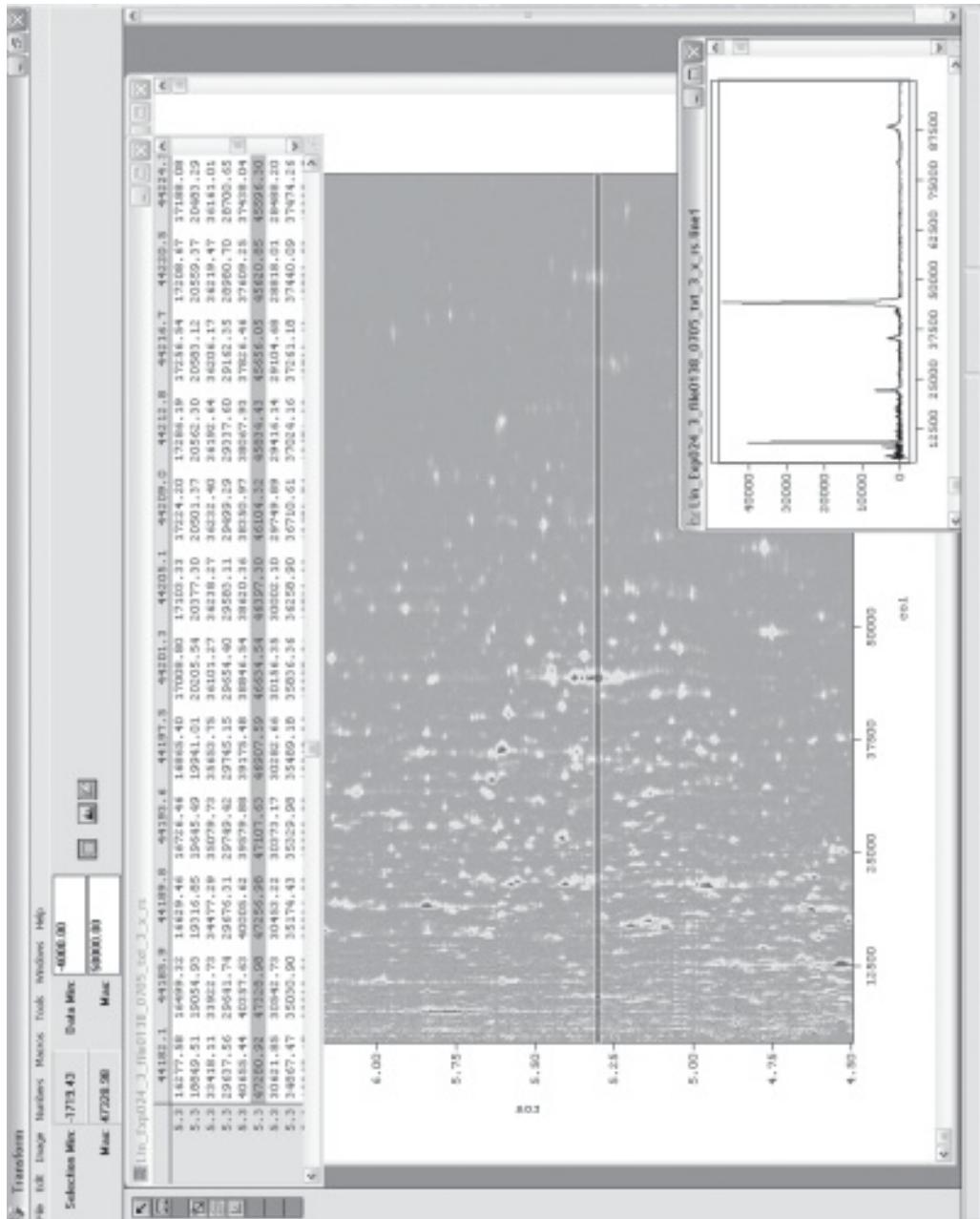
1. Spectra are converted to ASCII text files in which the m/z and intensity are listed as x,y data points (see **Note 15**).
2. Each spectrum is baseline corrected and subjected to a 51-point Savitsky-Golay smooth (see **Note 16**).
3. A single text file is created by catenating the individual ASCII spectra. This file consists of three columns: the first column is the m/z, the second column is the pI (or mm position) at which the spectrum was collected, and the third column is the intensity. Depending on the number of spectra acquired, this file is 500–800 Mb in size.
4. The text file is opened in Transform, which converts the data in the three columns into a matrix. It then uses this matrix to create a 2-D image in which one axis is the m/z, the second axis is the pI, and the intensities are visualized by the color of the spots in the image (**Fig. 2**). The m/z values may also be converted into their log values so that the image will be more similar to a classical 2-D gel.
5. The image is then saved and rotated into the appropriate orientation for visualization as a 2-D gel image. Resizing may be performed in most image-analysis software.

4. Notes

1. The trans isomer (>98%) of sinapinic acid was used instead of the more common racemic mixture. The racemic form has not been tested in our laboratory for use with virtual gels, but no difference in spectrum quality has been observed between the two forms when used with protein standards. The sinapinic acid was recrystallized to remove impurities, but comparisons between the spectra acquired using recrystallized and nonrecrystallized matrix have yielded no significant differences.
2. The effect of the reducing agents DTT and tris(2-carboxyethyl) phosphine (TCEP) on spectrum quality have been compared. The use of DTT as reductant appears to result in increased signal intensities (**Fig. 3**).
3. The solvent conventionally used for the matrix for typical protein samples is 1:1 acetonitrile:0.2% TFA. Using this solvent, the number of proteins observed from the gels is much lower than what would be expected from an *Escherichia coli* whole-cell preparation. The addition of 5% *n*-propanol to the matrix solvent mixture increased the number and intensity of peaks in the mass spectra, presumably a result of increased protein solubility (**Fig. 4**).
4. The effect of solvent composition on the quality of spectra acquired from IPG gels was investigated further. Cohen and Chait (9) studied the effects of matrix solvent composition on the selectivity of peptides in a complex mixture using MALDI-MS. Significant differences in the number and identity of peptides were observed, depending on the solvent mixture composition. In our studies, 2:3:3:2 FWAI has been found to be the most effective solvent for observing proteins in the 5- to 100-kDa mass range (**Fig. 5**). Substitution of *n*-propanol for isopropanol also works well, but fewer proteins above 30 kDa are observed, whereas the number and intensities of the proteins below approx 20 kDa significantly increases (**Fig. 6**).
5. A number of wash conditions have been evaluated, including rinsing vs soaking, solvent composition, and wash time. Rinsing with DI H₂O has proven insufficient to remove the

Fig. 1. (opposite page) The Applied Biosystems Voyager Control Panel software generates a graphical representation of the wells in a gel strip on the matrix-assisted laser desorption/ionization (MALDI) plate. This diagram is created from the parameters in the plate definition file (.plt), which defines the size shape, and coordinates of the individual wells on the gel strips. A custom .plt file must be created each time a gel is analyzed.





surface oil. Extensive wash times and the use of solvents such as 1:1 MeCN:0.2% TFA result in fewer proteins observed in the mass spectrum, presumably due to lateral diffusion of proteins within the gel and/or diffusion of proteins out of the gel and into the wash solution.

6. The duration of the matrix soak time has a significant effect on the quality and homogeneity of the matrix crystals that form on the surface of the gel. Gels on which the matrix has formed fine, homogenous crystals on the surface of the gel provide the most reproducible and highest-quality spectra. IPG gels on which the matrix has formed large visible crystals may yield spectra, but the results tend to be irreproducible, and it becomes necessary to find "sweet spots" from which to acquire spectra. Shorter soaks (approx 7 min or less) result in very sparse crystallization of the matrix on the surface of the gel. Longer soaks (15 min or more) typically result in the formation of large matrix crystals and/or fewer proteins observed in the spectrum. The reduction of proteins observed in the spectrum with the longer soak times may be attributed to either the formation of large matrix crystals or from diffusion of proteins out of the gel. A soak time of 10 min has produced the best-quality crystals, but the results are not highly reproducible unless an additional matrix layer is used (see **Note 7**).
7. The matrix soak was eliminated and matrix solution was layered onto the surface of the gel to determine whether diffusion of proteins into the matrix solution was occurring. As with short matrix soaks, layering did not result in sufficient matrix crystallization on the gel surface. Layering matrix on the surface of the gel following a 10-min matrix soak results in the formation of a very homogenous layer of fine matrix crystals, and crystallization is more reproducible than with the matrix soak alone.
8. Gel instability in the high vacuum of the mass spectrometer can be a severe problem. In early work, gels cracked and peeled off of their backing in the instrument. The addition of glycerol to the matrix was evaluated to improve gel stability, but poor matrix crystallization and low-quality mass spectra were observed.
9. Gel instability was initially attributed to over-drying of the polyacrylamide in the high vacuum of the mass spectrometer. However, for gels dried at ambient temperature for 3–5 d prior to loading into the mass spectrometer, no cracking has been observed. This suggests that the cracking is not caused by the gels becoming too dry, but rather from drying too rapidly in the high vacuum of the instrument.
10. Gels should be dried loosely covered. In a sealed container, very large matrix crystals form and poor-quality mass spectra are acquired.
11. Conductive adhesive was initially used to reduce the likelihood of charge buildup on the gel surface when high voltage is applied to the MALDI plate. More recent studies have used a thin layer of glue from commercially available glue sticks, which provides greater visibility of the gel and the crystals. No significant problems have resulted from the use of glue sticks, though systematic evaluations have not been carried out.
12. In order to achieve a gel surface flush with that of the plate, the center portion of an ABI MALDI plate was routed out to a depth of approx 1 mm. This eliminates the need for changes to the laser incidence angle and minimizes changes to instrument settings.
13. A program has been written using LabVIEW (National Instruments, Austin, TX) that takes as input the coordinates of the edges of each of the gel strips and the desired number of

Fig. 2. (opposite page) The software package Transform creates a matrix containing the m/z and intensity values from all of the spectra acquired from the immobilized pH gradient (IPG) gel. It uses this matrix to create a "virtual" two-dimensional gel image. The spectrum acquired from any given position on the IPG gel can also be viewed.

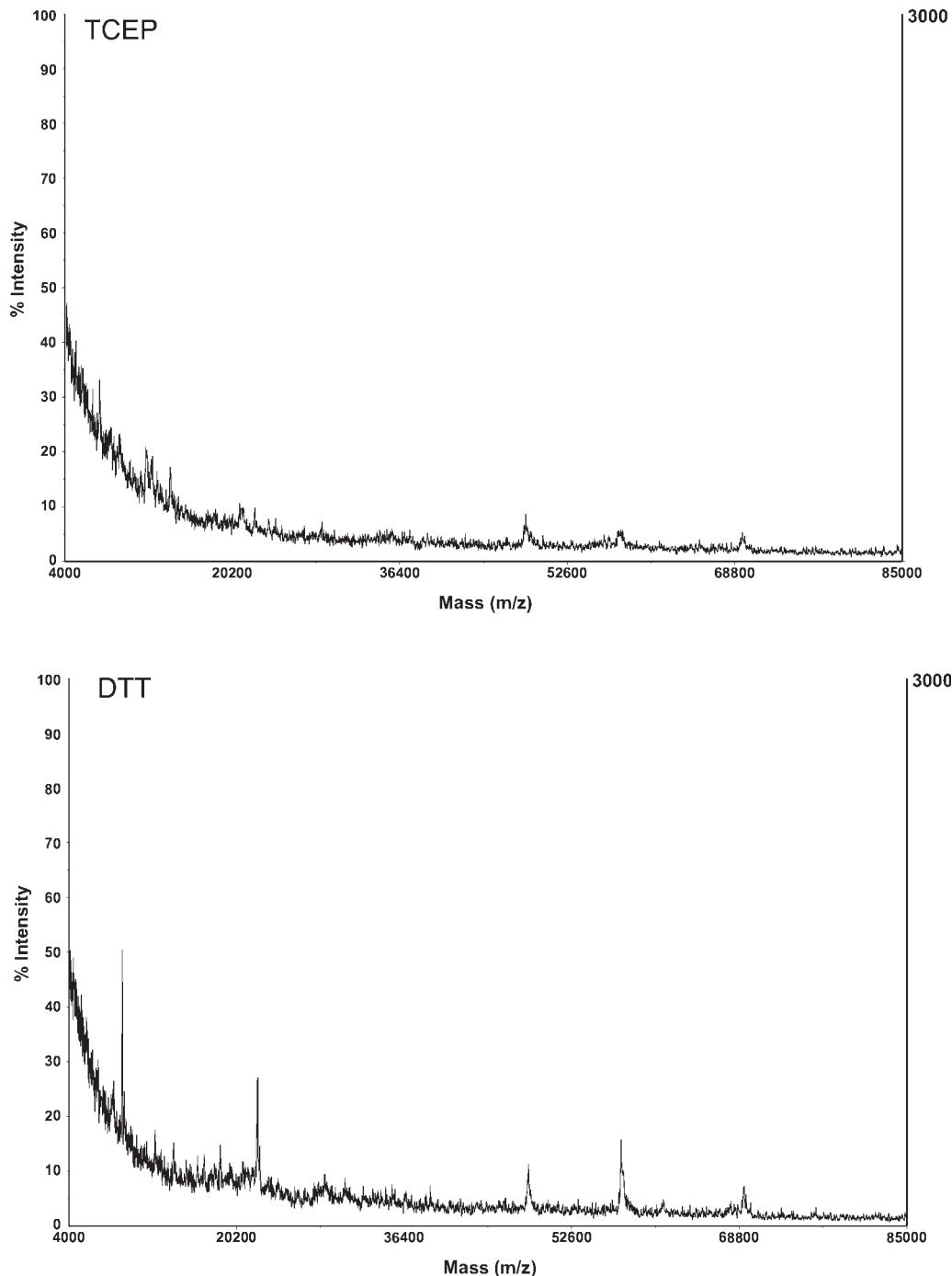


Fig. 3. Spectra illustrating the effects of using dithiothreitol vs tris(2-carboxyethyl) phosphine as a reductant in the sample preparation. The spectra were collected at the same mm position on identically run immobilized pH gradient gels.

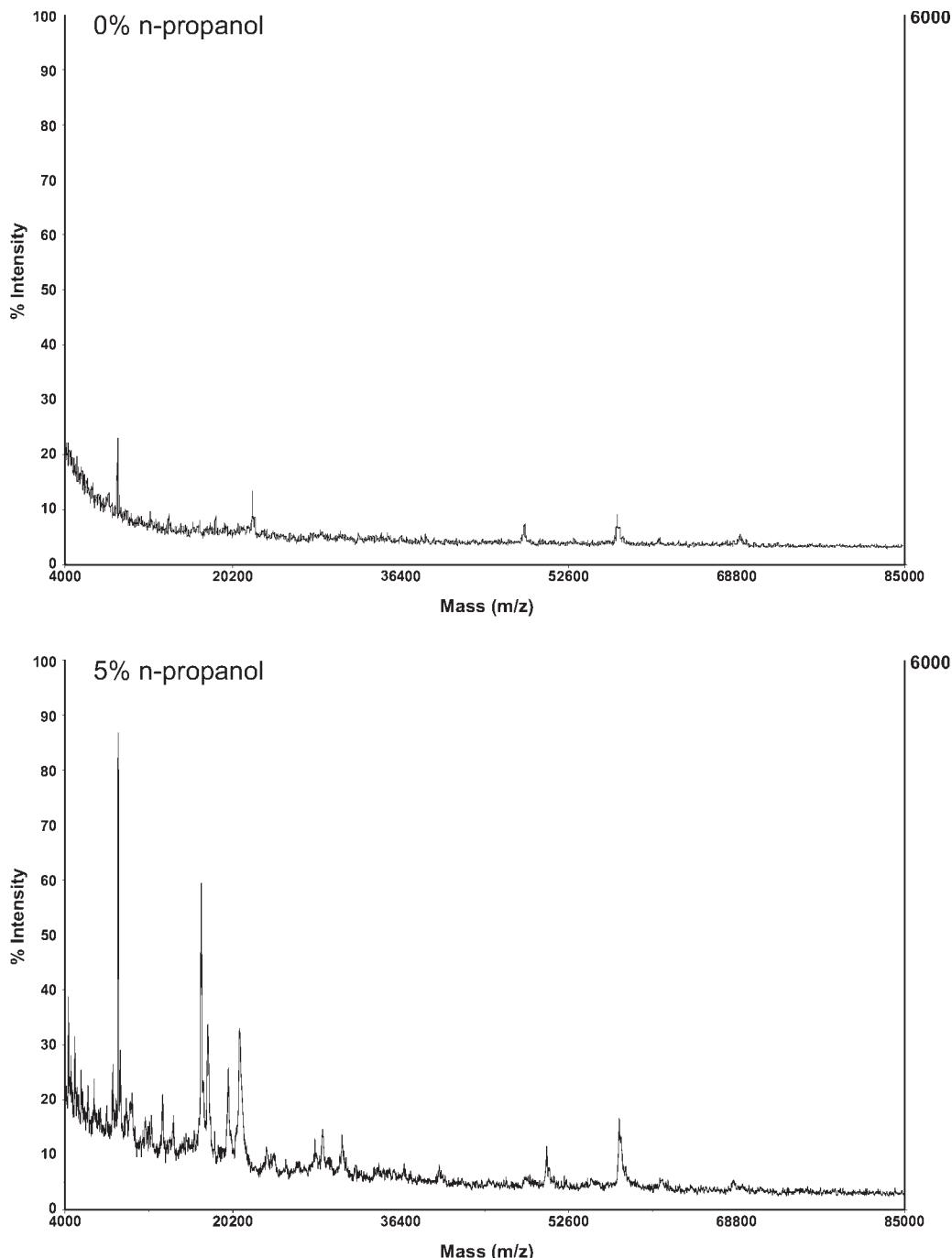


Fig. 4. Spectra illustrating the effects of adding 5% *n*-propanol to the 1:1 acetonitrile:0.2% trifluoroacetic acid solvent mixture for the matrix. The spectra were collected at the same mm position on identically run immobilized pH gradient gels.

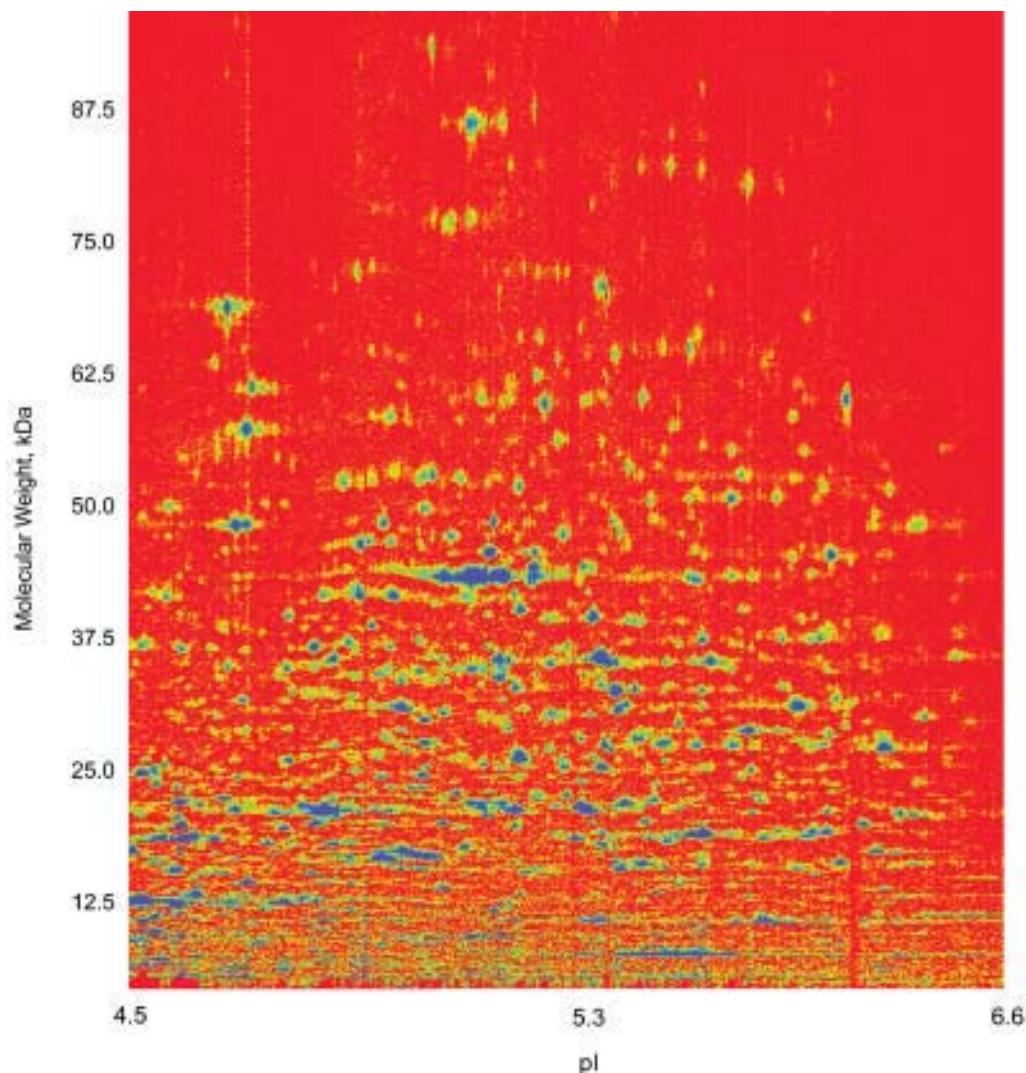


Fig. 5. A virtual two-dimensional gel produced using 2:3:3:2 formic acid:H₂O:MeCN:-isopropanol (FWAI) as the solvent mixture for the matrix.

- wells per gel and creates a properly formatted ABI .plt file from the information provided. Alternatively, this file can be created in Excel (or a text editor) and saved in the appropriate format. (All LabVIEW programs are available upon request.)
14. V2DGs assembled from manually acquired spectra often appear streaky (Fig. 7) because the acquired spectrum from a given well is often the best spectrum rather than an average spectrum. This occurs because a well that has very little protein can often yield a reasonable spectrum if a sweet spot is located manually. In an automated acquisition, the software is set to accumulate a set number of acceptable spectra from several spots within a well. This results in a random sampling of spectra, and V2DGs assembled in this fashion appear less streaky and proteins appear as discrete spots in the image (Figs. 5 and 6).
 15. The data format of the ABI *.dat spectrum file is proprietary and can be viewed only by using ABI's Data Explorer software. It is necessary to convert the files from the propri-

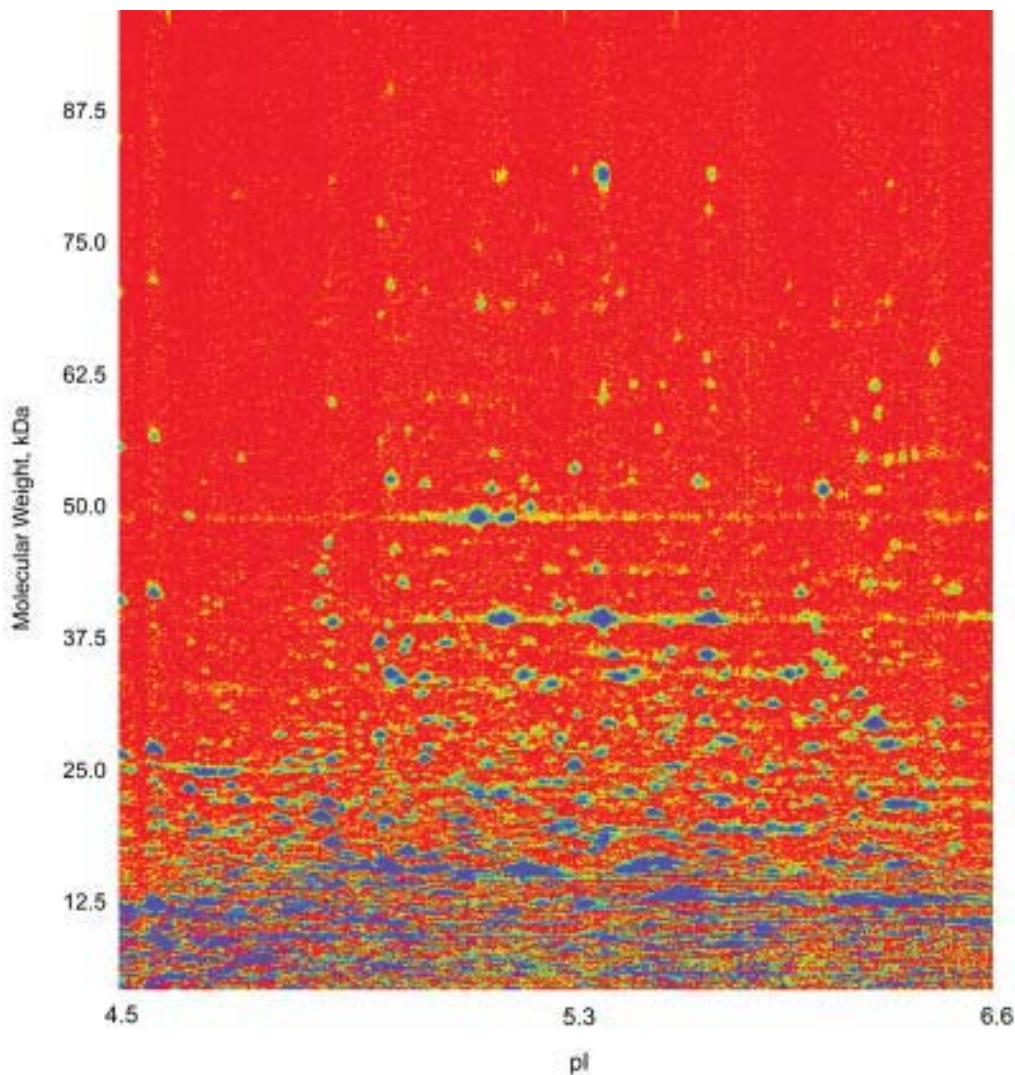


Fig. 6. A virtual two-dimensional gel produced using 2:3:3:2 formic acid:H₂O:acetonitrile:-*n*-propanol (FWAN) as the solvent mixture for the matrix.

etary *.dat format to ASCII format so the data can be utilized by third-party software. A program written using LabVIEW drives Data Explorer by taking advantage of the Common Object Model (COM) interface that Data Explorer exposes for purposes of automation. It calls Data Explorer and instructs it to open each spectrum and perform the conversion to an ASCII text file in a batch process.

16. A program written in LabVIEW performs peak detection, smoothing, and baseline correction on the ASCII spectrum files. Alternatively, Data Explorer can be driven to perform these tasks prior to converting the spectra to ASCII text files.

References

1. Wirth, P. J. and Romano, A. (1995) Staining methods in gel electrophoresis, including the use of multiple detection methods. *J. Chrom. A* **698**, 123–143.

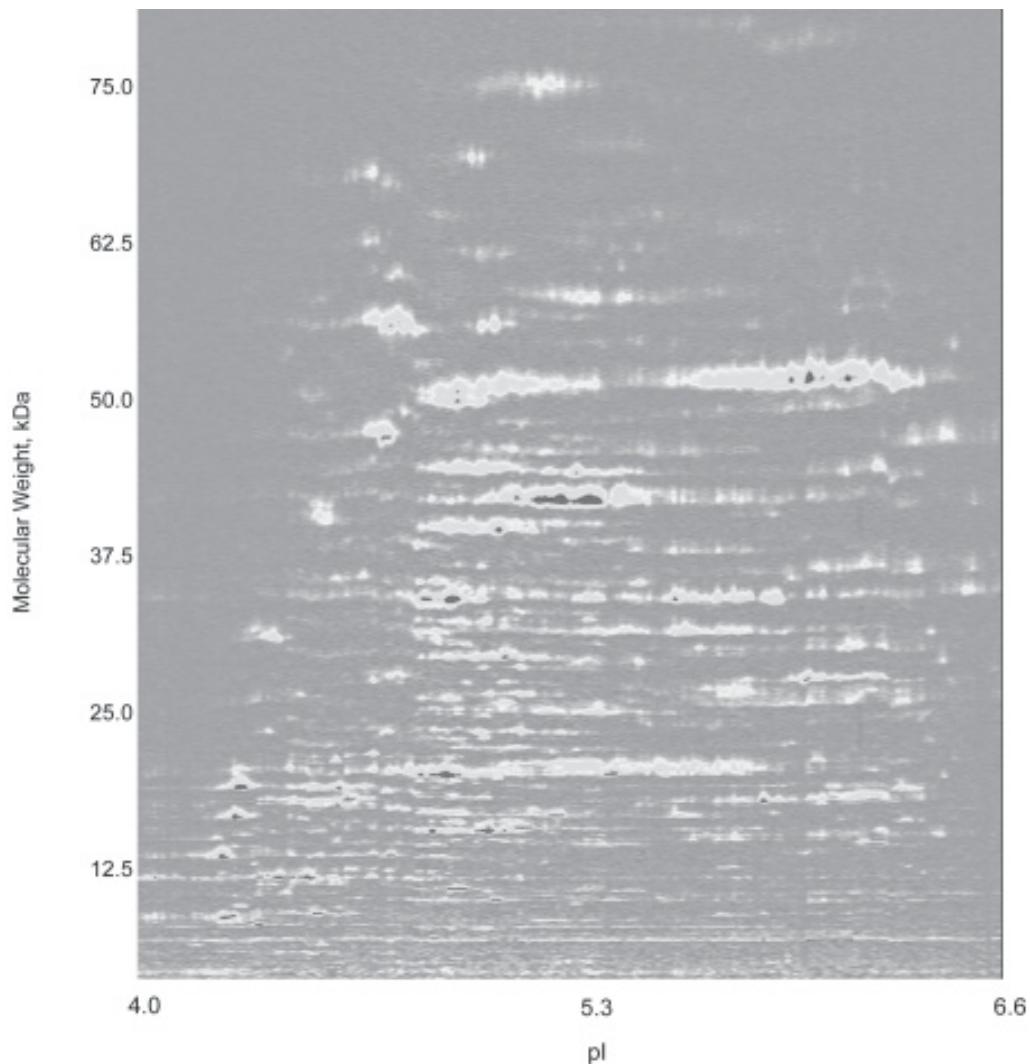


Fig. 7. Virtual two-dimensional gels acquired manually often result in images in which proteins appear as streaks rather than discreet spots on the gel.

2. Haynes, P. A. and Yates III, J. R. (2000) Proteome profiling—pitfalls and progress. *Yeast* **17**, 81–87.
3. Patton, W. F. (2000) A thousand points of light: The application of fluorescence detection technologies to two-dimensional gel electrophoresis and proteomics. *Electrophoresis* **21**, 1123–1144.
4. Ogorzalek Loo, R. R., Andrews, P. C., and Loo, J. A. (1993) Gels in vacuo? A minimalist approach for combining mass spectrometry and polyacrylamide gel electrophoresis. In: (Larsen, B. S. and McEwen, C. N., eds), *Mass Spectrometry of Biological Materials*, Marcel Dekker, New York: pp. 325–343.
5. Loo, J. A., Brown, J., Critchley, G., Mitchell, C., Andrews, P. C., and Ogorzalek Loo, R. R. (1999) High sensitivity mass spectrometric methods for obtaining intact molecular weights from gel-separated proteins. *Electrophoresis* **20**, 743–748.

6. Ogorzalek Loo, R. R., Stevenson, T. I., Mitchell, C., Loo, J. A., and Andrews, P. C. (1996) Mass spectrometry of proteins directly from polyacrylamide gels. *Anal. Chem.* **68**, 1910–1917.
7. Walker, A. K. and Andrews, P. C. (2001) Mass spectrometric imaging of immobilized pH gradient gels and creation of “virtual” two-dimensional gels. *Electrophoresis* **22**, 933–945.
8. Ogorzalek Loo, R. R., Cavalcoli, J. D., VanBogelen, R. A., et al. (2001) Virtual 2-D gel electrophoresis: Visualization and analysis of the *E. coli* proteome by mass spectrometry. *Anal. Chem.* **73**, 4063–4070.
9. Cohen, S. L. and Chait, B. T. (1996) Influence of matrix solution conditions on the MALDI-MS analysis of peptides and proteins. *Anal. Chem.* **68**, 31–37.

Identification of Posttranslational Modification by Mass Spectrometry

Alastair Aitken

1. Introduction

Many chemically distinct types of posttranslational modification of proteins are known. These include the wide variety of prosthetic groups at enzymes' active sites, acylations at the N- and the C-terminus and at internal sites in proteins. In this chapter, the examples of mass-spectrometry techniques employed for analysis of glycosylation and disulphide bonds are given. Techniques for identification of sites of phosphorylation are in Chapter 44.

The great majority of proteins analyzed by mass spectrometry are modified to some extent. In addition to the endogenous (i.e., physiological) posttranslational modifications, there are many artifacts due to sample handling, such as oxidation of cysteine and methionine. Modifications may also be deliberately introduced during sample work-up, such as cysteine derivatization to avoid partial conversion to the acrylamide adduct (as well as oxidation) if the sample is separated by gel electrophoresis. Certain modifications are unavoidable owing to the digestion protocol—e.g., homoserine and its lactone at the C-terminus of internal peptides as a consequence of cyanogen bromide (CNBr) digestion (1).

A list of the broad chemical diversity of known modifications and the side chains of the amino acids to which they are attached (2,3) is in the website “Delta Mass” (*see* list at end of this chapter). Many modifications have the same unit mass, and a particular problem may be the inability to distinguish between phosphorylation (79.966 Da) and sulphation (79.957 Da). This difference in exact mass may be distinguished by the very high mass accuracy obtainable in Fourier transform mass spectrometry (FT MS, *see Note 1*).

Posttranslational modifications can also be detected by matrix-assisted laser desorption/ionizaton (MALDI)-time-of-flight (TOF) measurement of the differences in mass of particular peptides before and after the removal of the modification, by addition of phosphatases (*see* Chapter 44) and glycosidases (*see Note 2*), for example.

The presence of a modification in a peptide may be suspected due to a change in expected relative high-performance liquid chromatography (HPLC) elution position. This is particularly applicable to on-line LC-MS (Chapter 37). A number of examples of some of these changes in elution position caused by a wide variety of modifications such as phosphorylation (Chapter 44), myristylation, acetylation, and cyclization of

Gln to pyro-Glu are listed in (1). The stability of the modifications varies enormously, but under suitable conditions, mass spectrometry can identify almost all of these, even the very labile modifications. This is in contrast to amino acid analysis after hydrolysis (normally using strong acid), which can nevertheless identify a large number of modified amino acids by their unique elution position in an amino acid analyzer (1). Many modifications can also be directly identified by the characteristic elution of their phenylthiohydantoin (PTH) derivatives during Edman degradation (1).

2. Materials

1. Mass spectrometer(s).
2. MALDI plates.
3. MALDI matrix such as 2,5-dihydroxybenzoic acid (DHB, Aldrich).
4. Low-volume pipet such as a 2 μ L Gilson or FinnPipette.
5. High-purity water.
6. 30% aq acetonitrile (ACN) containing 0.1% trifluoroacetic acid (TFA).
7. Microcentrifuge.
8. 0.5-mL Eppendorf tubes.
9. Vortex mixer.
10. Pepsin, Glu-C, and cyanogen bromide for proteolysis.
11. *N*-glycosidase, PNGase F (New England Biolabs).
12. Networked personal computer terminal with fast configuration.
13. Mass spectrometer data-handling software.
14. Other materials are listed in Chapters 30, 37, and 44.

3. Method

Methods for protein digestion, extraction of peptides, and obtaining peptides for mass fingerprinting and/or sequencing by tandem electrospray MS or MALDI-TOF-MS are described in Chapters 30–33, and 37. When carrying out mass fingerprint database searches (see Chapter 35), one can allow for the possibility of a wide range of modifications. For example, one can input any “user defined modification” on the list on the form page for the MS-Fit Website search engine (e.g., esterification of Asp and Glu; pyridoxyl-Lys; myristylation of Lys and N-terminal Gly; acetylation; mono-, di-, and trimethylation of Lys; nitration and sulphation of Tyr; phosphorylation of any combination of Ser, Thr, or Tyr). For example, if phosphorylation of Ser and Thr is chosen from the list, then the occurrence of any serine or threonine is considered as possibly either serine, threonine, phosphoserine, or phosphothreonine (see Note 3).

3.1. Mass Spectrometry of Glycosylation Sites and Structures of the Sugars

The attachment points of glycosylation sites *N*-linked (through asparagine) and *O*-linked (through serine), and the structures of the complex carbohydrates, can be determined by mass spectrometry (4).

1. Adjust the concentration of the sample to the optimum. This is probably between 20 and 100 pmol/ μ L for carbohydrates and glycosylated polypeptides (this will give a final concentration of 10–50 pmol/ μ L after mixing 1:1 with the matrix).
2. If the concentration is unknown, make a series of dilutions in water.
3. Prepare the stock DHB matrix solution (10 μ g/ μ L) in 30% aq ACN containing 0.1% TFA (see Note 4).

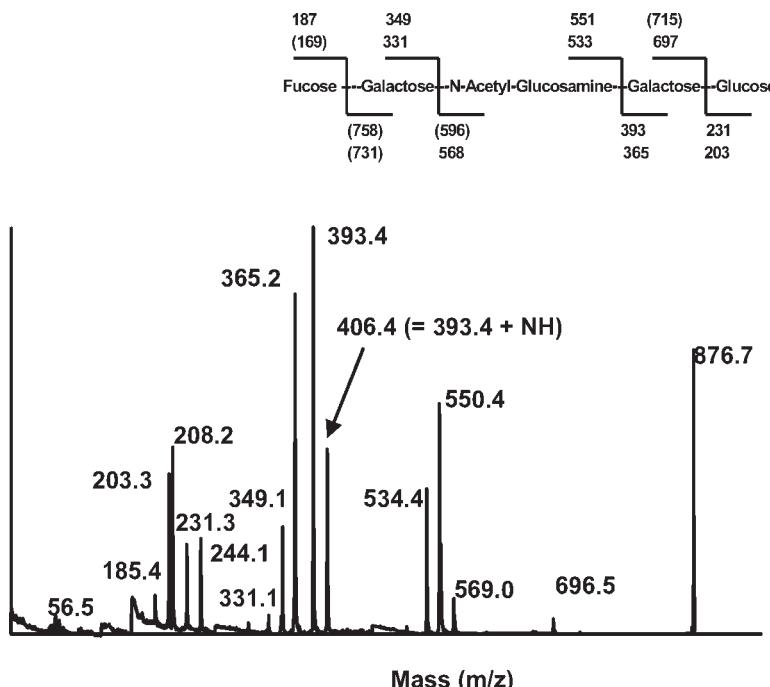


Fig. 1. The post-source decay (PSD) mass spectrometry spectrum of a glycan side chain. The compound Fuc1-2Gal-3GlucNAc1-3Gal1-4Glc was run by Dr Andy Cronshaw (EPIC, Edinburgh University) on an ABI DE-STR matrix-assisted laser desorption/ionization (MALDI)-time-of-flight (TOF) with 2,5-dihydroxybenzoic acid as matrix. The experimentally derived fragment masses are mainly within 1 Da of the theoretical. The masses in parentheses were not seen in this experiment. A number of abrupt changes in baseline correspond to where the PSD spectra have been “stitched” together (see Chapters 32 and 33). The parent ion peak at 876.7 Da is a result of the mass of the intact molecule as a sodium adduct, $[M + Na]^-$.

4. Apply equal volumes of the sample ($0.5 \mu\text{L}$) and matrix solution ($0.5 \mu\text{L}$) to the MALDI plate and analyze by MALDI-TOF (see Chapter 32).
5. To verify the identity of the glycosylated peptides, compare the mass spectra of the digest mixture before and after deglycosylation of the peptides. Use (typically) 100 U PNGase F in 25 mM phosphate buffer, pH 7.4. Incubate at 37°C for 18 h (see Note 2).
6. Post source-decay (PSD)-MS (see Chapter 32) can produce sequence information, and the loss of each monosaccharide unit of distinct mass can be interpreted from the spectra using the “glycomod” Website (see list at end of chapter) to reconstruct the glycosylation pattern (see example in Fig. 1).

3.2. Identification of Disulphide Linkages by Mass Spectrometry

1. Fragment the protein into peptides under low pH conditions to minimize disulphide exchange using pepsin, Glu-C, or cyanogen bromide (see Note 5).
2. Separate one-half of the digest and identify the disulphide-linked peptide fragments under mild oxidizing conditions (i.e., no reducing agents) by HPLC mass spectrometry (LC-MS).
3. Reduce an identical aliquot of the digested peptides with dithiothreitol (DTT) (100 mM) for 2 h at 37°C to cleave $-S-S-$ bonds
4. Re-analyze the digest mixture by on-line LC-MS as before (5).
5. Identify the peptide(s) that disappear from the spectrum. These were disulphide linked.

6. Examine the peptides that appear at the appropriate positions for the individual components. The sum of their masses should correspond to the disulphide linked-peptides (see Note 6).

3.3. Selected Ion Monitoring (SIM)

Selected ion monitoring (SIM) is used to look for ions that are characteristic of a particular posttranslational modification in a peptide (5). This technique has particular application for on-line LC-MS, where the instruments can be set up to monitor selected ion masses as the components elute successively from the capillary LC or reverse-phase HPLC column. The precise parameters depend on the type of instrument and the software program available (see Chapter 37).

1. Use the instrument software to set up detection algorithms to carry out tandem MS on each component as it elutes from the chromatography column.
2. Adapt this to enable selective detection of the low-mass fragment ion(s) that are characteristic markers of the presence of particular posttranslational modifications (see Note 7).
3. Print out the chromatogram of the total ion current and, if measured, the ultraviolet (UV) trace (normally at 215 nm).
4. Overlay or superimpose the fragment ion chromatogram to identify the modified peptides.

4. Notes

1. Set the mass tolerances to be consistent with the mass accuracy of the instrument that has generated the data. It is usually better to use units of parts per million (ppm) or % rather than Da, since mass spectrometers usually have an error that is mass dependent and cannot be uniformly expressed in Da over the whole mass range.
2. Glycoproteins will show multiple spots on two-dimensional gel electrophoresis, due to glycan heterogeneity (as will phosphoproteins if these have varying degrees of modification). The MALDI-MS spectrum of the untreated sample peptides may therefore be complex. To verify the glycosylated peptides, the *N*-glycans may be released by PNGase F treatment. This is accompanied by deamidation to Asp (an increase of 1 Da over the expected mass of the unmodified peptide in a sequence database).
3. To search for the possible presence of modified residues, some Protein Prospector programs allow the use of up to four user-specified amino acids, for which the user supplies the elemental compositions from which the mass is automatically calculated. A letter (lower-case *u*, *v*, *w*, or *x*) is given to specify each user-defined amino acid in a peptide sequence. The default elemental composition for the user-defined amino acids is that of glycine. In the Protein Prospector suite, “MS-Comp” and “MS-Tag UnKnome” algorithms consider the 20 naturally occurring amino acids as a default and can include options such as: *q*-pyroglutamic acid; *h*-homoserine lactone; *s*-phosphorylated serine, threonine, or tyrosine; and so on. The list of modifications used by Mascot is taken from the Unimod database <http://www.unimod.org/>.
4. DHB appears to be the most popular matrix for carbohydrates. DHB is less soluble in water than in some organic solvents, such as methanol or acetonitrile. Some protocols suggest preparing a saturated solution of DHB in water, vortexing the mixture for 60 s, centrifuging for 20 s, and using the resulting supernatant. The resolution and sensitivity of DHB can be improved with additives such as the mixture of 10% of 5-methoxysalicylic acid (2-hydroxy-5-methoxybenzoic acid) with DHB (called “super DHB”). This is a good matrix for glycoproteins as well as for carbohydrates, but is less tolerant of salts and detergents than other matrices; therefore, sample cleanup may be necessary. Glycoproteins also work well with α -cyano-4-hydroxy-cinnamic acid (HCCA) and sinapinic acid

(SA), depending on the mass (see Chapter 32). There is no particular sample preparation method that works well for all glycoproteins/glycopeptides, and each sample has to be treated individually. 2,4,6- trihydroxyacetophenone (THAP) also works well for the analysis of glycopeptides, especially those with acidic glycans, since it is a neutral matrix. The $[M + Na]^+$ is frequently the major ion observed in the MALDI spectra of carbohydrates (see Fig. 1).

5. It is important to fragment proteins under low pH conditions to prevent disulphide exchange. Partial acid hydrolysis, although nonspecific, has been successfully used in a number of instances. Proteases with active-site thiols should be avoided (e.g., papain, bromelain). Typical conditions for pepsin are 25°C for 1–2 h at pH 2.0–3.0 (i.e., 10 mM HCl or 5% acetic or formic acid) with an enzyme:substrate ratio of about 1:50. Endoproteinase Glu-C has a pH optimum at 4.0 as well as an optimum at pH 8.0. Typical conditions are therefore 37°C overnight in ammonium acetate at pH 4.0 with an enzyme:substrate ratio of about 1:50. CNBr digestion in guanidinium HCl (6 M) in 0.1 to 0.2 M HCl may be a more suitable acid medium, due to the inherent redox potential of formic acid (the most commonly used protein solvent), which could cause formylation of the polypeptides. Guanidinium HCl is recommended rather than urea, since the latter can readily degrade to form cyanates that will react with thiol (and amino) groups. Urea can be used if preferable for a particular protein, but the highest grade of urea should always be used, and solutions should be prepared immediately before use. Urea should be deionized immediately before use to remove cyanates, by filtration of the solution through a mixed-bed Dowex or Amberlite resin (e.g., AG 501-X8 [Biorad]) in a filter flask. When analyzing proteins containing multiple disulphide bonds, it may be appropriate to carry out an initial chemical cleavage (CNBr is particularly useful) followed by a suitable proteolytic digestion. The initial acid chemical treatment will cause sufficient denaturation and unfolding as well as peptide bond cleavage to assist the complete digestion by the protease. If a protein has two adjacent cysteine residues, this peptide bond will not be readily cleaved by specific endopeptidases. The use of thiol and related compounds should be avoided for obvious reasons. Despite this, it is possible that disulphide bonds will be partially reduced during the analysis, and peaks corresponding to the individual components of the disulphide-linked peptides may be observed. Control samples with the reagents (including the same matrix compounds) are essential to avoid misleading results due to additional matrix-derived peaks.
6. In the positive-ion mode, the mass (M) of disulphide-linked peptides (of individual masses A and B) will be detected as the pseudomolecular ion at $(M + H)^+$ and after reduction this will be replaced by two additional peaks for each disulphide bond in the polypeptide at masses $(A + H)^+$ and $(B + H)^+$. Remember that A + B equals M + 2, since reduction of the disulphide bond will convert cystine to two cysteine residues, i.e., $-S-S- \rightarrow -SH + HS-$. Peptides containing an intramolecular disulphide bond will not break into two smaller peptides, but will appear at 2 Da higher. Peptides in the reduced state can normally be readily reoxidized to form intra-molecular disulphide bonds by bubbling a stream of air through a solution of the peptide for a few minutes.
7. Examples include phosphorylation, glycosylation, sulphation, and acylation (6,7). Phosphopeptides can be identified by production of phosphate-specific fragment ions of 63 Da (PO_2^-) and 79 Da (PO_3^-) by collision-induced dissociation during negative-ion LC-MS. Glycopeptides can be identified by characteristic fragment ions, including hexose' (163 Da) and hexose-*N*-acetyl' (204 Da) oxonium ion. The most specific method, by tandem MS on a triple quadrupole, ion-trap, or hybrid instrument, involves monitoring of fragment ions during precursor-ion scan electrospray MS/MS, resulting in a total-ion current (tic) chromatogram of just the modified peptides present in the digest. If only a single quadrupole instrument is available, a similar technique, termed *collisional-excitation scan*

ning, can be performed. For example, glycopeptides can be fragmented to sugar oxonium ion fragments without a mass-selection process, resulting in a total ion chromatogram similar to the UV trace (8). A printout of the selected-ion chromatogram for these carbohydrate-specific ions, such as the *N*-acetylhexosamine oxonium ion (at m/z 204) produces a trace specific for the glycopeptides. The methods detect glycopeptides with oligosaccharides *N*- and *O*-linked to the peptide. Phosphoserine- and phosphothreonine-containing peptides can also be identified by the process known as *neutral loss scanning*, where these peptides show loss of 98 Da by β -elimination of H_3PO_4 (see Chapter 44). Techniques employed during SIM can include peak parking to rapidly lower the flow rate, to provide more time for the mass spectrometer to analyze a particular peak and increase the MS/MS acquisition (see Chapter 37).

References

1. Aitken, A. (1995) Protein chemistry methods, posttranslational modification, consensus sequences. In Price, N. C. (ed.) *Proteins Labfax*, Bios Scientific Publishers, Oxford, Academic Press, San Diego, pp. 253–285.
2. Turner, J. P. (2003) Letter codes, structures, masses and derivatives of amino acids, Appendix 1, vol. 211. In Smith, B. (ed.), *Protein Sequencing Protocols*, 2nd edition, Humana Press, Inc., Totowa, NJ, pp. 431–464.
3. Aitken, A. (2003) A Library of consensus sequences. In: (Smith, B., ed.) *Protein Sequencing Protocols*, Appendix 2, 2nd edition, Humana, Totowa, NJ: pp. 465–485.
4. Zeng, C. and Biemann, K. (1999) Determination of N-linked glycosylation of yeast external invertase by matrix-assisted laser desorption ionization time-of-flight mass spectrometry. *J. Mass Spectrom.* **34**, 311–29.
5. Nicola, N. A., Cross, B., and Sirnpson, R. J. (1993) The disulfide bond arrangement of leukemia inhibitory factor: homology to oncostatin M and structural implications. *Biochem. Biophys. Res. Commun.* **190**, 20–26.
6. Swiderek, K. M., Davis, M. T., and Lee, T. D. (1998) The identification of peptide modifications derived from gel-separated proteins using electrospray triple quadrupole and ion trap analyses. *Electrophoresis* **19**, 989–997.
7. Annan, R. S. and Carr, S. A. (1997) The essential role of mass spectrometry in characterizing protein structure: mapping posttranslational modifications. *J. Protein Chem.* **16**, 391–402.
8. Huddleston, M. J., Bean, M. F., and Carr, S. A. (1993) Collisional fragmentation of glycopeptides by electrospray ionization LCMS and LCMSMS: methods for selective detection of glycopeptides in protein digests. *Anal. Chem.* **65**, 877–884.

Websites

GlycoMod is a tool that can predict the possible oligosaccharide structures from their experimentally determined masses. The program can be used for free or derivatized oligosaccharides and for glycopeptides. Another algorithm, GlycanMass, also part of the ExPaSy suite, can be used to calculate the mass of an oligosaccharide structure from its oligosaccharide composition. This can be used for free or derivatized oligosaccharides and for glycopeptides. Glycomod and GlycanMass are found at <http://us.expasy.org/tools/glycomod/> and <http://us.expasy.org/tools/glycomod/glycanmass.html>, respectively.

Deltamass is a database of protein posttranslational modifications at <http://www.abrf.org/index.cfm/dm.home>. There are hyperlinks to references to the literature on the modifications.

The program FindMod predicts potential protein posttranslational modifications and

potential single amino acid substitutions in peptides, and is in the ExPASy suite at <<http://expasy.org/tools/findmod/>>. Experimentally measured peptide masses are compared with the theoretical peptides calculated from a specified SwissProt entry or from a user-entered sequence.

FindPept—Identifies peptides that result from unspecific cleavage of proteins from their experimental masses, taking into account artifactual chemical modifications, posttranslational modifications, and protease autolytic cleavage.

PeptideMass—Calculates masses of peptides and their posttranslational modifications for a SwissProt or TrEMBL entry, or for a user sequence.

Approaches to the *O*-Glycoproteome

Franz-Georg Hanisch and Stefan Müller

1. Which Proteins in a Given Proteome Carry *O*-Linked Glycans?

The term *proteome* refers to the highly fluctuating totality of expressed proteins in a particular cell, in subcellular fractions, or in body fluids, which varies qualitatively and quantitatively with the cellular state and the active functional networks. The definition of a functional proteome is performed in differential approaches and requires the identification of proteins and the quantification of the relative state-dependent abundances of each protein. In many cases, identification refers to posttranslationally modified versions of the proteins, which actually represent the functionally active or inactive isoforms in a regulated system (i.e., phosphorylation, β -glucosaminylation, ubiquitylation, and so on). Hence, not only the protein itself, but the “frozen state” of its fluctuating modification pattern (type and site) need to be defined on the molecular level and in the respective functional context. An example of this type can be seen in notch signaling, which is regulated by β -6-glucosaminylation of Fuc-*O*-Ser in the epidermal growth factor (EGF) domain of the protein (1).

About one-third of all proteins has been claimed to carry transitionally at least one phosphate group. Similar estimations were made with respect to the relative frequencies of glycosylated proteins, which range above 50% of all proteins and cover *N*-linked and *O*-linked glycan modifications. While these extrapolations from available structural data refer mainly to *N*-glycosylations, where the sequence motif N-X-S/T is known to form a specific glycosylation target (sequon), no reliable predictions of mucin-type *O*-glycosylation can be made on the basis of peptide sequence information. This is because no sequon can be defined for the addition of GalNAc to Ser or Thr of a peptide core, but instead more than 15 distinct polypeptide GalNAc transferases with different expression patterns and partially overlapping substrate specificities were claimed to be involved in initial *O*-glycosylation. This unpredictability necessitates structural analysis, both with respect to the presence and to the localization of *O*-linked glycans within a peptide. In this chapter, we refer to a variety of methodological approaches to the *O*-glycosylated subproteomes comprising mucin-type glycans with an α -GalNAc core and those based on β -GlcNAc. Each of these modifications brings along a variety of specific methodological problems, which have to be considered in advance. Mucin-type *O*-glycosylation of proteins can be expected to hamper their identification as a result of its structural complexity, the size of the glycans, and their clustering, which interferes with several stages of conventional proteomic approaches. Because of their

regulatory function, β -GlcNAc modifications, on the contrary, are generally substoichiometric and hence of very low abundance, which necessitates selective enrichment of β -GlcNAc-*O* peptides prior to mass spectrometry (MS) analysis. In case of α -Man- and α -Fuc-based *O*-glycosylation, which are not referred to in this chapter, a discrimination of proteins modified by these rare glycans from the bulk of mucin-type *O*-glycosylated or *N*-glycosylated proteins might represent the major problem.

1.1. Mucin-Type *O*-Glycosylated Proteins

In contrast with the case with insects (2) and other invertebrates (3), mucin-type *O*-linked glycans in vertebrates are generally rather complex in structure, because they are built up by elongation of up to eight distinct di-/trisaccharide cores with linear or branched polylactosamine-type backbones terminating in a great variety of sialic acid- and/or fucose-containing peripheral glycosides (4). This structural complexity of oligosaccharides, reaching 20 monosaccharides in size, represents the major problem in attempts to isolate the subproteome of mucin-type *O*-glycoproteins. Because most peripheral structures are common to mucin-type *O*-linked glycans and to *N*-linked glycans, a ready, reliable and convenient discrimination of native *N*- and *O*-glycoproteins by lectin-affinity chromatography is not possible. To solve this problem technically, the complexity of *O*-glycans has to be reduced, either by replacement of the glycan with tags or by controlled partial degradation of glycans to the level of the core GalNAc, a specific structural marker of mucin-type *O*-glycosylation.

All attempts based on the β -elimination Michael addition (BEMAD) replacement of *O*-linked glycans by specific tags encounter a multitude of problems. First of all, mucin-type glycans, as derivatives of α -linked GalNAc, are more stable to alkali-catalyzed β -elimination than β -linked GlcNAc. Hence, approaches reported for enrichment of the GlcNAc β -*O* proteome by β -elimination and Michael addition of tags (discussed later) cannot be easily transferred to mucin-type *O*-glycoproteins. Even under conventional β -elimination conditions for mucin-type *O*-glycans (50 mM NaOH, 50°C, 18 h), nonglycosylated peptides also undergo dehydration reactions at hydroxylated amino acids (Ser/Thr), and the resulting dehydro derivatives will be finally tagged. Using ammonia for liberation of *O*-linked glycans dehydration at unsubstituted Ser/Thr positions is much less pronounced; however, in this case primary amino groups are not only introduced by Michael addition into previously glycosylated sites, but also at Asn and Gln. These unavoidable side reactions make a reliable discrimination of *O*-glycosylated proteins/peptides impossible.

Instead of tagging the glycoproteins for their ready isolation, other approaches for their selective enrichment are based on the discrimination of glycosylated vs unglycosylated proteins and the specific elution of glycosylated peptides. A protocol for the isolation of *N*-glycosylated peptides published by the group of Aebersold (5) starts with the immobilization of the total glycoproteins to a hydrazide-activated solid support via their periodate-oxidized glycans. The covalently bound glycoproteins are treated with trypsin, and the peptide products are eluted successively with urea-containing buffer (nonglycosylated peptides) and by PNGase-F treatment (*N*-glycosylated peptides). Adaptations of this protocol to the selective liberation of *O*-glycosylated peptides need to be established. But a simple transfer of the protocol for *N*-glycopep-

tides is not possible, since an endo-glycosidase, which would cleave^a *O*-linked oligosaccharides irrespective of their structures, is currently not available.

In summary, *O*-glycoproteins, in particular real mucins with clustered mucin-type glycans, still represent a challenge for scientists interested in the development of robust methods that yield reliably quantitative data in differential approaches and allow a highly parallel analysis of complex proteomes (comprising glycosylated and nonglycosylated proteins). Proteomic approaches based on two-dimensional gel electrophoresis still encounter a variety of problems. These are caused by the charge heterogeneity of highly sialylated or sulfated glycoproteins and by their unusually high molecular masses. There is still no sensitive stain for the selective detection of glycoproteins in gels, and in particular no stain that would work equally well with both glycosylated and nonglycosylated proteins, and over a broad dynamic range. Densely *O*-glycosylated proteins are highly resistant to proteolysis, in particular to proteolysis with specific enzymes like trypsin. Finally, the identification rates in proteome analyses via mass-spectrometric peptide mapping are strongly affected by the reduction of protein coverage caused by complex *O*-glycosylation. The way out may be found in the development of two-dimensional liquid chromatographic approaches, which circumvent some of the problems and could be of particular value in the analysis of subproteomes like the *O*-glycosylated proteins.

1.2. O-Glucosaminylated Cytosolic and Nuclear Proteins

O-GlcNAc is a dynamic posttranslational modification occurring on a variety of nucleocytoplasmic proteins with a substoichiometric incidence at adjacent or identical sites as phosphorylation (6,7). A regulatory function of *O*-GlcNAc antagonistic to phosphorylation has been claimed (6). But despite its widespread occurrence and potential regulatory function, less than 50 sites of *O*-GlcNAc modification have been published (8). Thus, until recently, the only applied technical approach to map the sites of modification was based on the enzymatic tagging of *O*-GlcNAc with radiolabeled galactose (9). The low sensitivity of this laborious technique was overcome by a novel approach combining the mild β -elimination of the sugar with a Michael addition of nucleophilic tags, like dithiothreitol (DTT) or biotin pentylamine (10). Both tags enable the affinity-chromatographic isolation of previously *O*-GlcNAc-modified proteins/peptides. The alkali-catalyzed β -elimination with 1% triethylamine, 0.1% NaOH at 50°C for 2.5 h, is mild enough to avoid the massive liberation of phosphate groups, which are more resistant to alkali. Mucin-type *O*-glycans are also β -eliminated under these conditions, but at a low level. DTT addition to the dehydro derivatives introduces a free sulfhydryl group at the previously glycosylated sites, which can be used for the specific enrichment of the tagged proteins/peptides by covalent chromatography on activated thiol-sepharose. Alternatively, a biotinylated tag (biotin pentylamine, BAP) can be introduced into dehydro amino acids of the protein/peptide to enable affinity isolation on monoavidin columns. However, the DTT-based approach performed much better than BAP in the MS/MS analysis to map sites on the modified peptides.

^aNote added in proof: after submission of this manuscript, a paper was published that describes the metabolic labeling of *O*-glycoproteins with *N*-azidoacetylgalactosamine for their detection and isolation via phosphine conjugated affinity tags (53).

The utility of the method was demonstrated by the authors, who were able to use automated data-dependent scanning for MS/MS and Turbosequest searching against nonredundant databases to identify sites of *O*-GlcNAc modification. Accordingly, it can be stated that it has potential for high throughput and automation. Moreover, the method is claimed to be amenable for performing quantitative mass spectrometry using a differential isotopic labeling with DTT (light) and deuterated DTT (heavy).

There are two technical objections to the approach: First of all, there is a limited but considerable dehydration at unglycosylated Ser/Thr residues, resulting in an undesired tagging. Secondly, β -elimination occurs also at alkylated cysteines, and accordingly Cys-containing peptides are tagged with DTT and finally trapped on activated thiopharose. A further restriction is given by the fact that some *O*-GlcNAc-modified residues, in particular *O*-GlcNAc-Thr followed by a Pro, are more resistant to β -elimination and are not likely to be detected by this approach.

2. Which Site in a Peptide Is Modified by *O*-Glycosylation?

2.1. Mass Spectrometry-Based Protocols

As a result of the lack of sequons, the initial addition of GalNAc to serine or threonine in a peptide is largely unpredictable. Attempts were made on the basis of available structural data and trained neuronal network algorithms to establish a sequence-based tool for the prediction of *O*-glycosylation sites (Net-*O*-Glyc) (11)^b. However, this tool still suffers from insufficient reliability, since the actual substitution sites are often in disagreement with results from calculation. Reasons for this could be seen in the insufficient availability of solid chemical data on *O*-glycosylated sites in peptides, and in particular unique features of initial *O*-glycosylation. First of all, there are more than 15 different polypeptide GalNAc-transferases with distinct but partially overlapping substrate site specificities (12,13). Not all of these are expressed simultaneously and in the same cell, but instead show some degree of tissue specificity, and some of them seem to be developmentally regulated. Hence, the same protein target can display different site-specific glycosylation patterns in different organs. Moreover, there is a competition between polypeptide GalNAc-transferases and other glycosyltransferases acting on GalNAc to build up the di/trisaccharide cores (core 1 to core 8) (14,15). Such dynamic epigenetic regulatory mechanisms may underlie the observed microheterogeneity in the substitution of *O*-glycosylation sites—i.e., partial substitution of sites in human glycophorin A (16) or the random substitution with one to five glycans of the MUC1 tandem repeat (17).

There is evidence that *O*-glycosylation may have fundamental functional impact in the development of organisms, as demonstrated recently in a lethal *Drosophila* mutant with a defective ppGalNAc-transferase gene (18). Other examples are *O*-glycosylated receptors, like the neuropeptide receptor, where *O*-linked glycans serve as an apical targeting signal (19); human MUC1, where clathrin-mediated endocytosis of membrane exposed mucin is modulated by its glycosylation state (20); and antibacterial peptides, like Drosocin (21), which were demonstrated to be active only if specific sites carry *O*-linked glycans. These and many other reports make clear that there is a

^bNote added in proof: after submission of this manuscript, an improved version of the prediction tool for *O*-glycosylation sites became accessible under <http://www.cbs.dtu.dk/services/NetOGlyc3.1> and was described in ref. 54.

demand for chemical strategies allowing for a reliable and convenient site identification of *O*-glycosylation.

2.2. How to Fragment a Heavily O-Glycosylated Protein

There is an intrinsic problem for site identification of mucin-type *O*-glycosylation, which is caused by clustering or high density of *O*-linked glycans: the carbohydrate cover hampers effective proteolysis of the glycoprotein, a prerequisite for the isolation and sequencing of *O*-glycosylated peptide fragments. Whereas frequently used standard glycoproteins, like glycophorin A, despite their dense *O*-glycosylation, show ready fragmentation on trypsin treatment, real mucins, with their densely substituted repeat domains, are hardly digestable. A well-studied example represents the human mucin MUC1, where 20 to 120 repetitions of a 20-meric peptide (HGVTSAPDTRPA-PGSTAPPA) with five *O*-glycosylation sites can make up between 50 and 80% of the entire protein. The nonglycosylated protein core offers a variety of target sites for specific endopeptidases, like clostripain (Arg-C), protease IV from papaya (Gly-C), endopeptidase from *Flavobacterium meningosepticum* (Pro-C), *Staph. aureus* V8 protease (Glu-C/Asp-C), and endoproteinase from mutated *Pseudomonas fragi* (Asp-N). However, not even one of these enzymes is able to fragment the repeat domain of native, heavily *O*-glycosylated MUC1. There are two strategies that can solve the problem: (1) extensive digestion of native mucin with unspecific endopeptidases or (2) limited proteolysis of partially deglycosylated mucin with some of the above-mentioned specific enzymes.

To start with the latter, it was demonstrated that MUC1 from milkfat globule membranes can be effectively fragmented with clostripain (17), after the mucin had been partially deglycosylated to the level of core-GalNAc residues by sequential desialylation with dilute trifluoroacetic acid and treatment of the extensively dried sample with trifluoromethane sulfonic acid (0°C, 30 min). GalNAc substitution, even at Thr adjacent to the Arg-C cleavage site, does not prevent proteolysis by this enzyme. On the other hand, the already low activity of trypsin at the weak substrate position Arg-Pro is completely abolished by GalNAc-substitution of the vicinal Thr. Other specific endopeptidases with proven activity on GalNAc-substituted MUC1 repeat peptide cores are the Gly-C- (22) and the Pro-C-specific enzymes (15). An advantage of limited proteolysis with specific endopeptidases is the fact that parallel treatments of a glycoprotein with two or three of these enzymes can yield maps of overlapping *O*-glycosylated (GalNAc-modified) fragments. The mass increment of 203 Da corresponding to a GalNAc residue can be used in matrix-assisted laser desorption/ionization (MALDI)-MS to localize these *O*-glycosylated sites.

The other strategy, which is based on “unspecific” endopeptidases, expectedly will give less predictable results, but occasionally can be the last choice to cleave a proteolysis-resistant glycoprotein. Surprisingly, endopeptidases with a broader substrate specificity do not always cleave their substrate at multiple sites, yielding complex mixtures of peptide fragments. Even the unglycosylated peptide cores of the MUC1 repeat domain yield a restricted pattern of fragments when digested under controlled conditions with papain, proteinase K, or pronase (unpublished). There are preferential cleavage sites in the repetitive peptide core on the one hand, while some fragments resist further hydrolysis during periods of 6 h on the other. Oligomeric MUC1 repeats with dense and complex *O*-glycosylation, when digested with pronase for 16 h, yield only

one major product, a 21-meric glycopeptide, nevertheless with an extremely heterogeneous glycosylation.

2.3. Sequencing Strategies of GalNAc-O Peptides

To localize *O*-linked glycans within a peptide sequence, it is generally advantageous (1) to reduce the size and structural complexity of the oligosaccharides bound to a protein or (2) to replace the glycan by a tag. In both cases, structural information on the oligosaccharide is lost; however, a considerable facilitation in the analytical performance is achieved by this simplification. Starting with the first alternative, the partial deglycosylation by chemical or enzymatic methods yields GalNAc-modified peptides/proteins, which offer a series of advantages. The most reliable way to achieve partial de-*O*-glycosylation is the chemical cleavage of the peripheral and backbone sequences by a limited trifluoromethane sulfonic acid treatment (17). The remaining sugar residue, a GalNAc, can be used as a tag during chemical peptide degradation by partial acid hydrolysis (ladder sequencing) or by conventional Edman degradation. GalNAc-*O* peptides/proteins are also more easily digestible with endopeptidases, because of the reduction of steric hindrances caused by the bulky glycans. Finally, mass spectrometric sequencing of glycopeptides by postsource-decay (PSD)-MALDI-MS or electrospray ionization (ESI)-MS/MS is greatly facilitated in the case of GalNAc-*O* peptides, as a result of the uniform mass tag with a comparatively small mass increment.

2.3.1. Postsource-Decay MALDI Mass Spectrometry

MALDI-MS, which was introduced by Hillenkamp and Karas in 1988 (23), is the most sensitive mass spectrometric method reaching the attomol range in the analysis of peptides. This is a result of major technical progress made during the early 1990s, particularly when the first reflectron instruments became available and when the delayed extraction technology was invented. Both technical improvements resulted in increased resolution (>10.000) and mass accuracy of MALDI-MS measurements (10–100 ppm), the latter to some extent also with respect to sensitivity. The reflectron technology opened the gates to new applications for MALDI instruments, which were based on time-of-flight (TOF) analyzers: polymer sequencing by analysis of product ions arising from metastable decay of parent ions (postsource decay) in the first field-free drift path of the analyzer before they enter the electronic mirror (24). Kaufmann et al. showed that a complete sequence analysis for peptides up to 26 amino acids can be achieved in the femtomolar range (25). In 1997, this approach was adapted to the first analyses of *O*-glycosylated peptides (17,26,27). The general advantages of the method is that it does not require purification of the glycopeptide samples and that the sequencing can be performed in the upper femtomolar range. The blanking of nonrelevant ion species before the ion beam enters the flight tube enables the analysis of glycopeptides in the presence of other (glyco)peptides if their MH⁺ ions differ from the ion of interest by a certain mass increment. Together with the 10- to 100-fold higher sensitivity, this feature makes PSD-MALDI-MS a method that could be superior to Edman sequencing.

Generally, more than 90% of the fragment ions can be assigned as N-terminal (a_m, b_m, c_m) or C-terminal ions (x_n, y_n, z_n) according to the Biemann modification (28) of the Roepstorff and Fohlmann nomenclature (29). Fragment ions originating from the nonreducing end of the sugar are labeled as B⁺ or C⁺ according to the nomenclature of Domon and Costello (30). In the case of the *O*-glycosylated MUC1 repeat peptides

analyzed in the above-cited studies (17,26,27), nearly complete fragment series can be obtained from both ends of the molecules, allowing the unambiguous assignment of the glycosylation sites by considering the respective mass increments for GalNAc (+203 u) or Gal-GalNAc (+365 u), respectively. The most dominant fragment ions are those of the b_m - and y_n - series, but a_m and x_n ions and their companion ions at -17 and -18 are also particularly prominent. A further increase in complexity of the mass spectra comes from the cleavage at X-Pro, resulting in subfragmentation and the formation of "proline fragment ions." PSD fragments containing glycan residues are also detected as nonglycosylated ion species, but at much lower intensity (10–15%). This demonstrates the stability of the *O*-linked glycan chains under the described conditions.

There are several arguments against general applicability for this method. One severe drawback is that metastable fragmentation of the peptide is largely dependent on intrinsic properties of the analyte and, except for a variation of the matrix compound, this cannot be influenced by the mass spectrometric set up. There are glycopeptides, like those derived from human MUC1 repeats, which decompose readily and with high yields of daughter ions, while others, like the MUC2 repeat peptide with clusters of threonine, do not show any fragmentation, irrespective of *O*-glycosylation (unpublished).

Another major drawback became evident with the finding that increasing polarity of the analyte (which is associated with higher carbohydrate contents) can drastically reduce MH^+ ion intensities. An equimolar mixture of *O*-glycosylated peptides with identical amino acid sequence, but differing in the number of Gal-GalNAc moieties (1 to 5 mole per mole of peptide), gives rise to exponentially decreasing MH^+ ion intensities (the penta-substituted species being nearly undetectable in measurements using HCCA as the matrix).

2.3.2. Nano-Electrospray (qTOF) Mass Spectrometry

The above cited objections to a broad applicability of PSD-MALDI-MS in the analysis of *O*-glycosylated peptides gave rise to a search for alternative mass spectrometric methods that would permit the sequencing of high-mass, heavily *O*-glycosylated peptides. Electrospray ionization offers three principal advantages compared with MALDI ionization: (1) it is not dependent on the desorption efficiency of the analyte from a solid target, and hence will not show a reduced ion yield for analytes of higher polarity; (2) the fragmentation efficiency in collision-induced-dissociation (CID) experiments can be influenced via mass spectrometric parameters and is accordingly much less dependent on structural properties of the analyte; and (3) the ionization process is compatible with online liquid chromatography of the analytes.

The first demonstration that ESI-MS of *O*-glycopeptides could be a reasonable approach came from Carr et al., who were able to characterize recombinant soluble CD4 (31) and tryptic glycopeptides from fetuin (32). Also, another group, Peter-Katalinic et al., has demonstrated the potential of (+)- and (-)-ESI-MS for the analysis of nonderivatized *O*-glycosylated amino acids (33) and short peptides with clustered *O*-linked NeuAc-GalNAc (34). The potential of (+)-ESI-MS on triple-quadrupole and ion-trap mass spectrometers was later evaluated for sequencing of synthetic *O*-glycopeptides corresponding to the MUC4 tandem repeat (Alving and Peter-Katalinic, unpublished).

With the advent of hybrid-type systems combining a quadrupole analyzer for parent ion selection and a TOF analyzer for fragment ion separation after CID, a further improvement with respect to resolution and mass accuracy was made. Two reports

referring to the site identification by nanospray-ESI-qTOF-MS of *O*-linked glycans in synthetic MUC1 and MUC2 peptides were published (35,36).

In MS1, the native *O*-glycopeptides are represented by a series of positively charged ion species in the (+)-ESI mass spectra: as singly, doubly, and triply charged ion species corresponding to the singly charged deconvoluted ions $(M + H)^+$, $(M + Na)^+$, $(M + K)^+$, $(M + 2Na - K)^+$, and $(M + Na + K - H)^+$.

In MS2, it can be shown that y_n and b_m fragment ions are the most abundant in the spectrum, whereas z_n , x_n , and also a_m and c_m ions were only rarely detected. Similarly, the companion ions of singly charged y_n fragments (y_n-17 , y_n-18) or b_m fragments (b_m-17 , b_m-18) were rare and of low abundance. Facile losses of GalNAc residues are registered as indicated by ions at m/z 204 (HexNAc $+$) and a nearly complete series of nonglycosylated y_n and b_m fragment ion species. Internal fragmentation of y_n ions at X-Pro, as observed in PSD-MALDI-MS, was insignificant. In general, the MS/MS spectra are much less complex than the corresponding PSD-MALDI mass spectra. Nano-electrospray-qTOF mass spectrometry can be regarded as a sensitive, generally applicable method for determination of *O*-glycosylation sites.

2.3.3. Partial Gas-Phase Hydrolysis and Mass-Spectrometric Ladder Sequencing

The method introduced by Mirgorodskaya et al. (37) is based on the partial vapor-phase acid hydrolysis of glycopeptides, combined with a ladder sequencing by mass spectrometry. Using aqueous solutions of pentafluoropropionic acid (PFPA) or hydrochloric acid (HCl) in an inert atmosphere under reduced pressure, the polypeptide backbones fragment within 1 to 2 h at elevated temperatures with minimal cleavage of the glycosidic bonds. The generated ladder of peptides/glycopeptides can be analyzed by positive-ion reflectron MALDI or by electrospray MS.

The method is rapid and simple with regard to its performance, and does not need expensive and sophisticated experimental setup. It simply requires a 22-mL glass vial with a Mininert valve (Pierce) that allows evacuation of the reaction chamber. The sample (up to 20 pmol) is dried in the bottom of a 0.5-mL Eppendorf vial by vacuum centrifugation and placed in the glass vial containing 100 μ L of the diluted acids (20% aq HCl or PFPA). After flushing with argon and evacuation (1 mbar), the glass vial is heated for approx 1.5 h to 90°C (PFPA) or 50°C (HCl), respectively. After removal of remaining traces of acid by vacuum centrifugation, the generated peptide fragments are taken up in solvents compatible with application on MALDI mass-spectrometry targets (0.1% trifluoroacetic acid in mixtures with acetonitrile).

The method works at the lower nanogram level and avoids—by using gas-phase hydrolysis—the contamination of the sample with impurities. It is applicable not only for the analysis of mucin-type *O*-glycopeptides, but also for *O*-mannosylated peptides. In the case of mucin-type *O*-glycosylation, application of the method is not restricted to GalNAc-*O* peptides; it can be used also for glycopeptides with short, core-type glycans, like the core 1 disaccharide (Fig. 1). More labile glycosidic bonds, like those of sialic acid or fucose, do not survive gas-phase acid hydrolysis. The N- or C-terminal sequence ladders allow for a discrimination between potential glycosylation sites, because the glycosylated serines or threonines are indicated by the presence of an ion 203 Da (GalNAc) or 365 Da (Gal-GalNAc) above a MH^+ calculated for the nonglycosylated peptide. In general, the interpretation of the mass spectra implies knowledge of

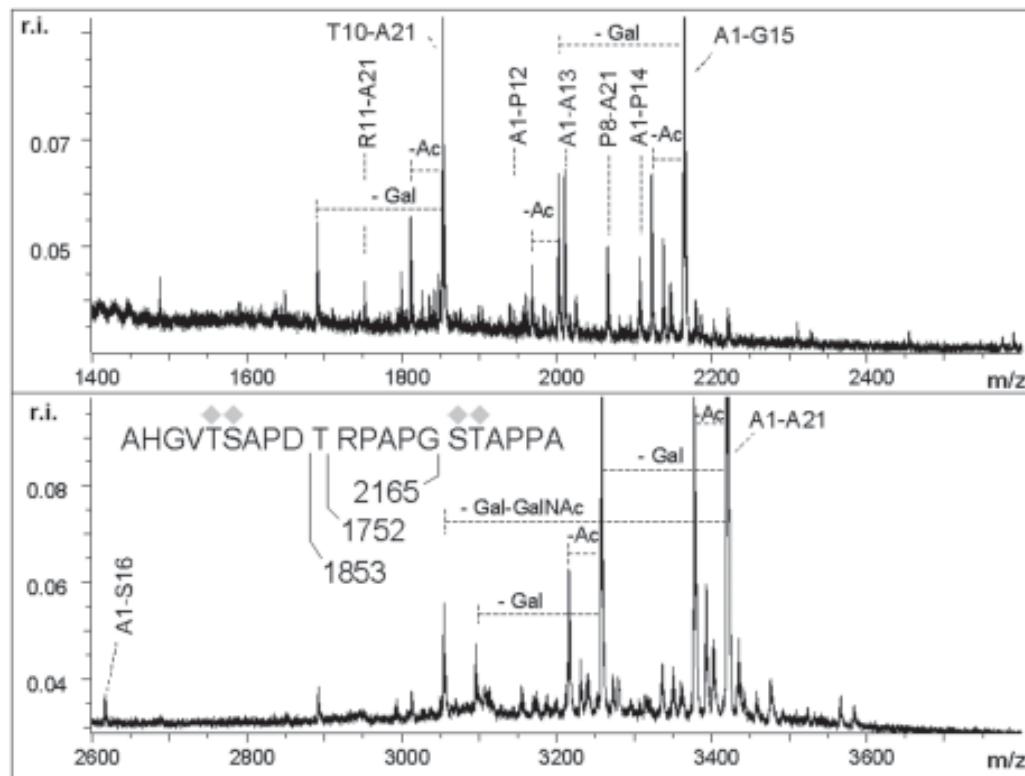


Fig. 1. Partial gas-phase hydrolysis and mass spectrometric ladder sequencing of a 21-meric *O*-glycopeptide substituted with four Gal-GalNAc moieties. A synthetic 21-meric MUC1 glycopeptide AHGVTSAPDTRPAPGSTAPPA with four Gal-GalNAc disaccharides substituted at Thr5, Ser6, Ser16, and Thr17 was treated with pentafluoropropionic acid (20%, v/v) in the gas phase at 50°C for 18 h to partially hydrolyze the peptide backbone. Peptide ladders were analyzed by positive-ion matrix-assisted laser desorption/ionization (MALDI)-time-of-flight (TOF) mass spectrometry of the sample mixed with matrix α -cyano-4-hydroxy cinnamic acid. The relevant peptide and glycopeptide ions allowing the unambiguous assignment of *O*-glycosylation sites are indicated. ◆ = Gal-GalNAc

the peptide sequence. The peptide mixtures generated by PFPa hydrolysis are often very complex. This can be due to several glycosylation sites in a peptide, but mostly results from intermediate reaction products at -18 Da (oxazolone intermediates) or -1 Da (cleavage of the bond between the amido group and the α -carbon, preferentially in the C-terminal peptide ladder) and from the de-*N*-acetylation of GalNAc (-42 Da). To reduce complexity of the spectra, the reaction intermediates at -18 Da can be converted into the final products by a brief incubation with 25% aqueous ammonia. De-*N*-acetylation of GalNAc can be minimized by using HCl hydrolysis or, if PFPa is used, can be reverted by re-*N*-acetylation with acetic anhydride in pyridine-methanol (1:1:5) for 15 min at ambient temperature (unpublished). However, the presence of ions 42 Da less than the MH^+ ion can be of diagnostic value, since they indicate a glycosylated (GalNAc) peptide.

2.4. Sequencing Strategies Based on Glycopeptide Derivatives Generated by β -Elimination/Michael Addition

The size, polarity, and charge of *O*-linked chains, and the heterogeneity of their structures, cause a variety of problems in mass spectrometry, particularly in attempts aiming at the site identification of glycosylation. In **Subheadings 2.3.1.–2.3.3.**, methods were presented that have a proven capacity to solve these problems by reducing complex chains to the level of core-type glycans, like GalNAc or Gal-GalNAc. A straightforward alternative to this strategy are methods based on the replacement of sugars by simple and uniform tags. The replacement is easily performed by a combination of base-catalyzed β -elimination of *O*-linked glycans and the Michael addition (BEMAD) of nucleophilic tags to the dehydro intermediates. There are three reports referring to strategies based on this principle, which have the feature in common of using amines as the base and nucleophilic tag (44–46).

2.4.1. BEMAD of Mucin-Type *O*-Linked Glycans With Ammonia and Alkylamines

The method by Rademaker et al. combines base-catalyzed β -elimination with the addition of ammonia to the unsaturated amino acid derivatives (38). In the final peptide products, all *O*-linked chains should be replaced by a primary amino group at the respective Ser/Thr positions. Hence, the approach fulfills the criterion of introducing a simple and uniform tag into previously glycosylated sites.

A brief description of the protocol: The glycopeptide (10–1000 pmol in water) is dried by vacuum centrifugation in the bottom of a 1.5-mL Eppendorf vial. Up to 300 μ L of 25% (v/v) aq ammonia is added, and the sample is heated at 45°C for at least 18 h. The reaction mixture is dried by vacuum centrifugation, and the aminylated product is taken up in solvents appropriate for each type of mass spectrometric measurement.

The protocol is simple and convenient, by avoiding sample cleanup after the reaction. Conversion into the final product is comparatively rapid and quantitative. The method offers advantages in the context of mass-spectrometric analysis of the peptide products due to the introduction of primary amines, which facilitates positive ion formation and detection by increasing the cationic properties of the analyte.

There are also several drawbacks to be mentioned, which clearly have to be considered. (1) On longer incubation times (exceeding 18 h), a significant proportion of nonglycosylated Thr/Ser positions is tagged, making the discrimination of glycosylated vs unglycosylated sites less reliable. (2) According to our experience, peptides with clustered *O*-linked glycans are not converted into the expected products. In these cases, the dehydro intermediates form extended resonance-stabilized structures, which resist addition of the nucleophile. (3) The discrimination of a glycosylated vs unglycosylated site is difficult, because the mass difference is only –1. High mass accuracy is needed, accordingly, which is achievable on a few instrumental configurations, such as ESI-QTOF mass spectrometers.

Two more recent reports refer to related strategies: one evaluating the potential of β -elimination/alkylaminylation in solution (39), the other a gas-phase β -elimination methylaminylation (40).

The former study introduces ethylamine and methylamine as substitutes for ammonia. Both reagents overcome the problem of a small mass difference introduced by the tag, since they add mass increments of +27 (ethylamine) or +13 (methylamine) to the

MH^+ ion of the peptide. The protocol presented in the paper was established to facilitate in-gel digestion of heavily *O*-glycosylated proteins and to enhance the coverage during mass-spectrometric protein identification. The usefulness of the strategy is demonstrated for human glycophorin (with 15 *O*-linked chains, partially in clusters) and human MUC1 from human milkfat globule membranes (with multiple repeats, each containing five *O*-glycosylation sites).

A brief outline of the protocol (39): Picomolar amounts of the glycopeptide are dried in a 0.5-mL Eppendorf tube and heated with 20 μL of the alkylamine for over 24 h (70% aq ethylamine) or 6 h (40% aq methylamine) at 50°C. The reaction mixtures are dried by vacuum centrifugation and without further work-up dissolved in appropriate solvent systems for mass spectrometry.

Several drawbacks became evident on routine application of the method: (1) The alkylamines not only replace *O*-linked glycans at Thr/Ser, but also add to the amido side groups of Asn and Gln, which interferes with the identification of *O*-glycosylated peptides by peptide-mapping strategies. (2) The ethylamine reagent is less volatile and sometimes needs longer drying periods. (3) During longer incubation times, the reaction with methylamine results in extensive peptide fragmentation, dependent on the peptide sequence. (4) As observed with ammonia, both reagents do not add to dehydro amino acids on treatment of peptides with clustered *O*-linked glycans.

In a similar approach, the BEMAD technology is combined with partial gas-phase hydrolysis (Subheading 2.2.3.), allowing a ladder sequencing of methylaminylated peptides by mass spectrometry (40). An advantage of this variant may be seen in the gas-phase treatment of the glycopeptides, which avoids contamination by the reagent.

2.4.2. Mass-Spectrometric Sequencing of BEMAD Derivatives

A comparative analysis of glycosylated peptides with their corresponding ethylaminylated derivatives by positive-ion PSD-MALDI-MS (39) revealed that equal molar amounts of both species (50 pmol) yielded fragment spectra of comparable ion intensities. However, in case of the EA-modified peptides, a three- to fivefold increase in signal intensity was observed in the low-mass range. Fragments of higher masses dominated in the spectra of the corresponding glycopeptides. The spectra of the EA peptides were characterized by a consistent series of b- and b-18 ions from the N-terminal, which broke off in the middle of the peptides, preferentially beyond the site of modification. The position of EA modification could be unequivocally assigned in each case. Tryptic peptide fragments (40–61, 40–49, and 50–61) of human glycophorin A were detectable only after β -elimination alkylaminylation, and could be sequenced by PSD-MALDI-MS to localize the sites of modification. Native MUC1 mucin from human milkfat globule membranes, which is indigestible with clostripain, could be fragmented by the enzyme after β -elimination alkylaminylation, and the mono- to penta-substituted 20-meric peptides were analyzed by MALDI-MS.

MS/MS analysis of ethylaminylated peptides by ESI-qTOF mass spectrometry (39) revealed that the total sequence of the 21-meric peptide was covered within a single experiment. Both series, the y- and the b-ions, were nearly complete. Identification of the substitution sites was always possible on the basis of intense y-ion series, showing incremental mass increase of +27. As revealed also for the corresponding glycopeptides, the MS/MS spectra were less complex than those from the respective PSD-MALDI-MS experiments, showing only y- and b-ions. In summary, ESI-qTOF proved

to be a suitable method for defining the site of modification within the peptides. The average mass accuracy of fragment ions was approx 50 ppm.

3. *O*-Glycoprofiling of Mucin-Type *O*-Glycoproteins

3.1. Mass Spectrometry-Based Protocols

To characterize the *O*-glycoproteome in full, the profile of *O*-linked glycans needs to be defined with respect to the structural parameters of each species and to the relative abundance in a mixture. The definition of an oligosaccharide structure comprises not only information on the monomer composition and monomer sequence, but moreover the analysis of linkage sites and anomeric configurations. Generally, combinations of different technologies have to be applied on quite substantial amounts of oligosaccharide to achieve this goal, as for example the combination of mass spectrometry, methylation analysis, and sequential exoglycosidase treatments, or the combination of ¹H-NMR and methylation analysis. In many cases such amounts of glycans are not available, particularly not in proteomic approaches, where protein amounts in the low picomol or upper femtomol range have to be handled. Consequently, the analysis is often limited to some of the structural parameters, like the monomer composition and the sequence of the glycans, which can be achieved by mass spectrometry with sufficient sensitivity. More important than the loss of structural information is the absolute requirement for quantitative data, which cannot be provided by methods based solely on MS. Since the yields of detectable ions are structure dependent, and, particularly in desorption ionization, decrease with size and polarity of the analytes, MS cannot be regarded as a quantitative method. Already in the early 1980s, quantitative patterns of glycan alditols were measured by normal-phase high-performance liquid chromatography (HPLC) on amino-modified silica coupled with ultraviolet (UV) detection at 192 nm (41). This approach was later improved by using fluorescently labeled glycans, which increased sensitivity of detection by two to three orders of magnitude (42). A prerequisite for the introduction of fluorescing tags is the availability of a reactive aldehyde at the reducing end of the glycan. Hence, the classical reductive β -elimination (1 M NaBH₄, 50 mM NaOH, 50°C, 18 h) used for the liberation of alkali-labile *O*-linked glycans as their stable alditols (43) cannot be applied. Two alternatives were introduced, which both have the capacity to liberate the glycans as reducing sugars on a microscale and with minimal losses by degradative side reactions (“peeling”).

3.2. Analysis of Reducing Glycans

To date, hydrazinolysis is one of the few chemical methods for the nonselective release of *O*-linked glycans in their reducing state, which can be performed on the microscale and with minimal losses by “peeling” reactions (44). Other methods that have been published, such as cleavage with aq ethylamine (45) or aq hydrazine (46), cannot be recommended due to extensive degradation of the glycans from their reducing terminal. *O*-linked chains can be selectively cleaved from the protein by a β -elimination-like process using dry hydrazine. The sample, placed in a glass vial, should be as dry as possible, salt-free, and in particular free from metal ions. Also, the quality of the hydrazine is essential (metal-free, anhydrous). It is recommended to prepare hydrazine by a double partial distillation under anhydrous argon and to store it under argon

in flame-sealed ampoules. Hydrazinolysis of 50 µg glycoprotein is performed in about 50 µL anhydrous hydrazine under argon for 5–6 h at 60°C. Afterwards, the reagent is removed by evaporation in a desiccator, the glycans are re-*N*-acetylated with acetic anhydride in saturated sodium bicarbonate (15 min, 0°C), desalted on Dowex AG50-X12(H⁺), and dried by vacuum centrifugation. One critical objection has to be made with reference to a general applicability of the method. According to the authors (44) and our own experience, mucin-type *O*-glycoproteins and in particular real mucins with clustered *O*-linked glycans will not give the expected yields of reducing glycans.

An alternative to hydrazinolysis was recently reported by Karlsson and Packer (47), who used nonreductive alkaline β -elimination for liberation of the glycans in a flow system. The principle of the method can be described by three essential steps: (1) the glycoprotein under study is immobilized to an alkali-stable reversed-phase support, (2) the glycans are released by alkali-catalyzed (50 mM KOH) β -elimination at 45°C and immediately removed from the hot reaction chamber, and (3) the glycans, neutralized on a protonated cation exchanger, are trapped on graphitized carbon, from where they can be finally eluted in a small volume. Rapid removal of the released glycans from the hot alkali prevents their ready degradation by “peeling.” This is achieved by an in-line flow system, which pumps the alkaline solution through the heated Poros R2 bead column at a flow rate of 0.1 mL/min. Hence the β -eliminated glycans leave the reaction chamber within seconds, are cooled to room temperature, and immediately pass through the AG50WX8(H⁺) resin. The protocol can be adapted to the analysis of *O*-glycoproteins that have been separated by one- or two-dimensional sodium dodecyl sulfate (SDS)-gel electrophoresis, since the macromolecular analytes are retained in the polyacrylamide during inline flow β -elimination of the glycans. A further advantage can be seen in the fact that no anhydrous conditions have to be maintained and that the generation and handling of hazardous hydrazine is avoided. Moreover, the method has proven to be applicable also in the analysis of real mucins.

Irrespective of the method of their release, the reducing glycans can be labeled with a fluorescing dye, as for example with 2-aminobenzamide (2-AB), according to Bigge et al. (48). Labeling is performed by mixing two volumes of 0.35 M 2-AB in glacial acetic acid with three volumes of 1 M Na(CN)BH₃ in dimethyl sulfoxide (DMSO). The reaction mixture is incubated for 2 h at 60°C, and the 2-AB labeled glycans are then separated from excessive reagent by ascending paper chromatography in 1-butanol-ethanol-water, 4:1:1 (v/v). The 2-AB-labeled glycans are eluted with water and filtered over 0.45-µm membranes prior to HPLC analysis. One disadvantage in the protocol is the partial loss of GalNAc and Gal-GalNAc during paper chromatography.

Analysis of the fluorescently labeled glycans can be performed by a successive anion-exchange chromatography to separate the oligosaccharides according to their sialic acid contents, followed by a normal-phase chromatography on polymer-based amino columns, which separates the glycans according to their sizes, if ionic interactions with the column matrix are suppressed. Identification of the glycans via their retention on HPLC columns is based on external standards of 2-AB-labeled glycans from well-characterized glycoproteins, in conjunction with successive exoglycosidase degradation (49,50). Further confirmation of the glycan structures is achievable by mass-spectrometric analyses.

We have evaluated the applicability of nanoflow normal-phase LC with on-line analysis of the eluate by ESI-qTOF mass spectrometry (Fig. 2 A,B). The glycans in

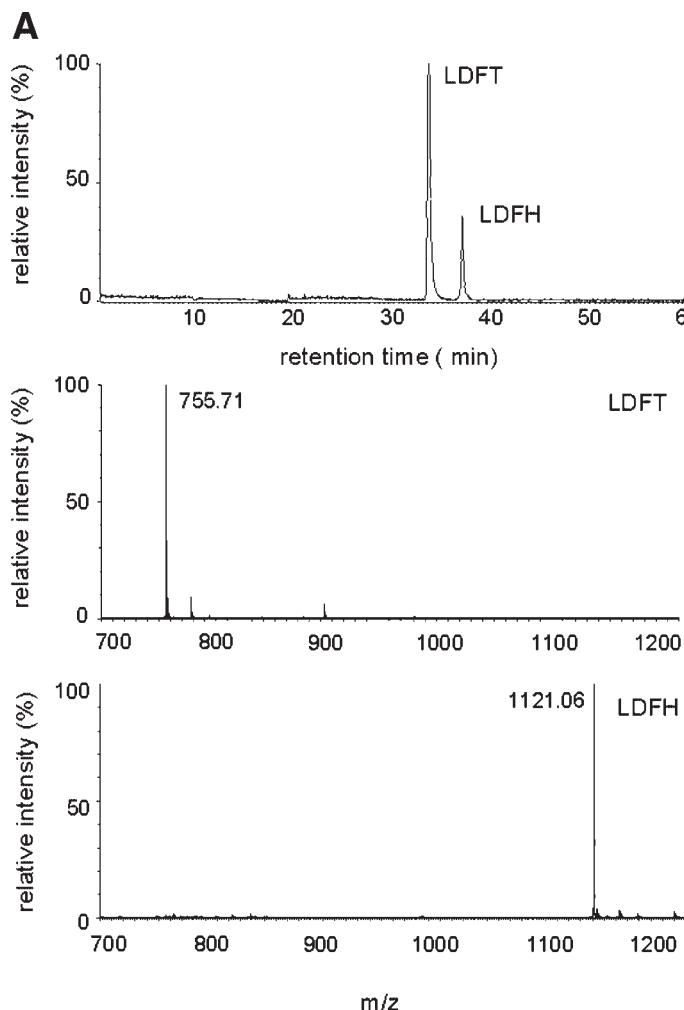


Fig. 2. Normal-phase nanoflow liquid chromatography combined with electrospray ionization (ESI)-q-time-of-flight (TOF) mass spectrometry (MS) of 2-aminobenzamide labelled glycans. 2AB-labelled glycans (1–5 pmol) were run successively over a graphitized carbon-trap column (HyperCarb, 5 μ m, 300 μ m \times 5 mm) followed by the analytical amino-modified silica column (Grom-Sil 80 Amino, 3 μ m, 75 μ m \times 15 cm, both from Grom, Herrenberg, Germany). The trap was equilibrated with water for loading and washing of the sample. Solvent A was a 9:1 mixture of acetonitrile and 50 mM ammonium formate, pH 4.5; solvent B was a 2:8 mixture of acetonitrile and 50 mM ammonium formate, pH 4.5. The trap was run at 30 μ L/min, the

water are first trapped on a 5-mm microcapillary column of graphitized carbon and then eluted with 10% aqueous buffer in acetonitrile onto the analytical column, which is a 15-cm microcapillary filled with amino-modified silica. Elution of the glycans and transfer into the ion source is performed by application of a gradient with an increasing content of the aqueous component (10–80%) within 30 min. Because of the aromatic label elution of the 2-AB, glycans can be registered and quantitated by UV detection at 254 nm. Positive-ion detection in ESI-qTOF mass spectrometry reveals the singly

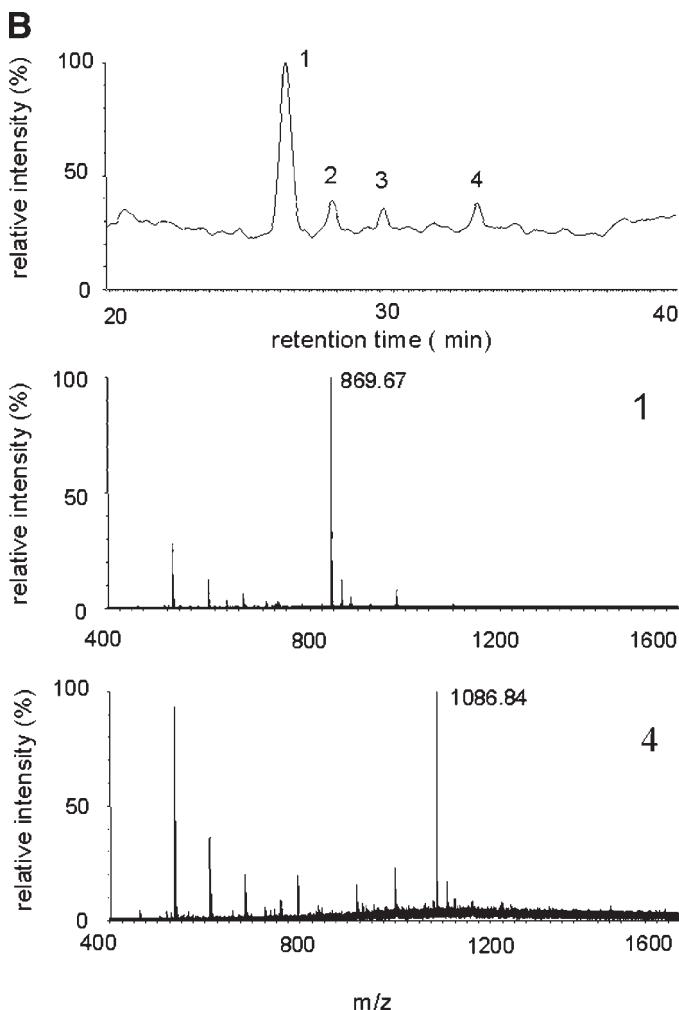


Fig. 2. (continued from opposite page) analytical column with a flow rate of 200 nL/min using a gradient of 0–100% solvent B for 30 min. The liquid chromatography (LC)-MS interface was a Micromass Nanoflow stage with a PRO-ADP PicoTip mounting block and distally coated PicoTips (10 μ m Tip i.d.) (New Objective, Woburn, MA). ESI-qTOF MS was performed in the positive ion mode. (A) LDFT, lacto-difucotetraose; LDFH, lacto-difucohexaose. (B) Mucin-type O-linked glycans from MUC1 fusion protein expressed in the embryonic kidney cell line EBNA-293: Peak 1, core2 tetrasaccharide Gal1-4GlcNAc1-6(Gal1-3)GalNAc; Peak 4, disialylated core1 tetrasaccharide NeuAc2-3Gal1-3(NeuAc2-6)GalNAc.

charged pseudomolecular ions MH^+ , which allow identification of the glycans via their monomeric compositions in conjunction with their retention times. The advantage of normal-phase chromatography over the previously published reversed-phase chromatography on graphitized carbon (51) is the better predictability of the elution behavior, according to the glycan size and structure. A drawback may be found in the fact that the glycans have to be eluted with solvents containing increasing proportions of aqueous buffer. The method can reach the low picomolar or even subpicomolar range, and hence

is similar in sensitivity to a recently described micro-scale approach based on the liberation of *O*-linked glycans by reductive β -elimination from *O*-glycoproteins separated by gel electrophoresis and blotted onto polyvinylidene fluoride membranes (51).

Finally, glycans can be released from *O*-glycoproteins in a derivatized form by combining the alkali-catalyzed β -elimination with rapid methylation of the hydroxy groups. The latter process, which occurs within a few minutes, saves the reducing glycans from degradation by “peeling.” The protocol follows principally the methylation procedure described by Anumula and Taylor (52). Approximately 25 μ g of the *O*-glycoprotein are dried in an Eppendorf tube and resolubilized in dry DMSO (20 μ L). The reaction is started by rapidly adding first methyl iodide (10 μ L) and, within 1 min, a finely powdered suspension of 2 *M* NaOH in dry DMSO (20 μ L). The reaction mixture is left at room temperature for 18 h and is then extracted with chloroform, which is washed three times with water. The chloroform phase is dried down in a stream of nitrogen and solubilized in organic solvents (methanol or acetonitrile) for mass spectrometric analysis. The methylated glycans in DHB matrix can be detected as their sodium adduct ions by positive-ion MALDI mass spectrometry (50). This approach cannot be regarded as being quantitative, but has the advantage of giving insight into the *O*-glycosylation profile of a glycoprotein by a rather simple, rapid, and convenient method.

References

1. Bruckner, K., Perez, L., Clausen, H., and Cohen, S. (2000) Glycosyltransferase activity of Fringe modulates Notch-Delta interactions. *Nature* **27**, 411–415.
2. Seppo, A. and Tiemeyer, M. (2000) Function and structure of *Drosophila* glycans. *Glycobiology* **10**, 751–760.
3. Wilson, I. B. H. (2002) Glycosylation of proteins in plants and invertebrates. *Curr. Opin. Struct. Biol.* **12**, 569–577.
4. Hanisch, F. G. (2001) O-Glycosylation of the mucin-type. *Biol. Chem.* **382**, 143–149.
5. Zhang, H., Li, X. J., Martin, D. B., and Aebersold, R. (2003) Identification and quantification of N-linked glycoproteins using hydrazide chemistry, stable isotope labelling and mass spectrometry. *Nat. Biotechnol.* **21**, 660–666.
6. Hart, G. W. (1997) Dynamic O-linked glycosylation of nuclear and cytoskeletal proteins. *Annu. Rev. Biochem.* **66**, 315–335.
7. Comer, F. I. and Hart, G. W. (2000) O-Glycosylation of nuclear and cytosolic proteins. Dynamic interplay between O-GlcNAc and O-phosphate. *J. Biol. Chem.* **275**, 29179–29182.
8. Vosseller, K., Wells, L., and Hart, G. W. (2001) Nucleocytoplasmic O-glycosylation: O-GlcNAc and functional proteomics. *Biochimie* **83**, 575–581.
9. Roquemore, E. P., Chou, T. Y., and Hart, G. W. (1994) Detection of O-linked N-acetylglucosamine (O-GlcNAc) on cytoplasmic and nuclear proteins. *Meth. Enzymol.* **230**, 443–460.
10. Wells, L., Vosseller, K., Cole, R. N., Cronshaw, J. M., Matunis, M. J., and Hart, G. W. (2002) Mapping sites of O-GlcNAc modification using affinity tags for serine and threonine post-translational modifications. *Mol. Cell. Proteomics* **1**, 791–804.
11. Gupta, R., Birch, H., Rapacki, K., Brunak, S., and Hansen, J. E. (1999) O-GLYCBASE version 4.0: a revised database of O-glycosylated proteins. *Nucleic Acid Res.* **27**, 370–372.
12. Ten Hagen, K. G., Fritz, T. A., and Tabak, L. A. (2003) All in the family: the UDP-GalNAc: polypeptide N-acetylgalactosaminyltransferases. *Glycobiology* **13**, 1R–16R.
13. Wang, H., Tachibana, K., Zhang, Y., et al. (2003) *Biochem. Biophys. Res. Commun.* **300**, 738–744.

14. Hanisch, F.-G., Müller, S., Hassan, H., et al. (1999) Dynamic epigenetic regulation of initial O-glycosylation by UDP-GalNAc: polypeptide N-acetylgalactosaminyl- transferases. *J. Biol. Chem.* **274**, 9946–9954.
15. Hanisch, F.-G., Reis, C. A., Clausen, H., and Paulsen, H. (2001) Evidence for glycosylation-dependent activities of polypeptide N-acetylgalactosaminyl- transferases rGalNAc-T2 and -T4 on mucin glycopeptides. *Glycobiology* **11**, 731–740.
16. Pisano, A., Redmond, J. W., Williams, K. L., and Gooley, A. A. (1993) Glycosylation sites identified by solid-phase Edman degradation: O-linked glycosylation motifs on human glycopohorin A. *Glycobiology* **3**, 429–435.
17. Müller, S., Goletz, S., Packer, N., Gooley, A. A., Lawson, A. M., and Hanisch, F.-G. (1997) Localization of O-glycosylation sites on glycopeptide fragments from lactation-associated MUC1. *J. Biol. Chem.* **272**, 24,780–24,793.
18. Schwientek, T., Bennett, E. P., Flores, C., et al. (2002) Functional conservation of sub-families of putative UDP-GalNAc: polypeptide N-acetylgalactosaminyltransferases in *Drosophila*, *Caenorhabditis elegans*, and mammals. One subfamily composed of 1(2)35Aa is essential in *Drosophila*. *J. Biol. Chem.* **277**, 22,623–22,638.
19. Yeaman, C., Le Gall, A. H., Baldwin, A. N., Monlauzeur, L., Le Bivic, A., and Rodriguez, Boulan, E. (1997) The O-glycosylated stalk domain is required for apical sorting of neurotrophin receptors in polarized MDCK cells. *J. Cell Biol.* **139**, 929–940.
20. Altschuler, Y., Kinlough, C. L., Poland, P. A., et al. (2000) Clathrin-mediated endocytosis of MUC1 is modulated by its glycosylation state. *Mol. Biol. Cell* **11**, 819–831.
21. Bulet, P., Dimarcq, J. L., Hetru, C., et al. (1993) A novel inducible antibacterial peptide of *Drosophila* carries an O-glycosylated substitution. *J. Biol. Chem.* **268**, 14,893–14,897.
22. Stadie, T., Chai, W., Lawson, A. M., Byfield, P., and Hanisch, F.-G. (1995) Studies on the order and site-specificity of GalNAc-transfer to MUC1 tandem repeats by UDP-GalNAc: polypeptide N-acetylgalactosaminyltransferases from milk or mammary carcinoma cells. *Eur. J. Biochem.* **229**, 140–147.
23. Hillenkamp, F. and Karas, M. (1990) Mass spectrometry of peptides and proteins by matrix-assisted ultraviolet laser desorption/ionization. *Meth. Enzymol.* **193**, 280–295.
24. Spengler, B., Lützenkirchen, F., and Kaufmann, R. (1993) On-target deuteration for peptide sequencing by laser mass spectrometry. *Org. Mass Spectrom.* **28**, 1482–1490.
25. Kaufmann, R., Kirsch, D., and Spengler, B. (1994) Sequencing of peptides in a time-of-flight mass spectrometer: evaluation of post-source-decay following matrix-assisted laser desorption ionization (MALDI). *Int. J. Mass Spectrom. Ion. Processes* **131**, 355–385.
26. Goletz, S., Thiede, B., Hanisch, F.-G., et al. (1997) A sequencing strategy for the localization of O-glycosylation sites of MUC1 tandem repeats by PSD-MALDI mass spectrometry. *Glycobiology* **7**, 881–896.
27. Goletz, S., Leuck, M., Franke, P., and Karsten, U. (1997) Structure analysis of acetylated and non-acetylated O-linked MUC1 glycopeptides by post-source-decay matrix-assisted laser desorption ionization mass spectrometry. *Rapid. Commun. Mass Spectrom.* **11**, 1387–1398.
28. Biemann, K. (1988) Contributions of mass spectrometry to peptide and protein structure. *Biomed. Environ. Mass Spectrom.* **16**, 99–111.
29. Roepstorff, P. and Fohlmann, J. (1984) Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomed. Mass Spectrom.* **11**, 601.
30. Domon, B. and Costello, C. (1988) A systematic nomenclature for carbohydrate fragmentations in FAB-MS/MS spectra of glycoconjugates. *Glycoconjugate J.* **5**, 397–409.
31. Carr, S. A., Hemling, M. E., Folena-Wasserman, G., et al. (1989) Protein and carbohydrate structural analysis of a recombinant soluble CD4 receptor by mass spectrometry. *J. Biol. Chem.* **264**, 21,286–21,295.

32. Carr, S. A., Huddleston, M. J., and Bean, M. F. (1993) Selective identification and differentiation of N- and O-linked oligosaccharides in glycoproteins by liquid chromatography-mass spectrometry. *Protein Sci.* **2**, 183–196.
33. Peter-Katalinic, J., Williger, K., Egge, H., Green, B., Hanisch, F.-G., and Schindler, D. (1994) The application of electrospray mass spectrometry for structural studies on a tetrasaccharide monopeptide from the urine of a patient with α -N-acetyl-hexosaminidase deficiency. *J. Carbohydr. Chem.* **13**, 447–456.
34. Peter-Katalinic, J., Ashcroft, A., Green, B., et al. (1994) Potential of electrospray mass spectrometry for structural studies of O-glycoamino acids and -peptides with multiple α 2,6-sialosyl-T_n glycosylation. *Org. Mass Spectrom.* **29**, 747.
35. Hanisch, F.-G., Green, B. N., Bateman, R., and Peter-Katalinic, J. (1998) Localization of O-glycosylation sites of MUC1 tandem repeats by Qtof ESI mass spectrometry. *J. Mass Spectrom.* **33**, 358–362.
36. Alving, K., Paulsen, H., and Peter-Katalinic, J. (1999) Characterization of O-glycosylation sites in MUC2 glycopeptides by nanoelectrospray Qtof mass spectrometry. *J. Mass Spectrom.* **34**, 395–407.
37. Mirgorodskaya, E., Hassan, H., Wandall, H. H., Clausen, H., and Roepstorff, P. (1999) Partial vapor-phase hydrolysis of peptide bonds: a method for mass spectrometric determination of O-glycosylated sites in glycopeptides. *Anal. Biochem.* **269**, 54–65.
38. Rademaker, G. J., Pergantis, S. A., Blok-Tip, L., Langridge, J. I., Kleen, A., and Thomas-Oates, J. E. (1998) Mass spectrometric determination of the sites of O-glycan attachment with low picomolar sensitivity. *Anal. Biochem.* **257**, 149–160.
39. Hanisch, F.-G., Jovanovic, M., and Peter-Katalinic, J. (2001) Glycoprotein identification and localization of O-glycosylation sites by mass spectrometric analysis of deglycosylated/alkylaminylated peptide fragments. *Anal. Biochem.* **290**, 47–59.
40. Mirgorodskaya, E., Hassan, H., Clausen, H., and Roepstorff, P. (2001) Mass spectrometric determination of O-glycosylation sites using beta-elimination and partial acid hydrolysis. *Anal. Chem.* **73**, 1263–1269.
41. Boersma, A., Lamblin, G., Degand, P., and Roussel, P. (1981) Separation of a complex mixture of oligosaccharides by HPLC on bonded-primary amine using a linear-gradient solvent system. *Carbohydr. Res.* **94**, C7–C9.
42. Merry, A. H., Neville, D. C., Royle, L., et al. (2002) Recovery of intact 2-aminobenzamide-labelled O-glycans released from glycoproteins by hydrazinolysis. *Anal. Biochem.* **304**, 91–99.
43. Carlson, D. M. (1968) Structures and immunochemical properties of oligosaccharides isolated from pig submaxillary mucins. *J. Biol. Chem.* **243**, 616–626.
44. Patel, T. P. and Parekh, R. B. (1994) Release of oligosaccharides from glycoproteins by hydrazinolysis. *Meth. Enzymol.* **230**, 57–66.
45. Chai, W., Feizi, T., Yuen, C. T., and Lawson, A. M. (1997) Non-reductive release of O-linked oligosaccharides from mucin glycoproteins for structure/function assignments as neoglycolipids: application in the detection of novel ligands for E-selectin. *Glycobiology* **7**, 861–872.
46. Cooper, C. A., Packer, N. H., and Redmond, J. W. (1994) The elimination of O-linked glycans from glycoproteins under non-reducing conditions. *Glycoconj. J.* **11**, 163–167.
47. Karlsson, N. G. and Packer, N. H. (2002) Analysis of O-linked reducing oligosaccharides released by an in-line flow system. *Anal. Biochem.* **305**, 173–185.
48. Bigge, J. C., Patel, T. P., Bruce, J. A., Goulding, P. N., Charles, S. M., and Parekh, R. B. (1995) Nonselective and efficient fluorescent labeling of glycans using 2-aminobenzamide and anthranilic acid. *Anal. Biochem.* **230**, 229–238.

49. Royle, L., Roos, A., Harvey, D. J., et al. (2003) Secretory IgA N- and O-glycans provide a link between the innate and adaptive immune systems. *J. Biol. Chem.* **278**, 20140–20153.
50. Müller, S. and Hanisch, F.-G. (2002) Recombinant MUC1 probe authentically reflects cell-specific O-glycosylation profiles of endogenous breast cancer mucin. *J. Biol. Chem.* **277**, 26103–26112.
51. Schulz, B. L., Packer, N. H., and Karlsson, N. G. (2002) Small-scale analysis of O-linked oligosaccharides from glycoproteins and mucins separated by gel electrophoresis. *Anal. Chem.* **74**, 6088–6097.
52. Anumula, K. R. and Taylor, P. B. (1992) A comprehensive procedure for preparation of partially methylated alditol acetates from glycoprotein carbohydrates. *Anal. Biochem.* **203**, 101–108.
53. Hang, H. C., Yu, C., Kato, D. L., and Bertozzi, C. R. (2003) A metabolic labeling approach toward proteomic analysis of mucin-type O-linked glycosylation. *Proc. Natl. Acad. Sci. USA* **100**, 14,846–14,851.
54. Julenius, K., Molgaard, A., Gupta, R., and Brunak, S. (2004) Prediction, conservation analysis and structural characterization of mammalian mucin-type O-glycosylation sites. *Glycobiology*, in press.

Identification of Protein Phosphorylation Sites by Mass Spectrometry

Alastair Aitken

1. Introduction

In this chapter, methods will be reviewed for the identification of the type of modified amino acid residue (serine, threonine, or tyrosine) and position in the sequence of a phospho-protein/peptide.

Reversible phosphorylation of proteins on serine, threonine, and tyrosine residues is now established as a major intracellular regulatory mechanism. Different protein kinases show widely preferred substrate specificities in their target proteins, and knowledge of the amino acid sequence surrounding a site of phosphorylation is vitally important to establish the class of kinase that is responsible for regulating the activity of the protein or enzyme being studied. The protein kinases that regulate the intracellular processes are themselves subject to regulation in a wide variety of ways. Many kinases are regulated by second messengers, and a number are under control of calcium ions. Other mechanisms for regulation of protein kinases include the presence of pseudo-substrate sequences in the regulatory domain. In addition, many protein kinases undergo autophosphorylation.

Analysis of modified peptides where no ^{32}P or other radiolabel is present, may be particularly difficult. Even the identification of phosphorylation sites where ^{32}P -radioactivity is present may pose problems if one relies on phosphopeptide map analysis. Depending on exact type of column and gradient, phosphorylated peptides will in general elute slightly ahead (1 to 2% acetonitrile) of their unphosphorylated analogs on reverse-phase high-performance liquid chromatography (HPLC) (1) (but see Note 1).

Immobilized metal-ion affinity chromatography (IMAC, or MC) using metal-affinity chromatography pipet tips is effective for the enrichment of phosphopeptides (2,3). This technique is also used for the purification of His-tagged proteins. The method describes the use of ZipTipMC pipet tips containing immobilized metal-affinity chromatography media (IDA resin) for enriching phosphopeptides. Alternatively, one can prepare one's own metal-ion affinity columns. Adsorption is based on the complex between the phosphate and metal ions that forms under acidic conditions (pH 2.5 to 5.5) (see Note 2). The phosphopeptides are desorbed at alkaline pH. Since the phosphate is attached to a diverse polypeptide backbone, the adsorption may not be entirely selective. The carboxylic acid groups of aspartic and glutamic acid also bind to the metal ions under similar conditions. The background adsorption of acidic peptides can

be reduced by repeated washes with dilute acid (pH 2.5 to 3.5). In some cases, mono-phosphorylated peptides may not be captured if the phosphate is inaccessible. The selectivity for phosphopeptides is enhanced by using buffers such as 2-(*N*-morpholino)ethanesulfonic acid (MES) at pH 5.5, or weak acids such as formic or acetic acid. The choice of metal may also vary, depending on the peptide digest and application. The metal ions used are those of gallium (III), iron (III) (aluminium III has also been used), copper (II), and nickel (II).

Detection of phosphopeptides on matrix-assisted laser desorption/ionization (MALDI)-time-of-flight (TOF) mass spectrometry (MS) after prior dephosphorylation with phosphatases is now frequently used. The MALDI-MS spectra are compared before and after treatment with phosphatase, and the characteristic 80-Da loss due to removal of phosphate is diagnostic. Dephosphorylation with alkaline phosphatase goes to completion at the picomole level of phosphopeptides. Because of the complexity of the mixture of peptides in an unenriched digest of a protein or mixture of proteins, it is unlikely that one would unequivocally identify the masses of the phosphopeptides in this technique.

Because the recovery of serine, threonine, or tyrosine after complete acid hydrolysis of a peptide may be very low if one of these is present as a phosphoamino acid, tandem mass spectrometric analysis is essential to confirm the exact amino acid composition and number of phosphorylated residues. If this is not available, then subdigestion of a peptide containing multiple phosphorylatable peptides, with an appropriate proteinase, is advisable (see Chapter 35). Identification of either positive or negative ions may yield more information, depending on the mode of ionization and fragmentation of an individual peptide. Phosphopeptides may give better ions in the negative mode, since they have a strong negative charge due to the phosphate group. However, phosphopeptides can run well on electrospray MS and MALDI-TOF in positive ion mode, although particular problems may be associated with electrospray MS of phosphopeptides, where high levels of Na^+ and K^+ adducts are regularly seen on species at charge states above that predicted by the theoretical number of basic groups (see **Figs. 1** and **2**). A particularly useful method exists for the derivatization of phosphoserine to the stable *S*-ethylcysteinyl derivatives (**4**) by β -elimination of the phosphate followed by addition of ethanethiol. This provides excellent structural information on MS.

2. Materials

2.1. Metal-Affinity Chromatography

1. ZipTipMC pipet tips (Millipore).
2. Pipet such as Gilson or Finnpipette P-10 pipet.
3. 200 mM Metal ion solutions: copper sulphate, nickel chloride, or gallium nitrate in high-purity water; ferric chloride in high-purity water with 10 mM HCl.
4. 0.1% Acetic acid with 50% acetonitrile (acetonitrile concentration may be adjusted from 5 to 30%, depending on starting material) (see **Note 3**).
5. High-purity water.
6. 1.0% Acetic acid with 10% acetonitrile.
7. Acidic binding solutions: 50 mM MES buffer (Sigma or Fisher) with 10% acetonitrile, adjusted to pH 5.5 with high-purity ammonium hydroxide or 0.1% acetic acid/10% acetonitrile, or 0.01% formic acid with 10% acetonitrile.
8. Aqueous ammonium hydroxide, elution solution (0.02% to 2.0%; see **Note 4**).
9. Other materials are as listed in Chapter 29.

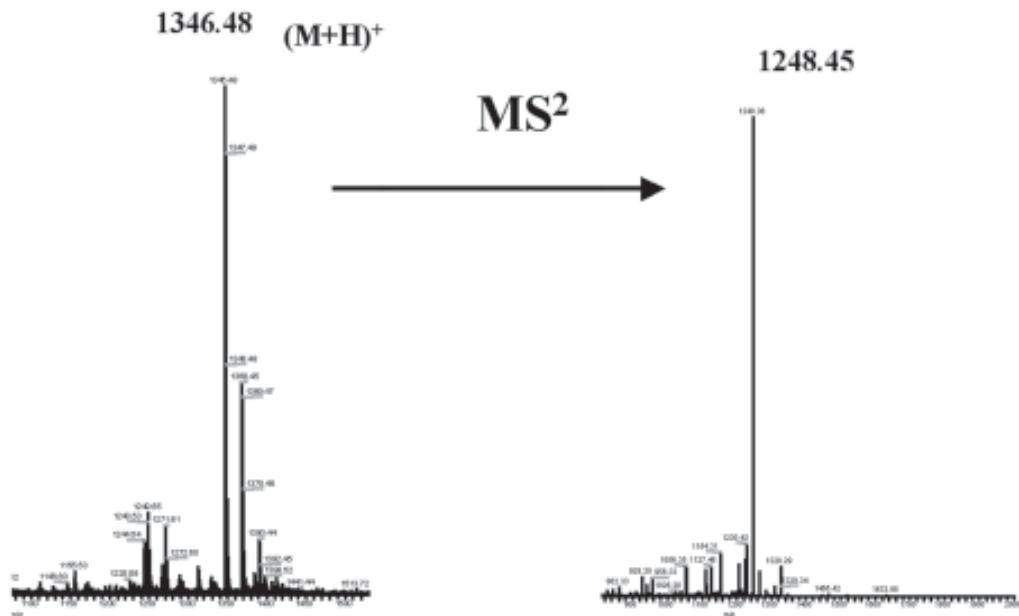


Fig. 1. Electrospray mass spectrum of phosphopeptide. The sample was analyzed on a ThermoFinnigan nanoelectrospray ion-trap mass spectrometer. The tandem mass spectrometry (MS)² shows loss of 98 Da, representing loss of H₃PO₄, verifying the presence of a phosphate group. (Sample analyzed by Dr. Rob Wakefield).

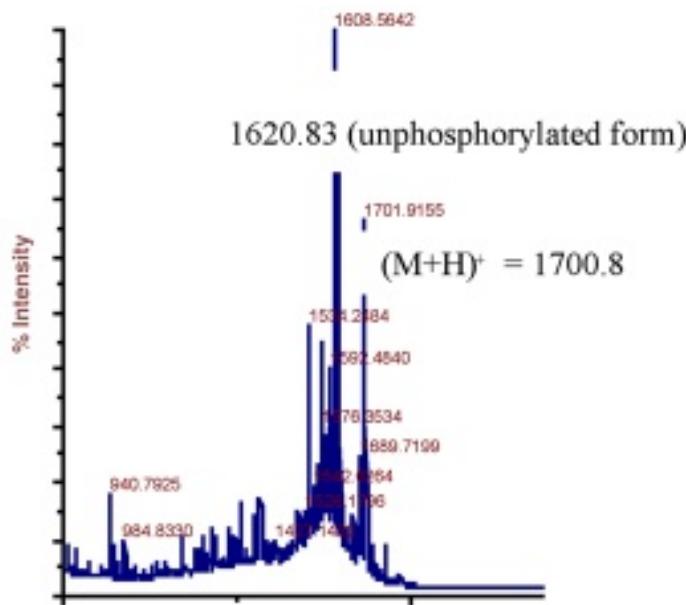


Fig. 2. Matrix-assisted laser desorption/ionization (MALDI)-time-of-flight (TOF) mass spectrometry (MS) of phosphopeptide in a mixture. The sample was analyzed on an Applied Biosystems Voyager-DE STR MALDI-TOF mass spectrometer. The peptide sequence is CFNRTS^PLPWQQLKA (M + H)⁺ = 1700.8 (1620.83 as the unphosphorylated form). S^P is phosphoserine. Sample analyzed by Dr. Caroline Johnstone.

2.2. Modification of Phosphoserine to S-Ethylcysteine

1. Incubation mixture: 0.2 mL water, 0.2 mL dimethylsulphoxide, 100 μ L ethanol, 65 μ L 5 M sodium hydroxide, and 60 μ L 10 M ethanethiol.
2. Glacial acetic acid.
3. Nitrogen gas source.

2.3. Detection of Phosphopeptides After Dephosphorylation With Phosphatases

1. Calf intestinal alkaline phosphatase stock solution of 20 U/ μ L (New England Biolabs or Roche Diagnostics).
2. 100 mM Tris-HCl, pH 8.0.
3. 50 mM NH_4HCO_3 .
4. 50% Acetonitrile.
5. High-purity water.

3. Methods

3.1. Enrichment of Phosphopeptides Before MALDI-TOF and Nanoelectrospray MS Using Metal-Affinity Chromatography Pipet Tips

3.1.1. Charging and Equilibration of the ZipTipMC

1. Prepare 1 mL metal ion solution.
2. Wash tip and dispense to waste three times with 10 μ L fresh 0.1% acetic acid with 50% acetonitrile.
3. Load the column with metal ions by charging the tip with multiple column volumes of this solution by carrying out 10 aspirate and dispense cycles with 10- μ L aliquots of metal ion solution.
4. Wash the tip and dispense to waste three times with 10 μ L of high-purity water. Wash the tip and dispense to waste three times with 10 μ L of fresh 1.0% acetic acid, 10% acetonitrile.
5. Equilibrate the tip by washing and dispensing to waste five times with 10 μ L of appropriate binding solution.
6. After use, the tips may be washed and regenerated for reuse up to three times.

3.1.2. Binding and Washing Using MES Buffer (pH 5.5)

1. Dilute sample in acidic binding solution (see Note 2), such as MES, formic acid, or acetic acid in aqueous acetonitrile.
2. Bind sample to ZipTip by fully depressing the pipet plunger to a dead stop using a 1-to 10- μ L volume.
3. Aspirate and dispense the sample 5 to 10 cycles for maximum binding.
4. Wash tip and dispense to waste three times with 10 μ L of binding solution.
5. Wash tip and dispense to waste three times with 10 μ L 0.1% acetic acid with 50% acetonitrile.
6. Wash tip and dispense to waste three times with 10 μ L high-purity water.

3.1.3. Alternative: Binding and Washing Using 0.1% Acetic or 0.01% Formic Acid

1. Dilute the sample in binding solution (0.1% acetic or 0.01% formic acid).
2. Bind the sample to the ZipTip by fully depressing the pipet plunger to a dead stop using a 1- to 10- μ L volume.
3. Aspirate and dispense the sample 5 to 10 cycles for maximum binding.
4. Wash tip and dispense to waste three times with binding solution.

5. Wash tip and dispense to waste three times with 10 mL 0.1% acetic acid (or 0.01% formic acid), 50% acetonitrile.
6. Wash tip and dispense to waste three times with high-purity water.

3.1.4. Eluting the Phosphopeptides for MALDI-TOF MS

1. Pipet 2 μ L of freshly prepared ammonium hydroxide elution solution (see Note 4) into a clean vial using a standard pipet tip.
2. Aspirate and dispense eluant through the ZipTip four to six times without introducing air.
3. For direct spotting onto a MALDI-TOF MS target plate, aspirate the desired volume of eluted phosphopeptide solution into the ZipTip and dispense directly onto target, let dry partially, then overspot with 1 μ L of matrix (see Note 5).

3.1.5. Eluting the Phosphopeptides for Nanoelectrospray MS (see Note 6)

1. Carry out steps 1 and 2 in Subheading 3.1.4.
2. For direct loading into a nanoelectrospray MS needle, elute the sample into a clean Eppendorf tube or vial, or use a GELoader tip to introduce directly into a nanospray needle.
3. Cut the GELoader tip approx 2–3 mm above where the tip is fused to its capillary (i.e., narrow) end.
4. Before final dispensing of the sample, press the cut-down GELoader tip firmly onto the ZipTip pipet tip with a slight twisting motion. The leak-free fit allows elution directly into the nanospray needle.

3.2. Modification of Phosphoserine to S-Ethylcysteine

1. In a fume cupboard, the peptide is dissolved in a capped tube containing 50 μ L incubation mixture in a capped Eppendorf tube. (Caution: ethanethiol has very strong and characteristic odor, reminiscent of a mains gas leak. Carry out this procedure in a properly maintained fume cupboard and alert your colleagues and security to the fact that you are using this compound.)
2. Flush the tube with nitrogen and incubate for 1 h at 50°C.
3. After allowing to cool, add 10 μ L of glacial acetic acid.
4. Apply the derivatized peptide either directly for analysis or concentrate first by vacuum centrifugation.
5. The β -elimination step during derivatization of a phosphoserine adjacent to a proline residue is slow; therefore, reaction time may be extended to 18 h at 50°C. Acetonitrile can be used instead of the normal solvents to minimize subsequent manipulations.
6. This procedure is also used for the selective isolation of phosphoseryl peptides (4). When the S-ethylcysteinyl peptides are applied to a reverse-phase HPLC column (e.g., Vydac C₁₈) and eluted with linear gradients of water/acetonitrile in 0.1% trifluoroacetic acid (TFA), the derivatized peptides emerge on average 4 to 5% acetonitrile later than the native phosphopeptide. A derivatized peptide from a doubly phosphorylated species will elute correspondingly later than the singly derivatized species, indicating the applicability of the method to multiple-phosphoseryl peptides. HPLC before and after derivatization by this “diagonal” technique should produce highly purified peptides even from a very complex mixture, since the elution position of all others will be unaffected.

3.3. Selective Detection of Phosphopeptides on Mass Spectrometry

1. Phosphate-specific fragment ions of 63 Da (PO₂⁻) and 79 Da (PO₃⁻) are produced by collision-induced dissociation during negative-ion LC-electrospray MS (5). This technique of selective detection of posttranslational modifications through collision-induced formation of low-mass fragment ions that serve as characteristic and selective markers for the

modification of interest, has been extended to identify other modifications such as glycosylation, sulphation, and acylation (see Chapter 42).

2. Ladder sequencing by MS (6) involves the generation of a set of nested fragments of a polypeptide chain followed by analysis of the mass of each component. Each component in the ragged polypeptide mixture differs from the next by loss of a mass that is characteristic of the residue weight (which may involve a modified side chain). In this manner, the sequence of the polypeptide can be read from the masses obtained in MS (see Chapters 31, 33, and 35 for details).

3.4. Detection of Phosphopeptides After Dephosphorylation With Alkaline Phosphatase (7)

1. Carry out MALDI-TOF mass analysis of aliquots of each sample that has been enriched in phosphopeptides after IMAC or after HPLC.
2. Dissolve separate samples in 0.5 μ L of 50mM NH_4HCO_3 , pH 8.9, containing 0.05 or 1 unit of calf intestinal alkaline phosphatase/ μ L.
3. Incubate samples for 2 h at 37°C.
4. For MALDI-TOF, the peptides may be desalted on C_{18} ZipTips (see Chapter 29) and eluted directly onto the target plate with 80% (v/v) aqueous acetonitrile, 0.1% TFA, containing the MALDI matrix, α -cyano-4-hydroxycinnamic acid (Aldrich).
5. A mass shift of 80 Da caused by the removal of the phosphate allows unambiguous determination of whether a particular peptide is phosphorylated.

3.4.1. On-Target Alkaline Phosphatase Treatment

1. Carry out MALDI-TOF mass analysis of samples enriched in phosphopeptides as in **Subheading 3.4.** using α -cyano-4-hydroxycinnamic acid matrix.
2. Dissolve the sample/matrix mixture on the target with 1–1.5 μ L of 50 mM ammonium bicarbonate containing 0.05 U/ μ L alkaline phosphatase.
3. Incubate the same samples *in situ* with the alkaline phosphatase on the MALDI target for 1–2 h at 37°C in a closed high-humidity chamber (a plastic box with a snap-on lid, containing wet tissue, is ideal) to prevent drying.
4. Stop the dephosphorylation reaction by the addition of 0.5 μ L of acetonitrile/TFA solution and dry samples immediately *in vacuo* to allow proper crystallization of the matrix.
5. Repeat the MALDI-TOF MS as before to determine which peptides have altered in mass by 80 Da (or multiples thereof).

4. Notes

1. If the HPLC separation is combined with mass spectrometric characterization, the level of TFA required for sharp peaks and good resolution of peptide (approx 0.1% v/v) results in near or complete suppression of signal. This does not permit true online HPLC-MS, as the concentration of TFA in the eluted peptide must first be drastically reduced. However, the new “low TFA,” 218MS54, reverse-phase HPLC columns from Vydac (300 Å pore size) are available in C_4 and two forms of C_{18} chemistries. They are also supplied in 1-mm diameter columns that are ideal for low levels of sample eluted in minimal volume. We have used as little as 0.005% TFA without major loss of resolution, and have observed minimal signal loss. There may be a difference in selectivity compared to “classical” reverse-phase columns—for example, we have observed phosphopeptides eluting approx 1% acetonitrile later than their unphosphorylated counterparts (the opposite to that conventionally seen). This is not a problem, but it is something of which one should be aware, and could be turned to advantage. All buffers should be prepared using freshly deionized water (such as Milli-Q), filtered before use through 0.45- μ m filters and degassed.

2. Enhanced binding of phosphopeptides to the ZipTipMC is achieved in the presence of a low-pH buffer such as 50 mM MES buffer (pH 5.5), or 0.1 % acetic acid or 0.01% formic acid. The sample containing phosphopeptide(s) could be subjected to a preliminary clean-up step with reverse-phase pipet-tip chromatography (see Chapter 29). If so, the elution buffer consisting of acetonitrile with 0.1%TFA will be compatible. Phosphopeptide binding to IMAC beads is essentially quantitative, provided the maximal binding capacity is not exceeded (around 20 pmol/μL resin).
3. Optimal performance of the ZipTipMC is dependent on the peptide digest, pH, or metal ion used. Maintaining a low pH (2.5 to 5.5) is critical for optimal phosphopeptide binding. Phosphopeptide binding and recovery differs depending on the metal ion used. The Millipore website protocol (www.millipore.com/ziptip) recommends considering first charging the ZipTipMC with copper ions. If results are unsatisfactory charge a new ZipTip with nickel, gallium, or iron, respectively. Most publications that involve use of custom-made metal-affinity tips appear to obtain optimal results with gallium (III) or iron (III), in that order. The protocols include 10% acetonitrile in the bind and wash steps to reduce nonspecific binding, but higher percentage acetonitrile can be used. A detergent such as 0.01% Tween or 30 mM imidazole can also be used in the binding and washing steps to minimize nonspecific binding. A similar system involving mini-spin columns, using a preequilibrated disc containing gallium (III), with a binding capacity for phosphopeptides of approx 150 μg, is available from Pierce (www.piercenet.com/).
4. The ammonium hydroxide elution solution may vary in strength from 2.0% to 0.02% w/v according to various authors. 0.1 M aqueous ammonium hydroxide is 0.17% w/v.
5. MALDI-TOF mass spectrometry is carried out in both linear and reflector-positive mode using α-cyano-4-hydroxycinnamic acid (saturated solution in 50% acetonitrile with 0.1% TFA). Alternatively, a matrix such as 2,4,6-trihydroxyacetophenone (THAP, Fluka) may be used to enhance detection of phosphorylated peptides on MALDI-TOF MS.
6. Some phosphopeptides may bind tightly to the beads and be recovered in low yield. Therefore, after washing the beads, such phosphopeptides can be released from aliquots of beads with 5 μL of 100 mM Tris-HCl (pH 8.0) (as a control), and with 5 μL of 100 mM Tris-HCl (pH 8.0) containing 0.5 units calf intestinal alkaline phosphatase and incubated for 15 min at 37°C. The disadvantage of this is that one may not be certain that a particular peptide had been phosphorylated. Beads are removed by centrifugation and the supernatant acidified with 0.5 μL of 10% (v/v) TFA before mass spectrometry.

References

1. Dubois, T., Howell, S., Zemlickova, E., Learmonth, M., Cronshaw, A., and Aitken, A. (2003) Novel in vitro and in vivo phosphorylation sites on protein phosphatase 1 inhibitor CPI-17. *Biochem. Biophys. Res. Commun.* **302**, 186–192.
2. Stensballe, A., Andersen, S., and Jensen, O. N. (2001) Characterization of phosphoproteins from electrophoretic gels by nanoscale Fe(III) affinity chromatography with off-line mass spectrometry analysis. *Proteomics* **1**, 207–222.
3. Raska, C. S., Parker, C. E., Dominski, Z., et al. (2002) Direct MALDI-MS/MS of phosphopeptides affinity-bound to immobilized metal ion affinity chromatography beads. *Anal. Chem.* **74**, 3429–3433.
4. Holmes, C. F. B. (1987) A new method for the selective isolation of phosphoserine-containing peptides. *FEBS Lett.* **215**, 21–24.
5. Annan, R. S., Huddleston, M. J., Verma, R., Deshaies, R. J., and Carr, S. A. (2001) A multidimensional electrospray MS-based approach to phosphopeptide mapping. *Anal. Chem.* **73**, 393–404.

6. Bartlet-Jones, M., Jeffery, W. A., Hansen, H. F., and Pappin, D. J. C. (1994) Peptide ladder sequencing by mass spectrometry using a novel, volatile degradation reagent. *Rapid Comm. Mass Spec.* **8**, 737–742.
7. Larsen, M. R., Sorensen, G. L., Fey, S. J., Larsen, P. M., and Roepstorff, P. (2001) Phosphoproteomics: evaluation of the use of enzymatic de-phosphorylation and differential mass spectrometric peptide mass mapping for site specific phosphorylation assignment in proteins separated by gel electrophoresis. *Proteomics* **1**, 223–238.

Quantitative Analysis of Protein Phosphorylation Status and Protein Kinase Activity on Microarrays Using Pro-Q™ Diamond Dye Technology

Karen Martin and Wayne F. Patton

1. Introduction

The human genome is estimated to contain 30,000 to 75,000 genes, but as a result of alternative mRNA splicing and protein posttranslational modifications, the human proteome may contain a million or more proteins. Of the potential posttranslational modifications, phosphorylation by protein kinases is of increasing importance and interest because of its involvement in many essential cellular processes such as metabolism, transcription, cell-cycle regulation, and apoptosis, as well as intercellular communication. The malfunction of these and other kinase-regulated processes is often manifested in clinical disease states.

This chapter outlines the use of a small-molecule fluorophore phosphosensor technology to provide global detection and quantitation of phosphorylated amino acids without a requirement for antibodies or radioactivity(1,2). Pro-Q™ Diamond phosphopeptide/phosphoprotein microarray stain is easy to use, readily reversible, and adaptable to many solid-phase applications (3). We outline the use of this technology with two-pad kinase substrate chips designed to profile the activity of AGC, CAMK, CMGC, CKI, and PTK families of protein kinases. **Figure 1** illustrates the results of incubating purified Abl tyrosine kinase (**Fig. 1A**) on the two-pad kinase-characterization arrays. The control reaction without enzyme was performed simultaneously (**Fig. 1B**). Kinase phosphorylation was characterized using Pro-Q Diamond microarray stain. Both results were quantitated using ImageGauge image analysis software (Fuji Photo Film, Tokyo, Japan) in profile mode. Profile mode quantitates the peak intensity of each spot. Profiles of intensity values are compared and illustrated (**Fig. 1C**). Abl tyrosine kinase specifically phosphorylated its peptide substrate without any cross-reactivity. Abl tyrosine kinase is a member of the PTK kinase family.

Combining protein microarray technology with a fluorescent sensor of phosphorylation status provides a robust platform for the screening of protein kinase and phosphatase inhibitors, as well as the identification of physiologically relevant substrates for protein kinases.

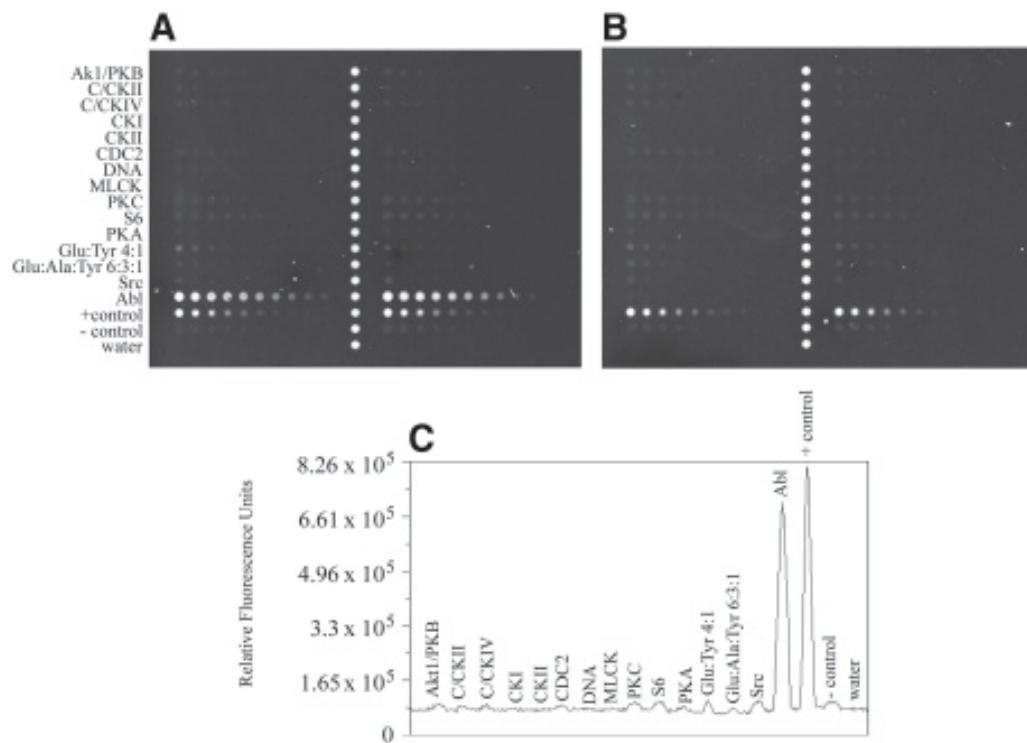


Fig. 1. Peptide microarray analysis of Abl tyrosine kinase selectivity. Two-pad kinase substrate chips designed to profile the activity of AGC, CAMK, CMGC, CKI, and PTK families of protein kinases were incubated with purified Abl tyrosine kinase (A). The control reaction without enzyme was performed simultaneously (B). Kinase phosphorylation was characterized using Pro-Q™ Diamond phosphoprotein/phospho-peptide microarray stain. Both results were quantitated using ImageGauge image analysis software (Fuji Photo Film, Tokyo, Japan) in profile mode. Profile mode quantitates the peak intensity of each spot. Profiles of intensity values of the first column of spots, of the reaction with kinase, are compared and illustrated (C). Abl tyrosine kinase specifically phosphorylated its peptide substrate without any cross-reactivity. Abl tyrosine kinase is a member of the PTK kinase family. Specific kinase peptide and polymer substrates are identified on the y-axis (A & B). Peak signals are labeled on the x-axis according to the kinase peptide substrate (spot) they were quantitated from (C). Abbreviations for peptide substrates are as follows: Akt1/PKB, Akt1/Protein Kinase B; C/CKII, Ca^{2+} /calmodulin-dependent protein kinase II; C/CKIV, Ca^{2+} /calmodulin-dependent protein kinase II- γ or Ca^{2+} /calmodulin-dependent protein kinase IV; CKI, casein kinase I; CKII, casein kinase II; Cdc2, Cdc2 protein kinase; DNA, DNA-dependent protein kinase; MLCK, myosin light-chain kinase; PKC, protein kinase C; S6, S6 kinase/Rsk-2; PKA, cAMP-dependent protein kinase (Protein Kinase A); Glu:Tyr 4:1, glutamic acid (4 residues): tyrosine (1 residue) polymer; Glu:Ala:Tyr 6:3:1, glutamic acid (6 residues): alanine (3 residues): tyrosine (1 residue) polymer; Src, Src tyrosine kinase; Abl, Abl tyrosine kinase.

2. Materials

1. 100 mM HEPES buffer with 1% BSA, pH 7.5.
2. Forceps.
3. Gloves.

4. 10 mM ATP and kinase reaction buffer.
5. Purified enzyme or cell lysate.
6. 10% Sodium dodecyl sulfate (SDS).
7. Distilled/deionized water.
8. Clay Adams Brand Nutator (Becton and Dickinson, San Jose, CA).
9. Labquake rotisserie (Barnstead/Thermolyne, Dubuque, IA).
10. Microarray high-speed centrifuge (Telechem International Inc., Sunnyvale, CA).
11. Lint-free tissue.
12. Paragon™ Protein Kinase Characterization Kit with Pro-Q Diamond microarray stain and kinase-substrate arrays (Molecular Probes, Inc., Eugene, OR).
13. Paragon Phosphoprotein/Phosphopeptide Microarray Stain Kit with Pro-Q Diamond microarray stain (Molecular Probes, Inc., Eugene, OR).

3. Method

The protocols described below are written for use both with preprinted content arrays or for investigators who have used their own content to print arrays onto the HydroGel®-coated slides provided in the Paragon Phosphoprotein/Phosphopeptide Microarray Stain Kit (*see Note 1*).

3.1. Using HydroGel-Coated Slides

HydroGel-coated slides are dried prior to printing. In its dried state, the HydroGel substrate supports both non-contact and contact printing. Contact printers may require an initial adjustment of the z-position depth or substrate thickness settings (for travel of the pin head).

3.1.1. Printing Buffers and Conditions

The HydroGel substrate is compatible with phosphate and borate buffer systems and glycerol concentrations up to 40%, but we do not recommend using phosphate buffer systems because Pro-Q Diamond microarray stain specifically labels phosphate residues. Contaminating phosphate molecules may have adverse affects on staining. The HydroGel substrate is stable from pH 5.0 to 9.0, and pH affects protein immobilization.

3.1.2. Immobilization of Proteins or Peptides on HydroGel-Coated Slides

The immobilization of proteins within the HydroGel substrate is a function of post-printing incubation time. Both the printing buffer used and the proteins or peptides printed will affect the efficiency of immobilization.

1. Incubate arrays in a humidified chamber (approx 65%). Efficient immobilization requires a postprinting incubation of at least 8 h in the humidity chamber (*see Note 2*). For thermally stable proteins, the incubation should be conducted at 30°C. Placing a saturated sodium chloride solution at the bottom of a standard incubator at 30°C yields the proper level of relative humidity. The incubation can be carried out at a lower temperature when using unstable or impure proteins. After printing, the slide can be stored dry at room temperature or 4°C for approx 6 mo.

3.1.3. Considerations for Slide Handling

For best results, all the steps described below should be performed while wearing gloves and using an instrument, such as forceps, to handle slides. The investigator should avoid handling the slides directly with gloves and especially with bare hands.

3.2. Kinase Reaction Standard Protocol

1. After printing and postprinting incubation, block the slides for at least 1 h at room temperature in 100 mM HEPES (pH 7.5), 1% BSA, with agitation (*see Note 3*). This step is usually performed in a large volume (25–30 mL) in a secondary container like component D of either Paragon kit.
2. After blocking, rinse the slides briefly in a steady stream of tap water, and then spin the slides briefly in a microarray high-speed centrifuge (Telechem International, Inc., Sunnyvale, CA) equipped with a slide-holder rotor, to remove excess water. Dry the glass surface, but avoid completely drying the HydroGel pad.
3. To the dry glass surface of the slide affix one of the included microarray staining chambers (Component C) to the area surrounding the pad by first lightly laying the gasket on the slide. Next, press only the edges of the gasket to firmly attach the adhesive to the glass. Use a flat-edged object to press on the adhesive. Avoid pressing on the center of the gasket (*see Note 4*).
4. Through the port holes, pipet your purified enzyme or cell lysate with reaction buffer and ATP (*see Note 5*) into the gasket. Be sure to perform no-enzyme or heat-inactivated negative controls for comparison.
5. Seal the port holes with seal tabs. The slide can then be placed in a secondary hybridization chamber, if one is available, to prevent evaporation. The secondary hybridization chamber or the slide alone can then be incubated at 30°C (or the necessary temperature for the kinase of interest) in an incubator or hybridization oven for 1 h (*see Note 6*). This step should be performed with some sort of mechanical agitation or rotation (*see Note 7*) to allow equal dispersion of the enzyme over the surface of the slide.
6. After incubation with the kinase, remove the gasket using forceps.
7. Move the slide to a secondary container (like the slide-holder tubes provided with the kit) prefilled (*see Note 8*) with 10% SDS, and incubate the slides in the 10% SDS for 5 min with agitation (*see Note 7*).
8. After 5 min, transfer the slides to a second container prefilled with 10% SDS and perform the 5-min SDS wash for a second time with agitation.
9. After the second SDS wash, transfer the slides to a container prefilled with water. Wash the slides for 5 min with agitation in the water. Repeat this water wash six more times for a total of seven washes, transferring between two containers prefilled with water before each wash.
10. Finally, rinse the slides briefly (20 s) in a steady stream of tap water, and then spin them briefly in a microarray high-speed centrifuge equipped with a slide-holder rotor, to remove excess water. Dry the glass surface, but avoid completely drying the HydroGel pad, and proceed to **Subheading 3.3**.

3.3. Microarray Stain Protocol

1. Using sterile forceps, peel off the printed liner starting from the tab end, and attach the adhesive of the microarray staining chamber (Component C) to the flat, dry surface of the array. If using a microscope-sized slide, align the edges of the slide with the edges of the chamber. To ensure a secure seal, press the surface of the cover over the adhesive area with a flat-edged object. Press only the edges of the chamber adhesive, avoiding pressing on the middle of the chamber (*see Note 4*).
2. To fill the microarray staining chamber, pipet 50 to 60 μ L of Pro-Q Diamond phosphoprotein/phosphopeptide microarray stain (Component A) through one port on the seal cover while allowing air to escape through the other port (*see Note 9*).
3. To seal the ports, first wipe excess reagent off the surface of the chamber using a lint-free tissue. Use sterile forceps to remove a seal-tab (Component C) from the liner strip. Place

a seal-tab over each filling port. Press on both seals simultaneously, using finger pressure, to assure a secure seal.

4. Incubate the slides for 45 min in Pro-Q Diamond phosphoprotein/phosphopeptide microarray stain, with agitation. If using the included gasket, we recommend using a nutator-type rotator—fluid dynamics are important (*see Note 7*). This step and all remaining steps should be treated as light sensitive. Slides and stain can be protected from light by covering with aluminum foil.
5. Prior to removing the microarray staining chamber, prefill a slide-holder tube (Component D) with approx 25–30 mL of Pro-Q Diamond microarray destain solution (Component B) and set aside (*see Note 8*).
6. To remove the microarray staining chamber, grasp the tab end along the top edge of the chamber seal and peel it away from the microscope slide. Do not allow the slide to dry. Immediately transfer the slide to the slide-holder tube filled with Pro-Q Diamond microarray destain solution.
7. Incubate the slide for 15 min, with agitation or rotation (*see Note 7*) in approx 25–30 mL of Pro-Q Diamond microarray destain solution. Close to the end of the 15-min incubation time, prefill the second slide-holder tube with approx 25–30 mL of Pro-Q Diamond microarray destain solution (*see Note 8*).
8. After the first 15-min incubation (**step 7**) in destain solution, transfer the slides from the first tube to the prefilled second tube. Use forceps to transfer the slides, and allow the destain solution to sheet off. Do not pour off the destain solution with the slides still in the tube (*see Note 8*).
9. Incubate the slides in Pro-Q Diamond microarray destain solution, in the second tube, for 15 min with agitation or rotation (*see Note 7*).
10. Remove the microarray destain solution from the first tube by pouring it off to a container for flammable waste (do not do this with slides still in the tube) and prefill the first tube again with approx 25–30 mL of microarray destain solution (*see Note 8*).
11. After the second 15-min incubation (**step 9**) in microarray destain solution, transfer the slides back to the first tube refilled with destain solution in **step 10**. Use forceps to transfer the slides, and allow the microarray destain solution to sheet off. Do not pour off the microarray destain solution with the slides still in the chamber.
12. Incubate the slides for a third time in Pro-Q Diamond microarray destain solution for 15 min, with agitation or rotation (*see Note 7*).
13. Prior to the end of the 15-min incubation in **step 12**, remove the microarray destain solution from the second tube by pouring it off to a container for flammable waste (do not do this with slides still in the tube), and wash the tube thoroughly with deionized/distilled water. Ten to 15 rinses with water should be sufficient.
14. After washing the tube with deionized/distilled water, fill the tube with approx 25–30 mL of deionized/distilled water (*see Note 8*). The pH of this final water wash could adversely affect destaining; the pH should be between pH 4.0 and pH 5.0 for optimal destaining.
15. After the third 15-min incubation (**step 12**), transfer the slides to the slide-holder tube filled with deionized/distilled water, and incubate the slides for a final 15 min in the water, with agitation or rotation (*see Note 7*).
16. After the final 15-min water wash, take the slides in the tube to a distilled water tap. Alternately, a squeeze bottle filled with deionized/distilled water can be substituted. Remove the slides from the tube filled with water using forceps, and allow the water to sheet off. Do not pour off the water in the tube, as you will place the slides back in the tube and water for transfer. Place the slides, one at a time, under a steady stream of water, holding them with forceps. Rinse the front and the backs of the slides for 20 s total (alternating between front and back). Transfer the slides back to the tube of water.

17. Transfer the slides from the tube filled with water, using forceps and allowing the water to stream off, to a bench-top centrifuge with a slide holder, or a microarray high-speed centrifuge equipped with a slide holder. Do not pour off the water in the tube with the slides in it. Spin the slides briefly (less than 10 s) to dry them, and then image.

3.4. Imaging

The peak absorbance and emission of Pro-Q Diamond phosphoprotein/phosphopeptide microarray stain are 555 nm and 580 nm, respectively. Any microarray scanner, such as the ScanArray instruments (PerkinElmer Protein Microarray Program, Meriden, CT) or the GenePix instruments (Axon, Union City, CA), equipped with a 532-nm or 543-nm laser can be used to excite the Pro-Q Diamond dye. Scanners utilizing a white light arc lamp source like the arrayWoRx^e scanner (Applied Precision, Issaquah, WA) can also be used with the appropriate bandpass excitation filter. Since the peak emission is at 580 nm, bandpass filters at or near 580 nm or long-pass filters allowing transmittance of light at wavelengths 555 nm and beyond are appropriate for capturing the fluorescence of Pro-Q Diamond dye.

Some microarray scanners that utilize a confocal scanning mode incorporate *z*-position calibration functions when scanning three-dimensional surfaces. It is appropriate and recommended to determine the optimal *z*-position for scanning when using the HydroGel-coated slides. Determining the optimal *z*-position will reduce standard deviations in signal strength.

4. Notes

1. Pro-Q Diamond phosphoprotein/phosphopeptide microarray stain may be used with other microarray surfaces using different attachment chemistries (1), but destaining protocols may need to be optimized to reduce nonspecific background staining.
2. The efficiency of immobilization increases with the length of time of postprinting incubation. We recommend an extended postprinting incubation time of at least 16 h. The investigator should empirically determine the optimal post-printing incubation time for the most efficient immobilization.
3. Other standard buffers, such as MOPS and Tris, may be used in place of HEPES, but phosphate-buffered systems should not be used. We do not recommend using phosphate buffer systems because Pro-Q Diamond microarray stain specifically labels phosphate residues. Contaminating phosphate molecules may have adverse affects on staining.
4. The height of the HydroGel-coated pad is 4 μm when dry and can swell to 20 μm when wet. The height of the microarray staining chamber from the surface of the slide is 150 μm . Although the height of the gasket should prevent the gasket from resting on the acrylamide pad, the center of the gasket sometimes presses on the center of the pad and inhibits the dispersion of pipetted solution. To avoid this frustration, gently place the gasket on the glass surface and press only on the adhesive edge to make a tighter seal to the glass.
5. 100 μM ATP was used for investigations described in **Fig. 1**. Depending on the type of assay being performed, the concentration of ATP may need to be adjusted closer to the actual, determined K_m for the kinase being used.
6. The time of incubation with the lysate or kinase may need to be determined empirically by the investigator if 1 h does not appear to be long enough.
7. Fluid dynamics are important for equal dispersion and even washing and staining. We recommend performing agitation of the small volumes (in microarray staining chambers) on a nutator-type rotator such as the Clay Adams Brand nutator (Becton and Dickinson,

- San Jose, CA). Large-volume steps should be performed on a rotisserie-type agitator such as the Labquake rotisserie (Barnstead/Thermolyne, Dubuque, IA).
8. Prefill containers for all large-volume steps (10% SDS, water, and destaining solution) and add the slides to the prefilled solution. Do not add solutions to the slides in the tube and do not pour off the solutions with the slides still in the tube. Use forceps to transfer slides, and allow solutions to sheet off. These handling recommendations will result in consistent, even staining.
 9. If alternate microarray surfaces are used for generating arrays and the array dimensions do not fit within the included microarray staining gasket (gasket dimensions are 22 mm × 22 mm × 0.150 mm), the volume of Pro-Q Diamond phosphoprotein/phosphopeptide microarray stain should be scaled up or down. We do not recommend using the microarray stain for staining volumes in excess of 1 mL. Other adhesives and plastics may interfere with dye labeling.

References

1. Martin, K., Steinberg, T. H., Cooley, L. A., Gee, K. R., Beechem, J. M., and Patton, W. F. (2003) Quantitative analysis of protein phosphorylation status and protein kinase activity on microarrays using a novel fluorescent phosphorylation sensor dye. *Proteomics* **7**, 244–255.
2. Steinberg, T. H., Agnew, B. J., Gee, K. R., et al. (2003) Global quantitative phosphoprotein analysis using Multiplexed Proteomics technology. *Proteomics* **7**, 1128–1144.
3. Martin, K., Steinberg, T. H., Goodman, T., et al. (2003) Strategies and solid-phase formats for the analysis of protein and peptide phosphorylation employing a novel fluorescent phosphorylation sensor dye. *Comb. Chem. High Throughput Screen* **4**, 331–339.

New Challenges and Strategies for Multiple Sequence Alignment in the Proteomics Era

Julie D. Thompson and Olivier Poch

1. Introduction

The postgenomic era is presenting new challenges for bioinformatics. High-throughput genome sequencing and assembly techniques, together with new information resources, such as structural proteomics, transcriptome data from microarray analyses, or light microscopy images of living cells, have led to a rapid increase in the amount of data available, ranging from complete genome sequences to cellular, structure, phenotype, and other types of biologically relevant information. Thus, genomic and proteomic research has transformed molecular biology from a “data poor” to a “data rich” science. The question now is whether or not we can make sense of all these data using bioinformatics approaches in such areas as genome annotation, comparative genomics, and gene expression analysis. In the face of this ever-increasing volume of complex and constantly evolving data, the integration of experimental data with bioinformatic comparative and predictive analyses will be crucial to the complete description of protein function, not only at the molecular level but also at the higher levels of the pathways, macro-molecular complexes, cells, or organs a protein belongs to.

As a result, advanced database techniques and sensitive data-mining systems are now required to manage and extract the knowledge that is potentially buried in the hundreds of terabytes of data distributed over the various Internet-based resources. The goal of data mining is to detect patterns or relationships in the data that might lead to hidden information, thereby enabling intelligent, knowledge-driven decision-making. New data-mining techniques are being developed in fields such as statistics, artificial intelligence, and rule-based approaches, as well as in clustering and classification methods. For example, decision trees are being used to identify possible targets in high-throughput structural proteomics (1). Association rule discovery is used for finding and describing relationships between different items in a large data set (2,3). Correlation analysis and clustering is used to determine local structural information such as the catalytic triad, metal-binding sites, and the *N*-linked glycosylation site (4). Clustering of gene expression profiles is another area of research attracting much effort (for a review, *see ref.* 5). In addition to these data-mining techniques, new analytical tools are needed to organize and interpret this large volume of data, ranging from data validation and refinement to the extraction of the pertinent information and decision support. Close integration of these software protocols into a fully automatic ensemble is

necessary to enable smooth operation, minimizing the necessity for the operator to have special knowledge of the underlying methods. Some efforts are now being made to develop software protocols and models to facilitate the automatic integration of different biological data and applications. The complexity of the systems encourages using object-oriented models and implementation technologies (for a review, *see ref. 6*). Proper integration will also require standard data formats, models, and ontologies to make information exchange as transparent as possible, and the eXtensible Markup Language (XML) has become the widely accepted Internet data-transfer protocol (7). The process is made even more complex by the need to exchange data among the distributed resources on a real-time basis in order to achieve optimal synchronization. Protocols such as Corba (e.g., 8–10) are therefore required that manage communication among distributed applications.

The volumes of data now being generated and the amount of computing needed to process them have also led to the application of new computational techniques, such as massively parallel supercomputers, or GRID technologies (11), which enable large-scale sharing of resources across geographically distributed groups. For instance, the European HealthGRID project covers a range of biomedical information from the molecular level (genetic and proteomic information) over cells and tissues, to the individual and finally the population level (social healthcare).

As a central concept in molecular biology, the gene and its related products represent an ideal basis for the integration of this mass of biological information in the context of the protein family, and multiple alignment of genomic and protein sequences is one example where a shift of thinking or focus is now leading to the development of new methods. The explosion of the sequence databases and today's advanced database search techniques mean that the alignment of large sets of highly complex proteins has become a standard requirement. Multiple alignment techniques are responding to the challenge, with current developments moving away from a single, all-encompassing algorithm towards co-operative, knowledge-based systems that exploit the new structure and functional data that are available.

2. The Central Role of Multiple Alignments in Proteomics

Sequence comparisons or alignments have been used since their introduction in the early seventies in a wide range of molecular biology applications. Alignments of two sequences, known as pairwise alignments, are mainly used to search the sequence databases in order to identify potential homologs, i.e., sequences that have evolved from a common ancestor. Generally, homologous proteins share the same three-dimensional (3-D) structure and have similar functions, active sites, or binding domains. Two main categories of pairwise alignments exist: global alignments, in which the sequences are aligned along their full length, and local alignments, which identify only the most conserved subsegments of the sequences. Pairwise alignments can be naturally extended to the alignment of more than two sequences. These multiple sequence alignments were originally used in the identification of conserved motifs or key functional residues in a family of proteins (12), and in evolutionary studies to define the phylogenetic relationships between organisms (13). Of course, in the current era of complete genome sequences, it is now possible to perform comparative multiple sequence analysis

at the genome level (e.g., 14–16). Global multiple sequence alignments now play a fundamental role in most of the computational methods used in proteomics, from gene identification and validation to the determination of the protein 3-D structure and the characterization of the molecular and cellular functions of the protein.

2.1. Gene Identification and Validation

One important aspect in biotechnology is gene discovery and target validation for drug discovery. At the time of writing, over 1000 genomes (from bacteria, archaea, and eukaryota, as well as many viruses and organelles) are either complete or being determined, but biological interpretation, i.e., annotation, is not keeping pace with this avalanche of raw sequence data. There is still a real need for accurate and fast tools to analyze these sequences and, especially, to find genes and determine their functions. Unfortunately, finding genes in a genomic sequence is far from being a trivial problem. The most widely used approach consists of employing heterogeneous information from different methods, including the detection of a bias in codon usage between coding and noncoding regions and *ab initio* prediction of functional sites in the DNA sequence, such as splice sites, promoters, or start and stop codons. Most current methods of detection of a signal that may represent the presence of a functional site use position-weight matrices (PWM), consensus sequences, or hidden Markov models (HMMs), and the reliability and accuracy of these methods depends critically on the quality of the underlying multiple alignments (for a review, *see ref. 17*). For prokaryotic genomes, these methods are highly successful, identifying over 95% of the genes (e.g., *ref. 18*), although the exact determination of the start site location remains more problematic because of the absence of relatively strong sequence patterns. The process of predicting genes in higher eukaryotic genomes is complicated by several factors, including complex gene organization, the presence of large numbers of introns and repetitive elements, and the sheer size of the genomic sequence (for a review, *see ref. 19*). It has been shown that comparison of the *ab initio* predicted exons with protein, expressed sequence tag (EST), or cDNA databases can improve the sensitivity and specificity of the overall prediction. In the re-annotation of the *Mycoplasma pneumoniae* genome (20), sequence alignments were used in the prediction of N/C-terminal extensions to the original protein reading frame. Multiple alignments can also be used to improve the detection of short or compositionally biased genes, which often escape annotation. This method was used to detect 24 potential ribosomal genes in a number of complete genomes, that were overlooked during the original gene prediction process (21).

But identifying genes alone will not be sufficient, and there has been growing interest in alternative splicing as a mechanism for expanding the repertoire of gene functions. For instance, a computational method was used to partition ESTs and other transcripts in the Unigene database into subclasses that represent splice variants of a common progenitor gene (22). Multiple alignments of each subclass were constructed, and the information contained in the annotation to each sequence allows one to determine, for example, whether the alternative splicing is tissue-, developmental stage, or disease-state specific. Alternative splicing patterns are also represented by multiple alignments in databases such as the AsMamDB (23), a database of alternatively spliced genes of mammals.

2.2. Functional Annotation

In most genome annotation projects, the standard strategy to determine the function of a novel protein is to search the sequence databases for homologs and to propagate the structural/functional annotation from the known to the unknown protein. Recent developments in database search methods have exploited multiple sequence alignments to detect more and more distant homologs (e.g., refs. 24–27). However, most automatic genome projects use information from only the top best hits in the database search, as sequence hits with higher expect values are considered unreliable. This has led to a certain number of errors in genome annotations. Two types of error have already been identified: those of under- and over-prediction. Underprediction implies that functional information is not transferred because the chain of propagation is broken—for example, because the top-scoring hits in the database search are all uncharacterized. Genome annotation systems such as Magpie (28), Imagene (29), GeneQuiz (30), and Alfresco (31) therefore use multiple alignments to reliably incorporate information from more distant homologs.

2.3. 3-D Structure Determination

Multiple alignments play an important role in a number of aspects of the characterization of the 3-D structure of a protein. For example, one of the most common ways to determine the domain structure of an unknown protein is to search domain databases, such as InterPro (32). These databases contain representations such as profiles or HMMs of individual protein domains, based on multiple alignments of known sequences. A number of *ab initio* methods have also been developed that exploit the information contained in multiple sequence alignments (e.g., refs. 33,34). Another important application of multiple alignments is homology structure modeling. Sequence similarity between proteins usually indicates a structural resemblance, and accurate sequence alignments provide a practical approach for structure modeling, when a 3-D structural prototype is available. Multiple sequence alignments are used to significantly increase the accuracy of both 2-D (35) and 3-D (36) structure prediction methods by taking into account the overall consistency of putative features. Similarly, multiple alignments are also used to improve the reliability of other predictions, such as transmembrane helices (for a review, see ref. 37). More detailed structural analyses also exploit the information in multiple alignments. For example, binding surfaces common to protein families were defined on the basis of sequence conservation patterns and knowledge of the shared fold (38). More recently, pathways of energetic connectivity through protein folds were identified on the basis of the degree of statistical coupling between positions in multiple sequence alignments (39).

2.4. Interaction Network

In the postgenomic view of cellular function, each biological entity is seen in the context of a complex network of interactions. New and powerful experimental techniques, such as the yeast two-hybrid system (40) or tandem-affinity purification and mass spectrometry (41), are used to determine protein–protein interactions systematically. In parallel with these developments, a number of computational techniques have been designed for predicting protein interactions. The performance of the Rosetta method, which relies on the observation that some interacting proteins have homologs

in another organism fused into a single protein chain, has recently been improved using multiple sequence alignment information and global measures of hydrophobic core formation (42). A measure of the similarity between phylogenetic trees of protein families has also been used to predict pairs of interacting proteins (43). Another approach involves quantifying the degree of co-variation between residues from pairs of interacting proteins (correlated mutations), known as the “*in silico* two-hybrid” method. For certain proteins that are known to interact, correlated mutations have been demonstrated to be able to select the correct structural arrangement of two proteins based on the accumulation of signals in the proximity of interacting surfaces (44). This relationship between correlated residues and interacting surfaces has been extended to the prediction of interacting protein pairs, based on the differential accumulation of correlated mutations between the interacting partners (interprotein correlated mutations) and within the individual proteins (intraprotein correlated mutations) (45).

3. The Evolution of Multiple Alignment Methods

The comparison or alignment of biological sequences began in the early seventies, with the first dynamic programming algorithm for the global (or full-length) alignment of two sequences, introduced by Needleman and Wunsch (46). The optimal local alignment between a pair of sequences involves a simple modification to the original method, defined by Smith and Waterman (47), in which only the highest-scoring subsegments of the two sequences are aligned. The pairwise dynamic programming algorithm was soon extended to the first formal algorithm for multiple sequence alignment (48). However, the optimal multiple alignment of more than a few sequences (more than 10) remains impractical due to the intensive computer resources required, despite more recent space and time improvements (e.g., ref. 49). In order to multiply align larger sets of sequences, most programs in use today employ some kind of heuristic approach to reduce the problem to a reasonable size. Traditionally, the most popular method has been the progressive alignment procedure (50). A multiple sequence alignment is built up gradually by aligning the two closest sequences first and successively adding in the more distant ones. A number of alignment programs based on this method exist; notably, ClustalW (51) and ClustalX (52) are based on the global Needleman-Wunsch algorithm. In contrast, the Pima program (53) uses the Smith-Waterman algorithm to find a local multiple alignment.

More recently, algorithms other than dynamic programming have been exploited in the search for more accurate multiple alignments in a wider variety of situations. While the above methods have proved relatively successful in providing multiple alignments of sequences that are related over their entire lengths or contain relatively well conserved regions, the multiple alignment problem is becoming more complex. Global alignment of complex, multidomain proteins, often containing large N/C-terminal extensions and/or internal insertions, is becoming a standard requirement. New developments that aim to resolve at least some of these problems include the use of HMMs in programs such as HMMT (54) or SAM (55), genetic algorithms in SAGA (56), segment-to-segment alignments in Dialign (57), and Gibbs sampling (58) or iteration techniques, notably in the prrp program (59). DCA (60) recursively divides the sequences to be aligned into sets of smaller subsequences until the sequence set is small enough to be aligned with the MSA algorithm. Taylor (61) uses a double dynamic

programming algorithm to avoid some of the pitfalls involved in the greedy progressive alignment methods. Many of these new programs, together with some of the more popular traditional methods, were compared in a recent study of multiple alignment programs (62). The comparison, based on the benchmark alignment database BAliBASE (63), showed that while global alignment methods in general performed better for sets of sequences that were of similar length, local algorithms were more successful at identifying the most conserved motifs in sequences containing large extensions and insertions. The same study of alignment programs also showed that the new iterative methods often produced more accurate alignments, albeit at the cost of a large time penalty. Although the time complexity of the algorithms has generally been low enough to provide a fast enough response time for interactive multiple alignment, the arrival of large numbers of genomic sequences requiring batch processing of thousands of multiple alignments now demands much faster throughput times. As a result of this comparison, some progress has recently been made by exploiting a combination of both local and global algorithms to produce a single multiple alignment (64–67).

3.1. Alignment Quality Analysis

In the face of this complexity, the assessment of the quality of a multiple alignment becomes a daunting task. The ideal solution would be to define a score that would describe the optimal or “biologically correct” multiple alignment. Several scoring systems have been proposed, including the sum-of-pairs, minimum entropy, and maximum likelihood scores. These measures, also known as objective functions, are currently used to evaluate and compare multiple alignments from different sources. They are also used in iterative alignment methods to improve the alignment by seeking to maximize the objective function. Some developments in this field have recently been reported, including global objective functions (e.g., refs. 68–71) and measures of local reliability or column conservation (e.g., refs. 72–75).

4. The Proteomics Era: New Challenges and New Approaches

Given the ever-increasing amount of genome sequence information (over half a million protein sequences), the size of data sets that need to be routinely analyzed is increasing. Many protein families have hundreds or even thousands of members. For example, the HIV GP120 glycoprotein has over 36,000 sequences in SwissProt/TrEMBL databases, and the 20 largest protein families in the Pfam database all contain over 6000 sequences. The construction of multiple alignments of such families is prohibitively expensive, and a subset of sequences must be selected for alignment. However, the question of which sequences should be chosen is far from academic and depends on the interests of the biologist. Requirements range from the alignment of the products of a single gene to the simultaneous analysis of alignments of several protein families linked by transcriptomic information, macro-molecular complex, and so on. Alignments of single gene-products are used in research concerning synonymous or nonsynonymous single-nucleotide polymorphisms (SNPs), in order to understand the genetics of phenotype variation or the genetic basis of complex diseases. Multiple alignments of a particular protein family within a single organism have important applications in many areas. For example, a detailed catalog of the 518 protein kinase genes encoded by the human genome was created from in-depth HMM profile analysis and multiple sequence alignments (76), providing significant information for the develop-

ment of a new generation of drugs to treat cancer and other diseases. Multiple alignments of complete protein families covering various organisms are routinely constructed during evolutionary studies, for the identification of protein domain organization, and so on. More recently, alignments of several protein families have been analyzed to predict physically interacting protein pairs and to identify the most likely sequence regions involved in the interactions. For many of these examples, the choice of sequences to be included in the multiple alignment is evident. In other cases, automatic classification or clustering methods are required to define relevant sequence subgroups. To highlight the diversity of the overall protein family, one simple solution is to align only a representative subset containing no pair of proteins with a sequence identity greater than a user-defined threshold (e.g., [ref. 77](#)). Much effort is now being applied to the development of other, more sophisticated automatic sequence clustering algorithms (e.g., [refs. 78–81](#)). Once the sequences have been clustered into subsets, the next task is to select a representative sequence (e.g., [ref. 82](#)) or to define a consensus sequence (e.g., [ref. 83](#)) for each subset, that will be included in the final multiple alignment.

4.1. Quality Control

In order to construct the accurate alignments needed to perform such in-depth analyses, quality control is required at all levels. The first task of today's multiple alignment methods must be to validate the sequences themselves. Many of the sequences in the protein databases are the result of predictions made by computer programs that have not been confirmed experimentally and thus may be unreliable. In a recent comparison of seven programs for the prediction of gene structures from genomic sequences ([84](#)), the programs still had a considerable proportion of incorrect and missed exons, with the best program achieving 76% exon accuracy. Exon validation, start codon validation, and so on, as described above, is essential at this point to ensure that only relevant information is passed for further analyses. Otherwise, errors will be propagated further and might generate misleading patterns and result in false hypotheses. These steps are usually termed “online quality control,” “online transaction processing,” or “online analysis processing.” As an example, [Fig. 1](#) shows the results of a sequence analysis using the program AliSplicer (Bianchetti, manuscript in preparation), which combines a hierarchical analysis of a multiple sequence alignment with Blast searches of genome and EST databases to detect and validate potential splicing errors.

4.2. Alignment Complexity

Although the accuracy of multiple alignment programs has improved significantly since their original introduction, a number of problems remain to be solved. In particular, proteins with nonlinear elements such as repeats, inversions, and circular permutations or low complexity regions, such as transmembrane proteins or coiled coils, cause particular problems. Large, multidomain proteins are becoming more and more prevalent, in particular with the arrival of a number of genome sequences from eukaryotic organisms. Repeated domains and short segments of repeated single amino acids have been found in higher proportions in eukaryotic proteins than in prokaryotic or archaeal ones ([85,86](#)), while transmembrane proteins account for 20–30% of all open reading frames (ORFs) encoded by a genome ([87](#)). It is clear that an accurate global multiple alignment can no longer be constructed from the primary sequence data alone. No single algorithm currently exists that can cope with the highly complex proteins

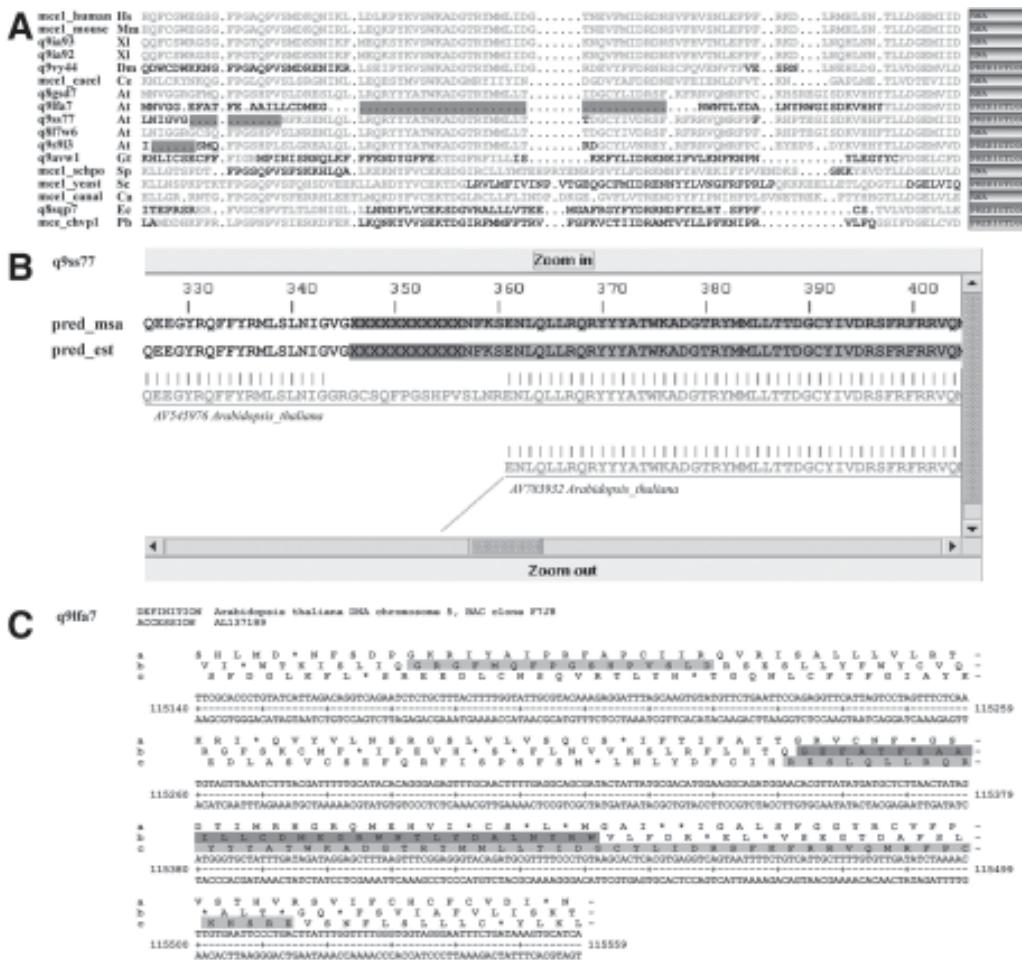


Fig. 1. Validation of sequences based on multiple sequence alignment and expressed sequence tags (ESTs), using the AliSplicer Web server. **(A)** Part of a multiple alignment of 17 mRNA capping enzyme sequences. Validated sequence segments are shown in gray. Potential sequence errors are indicated by shaded boxes. **(B)** Validation of a potential error in sequence q9ss77 by comparison with EST sequences. Pred_msa and pred_est refer to the predictions made by AliSplicer based on the multiple sequence alignment and available EST sequences, respectively. The character X indicates an erroneous deletion in the original sequence. EST sequences are shown in gray below the AliSplicer predictions. **(C)** Validation of the detected error in sequence q9fa7 by examination of the complete genome sequence. The original predicted exon is boxed in dark gray, and the proposed correction is boxed in light gray.

detected by today's database search programs. There is general agreement that complementary algorithms and/or other information is necessary. But the introduction of structural/functional data leads to new problems, particularly when the information is predicted. For example, many proteins are annotated automatically, based on the assumption that similar sequences will have similar functions, and the expected level of error varies from approx 5–8% for general functions to up to 30% for detailed assignments of enzymatic function (88).

4.3. New Strategies: Knowledge-Based Multiple Sequence Alignments

Recent developments in multiple alignment methods have tended towards an integrated system bringing together knowledge-based or text-mining systems and prediction methods, with their inherent unreliability. For example, the program T-Coffee (66) uses information from a precompiled library of different pairwise alignments, including local, global, or structural alignments. MAFFT (89) uses a fast Fourier transform to identify local homologous regions in a set of sequences. A different approach involves exploiting secondary structure information either from the available 3-D structures or from computational predictions to increase alignment sensitivity (e.g., refs. 90–92). The rationale of the development of the DbClustal program (67) was to exploit the information available in the public databases to improve the accuracy of global multiple alignments. Local alignment information from the Ballast program (93) is incorporated into a ClustalW global alignment in the form of a list of anchor points between pairs of sequences. DbClustal has been incorporated into a number of programs, including GScope (Ripp, unpublished data), a semiautomatic genome annotation and analysis software suite, which was used in the annotation of the *Pyrococcus abyssi* genome (94), where the number of alignments with totally misaligned sequences was reduced from 20% to less than 2%. Rascal (95) is an alternative, knowledge-based program designed to improve an existing multiple alignment constructed using any of the above methods. It uses information from clustering algorithms (96,97) and residue conservation analysis (71) in a two-step refinement process to detect and correct local alignment errors.

Other specialized data resources, such as motif/domain databases, mutation data, cellular location, and domain structure classifications, can also be exploited. The information mined from these databases will be used, not only to improve the accuracy of the multiple alignment, but also in the subsequent biological analyses. However, if this mass of data is not classified and presented in a well-structured format, the pertinent information risks being submerged in the flood. A major challenge will be the selection of the most descriptive and useful information and the presentation of this information in a suitable format for the biologist. Global sequence alignments provide an ideal framework for the integration and presentation of the complex information pertaining to a particular protein family (98). The extent of sequence variation at a given position in a set of aligned sequences can be useful information, and has been used to automatically identify sequence motifs (99) and to predict protein function and fold structure (100). Combining this with structural/functional information can lead to valuable biological insights (e.g., refs. 101,102). **Fig. 2** shows an example of a detailed integration of structural/functional information in a global multiple sequence alignment of 28 RNA triphosphatase/mRNA capping enzyme sequences. 5' capping of mRNA capping involves three enzymes: RNA 5'-triphosphatase, guanylyltransferase, and N7G-methyltransferase (103). These three activities are encoded separately in prokaryotes and most low eukaryotes, but in the higher eukaryotes, the first two steps of capping are catalyzed by bifunctional triphosphatase-guanylyltransferase enzymes. Unfortunately, the complete structural/functional information is generally not available for all protein sequences. Although the genome sequencing projects have led to a rapid increase in protein sequence information, experimentally verified information on protein function lags a long way behind. The missing information must be either transferred

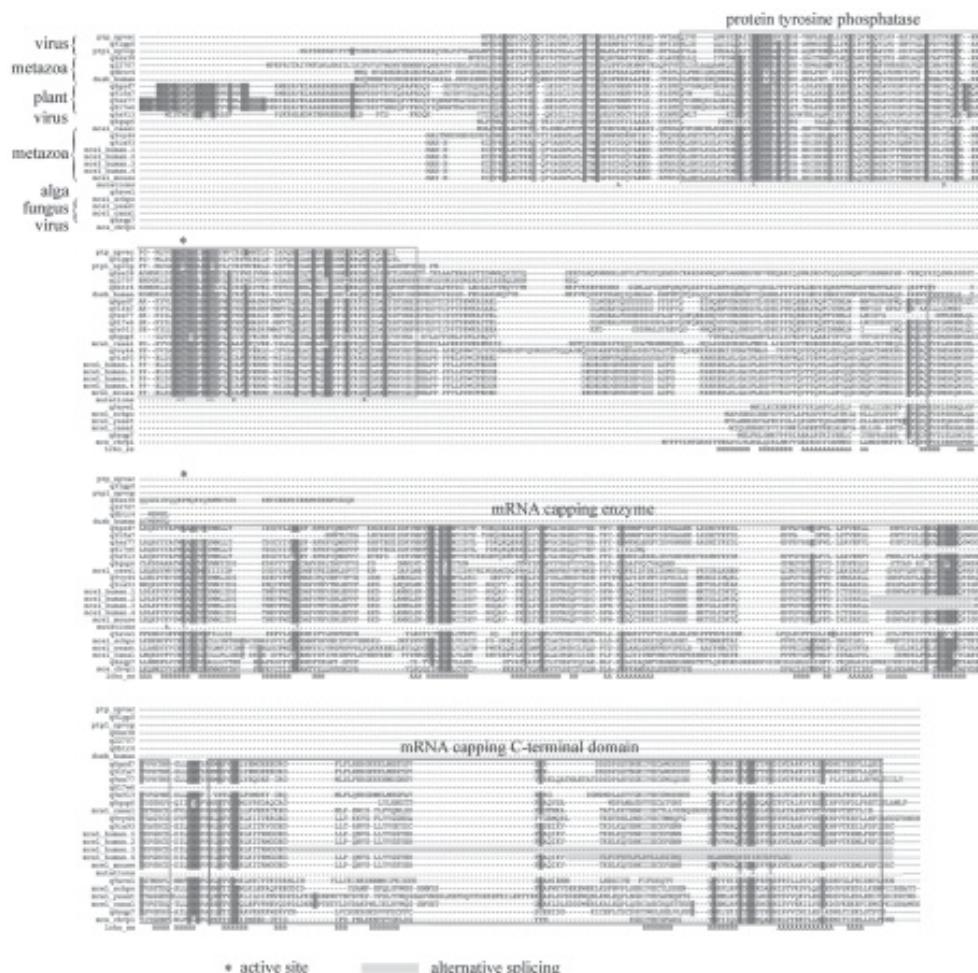


Fig. 2. Part of a global multiple alignment showing the domain organization of 28 protein tyrosine phosphatase/mRNA capping enzyme sequences. Residues conserved in at least 80% of the sequences are shaded. Four human isoforms are included in the alignment, with the alternative splicing indicated by shaded segments. The line entitled “mutations” indicates mutagenesis experiments in *Mus musculus* (SwissProt identity: mce1_mouse). Mutations leading to at least 50% loss of function are highlighted by black circles, the remaining mutations having little or no effect. Secondary structure elements derived from the three-dimensional structure of the chlorella virus protein are shown below the alignment (H, helix, B, beta sheet).

from other, homologous proteins or predicted *ab initio*. In the context of the multiple sequence alignment, homology can be more reliably determined (e.g., refs. 71,104) and predicted information, with its inherent unreliability, can be cross-validated to significantly increase the accuracy of the predictions.

Multiple alignment methods are consequently evolving from being stand-alone tasks to becoming integrated, interactive tools, such as that illustrated in Fig. 3. New state-of-the-art systems will include data- and text-mining components for rapid database querying and knowledge acquisition. The construction of accurate, high-quality align-

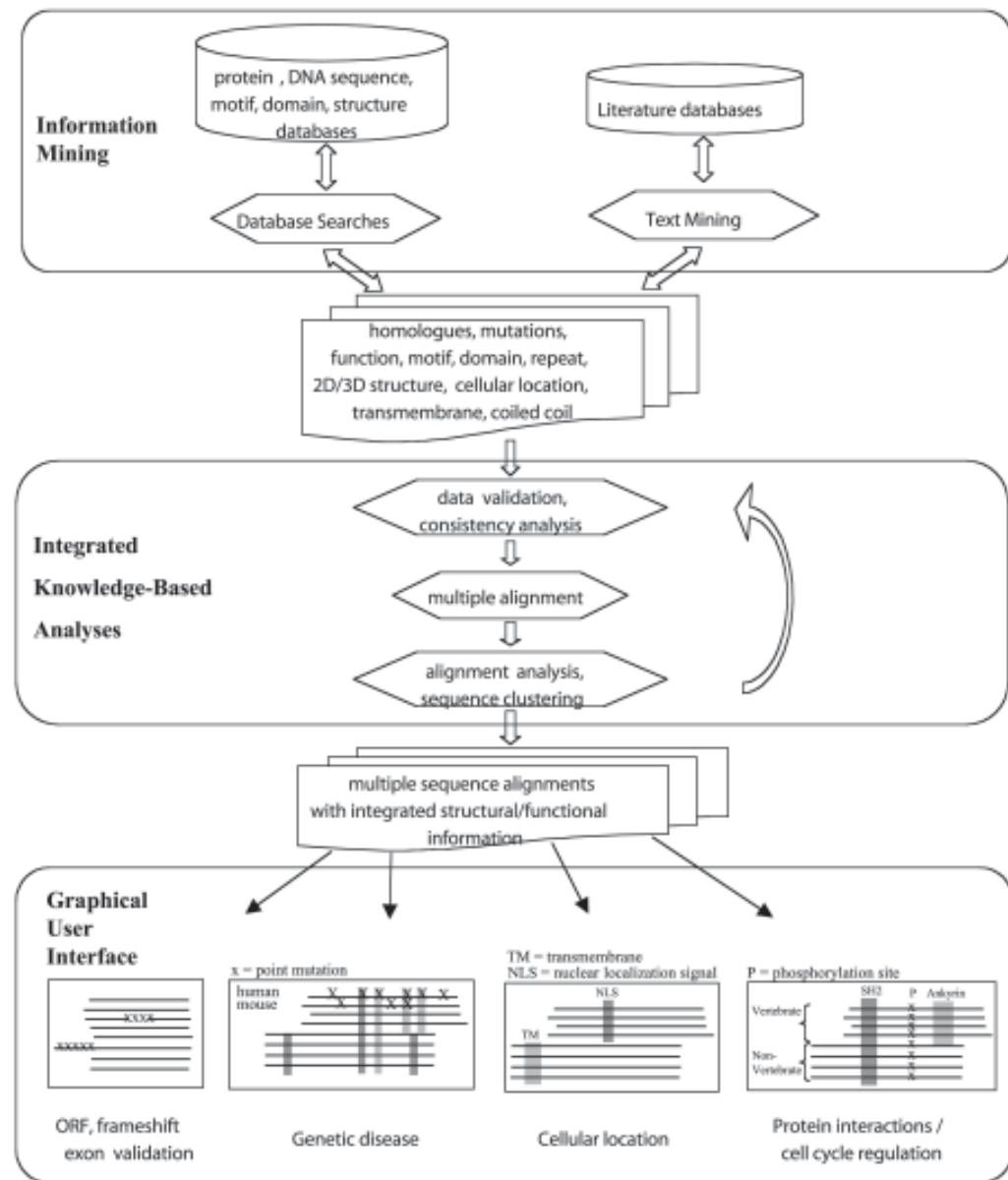


Fig. 3. Schematic architecture of an integrated multiple sequence alignment system. Three main tasks are identified: information mining, multiple alignment construction/validation, and a graphical user interface.

ments will also necessitate an iterative phase of data integration, validation, and refinement. As an example, **Fig. 4** shows an alignment of ribosomal L5/L18 proteins from various organisms, including eukaryotes and prokaryotes. Nuclear localization (NLS) and export (NES) signals have been identified experimentally in the human ribosomal protein L5 (105). Sequence conservation suggests that the N-terminal NLS and the NES are both present in the other eukaryotic sequences in the alignment. However,

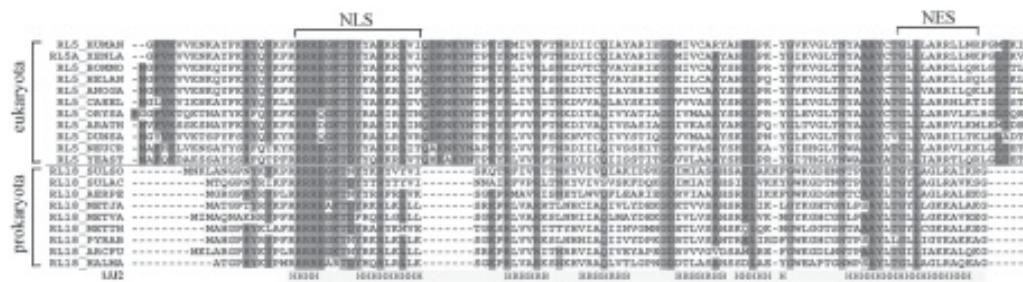


Fig. 4. Part of an alignment of 20 ribosomal L5/L18 proteins. Residues conserved in at least 80% of the sequences are boxed in gray. The positions of the experimentally verified nuclear localization (NLS) and export (NES) signals in RL5_HUMAN are indicated above the alignment. Secondary structure elements derived from the three-dimensional structure of the *Halobacterium marismortui* L18 protein (PDB identity: 1JJ2) are shown below the alignment (H, helix, B, beta sheet).

propagation of the NLS prediction to the other sequences in the alignment (all sharing more than 58% residue identity with the human protein in the N-terminal region) would result in the erroneous hypothesis of the presence of NLSs in prokaryotic organisms.

4.4. Automatic Multiple Alignment Pipelines

This growing complexity of the multiple alignment process clearly requires multiple analysis and investigation steps that need to be integrated into automatic pipelines. For example, the ANTHEPROT web server (106) allows the incorporation of secondary structure predictions within a multiple alignment, and full interactive editing of alignments in graphic windows. The different methods are provided in a single interface; tasks can be submitted to a remote server and can retrieve data from a remote Web server. ENDscript (107) takes a PDB file as input and performs a Blast search for homologs, followed by a multiple alignment with Multalin/ClustalW. The results are then displayed together with secondary structure elements, accessibility, hydropathy, and intermolecular contacts. PipeAlign (108) is a protein family analysis Web server, which takes either a protein sequence or a set of unaligned sequences, or a multiple alignment, as input and performs an automatic analysis, including a Blast search for sequence homologs, identification of locally conserved segments with Ballast, construction and refinement of a MACS, and definition of possible functional subfamilies.

4.5. User Access and Visualization

Another problem is the ease of use of the new complex systems software currently being developed. Some genomic software is difficult to operate for biologists with limited computer training. Programs that have nongraphical, command-line-driven interfaces are not intuitive, because they require the use of exact command syntax, including all possible options. In contrast, graphical interfaces, such as Jalview (109) or Modview (110), allow visualization of multiple protein sequences and structures, with highlighting of features such as conserved residues, active sites, fragments, or domains. In addition, some genomic programs are designed to run on specific plat-

forms with specific operating systems (e.g., Unix). Users who are not familiar with an operating system may have difficulty in installing and using these programs. One solution is to use a Web interface, which allows the user to access data files as well as analysis programs in an integrated fashion, regardless of client platforms. One example is W2H (111), a Web-based interface to the popular GCG Sequence Analysis Software Package (Wisconsin Package). Some systems running a number of different automatic bioinformatics analyses, e.g., Pfaat (112), also allow expert knowledge to be manually incorporated in the results.

5. Future Perspectives

The availability of fully sequenced genomes and the enormous amount of data from structural and functional proteomics projects opens the way to new methods of analyzing protein function. In order to fully understand the functions and molecular interactions of a particular protein, such diverse information as cellular location, degradation and modification, 2-D/3-D structures, mutations and their associated illnesses, the evolutionary context, and literature references must be assembled, classified, and made available to the biologist. Integrated multiple alignments of complete sequences provide an ideal workbench for the integration and presentation of the most vital and relevant aspects of all these sequence data. But cutting-edge computational infrastructure comprising both data storage and analysis algorithms is beneficial only as long as it is realized as a network of transparent and smoothly interacting components. Automation of the full multiple alignment construction process is essential, from searching the databases for homologous sequences to integration and presentation of the pertinent biological information.

Such an alignment network will facilitate information retrieval and knowledge discovery, with functionalities for interactive queries, combinations of sequence and text searches, and sorting and visual exploration of search results. Although few, if any, of these methods have reached the status of validated proteomic tools, the rapid pace at which they are developing suggests that the rich and varied sources of information contained in the proteome will become increasingly accessible. The tools will allow the validation, visualization, integration, and interpretation, in a biological context, of the vast amounts of diverse data generated by the application of proteomic and genomic discovery science tools.

References

1. Bertone, P., Kluger, Y., Lan, N., et al. (2001) SPINE: an integrated tracking database and data mining approach for identifying feasible targets in high-throughput structural proteomics. *Nucleic Acids Res.* **29**, 2884–2898.
2. Oyama, T., Kitano, K., Satou, K., and Ito, T. (2002) Extraction of knowledge on protein-protein interaction by association rule discovery. *Bioinformatics* **18**, 705–714.
3. Creighton, C. and Hanash, S. (2003) Mining gene expression databases for association rules. *Bioinformatics* **19**, 79–86.
4. Oldfield, T. J. (2002) Data mining the Protein Data Bank: residue interactions. *Proteins* **49**, 510–529.
5. Shannon, W., Culverhouse, R., and Duncan, J. (2003) Analyzing microarray data using cluster analysis. *Pharmacogenomics* **4**, 41–52.

6. Wiechert, W., Joksch, B., Wittig, R., Hartbrich, A., Honer, T., and Mollney, M. (1995) Object-oriented programming for the biosciences. *Comput. Appl. Biosci.* **11**, 517–534.
7. Achard, F., Vaysseix, G., and Barillot, E. (2001) XML, bioinformatics and data integration. *Bioinformatics* **17**, 115–125.
8. Achard, F. and Barillot, E. (1997) Ubiquitous distributed objects with CORBA. *Pac. Symp. Biocomput.* 39–50.
9. Campagne, F. (2000) Clustalnet: the joining of Clustal and CORBA. *Bioinformatics* **16**, 606–612.
10. Wang, L., Rodriguez-Tome, P., Redaschi, N., McNeil, P., Robinson, A., and Lijnzaad, P. (2000) Accessing and distributing EMBL data using CORBA (common object request broker architecture). *Genome Biol.* **1**, RESEARCH0010.
11. Foster, I. (2003) The grid: computing without bounds. *Sci. Am.* **288**, 78–85.
12. del Sol Mesa, A., Pazos, F., and Valencia, A. (2003) Automatic methods for predicting functionally important residues. *J. Mol. Biol.* **326**, 1289–1302.
13. Phillips, A., Janies, D., and Wheeler, W. (2000) Multiple sequence alignment in phylogenetic analysis. *Mol. Phylogenet. Evol.* **16**, 317–330.
14. Morgenstern, B. (2000) A space-efficient algorithm for aligning large genomic sequences. *Bioinformatics* **16**, 948–949.
15. Hohl, M., Kurtz, S., and Ohlebusch, E. (2002) Efficient multiple genome alignment. *Bioinformatics* **18**, S312–S320.
16. Brudno, M., Do, C. B., Cooper, G. M., et al. (2003) LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.* **13**, 721–731.
17. Mathe, C., Sagot, M. F., Schiex, T., and Rouze, P. (2002) Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res.* **30**, 4103–4117.
18. Aggarwal, G. and Ramaswamy, R. (2002) *Ab initio* gene identification: prokaryote genome annotation with GeneScan and GLIMMER. *J. Biosci.* **27** (Suppl 1), 7–14.
19. Zhang, M. Q. (2002) Computational prediction of eukaryotic protein-coding genes. *Nat. Rev. Genet.* **3**, 698–709.
20. Dandekar, T., Huynen, M., Regula, J. T., et al. (2000) Re-annotating the mycoplasma pneumoniae genome sequence: adding value, function and reading frames. *Nucleic Acids Res.* **28**, 3278–3288.
21. Lecompte, O., Ripp, R., Thierry, J. C., Moras, D., and Poch, O. (2002) Comparative analysis of ribosomal proteins in complete genomes: an example of reductive evolution at the domain scale. *Nucleic Acids Res.* **30**, 5382–5390.
22. Burke, J., Wang, H., Hide, W., and Davison, D. B. (1998) Alternative gene form discovery and candidate gene selection from gene indexing projects. *Genome Res.* **8**, 276–290.
23. Ji, H., Zhou, Q., Wen, F., Xia, H., Lu, X., and Li, Y. (2001) AsMamDB: an alternative splice database of mammals. *Nucleic Acids Res.* **29**, 260–263.
24. Altschul, S. F., Madden, T. L., Schaffer, A. A., et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402.
25. Eddy, S. R. (1998) Profile hidden Markov models. *Bioinformatics* **14**, 755–763.
26. Karplus, K., Barrett, C., and Hughey, R. (1999) Hidden Markov models for detecting remote protein homologies. *Bioinformatics* **14**, 846–856.
27. Yona, G. and Levitt, M. (2002) Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. *J. Mol. Biol.* **315**, 1257–1275.
28. Gaasterland, T. and Sensen, C. W. (1996) Fully automated genome analysis that reflects user needs and preferences. A detailed introduction to the MAGPIE system architecture. *Biochimie* **78**, 302–310.

29. Medigue, C., Rechenmann, F., Danchin, A., and Viari, A. (1999) Imagene: an integrated computer environment for sequence annotation and analysis. *Bioinformatics* **15**, 2–15.
30. Hoersch, S., Leroy, C., Brown, N. P., Andrade, M. A., and Sander, C. (2000) The GeneQuiz web server: protein functional analysis through the Web. *Trends Biochem. Sci.* **25**, 33–35.
31. Jareborg, N. and Durbin, R. (2000) Alfresco—A workbench for comparative genomic sequence analysis. *Genome Res.* **10**, 1148–1157.
32. Mulder, N. J., Apweiler, R., Attwood, T. K., et al. (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.* **31**, 315–318.
33. Bejerano, G., Seldin, Y., Margalit, H., and Tishby, N. (2001) Markovian domain finger-printing: statistical segmentation of protein sequences. *Bioinformatics* **17**, 927–934.
34. George, R. A. and Heringa, J. (2002) SnapDRAGON: a method to delineate protein structural domains from sequence data. *J. Mol. Biol.* **316**, 839–851.
35. Heringa, J. (2000) Computational methods for protein secondary structure prediction using multiple sequence alignments. *Curr. Protein Pept. Sci.* **1**, 273–301.
36. Al-Lazikani, B., Jung, J., Xiang, Z., and Honig, B. (2001) Protein structure prediction. *Curr. Opin. Chem. Biol.* **5**, 51–56.
37. Chen, C. P., Kernytsky, A., and Rost, B. (2002) Transmembrane helix predictions revisited. *Protein Sci.* **11**, 2774–2791.
38. Lichtarge, O., Bourne, H. R., and Cohen, F. E. (1996) An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257**, 342–358.
39. Lockless, S. W. and Ranganathan, R. (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* **286**, 295–299.
40. Ito, T., Ota, K., Kubota, H., et al. (2002) Roles for the two-hybrid system in exploration of the yeast protein interactome. *Mol. Cell Proteomics* **8**, 561–566.
41. Gavin, A. C., Bosche, M., Krause, R., et al. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147.
42. Bonneau, R., Strauss, C. E., and Baker, D. (2001) Improving the performance of Rosetta using multiple sequence alignment information and global measures of hydrophobic core formation. *Proteins* **43**, 1–11.
43. Pazos, F. and Valencia, A. (2001) Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng.* **14**, 609–614.
44. Pazos, F., Helmer-Citterich, M., Ausiello, G., and Valencia A. (1997) Correlated mutations contain information about protein-protein interaction. *J. Mol. Biol.* **271**, 511–523.
45. Pazos, F. and Valencia, A. (2002) In silico two-hybrid system for the selection of physically interacting protein pairs. *Proteins* **47**, 219–227.
46. Needleman, S. B. and Wunsch, C. D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453.
47. Smith, T. F. and Waterman, M. S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197.
48. Sankoff, D. (1975) Minimal mutation trees of sequences. *SIAM J. Appl. Math.* **28**, 35–42.
49. Gupta, S. K., Kececioglu, J. D., and Schaffer, A. A. (1995) Improving the time efficiency of the shortest-paths approach to sum-of-pairs multiple sequence alignment. *J. Comput. Biol.* **2**, 459–472.
50. Feng, D. F. and Doolittle, R. F. (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.* **25**, 351–360.
51. Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and matrix choice. *Nucleic Acids Res.* **22**, 4673–4680

52. Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F., and Higgins, D. G. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **22**, 4673–4680.
53. Smith, R. F. and Smith, T. F. (1992) Pattern-induced multi-sequence alignment (PIMA) algorithm employing secondary structure-dependent gap penalties for use in comparative protein modelling. *Protein Eng.* **5**, 35–41.
54. Eddy, S. R. (1995) Multiple alignment using hidden Markov models. *ISMB* **3**, 114–120.
55. Karplus, K., Barrett, C., and Hughey, R. (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics* **10**, 846–856.
56. Notredame, C. and Higgins, D. G. (1996) SAGA: sequence alignment by genetic algorithm. *Nucleic Acids Res.* **24**, 1515–1524.
57. Morgenstein, B., Dress, A., and Werner, T. (1996) Multiple DNA and protein sequence alignment based on segment-to-segment comparison. *Proc. Natl. Acad. Sci. USA* **93**, 12098–12103.
58. Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F., and Wootton, J. C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **262**, 208–214.
59. Gotoh, O. (1996) Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments. *J. Mol. Biol.* **264**, 823–838.
60. Stoye, J. (1998) Multiple sequence alignment with the Divide-and-Conquer method. *Gene* **211**, GC45–56.
61. Taylor, W. R., Saelensminde, G., and Eidhammer, I. (2000) Multiple protein sequence alignment using double-dynamic programming. *Comput. Chem.* **1**, 3–12.
62. Thompson, J. D., Plewniak, F., and Poch, O. (1999) A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res.* **27**, 2683–2690.
63. Thompson, J. D., Plewniak, F., and Poch, O. (1999) BaliBASE: A benchmark alignment database for the evaluation of multiple sequence alignment programs. *Bioinformatics* **1**, 87–88.
64. Brochieri, L. and Karlin, S. (1998) A symmetric-iterated multiple alignment of protein sequences. *J Mol Biol* **276**, 249–264.
65. Bucka-Lassen, K., Caprani, O., and Hein, J. (1999) Combining many multiple alignments in one improved alignment. *Bioinformatics* **15**, 122–130.
66. Notredame, C., Higgins, D. G., and Heringa, J. (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302**, 205–217.
67. Thompson, J. D., Plewniak, F., Thierry, J. C., and Poch, O. (2000) DbClustal: rapid and reliable global multiple alignments of protein sequences detected by database searches. *Nucleic Acids Res.* **28**, 2919–2926.
68. Notredame, C., Holm, L., and Higgins, D. G. (1998) COFFEE: an objective function for multiple sequence alignments. *Bioinformatics* **14**, 407–422.
69. Hertz, G. Z. and Stormo, G. D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* **15**, 563–577.
70. Gonnet, G. H., Korostensky, C., and Benner, S. (2000) Evaluation measures of multiple sequence alignments. *J. Comput. Biol.* **7**, 261–276.
71. Thompson, J. D., Plewniak, F., Ripp, R., Thierry, J. C., and Poch, O. (2001) Towards a reliable objective function for multiple sequence alignments. *J. Mol. Biol.* **314**, 937–951.
72. Pei, J. and Grishin, N. V. (2001) AL2CO: calculation of positional conservation in a protein sequence alignment. *Bioinformatics* **17**, 700–712.
73. Cline, M., Hughey, R., and Karplus, K. (2002) Predicting reliable regions in protein sequence alignments. *Bioinformatics* **18**, 306–314.

74. Fares, M. A., Elena, S. F., Ortiz, J., Moya, A., and Barrio, E. (2002) A sliding window-based method to detect selective constraints in protein-coding genes and its application to RNA viruses. *J. Mol. Evol.* **55**, 509–521.
75. Schlosshauer, M. and Ohlsson, M. (2002) A novel approach to local reliability of sequence alignments. *Bioinformatics* **18**, 847–854.
76. Manning, G., Whyte, D. B., Martinez, R., Hunter, T., and Sudarsanam S. (2002) The protein kinase complement of the human genome. *Science* **298**, 1912–1934.
77. Wang, L. and Xu, Y. (2003) SEGID: identifying interesting segments in (multiple) sequence alignments. *Bioinformatics* **19**, 297–298.
78. Holm, L. and Sander, C. (1998) Touring protein fold space with Dali/FSSP. *Nucleic Acids Res.* **26**, 316–319.
79. Trelles, O., Andrade, M. A., Valencia, A., Zapata, E. L., and Carazo, J. M. (1998) Computational space reduction and parallelization of a new clustering approach for large groups of sequences. *Bioinformatics* **14**, 439–451.
80. Kawaji, H., Yamaguchi, Y., Matsuda, H., and Hashimoto, A. (2001) A graph-based clustering method for a large set of sequences using a graph partitioning algorithm. *Genome Inform. Ser. Workshop Genome Inform.* **12**, 93–102.
81. Enright, A. J., Van Dongen, S., and Ouzounis, C. A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575–1584.
82. May, A. C. (2001) Optimal classification of protein sequences and selection of representative sets from multiple alignments: application to homologous families and lessons for structural genomics. *Protein Eng.* **14**, 209–217.
83. Keith, J. M., Adams, P., Bryant, D., et al. (2002) A simulated annealing algorithm for finding consensus sequences. *Bioinformatics* **18**, 1494–1499.
84. Rogic, S., Mackworth, A. K., and Ouellette F. B. (2001) Evaluation of gene-finding programs on mammalian sequences. *Genome Res.* **11**, 817–832.
85. Marcotte, E. M., Pellegrini, M., Yeates, T. O., and Eisenberg, D. (1999) A census of protein repeats. *J. Mol. Biol.* **293**, 151–160.
86. Huntley, M. and Golding, G. B. (2000) Evolution of simple sequence in proteins. *J. Mol. Evol.* **51**, 131–140.
87. Wallin, E. and von Heijne, G. (1998) Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci.* **7**, 1029–1038.
88. Devos, D. and Valencia, A. (2001) Intrinsic errors in genome annotation. *Trends Genet.* **17**, 429–431.
89. Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066.
90. Heringa, J. (1999) Two strategies for sequence comparison: profile-preprocessed and secondary structure-induced multiple alignment. *Comput. Chem.* **23**, 341–364.
91. Jennings, A. J., Edge, C. M., and Sternberg, M. J. (2001) An approach to improving multiple alignments of protein sequences using predicted secondary structure. *Protein Eng.* **14**, 227–231.
92. Shi, J., Blundell, T. L., and Mizuguchi, K. (2001) FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J. Mol. Biol.* **310**, 243–257.
93. Plewniak, F., Thompson, J. D., and Poch, O. (2000) Ballast: blast post-processing based on locally conserved segments. *Bioinformatics* **16**, 750–759.
94. Cohen, G. N., Barbe, V., Flament, D., et al. (2003) An integrated analysis of the genome of the hyperthermophilic archaeon *Pyrococcus abyssi*. *Mol. Microbiol.* **47**, 1495–1512.

95. Thompson, J. D., Thierry, J. C., and Poch, O. (2003) RASCAL: rapid scanning and correction of multiple sequence alignments. *Bioinformatics* **19**(9), 1155–1161.
96. Wicker, N., Perrin, G. R., Thierry, J. C., and Poch, O. (2001) Secator: a program for inferring protein subfamilies from phylogenetic trees. *Mol. Biol. Evol.* **18**, 1435–1441.
97. Wicker, N., Dembele, D., Raffelsberger, W., and Poch, O. (2002) Density of points clustering, application to transcriptomic data analysis. *Nucleic Acids Res.* **30**, 3992–4000.
98. Lecompte, O., Thompson, J. D., Plewniak, F., Thierry, J. C., and Poch, O. (2001) Multiple alignment of complete sequences (MACS) in the post-genomic era. *Gene* **270**, 17–30.
99. May, A. C. (2002) Definition of the tempo of sequence diversity across an alignment and automatic identification of sequence motifs: application to protein homologous families and superfamilies. *Protein Sci.* **11**, 2825–2835.
100. Kunin, V., Chan, B., Sitbon, E., Lithwick, G., and Pietrokovski, S. (2001) Consistency analysis of similarity between multiple alignments: prediction of protein function and fold structure from analysis of local sequence motifs. *J. Mol. Biol.* **307**, 939–949.
101. Mirny, L. A. and Shakhnovich, E. I. (1999) Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *J. Mol. Biol.* **291**, 177–196.
102. Ota, M., Kinoshita, K., and Nishikawa, K. (2003) Prediction of catalytic residues in enzymes based on known tertiary structure, stability profile, and sequence conservation. *J. Mol. Biol.* **327**, 1053–1064.
103. Shatkin, A. J. and Manley, J. L. (2000) The ends of the affair: capping and polyadenylation. *Nat. Struct. Biol.* **7**, 838–842.
104. Errami, M., Geourjon, C., and Deleage, G. (2003) Detection of unrelated proteins in sequences multiple alignments by using predicted secondary structures. *Bioinformatics* **19**, 506–512.
105. Rosorius, O., Fries, B., Stauber, R. H., Hirschmann, N., Bevec, D., and Hauber, J. (2000) Human ribosomal protein L5 contains defined nuclear localization and export signals. *J. Biol. Chem.* **275**, 12,061–12,068.
106. Deleage, G., Combet, C., Blanchet, C., and Geourjon, C. (2001) ANTHEPROT: an integrated protein sequence analysis software with client/server capabilities. *Comput. Biol. Med.* **31**, 259–267.
107. Gouet, P. and Courcelle, E. (2002) ENDscript: a workflow to display sequence and structure information. *Bioinformatics* **18**, 767–768.
108. Plewniak, F., Bianchetti, L., Brelivet, Y., et al. (2003) PipeAlign : a new toolkit for protein family analysis. *Nucleic Acids Res.* **31**(13), 3829–3832.
109. Clamp, M. E., Cuff, J. A., and Barton, G. J. (1999) Jalview—a java multiple sequence alignment viewer and editor. <<http://www.compbio.dundee.ac.uk/>>.
110. Ilyin, V. A., Pieper, U., Stuart, A. C., Marti-Renom, M. A., McMahan, L., and Sali, A. (2003) ModView, visualization of multiple protein sequences and structures. *Bioinformatics* **19**, 165–166.
111. Senger, M., Flores, T., Glatting, K., Ernst, P., Hotz-Wagenblatt, A., and Suhai, S. (1998) W2H: WWW interface to the GCG sequence analysis package. *Bioinformatics* **14**, 452–457.
112. Johnson, J. M., Mason, K., Moallemi, C., Xi, H., Somaroo, S., and Huang, E. S. (2003) Protein family annotation in a multiple alignment viewer. *Bioinformatics* **19**, 544–545.

The Clustal Series of Programs for Multiple Sequence Alignment

Julie D. Thompson

1. Introduction

Sequence comparison or alignment is one of the fundamental tools in molecular biology. Alignments are used to search the sequence databases for homologous proteins, in the identification of conserved regions in a set of related sequences, for two-dimensional (2-D)/3-D structure predictions, phylogenetic studies, and so on. The comparison of biological sequences began in the early seventies, with the first dynamic programming algorithm for the global (or full-length) alignment of two sequences (1). Dynamic programming is a rigorous mathematical technique that is guaranteed to find the maximal scoring alignment for any two sequences. The optimal local alignment between a pair of sequences, in which only the highest-scoring subsegments of the two sequences are aligned, involves a simple modification (2) to the Needleman–Wunsch method.

The first formal algorithm for the alignment of more than two sequences (3) was developed as a direct extension of the pairwise dynamic programming algorithm. However, the optimal multiple alignment of more than a few sequences (more than 10) remains impractical because of the intensive computer resources required, despite some recent time and space improvements. A different approach to the alignment problem therefore involves the use of heuristics or approximate methods, which do not guarantee an optimal alignment solution but are less time-consuming. In order to multiply align larger sets of sequences, most programs in use today employ some kind of heuristic approach to reduce the problem to a reasonable size. Traditionally, the most popular method has been the progressive alignment procedure (4). The alignment consists of three steps: first, all the sequences are compared to each other in a pairwise fashion; next, a guide tree is created from the pairwise sequence distances and written to a file; finally, the multiple alignment is built up by aligning larger and larger groups of sequences, following the order given by the guide tree. The rationale of the progressive technique is to try to build up the alignment, starting with the most closely related sequences. By the time the more distantly related sequences are included in the alignment, some information is already available about the conservation or lack of it at each position in the subalignments.

One of the most widely used multiple alignment programs in use today is the Clustal family of programs, based on a memory-efficient dynamic programming algorithm (5) and the progressive alignment technique. The first Clustal program was written in Fortran and was designed to be run on personal computers (PCs) (6,7). ClustalV (8) was the first version to be written in C, and featured the ability to produce phylogenetic trees from alignments, using the neighbor-joining method (9) and bootstrap confidence measures (10). ClustalW (11) was derived from ClustalV by the addition of numerous new features for improving the sensitivity of protein alignments and for extending the functionality of the interface. ClustalW supports a full command-line interface, which allows it to be used automatically as part of larger analyses (e.g., it can be run from scripts). ClustalX (12) was based directly on ClustalW, but the basic text menu user interface was replaced by a more user-friendly graphical interface, and extensive graphical features for annotating alignments were added. Both ClustalX and ClustalW programs are now actively maintained; the most recent version number (February 2003) is 1.83. In their simplest applications, the programs are used to take a set of homologous sequences (all DNA/RNA or all protein) and to produce a single multiple alignment. This multiple alignment procedure covers the vast majority of Clustal usage and will be sufficient for most cases. However, in difficult cases it may be advantageous for the user to intervene in the automatic process, either by changing some of the critical alignment parameters or by manually guiding the order of alignment using the profile alignment options. Clustal also provides basic tools for phylogenetic analyses based on the multiple alignment. The phylogenetic trees are written to a text file, which can then be viewed using a graphical display program such as TreeView (13) or the NJPlot program (14), which is supplied with Clustal.

2. Materials

Users may run Clustal remotely from several sites using the Worldwide Web (WWW), or the programs may be downloaded to be run locally on a UNIX workstation (SUN, Alpha, Silicon Graphics, and so on), a PC with either MSDOS or Windows, a Power Macintosh, or any other computer supporting a C compiler. The Clustal series of programs are available by anonymous ftp to [ftp-igbmc.u-strasbg.fr](ftp://ftp-igbmc.u-strasbg.fr) or [ftp.ebi.ac.uk](ftp://ftp.ebi.ac.uk). For ClustalW, executable programs are supplied for Power Macintosh computers and for PCs running either the Windows or DOS operating system. ClustalX uses the Vibrant multiplatform user interface development library, developed by the National Center for Biotechnology Information (NCBI) (Bldg 38A, NIH 8600 Rockville Pike, Bethesda, MD 20894) as part of their NCBI Software Development Toolkit. As executable programs are supplied for most major platforms, it is not usually necessary to download the Vibrant toolkit in order to use ClustalX. To compile ClustalX on an unsupported platform, the toolkit should be obtained by anonymous ftp to <ncbi.nlm.nih.gov>.

3. Methods

The programs ClustalW and ClustalX provide alternative user interfaces to the Clustal multiple alignment software. The alignments produced by the two programs are exactly the same; the only difference between ClustalW and ClustalX is the way in which the user interacts with the program. ClustalW is now mainly used as a command-line program by WWW servers and automatic batch systems, although the pro-

gram does provide text menus that can be used to input sequences and perform multiple alignments. Most users who run Clustal interactively now use the graphical interface provided by ClustalX. ClustalX is therefore used here to illustrate the basic alignment procedures.

3.1. Multiple Sequence Alignment

Sequences can be input to both ClustalW and ClustalX in one of seven file formats. All sequences must be in the same file. The formats that are automatically recognized are: NBRF/PIR, EMBL/SwissProt, Pearson (FastA), Clustal, GCG/MSF, GCG9/RSF, and GDE flat file. One of the simplest file formats for preparing sequences is the Pearson (FastA) format (Fig. 1). The sequences must be all nucleotide or all amino acid, and the program will attempt to guess which by the composition of the letters. Upper or lower case can be used, most symbols and numbers will be ignored (removed), and unrecognized residues will be counted as X or N. If the input file is prepared using a word processor, the data file should be saved as plain text with line breaks, i.e., as a simple ASCII file. ClustalX cannot deal with native word-processor formats.

3.1.1. Start a ClustalX Session

1. On PC and Macintosh computers, click on the ClustalX icon. On UNIX systems, type `clustalx &`.
2. The ClustalX window on UNIX or PC systems has a series of menu items across the top. For Macintosh users, the menu items are displayed at the top of the screen, separate from the ClustalX window itself.
3. Options can be selected by moving the mouse cursor to one of the menu items and clicking the left mouse button to display the list of menu options under that item, then moving the cursor to the appropriate option and clicking the mouse button again.

3.1.2. Load Sequences in ClustalX

1. Select the File menu, Load Sequences option in the ClustalX window.
2. In the new window displaying the user's subdirectories and files, select a file containing the unaligned sequences. Use the mouse cursor to highlight the filename in the file selection window, then click the OK button at the bottom of the window.
3. The unaligned sequences will be displayed in the ClustalX window with the sequence names on the left-hand side. The sequence alignment is for display only, it cannot be edited here. A ruler is displayed below the sequences, starting at 1 for the first residue position. The line above the alignment is used to mark strongly conserved positions. Sequence residues are colored to highlight conserved features in a multiple alignment, although at this stage the sequences are not yet aligned, and the residue coloring will not be informative.

3.1.3. Construct a Multiple Alignment of the Sequences

1. Select the Alignment menu, Do Complete Alignment option.
2. A new window will appear, displaying the default filenames for the output guide tree file and the output alignment file. If required, edit these filenames, before clicking on the Align button. For a description of the output format options, *see Note 1*.
3. ClustalX will perform the complete multiple alignment of the sequences shown in the window. The current status of the alignment process is continuously updated in the message area at the bottom of the ClustalX window. When the alignment is complete, the window display is updated to show the aligned sequences with gaps represented by “-” characters (Fig. 2). The multiple alignment is automatically written to the output file selected in step 2.

```

>trkb_human
SQLKPDTFVQHIKRHNIVLKRELGEAGFKVFLAECYNLCPEQDKILVAVKTLKDASDNA
RKDFHREAELLNLQHEHIVKFYGVCGVEGDPPLIMVFEYMKHGDLNKFRLAHGPDAVLMAE
GNPPTELTQSQMLHIAQQIAAGMVYLASQHFVHRLATRNCLVGENLLVKIGDFGMSRDV
YSTDYRVGGHTMLPIRWMPPESIMYRKFTIESDVWSLGVLWEIFTYGKQPWYQLSNNE
VIECITQGRVLQRPRTCPQEYELMLGCWQREPHMRKNIKGIHTLLQNLAKASPVYLDIL
>q24327
KKSahiyeqlalprsglseliqigrgegdvfvglklatlvtspsdkadtekqhsnse
GSGGSGSGSTTLSTLNKEKRSKTSMDIEEIKEEQEQQHNGSLEQLVLVKALNKVKDEQ
ACQEFRQLDLLRAISHKGVVRLFGLCREKDPMVLEYTDWGLKQFLLATAGKVNTAT
AGSSPPPLTTSQVLAVAYQIARGMDAIYRARFTHRLATRNCVISSEFIVKVSYPALCK
DKYSREYHKHRNTLLPIRWAPECIQEDEYTTKSDIFAYGVVVWELFNQATKLPEELTN
EQVVRQSQAGSLEWSVAEATPDSLREILLSCWVSNPKERPSFSQLGAALSKAMQSAEK
>kpro_maize
TSNFRRYSYRELVKATRKFKVELGRGESGTVYKGVLEDDRHVAVKLENVRQGKEVFQAE
LSVIGRINHMNLVRIWGFCSEGSHRLLVSEYVENGSLANILFSEGGNNILLDWEGRFNIAL
GVAKGLAYLHHECLEWVVIHCDVKPENILLDQAFEPKITDFGLVKLLNRGGSTQNVSHVRG
TLGYIAPEWVSSLPIATAKVDVSYGVVLELLTGTRVSELVGGTDEVHSLRKLVRLMSA
KLEGEEQSWIDGYLDKLNRPVNYVQARTLIKLAWSCLEEDRSKR
>kraf_drome
RDAKSSEENWNILAEELIGPRIGSGSGFTVYRAHWHPVVKTLNVKTPSPAQLQAFKN
EVAMLKTRHCNILLFMGCVSKPSLAIVTQWCEGSSLYKHVHSETKFKLNTLIDIGRQV
AQGMDYLHAKNIIHRDLKSNNIFLHEDLSVKIGDFGLATAKTRWSGEKQANQPTGSILWM
APEVIRMQELNPYSQSFQSDVYAFGIVMYELLAECIPLYGHISNQDQILFMVGRGLLRRPDMSQ
VRSDARRHSKRLAEDCIKYTPKDRPLFRPLLNMLENMLRTLPIKRSAS
>insr_human
PCSVYVPDEWEVSREKITLLRELGQGSFGMVYEGNARDIICKGEAETRAVAKTVNESASLR
ERIEFLNEASVMKGFTCHHVVRLLGVVSKGQPTLVMELMAHGLKSYLRSLRPEAENN
GRPPPTLQEMIQMAAEIADGMAYLNAKKFVHRLAARNCMVAHDFTVKIGDFGMTRDIYE
TDYYRKGGKGLLPRWMAPESLKDGVFTSSDMWSFGVVLWEITSLAEQPYQGLSNEQVL
KFVMDGGYLDQPDNCPERVTDLMRMCWQFNPKMRPTFLEIVNLLKDDLHPSFPEVFFH

```

Fig. 1. A sample sequence input file containing five protein kinase domain sequences in FASTA format.

3.1.4. Quality Control

The ClustalX graphical interface offers several methods of analyzing the multiple alignment.

1. Strongly conserved positions are indicated on the line above the alignment. The “*” character indicates positions which have a single, fully conserved residue. The “:” and “.” characters indicate that the column is strongly or weakly conserved, respectively. These symbols are also included in the output text file when CLUSTAL format is used.
2. The sequence residues are colored in order to highlight conserved features—for example, hydrophylic or hydrophobic positions in the alignment. For more details of the coloring scheme, *see Note 2*.

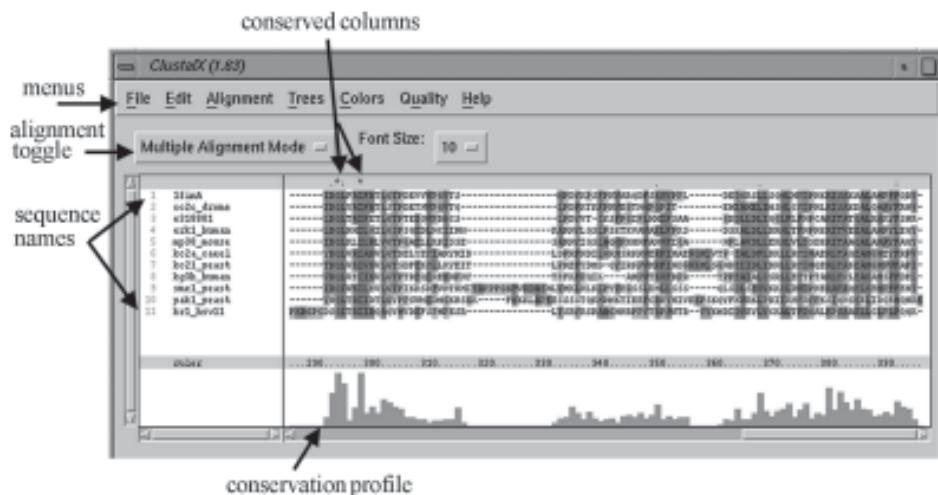


Fig. 2. The ClustalX window in multiple sequence alignment mode. The sequence display area shows part of a multiple alignment of the catalytic domain of 11 protein kinase sequences. Residues are shaded according to the default coloring scheme.

3. The quality curve displayed below the alignment plots a “conservation” score for each column in the alignment. A high score indicates a well-conserved column; a low score indicates low conservation.
4. Finally, there are extensive facilities for directly highlighting sections of sequences or blocks of alignment that appear to be very unreliable or poorly aligned, or where the alignment is very ambiguous. These facilities are found under the Quality item of the main menu at the top of the ClustalX window.

3.1.5. Change the Alignment Parameters

If the alignment that is obtained using default settings is not optimal, the user can modify a large number of alignment parameters. Some typical alignment problems and possible solutions are described in the Notes section.

1. Select the Alignment menu, Alignment Parameters option. Then, select either Pairwise Alignment Parameters or Multiple Alignment Parameters. Pairwise alignment parameters will mainly affect the speed/sensitivity of the initial alignments that are used to construct the guide tree, but will not normally have a great effect on the final multiple alignment (*see Note 3*). In contrast, the multiple alignment parameters control exactly how the final multiple alignments are carried out (*see Note 4*).
2. Rebuild the multiple alignment. If the pairwise parameters have been changed, it will be necessary to rebuild the complete multiple alignment, as in **Subheading 3.1.3**. If only the multiple alignment parameters have been changed, the first stages (pairwise alignments, guide tree) can be re-used by using the File menu, Do Alignment from Guide Tree option. A window appears with the default filenames of the input guide tree (written during the multiple alignment process in **Subheading 3.1.3**), and the output alignment file. ClustalX will perform only the final multiple alignment of the sequences shown in the window. When the alignment is complete, the window display is updated to reflect the new multiple alignment.

3.1.6. Save the Alignment

During the alignment process, the final multiple alignment is automatically written to the output file. This file may be specified by the user, or the default name and format type may be used. In addition, after the multiple alignment is completed, the user has the option of changing the output file format, or of saving only a selected part of the whole alignment and of getting the output alignment written out to a file again.

1. Select the File menu, Save Sequences As option.
2. A window will appear offering the user a choice of one of the six output formats (see **Note 1**). In addition, the output filename may be specified by the user.
3. Click on the OK button to save the sequence alignment to the selected file.

3.1.7. Create a Postscript Image of the Alignment

The ClustalX alignment display can be saved in a postscript file, which can then be either sent directly to a printer, or loaded into a graphics editing program. The file will automatically include the colored sequences, and the consensus and ruler lines. The Alignment Quality curve can be optionally included in the output file.

1. Select the File menu, Write alignment as postscript option. A window will appear with a number of options for customizing the postscript output.
2. A number of options are available to allow you to configure your postscript output file. For example, the exact RGB values required to reproduce the colors used in the alignment window are specified in a file that can be edited by the user. The alignment can be displayed on A4, A3, or US Letter size pages, and as either a landscape or portrait page.
3. Click on the OK button to save the sequence alignment to the selected file.

3.2. Merge Two Existing Alignments

ClustalW and ClustalX allow the user to re-use an old alignment and add new sequences to it, or even merge two alignments together. This is known as profile alignment (the term *profile analysis* was first used by Gribskov et al. [15]). This is useful in any ongoing project where new sequences are being generated and alignments need updating. Adding new sequences to an old alignment has some advantages. Firstly, it is much faster than redoing the alignment from scratch each time. Secondly, the original sequence alignment is kept intact, which is especially useful if the alignment had been hand edited. A profile is simply an alignment of one or more sequences (e.g., an alignment output file from Clustal). One or both sets of input sequences may include secondary structure assignments or gap penalty masks to guide the alignment (see **Note 5**). Sequences and existing alignments can be input to both ClustalW and ClustalX in one of the seven file formats described in **Subheading 3.1.**

3.2.1. Start ClustalX in Profile Alignment Mode

1. Start a ClustalX session (as in **Subheading 3.1.1.**) and switch to Profile Alignment Mode by clicking on the Multiple Alignment Mode toggle button just above the sequence display area. The single sequence display area will be replaced by two display areas. Initially, both areas are empty.

3.2.2. Load the Profiles

1. Load the first profile by selecting the File menu, Load Profile 1 option and then selecting a file in the file selection window. The procedure is similar to that used in **Subheading 3.1.2.** Profile 1 should contain a single sequence or an existing alignment of two or more

sequences—e.g., an alignment file that was produced by ClustalX at an earlier stage. The selected alignment will be displayed in the top half of the ClustalX window (*see Sub-heading 3.1.2.* for a description of the display).

2. Load the second profile, by selecting the File menu, Load Profile 2 option. The procedure is the same as that used for loading the first profile. Profile 2 should contain a single sequence or several aligned sequences. The selected alignment will be displayed in the bottom half of the ClustalX window.

3.2.3. Align the Two Profiles

1. Select the Alignment menu, Align Profile 2 to Profile 1.
2. A window will appear that displays the default filenames for the output guide tree files and the output alignment file. If required, edit these filenames, before clicking on the Align button. For output format options, *see Note 1.*
3. ClustalX will align the two profiles together to form a single multiple alignment. The original alignments are not altered. The two profiles are simply aligned together by introducing complete columns of gaps into one or both of the profiles. The current status of the alignment process is continuously updated in the message area at the bottom of the ClustalX window. When the alignment is complete, the window display areas are updated to show the aligned profiles.
4. Clicking on the Lock Scroll button just above the top display area will remove the horizontal scroll bar from the top display area (**Fig. 3**). The single remaining scroll bar at the bottom of the window will then allow both profile display areas to be scrolled together.
5. Merge the two profiles by switching back to multiple alignment mode, using the toggle button just above the top sequence display area. The sequences from both profiles are merged into a single alignment.

3.3. Align New Sequences to an Existing Alignment

A second option is to align the sequences from the second profile, one at a time, to the first profile. This is useful for incorporating a set of new sequences (not aligned) into an older alignment. The procedure to follow is very similar to that used above to merge two existing alignments. In this case, however, the second profile should contain one or more unaligned sequences. The sequences can be aligned to Profile 1 by selecting the Alignment menu, Align Sequences to Profile 1 option. The set of aligned sequences (Profile 1) may include secondary structure assignments or gap penalty masks to guide the alignment (*see Note 5.*)

4. Notes

1. By default, the output file of the program is produced in CLUSTAL format, which can be read by many other sequence analysis packages. To change this, select the output format using the Alignment menu, Output Format Options window. The final multiple alignment can be saved in one (or more than one) of seven file formats: CLUSTAL, NBRF/PIR, GCG/MSF, PHYLIP, NEXUS, FASTA, or GDE. The different output file formats are provided for compatibility with a wide range of multiple alignment analysis programs. You can also change the default case of the residues from lower case to upper for GDE output, by clicking the appropriate button in this window. Residues are not normally numbered in the output, but an option is provided to use numbers here. By default, the order of the sequences is changed to reflect the order of alignment. In this way, the sequences in the alignment are clustered into similar groups. This can be changed by setting the output order to be the same as the input order. Finally, the values of the parameters (gap penalties, amino acid weight matrix, and so on) can be printed out in the output file by clicking

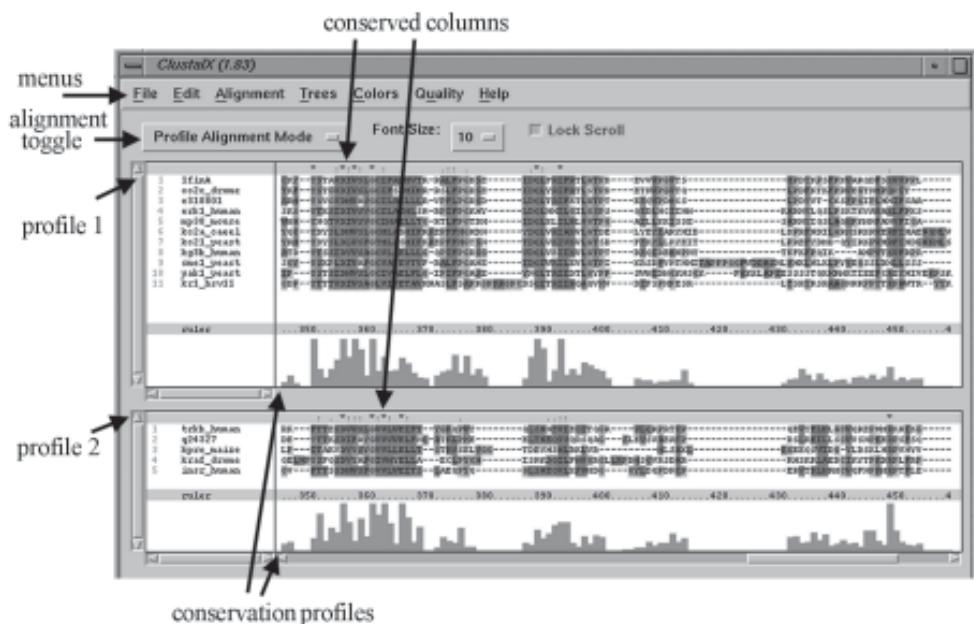


Fig. 3. The ClustalX window in profile alignment mode. The top display area shows part of a multiple alignment of 11 protein kinase sequences. The bottom display area shows another multiple alignment consisting of 5 protein kinase sequences. The two alignments, or profiles, have been aligned together using the Profile Alignment option. Residues in both display areas are shaded according to the default coloring scheme.

the lowest button here. The output files are produced as plain text or ASCII, and if you wish to view these using a word-processing package, you must use a fixed space font such as Courier. This ensures that the aligned residues from the different sequences will be placed neatly in columns.

2. Clustal X provides a versatile coloring scheme for the sequence alignment display. The sequences (or profiles) are colored automatically, when they are loaded. Sequences can be colored either by assigning a color to specific residues, or on the basis of an alignment consensus. In the latter case, the alignment consensus is calculated automatically, and the residues in each column are colored according to the consensus character assigned to that column. The rules used to color the alignment are specified in a color parameter file, which can be edited by the user. These colored alignments cannot be seen in the normal alignment output files. If you wish to print these out using the colors, you need to produce a Postscript file (*see Subheading 3.1.7.*) and print it out with a Postscript-capable printer.
3. Pairwise parameters. The most important choice here is that between Slow-Accurate and Fast-Approximate pairwise alignments. The Accurate alignments are carried out using a dynamic programming method (5) to align every pair of sequences. This may be too slow for large numbers (e.g., >100) of long (e.g., >1000 residue) sequences. In this case, the Fast/Approximate alignments using the method of Wilbur and Lipman (16) may be more suitable. These are several orders of magnitude faster to construct than the former, and allow huge data sets to be aligned. The effects on the accuracy of the final alignments are minor except in cases where the alignment is especially difficult.
4. Multiple parameters. Each step in the final multiple alignment consists of aligning two alignments or sequences. This is done progressively, following the branching order in the

guide tree. The multiple alignment parameters window allows the user to change the scoring matrices and the penalties for opening and extending gaps in the sequences. Gap penalties usually need to be altered for aligning nucleic acids, e.g., they are likely to require reduction if divergent sequences are present in the set. For proteins, this is not so often the case, as these parameters are modified in various complicated ways by another set of parameters (the protein gap parameters). However, it should be noted that the default parameters have been optimized for shorter sequences (less than 300 residues). For longer sequences, it may be useful to lower the gap opening and extension penalties. The Delay Divergent Sequences option delays the alignment of the most distantly related sequences until after the most closely related sequences have been aligned. For small numbers of sequences (<100), this generally improves the alignment of the most divergent sequences. For very large alignments, however, it may be better to follow the branching order of the guide tree exactly by setting the Delay Divergent Sequences option to zero. One parameter that is very important is the Use Negative Matrix option. This controls whether the amino acid weight matrix will contain positive values only or positive and negative. By default, an all-positive matrix is used, and for fully co-linear sequences, the default gives slightly (but clearly) better alignments. But it is sometimes necessary to choose the negative matrix option, especially if you have large terminal extensions/deletions or fragments of sequences, or when the sequences are only partially related, as often occurs when a sequence set is taken directly from a database search output.

5. The use of secondary structure-based penalties has been shown to improve the accuracy of sequence alignment. ClustalX allows secondary structure/gap penalty masks to be supplied with the input sequences used during profile alignment. (NB: The secondary structure information is not used during multiple sequence alignment.) If a solved structure is known, the secondary structure information can be used to guide the alignment by raising gap penalties within secondary structure elements, so that gaps will preferentially be inserted into unstructured surface loop regions. Alternatively, a user-specified gap penalty mask can be supplied for a similar purpose. A gap penalty mask is a series of numbers between 1 and 9, one per position in the alignment. Each number specifies how much the gap opening penalty is to be raised at that position—i.e., a mask figure of 1 at a position means no change in gap opening penalty; a figure of 4 means that the gap opening penalty is four times greater at that position, making gaps four times harder to open. For many 3-D protein structures, secondary structure information is recorded in the feature tables of Swiss-Prot database entries. Clustal X looks for Swiss-Prot HELIX and STRAND assignments. The structure and penalty masks can also be read from the CLUSTAL alignment format or the GDE flat file format.

References

1. Needleman, S. B. and Wunsch, C. D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453.
2. Smith, T. F. and Waterman, M. S. (1981) The identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197.
3. Sankoff, D. (1975) Minimal mutation trees of sequences. *SIAM J. Appl. Math.* **28**, 35–42.
4. Feng, D. F. and Dolittle, R. F. (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.* **25**, 351–360.
5. Myers, E. W. and Miller, W. (1988) Optimal alignments in linear space. *CABIOS* **4**, 11–17.
6. Higgins, D. G. and Sharp, P. M. (1988) CLUSTAL : a package for performing multiple sequence alignments on a microcomputer. *Gene* **73**, 237–244.
7. Higgins, D. G. and Sharp, P. P. (1989) Fast and sensitive multiple sequence alignments on a microcomputer. *CABIOS* **5**, 151–153.

8. Higgins, D. G., Bleasby, A. J., and Fuchs, R. (1992) CLUSTAL V: improved software for multiple sequence alignment. *Comput. Appl. Biosci.* **8**, 189–191.
9. Saitou, N. and Nei, M. (1987) The neighbor-joining method : a new method for reconstructing phylogenetic trees . *Mol. Biol. Evol.* **4**, 406–425.
10. Felsenstein, J. (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**, 783–791.
11. Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680.
12. Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F., and Higgins, D. G. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25**, 4876–4882.
13. Page, R. D. (1996) TreeView: an application to display phylogenetic trees on personal computers. *Comput. Appl. Biosci.* **4**, 357–358.
14. Perrière, G. and Gouy, M. (1996) WWW-Query: an on-line retrieval system for biological sequence banks. *Biochimie* **78**, 364–369.
15. Gribskov, M., McLachlan, A. D., and Eisenberg, D. (1987) Profile analysis : detection of distantly related proteins. *Proc. Natl. Acad. Sci. USA* **84**, 4355–4358.
16. Wilbur, W. J. and Lipman, D. J. (1983) Rapid similarity searches of nucleic acid and protein data banks. *Proc. Natl. Acad. Sci. USA* **80**, 726–730.

FASTA Servers for Sequence Similarity Search

Biju Issac and Gajendra P. S. Raghava

1. Introduction

In the last few years, many eukaryotic (including human and mouse) and prokaryotic genomes have been either completely sequenced or are under sequencing (1–3). In the coming 5–10 yr, most of the known organisms will have been sequenced. This has and will lead to exponential growth in nucleotide and protein databases over the years; for example, International Nucleotide Sequence Databases (INSD), composed of DDBJ (<http://www.ddbj.nig.ac.jp/>), EMBL Bank (<http://www.ebi.ac.uk/embl/>), and GenBank (<http://www.ncbi.nlm.nih.gov/>), had released more than 30 million entries by the end of 2003 (4). The availability of these increasingly expanding databases poses a major challenge to bioinformatics experts for developing effective programs or Web servers that extract maximum information from these databases. Database similarity search is perhaps the fastest, cheapest, and most powerful such experiment a biologist can conduct. As the databases become more complete, a sequence similarity search is more likely to reveal database sequences with statistically significant similarity, and thus inferred homology, to a query sequence. Though sharing significant sequence similarity is no guarantee of shared function, the availability of similar sequences is proving useful in discovering relationships between newly sequenced proteins or genes and various classes in the databases (5–7).

1.1. DOT Matrix and Dynamic Programming

The DOT matrix is one of the oldest and most useful techniques to detect the similar regions between two sequences. It also allows detecting insertions and deletions (indels) between sequences, because they shift the diagonal horizontally or vertically by the amount of change (8). But the major limitation of this technique is that it is manual and subjective in nature. In the past, attempts have been made to develop sequence alignment methods to measure the similarity between sequences quantitatively. Needleman and Wunsch (9) developed a global alignment method using dynamic programming (DP). Though one can get optimized alignment between two sequences using DP, it is not suitable for database scanning. This is because a database consists of sequences of variable lengths, and this method is suitable only when the length of two sequences is of the same order. DP is also very slow and extraordinarily memory intensive. The time it takes to execute a job, and the memory required to handle such data, is directly proportional to the length of the query sequence (N) and the cumulative length (M) of the database sequences. Smith and Waterman (10) modified the DP to develop the

Smith-Waterman algorithm, to make it suitable for local alignment. Local alignments are suitable for database scanning because these allow matching similar regions between two sequences.

1.2. Approximation Algorithms

Although the Smith-Waterman method is a fast implementation of DP, it too takes a long time to scan big databases. Subsequently, approximation methods that are nearly as sensitive but much faster than DP-based methods have revolutionized the concept of similarity searches. These methods search for words or k-tuples from the query and then join these words into an alignment by the DP method. FASTA and Basic Local Alignment Search Tool (BLAST) are two approximation methods that are at least 50 times faster than the Smith-Waterman DP algorithm (11-15).

While these two programs offer many of the same search capabilities, BLAST is usually faster than FASTA. Both these programs have their advantages and disadvantages. For certain translated DNA-protein comparisons and using specific matrices, FASTA is more sensitive than its BLAST counterparts (16). Both these methods are heuristic—i.e., an empirical method of computer programming in which rules of thumb are used to find a solution and feedback is used to improve performance.

2. FASTA Package

The FASTA program package has evolved significantly since its introduction 20 yr ago (16). It is a program for rapid alignment of pairs of protein and DNA sequences. Rather than comparing individual residues in the two sequences, FASTA instead searches for matching sequence patterns or words, called *ktups* (11,12,16). It compares an input DNA or protein sequence to all of the sequences in a target sequence database, and then reports the best-matched sequences and local alignments of these matches with the input sequence. The total number of programs offered today in the package is much more than the original four—fasta, tfasta, lfasta, and rdf. The latest version of the program includes the rigorous Smith-Waterman searches (sssearch3), searches with mixed peptide sequences (fastx3, fasty3, tfastx3, and tfasty3), and the program for estimating statistical significance from shuffled-sequence similarity scores (prss3).

3. FASTA Algorithm

FASTA uses four steps to calculate three scores that characterize sequence similarity; these steps are outlined in Appendix I. The first step uses a rapid technique for finding identities shared between two sequences; the method is similar to an earlier technique described by Wilbur and Lipman (16). FASTA programs achieve much of their speed and selectivity in the first step of the comparison by using a lookup table to locate all identities or groups of identities between two DNA or amino acid sequences (17).

In conjunction with the lookup table, the “diagonal” method finds all regions of similarity between the two sequences, counts ktup matches, and penalizes for intervening mismatches. A simple formula identifies portions of a diagonal with a high density of identities, referred to as a *local region of similarity*, or simply a *region*. This more sensitive formula is used for protein sequence comparisons; the constant value for ktup matches is used for DNA sequence comparisons. FASTA then saves the ten best local regions, regardless of whether they are on the same or different diagonals.

After the ten best local regions are found in the first step, they are rescored using a scoring matrix that allows runs of identities shorter than ktup residues and conservative replacements to contribute to the similarity score. For each of the best diagonal regions rescanned with the scoring matrix, a subregion with the maximal score is identified. The FASTA program uses the single best-scoring initial region to characterize pairwise similarity; the initial scores are used to rank the library sequences. Then it checks to see whether several initial regions can be joined together in a single alignment to increase the initial score. Given the locations of the initial regions, their respective scores, and a “joining” penalty (analogous to a gap penalty), it calculates an optimal alignment of initial regions as a combination of compatible regions with maximal score. This optimal alignment of initial regions is rapidly calculated using a DP algorithm. FASTA uses the resulting score, termed *inith score*, to rank the library sequences. In the optimization step, only those initial regions are included whose scores are above an empirically determined threshold. FASTA joins an initial region only if its similarity score is greater than the cutoff value, a value that is approximately one standard deviation above the average score expected from unrelated sequences in the library (18).

After a complete search of the library, FASTA plots the initial scores of each library sequence in a histogram, calculates the mean similarity score for the query sequence against each sequence in the library, and determines the standard deviation of the distribution of initial scores. The initial scores are used to rank the library sequences, and, in the fourth and final step of the comparison, the highest-scoring library sequences are aligned using a modification of the standard Needleman–Wunsch optimization method (10). The optimization employs the same scoring matrix used in determining the initial regions; the resulting optimized alignments are calculated for further analysis of potential relationships, and the optimized similarity score is reported. Because FASTA calculates an initial similarity score based on an optimization of initial regions during the library search, the initial score is much closer to the optimized score for many sequences. The FASTA method may yield initial scores that are higher than the corresponding optimized scores.

4. Installation of FASTA Program

Installing and running a command-line version of the FASTA package is much easier than installing a Worldwide Web (WWW) site for FASTA searches. Different versions of the package are available, and users can download the files that are best suited for their operating system (OS). Basically, the FASTA package can be downloaded from <ftp://ftp.virginia.edu/pub/fasta/>. Detailed installation procedures for the different operating systems are provided here in **Table 1**.

5. How to Execute FASTA

All FASTA sequence comparison programs require similar information. A query file, library file, and ktup parameters are the minimum requirement for all programs. Input files can be either of two types: (1) plain sequence files (files that contain nothing but sequence residues) and (2) FASTA format files. FASTA files are the same as plain sequence files with each sequence preceded by a comment line with a “>” in the first column. Formats of the library sequences and their corresponding numbers that are currently used by FASTA program are:

Table 1
Download and Installation Instructions of FASTA Package

Operating system	Installation procedures
UNIX/LINUX	<ol style="list-style-type: none"> 1. Download latest FASTA files for Mac or Windows Operating System from site ftp://ftp.virginia.edu/pub/fasta/mac_fasta/ or ftp://ftp.virginia.edu/pub/fasta/win32_fasta/. 2. Unpack the file in an empty directory by typing “uncompress fasta34t23bl.shar.Z.” 3. Type “sh fasta34t23bl.shar.” 4. Type “make all.” 5. Type “make -f Makefile.<machine type>.”
Mac/Windows	<ol style="list-style-type: none"> 1. Download latest FASTA file fasta34t23bl.shar.Z from site ftp://ftp.virginia.edu/pub/fasta/. 2. Unpack the self-extracting file in an empty directory. 3. Start using the executable binaries.
WWW server	<ol style="list-style-type: none"> 1. Download latest FASTA file fasta34t23bl.shar.Z from site ftp://ftp.virginia.edu/pub/fasta/. 2. Uncompress and unshar it in a system directory. 3. Compile the programs using appropriate machine type. 4. Edit and save the “fastlibs” file along with your databases and set the environment variable “FASTLIB.” 5. Run and test FASTA commands 6. Edit “search_frm.cgi” and “search_run.cgi” and replace “/home/wrp/public_html/tmp/errors.log” with the location of the file you want errors sent to. 7. Edit my-cgi.pl to indicate where various files are on your system. 8. Execute search_frm.cgi from command line to generate a WWW page (html) and does not have any error messages.

- 0 Pearson/FASTA (>SEQID - comment/sequence)
- 1 Uncompressed Genbank (LOCUS/DEFINITION/ORIGIN)
- 2 NBRF CODATA (ENTRY/SEQUENCE)
- 3 EMBL/SWISS-PROT (ID/DE/SQ)
- 4 Intelligenetics (;comment/SEQID/sequence)
- 5 NBRF/PIR VMS (>P1;SEQID/comment/sequence)
- 6 GCG (version 8.0) Unix Protein and DNA (compressed)
- 11 NCBI Blast1.3.2 format (unix only)
- 12 NCBI Blast2.0 format (unix only, fasta32t08 or later)

5.1. A Sample FASTA Search

Use a ktup of 1 or 2 for protein sequences, or 1 to 6 for DNA sequences; a ktup of 2 is about 5 times faster than a ktup of 1. For example, let us consider a file XM_352397.fa that contains the mRNA sequence XM_352397 (human ribosome biogenesis protein BMS1 homolog; GI: 37559472) in FASTA format, as shown in [Fig. 1](#). Search this sequence with the fasta34 program against the *Escherichia coli* nucleotide sequence database file (ecolint) in FASTA format, with ktup = 6. Using the command line

```
>gi|37559472|ref|XM_352397.1| Homo sapiens similar to Ribosome biogene
ATGAAGAAAGCGGGGGCCCGCGCAGCATGTTGGTCCGGGCAGGGAGCCCGGCCGCGCTCGTGCCTTCTTC
TCGCCCCACGACGGTGGCCGGCGGGCGGGCTCTGCTTCCCGGACCCATGAACAAAGCTGGAGATGCGGCTTT
GGAGCTGGCCGGCTGGGACCTGCTTCTTGGCTTCGGGGACCCGAGCCCAGGGCGTTCCCTCCGGATC
GGTGGCCAGGGCTTGGGGCGTGTGCGACTCGGGACCCCGAGCCGGCAGCAGCGGGCCAGCATGG
CAAGGGAGCAGGCCAGCTCTCCCCACCGCGTCCCCGGATGGGGGCTTCGGGTGGGAGCCGGCCGGGA
CATCCAGGGAAATGTTAAATTCTGCCTTGGAAAGTGGCATATTGAAAGTGCTGTGATTCAAACGTGCAAG
GAGATAAGGCAGCAGATCAAGAAAGCACTCCGGCTCCAGAAGGAGCCTTCAGGCCAGCTTGAGCATA
AGCTGCTGATGAGCAATACTGTCTTCATGCGAAACTGGTATCTGTGTTCCATCCCAGCCTTCTATAACCC
AGTAACATCTTTTTGAAACCAGTGGGTGAGAAAGACACCTGGTCAGGAACCGGGACCACAGGACAACTC
AGGCTCACCCACGGCATCAGACTAAAGGCAAAACAGGACTCTGTATAA
```

Fig. 1. FASTA formatted human sequence of a homolog of Ribosome biogenesis protein (BMS1) with GenBank Accession number XM_352357 (gi:37559472).

“fasta34 XM_352397.fa ecolint 6” will produce an output with a histogram as shown in **Fig. 2A**. The program also produces an alignment between the matched *E. coli* sequence from the database and the query sequence (**Fig. 2B**). Options available for different programs can be changed by the user. These change the scoring matrix and gap penalties, use alternate statistical estimation methods, and change the format of the alignment output. Though most options apply to all of the programs in the package, other options are specific to fasta34 or tfastrx/y34. Generally, command-line options can be divided into five general categories: scoring parameter, statistics, algorithm-specific, file specification, and output options (**Table 2**).

6. Performance of FASTA

FASTA programs have all undergone recent enhancements that have improved detection of more remotely related sequences. Theoretically, it is as good as its contemporary, the popular search program BLAST. Both the local alignment programs have their own advantages and disadvantages; however, this chapter is not a review of their comparative performances. In general, the BLAST family of sequence-comparison programs offers many of the same search capabilities as the FASTA programs but is faster, with some loss in sensitivity. On the other hand, the FASTA programs provide more accurate and full-length alignments between pairs of sequences, which prove more useful than the several short high-segment pairs produced by BLAST. FASTA has been found to have 43.2% coverage on default parameters, compared to 21.6% for BLAST (*18*).

The current BLAST package will identify an unknown protein as effectively as the fasta or ssearch programs, and produce an alignment similar to that of a rigorous Smith-Waterman search. The FASTA programs perform much better than BLAST for translated DNA-protein comparison and DNA database searches (*15*). They treat all three forward reading frames as a single sequence, to easily produce high-quality alignments that extend across the length of the matched protein sequence, and allow similarities from different reading frames to be combined in a natural way to improve sensitivity. FASTA programs allow searching of short sequences with smaller ktups, while matrices with high mismatch penalties can be used to identify long identities in sequences. For reviews on FASTA, see Pearson (*15,17,19–21*). The information provided above is largely based on these reviews, as a thorough investigation into the comparative performance of BLAST and FASTA is lacking.

```

FASTA searches a protein or DNA sequence data bank
version 3.4t07 Nov 21, 2001
Please cite:
W.R. Pearson & D.J. Lipman PNAS (1988) 85:2444-2448

XM_352397.fa, 678 nt
vs /usr3/blastpr/bakup_db/MIRROR/fasta/ecolint library

      opt      E()
< 20      0      0:
22      0      0:          one = represents 1 library sequences
24      0      0:
26      0      0:
28      1      0:=
30      2      1:*=*
32      2      2:*=*
34      12      6:====*=====
36      14      12:=====*===
38      20      20:=====
40      34      27:=====
42      27      33:=====
44      24      37:=====
46      41      38:=====
48      37      36:=====
50      29      33:=====
52      27      29:=====
54      26      25:=====
56      19      21:=====
58      16      17:=====
60      10      14:=====
62      16      11:=====
64      13      9:=====
66      6      7:=====
68      7      5:=====
70      5      4:=====
72      4      3:=====
74      1      2: - *

```

Fig. 2. Typical FASTA report generated after search of XM_352357 sequence against *Escherichia coli* nucleotide sequence database with histogram.

7. Types of FASTA Programs

Users new to FASTA are unfamiliar with the type and function of different FASTA programs available for these kinds of searches and when to use them. For these users, **Table 3** provides some valuable insights on what some of the popular FASTA programs do and when to use them.

8. Interpretations of Expectation (E) Values

The E value, or expectation value, calculated by FASTA is the number of times you would expect to see an equal or greater score by chance in a search of the database.

```

>>gi|1788045|gb|AE000270.1|AE000270 Escherichia coli K-1 (10106 nt)
rev-comp initn: 142 initl: 96 opt: 114 Z-score: 89.4 bits: 30.8 E(): 1.7
banded Smith-Waterman score: 114; 63.889% identity (67.647% ungapped) in 72 nt overlap (496-429:8011-8082)

      520      510      500      490      480      470
gi|37- AAGTTCGCATGAAAGACAGTATTGCTCATCAGCAGCTTATGCTC&gt;AGCTGGCCTGGAAAG-
      :::: ::::: ::::: ::::: :::::
gi|178 GCGTGACGTATAACGCTGCAGCGCGTGTGGCGCAGATTATGCTTTATGCTGGCGTGAAAAGA
      7990      8000      8010      8020      8030      8040

      460      450      440      430      420
gi|37- ---GCTCCTTCTGGAGCCCGGAGTGCTTTCTTGATCTGCTGCCTTATCTCGCTGACAGTT
      :::: ::::: :::: ::::: :::::
gi|178 TGTGCGCTGCTGGATGGCGCGTGGCAAACCTGTCGACGCGGGACTGCCTGTTGAGCG
      8050      8060      8070      8080      8090      8100

      410      400      390      380      370      360
gi|37- TGAATCACAGCACCTTCAAATATGGCCACTTCCAAGGCAGAATTAAACATTCCCTGGATG
      8110      8120      8130      8140      8150      8160

>>gi|1787058|gb|AE000186.1|AE000186 Escherichia coli K-1 (9963 nt)
rev-comp initn: 163 initl: 78 opt: 109 Z-score: 84.1 bits: 29.8 E(): 3.3
banded Smith-Waterman score: 109; 74.000% identity (84.091% ungapped) in 50 nt overlap (403-357:8841-8887)

      430      420      410      400      390      380
gi|37- GCTGCCTTATTCGCTGACAGTTGAATCACAGCACCTTCAA---ATATGCCACTTCA
      ::::: ::::: ::::: ::::: ::::: :::::
gi|178 AAACCGTCTTATTCTTCGCTCAACGGCAGCACCTTCAACGTTTATCGCCACTTCC
      8820      8830      8840      8850      8860

      370      360      350      340      330      320
gi|37- AGGCAGAATTAAACATTCCCTGGATGTCCCGGGCGCTCCACCCGAAGGCCGCCATCC
      ::::: ::::: :::::

```

Fig. 2. (continued) Alignment between regions in database sequences as provided by the FASTA report.

Whereas a sensible E-value threshold (0.001–0.01) can ensure that researchers avoid false-positive errors, little can be done to avoid false-negatives, i.e., labeling a sequence as unrelated to anything in the database when in fact a homolog is present. $E < 0.01$ says that you expect to see a score that high or higher once by chance, every time you do a search. The E value reported by the FASTA programs ranges from 0 to D , where D is the number of entries in the database. $E < 0.001$ says once in 1000 searches, and so on. An E of approx 1 says that one would expect a score that high, simply by chance, every time you do a search. However, as the sequence databases grow more complete and protein families expand, the rate of false-negatives should decrease.

E increases linearly with the number of database entries. Therefore, a similarity found in a search of a bacterial genome with 1000–5000 entries will be 50- to 250-fold more significant than an alignment with exactly the same score found in the nonredundant protein database with 250,000 entries. Thus, when searching for very distant relationships, one should always use the smallest database that is likely to contain the homolog of interest. Whereas size reduces search sensitivity, larger databases can be effective when they provide more diverse members of a protein family.

9. Internet Resources on FASTA

The availability of newly sequenced genomes from the mega-genome sequencing projects has provided a unique opportunity for finding homologous sequences in the known sequence databases. Such an effort will give a direction for correct identification of the functionality of the new sequences using homologous sequences from the databases. Though initial annotations for homolog detection use faster algorithms, FASTA is one of the important useful algorithms that are used for detecting homologs.

Table 2

Options for Changing Scoring Parameters and Output Display in Command-Line Version of FASTA (adapted from help documentation of the FASTA package)

Options	Parameters changed
-a	Show both sequences in their entirety (fasta34, ssearch34 only).
-A	Force Smith-Waterman alignments for fasta34 DNA sequences.
-b #	Number of sequence scores to be shown on output
-B	Show normalized score as a z score, rather than a bit score.
-c #	Threshold score for optimization.
-d #	Maximum number of alignments to be displayed (ignored if “-Q” not used).
-E #	Sets the highest E value.
-f	Penalty for the first residue in the gap.
-F #	Sets the lowest E value.
-g	Penalty for additional residues in a gap.
-h	Penalty for frameshift (only for fastx34/y34, tfastx34/tfasty34).
-H	Omit histogram.
-I	Reverse complement (invert) query DNA sequence.
-j #	Penalty for frameshift within a codon (only for fasty34/tfasty34).
-l	Location of library menu file
-L	Display more library sequence information in the alignment
-m #	Specify output alignment type: 0-6,9,10.
-M	low - high. Include library (protein) sequences of length between low and high.
-n	Force query to be treated as DNA.
-o	Turn off default optimization of all scores greater than OPTCUT.
-O	Output file name
-p	Force query to be treated as protein.
-Q, -q	Quiet
-r	Specify match-mismatch scores for DNA comparisons.
-R	Summary result file name for every sequence in library.
-s	Specify the scoring matrix file
-S	Treat lower-case characters in the query or library sequences as “low-complexity” residues.
-w #	Line length (width) = number (<200).
-x #	Penalty for a match to an “X,” independently of PAM matrix.
-X	Specifies offsets for the beginning of the query and library sequence.
-y	Set the width of the band used for calculating “optimized” scores.
-z	-1 turns off statistical calculations 0 estimates the significance of match from the mean and Std. Dev of library scores. 1 uses a weighted regression of average score vs library sequence length 2 uses maximum likelihood estimates of Lambda and K 3 uses Altschul-Gish parameters. 4 and 5 uses two variations of option 1.

Therefore, quite a few Internet-based servers are available for different DNA and protein sequence comparisons. A list of important FASTA Web servers available on the Internet is given in Appendix II.

Table 3
Explanatory Strategy for Simple Investigative Problems in Similarity Searches Using FASTA

Investigative problems	Explanatory strategy
To identify expressed sequence tag (EST) sequence from newly sequenced genomic DNA.	Use fastx34 or fasty34 to first check whether the EST sequence in question encodes a product homologous to a known protein.
To identify new orthologs from EST database.	Use tfastx34 or tfasty34 to search sequences from the same species utilizing low scoring matrices to detect close relationships and avoid distant relationships.
To identify structural elements (e.g., protein coding genes) from newly sequenced genomic DNA.	Use fasta34 for DNA–DNA comparisons. However, use fastx34 to detect protein-encoding region by protein sequence comparison first and then trying translated protein sequence comparison using fasty34.
To identify unknown protein	For general protein to protein comparisons use fasta34 with ktup = 2 for speed or ktup = 1 for a more sensitive search. Maximum sensitivity search can be achieved by using ssearch34 though DNA comparisons offer no additional advantage. If a homolog cannot be found in protein databases, check the DNA databases with tfastx34 or tfasty34.

10. Genome-Wide FASTA Search Using GWFasta

The growing number of genomic sequences in the databases signals an opportunity for significant advancement in functional annotation using similarity search programs. Existing search programs have a limit to the speed and sensitivity for a typical search against these huge databases. Since the redundancy in the present databases ensures that rich information content of the sequences is lost as a result of multiple hits to identical database sequences, albeit with minor variations, parsing the report needs manual input. However, voluminous data has ensured that manual parsing of a typical FASTA report is increasingly difficult. The subsequent sequence analysis of the reported hits also proves to be a difficult task, because of the high risk of errors on repeated transfer of data from one sequence analysis program to another. Moreover, the generated reports lack a genomic perspective in their presentation of results.

A reliable tool that can combine evidence from genomic sequence comparisons with clues from intrinsic sequence properties will be of importance for ordinary users in the future (22). However, maintaining such a system on personal computers is difficult, owing to the increase in database sizes and the requirement for installation of various programs on specific machines. Such a complex system can be installed only on specific hardware, and will need specialized software. GWFasta offers a flexible and convenient user interface that supports searches against user-selected multiple genome and proteome databases; fully automated batch submission of multiple sequences; searches with various FASTA programs; and convenient post-processing of FASTA output (23).

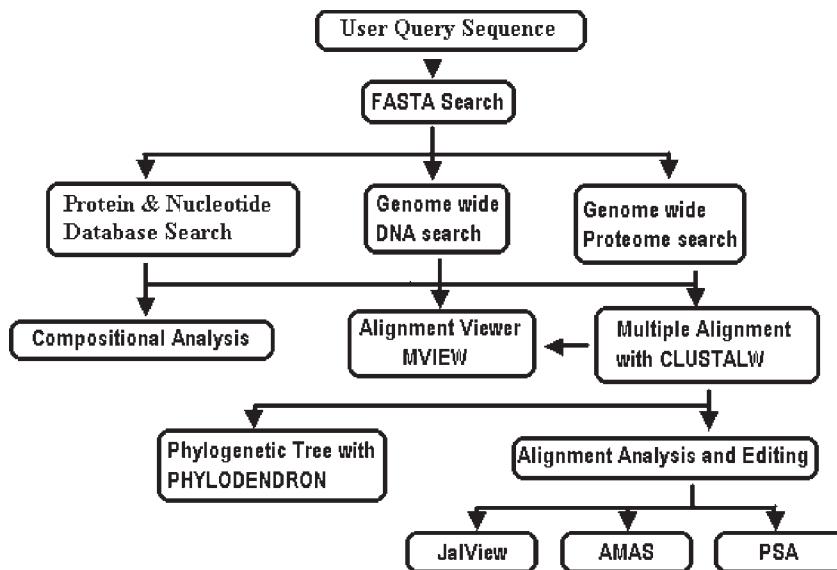


Fig. 3. General architecture of GWFasta server represented as a flow diagram.

GWFasta is a unique server, whose main objectives are (1) to give genomic perspective to the FASTA output; (2) to assist users in post-processing of FASTA reports in areas like compositional analysis, sequence alignment, and phylogenetic analysis; and (3) to automate submission for multiple sequences. While allowing standard FASTA searches, the GWFasta server provides for post-processing of FASTA search output that includes (1) viewing of FASTA alignment using MVIEW; (2) ClustalW for multiple sequence alignment; (3) generation of phylogenetic trees; (4) compositional analysis of user-selected sequences; and (5) editing and analysis of multiple sequence alignment. An advantage of using the GWFasta server is its batch processing capability, where the multiple query submission can be processed and individual outputs are related to the user's e-mail ID. A general representation of the GWFasta architecture is available in [Fig. 3](#).

10.1. Selection of Databases

The server maintains 65 microbial genomes (11 archaea and 54 bacteria) and 8 eukaryotic genomes ([Fig. 4](#)). Among the proteomes available for searching in the server, 53 are microbial (11 archaea and 42 bacteria). A mirror database package installed in the GWFasta server ensures that the different useful databases are updated weekly. The number of databases available for similarity searches permits the GWFasta server to provide users with three types of searches. Standard FASTA searches can be done using the search against the nonredundant databases. On the other hand, the server also allows genome- and proteome-wide searches where prokaryote and eukaryote databases can be searched separately.

10.2. Selection of Input Parameters

One of the important criteria for users regarding software packages is the selection of those values that give correct results for any option in a multi-choice environment

<input checked="" type="checkbox"/> All Microbial Genomic Sequences	
<input type="checkbox"/> All Archaeal Genomes	<input type="checkbox"/> All Bacterial Genomes
Archaeal Genomes	
<input type="checkbox"/> <i>Aeropyrum pernix</i> K1	<input type="checkbox"/> <i>Archaeoglobus fulgidus</i>
<input type="checkbox"/> <i>Halobacterium</i> sp. NRC-1	<input type="checkbox"/> <i>Halobacterium</i> sp. NRC-1 pNRC100 plasmid
<input type="checkbox"/> <i>Halobacterium</i> sp. NRC-1 pNRC200 plasmid	<input type="checkbox"/> <i>Methanobacterium thermoautotrophicum</i>
<input type="checkbox"/> <i>Methanococcus jannaschii</i>	<input type="checkbox"/> <i>Methanococcus jannaschii</i> large ECE
<input type="checkbox"/> <i>Methanococcus jannaschii</i> small ECE	<input type="checkbox"/> <i>Pyrococcus abyssi</i>
<input type="checkbox"/> <i>Pyrococcus horikoshii</i> OT3	<input type="checkbox"/> <i>Sulfolobus tokodaii</i>
<input type="checkbox"/> <i>Sulfolobus solfataricus</i>	<input type="checkbox"/> <i>Thermoplasma acidophilum</i>
<input type="checkbox"/> <i>Thermoplasma volcanium</i>	<input type="checkbox"/>
Bacterial Genomes	
<input type="checkbox"/> <i>Agrobacterium tumefaciens</i> C58 circular	<input type="checkbox"/> <i>Agrobacterium tumefaciens</i> C58 linear
<input type="checkbox"/> <i>Agrobacterium tumefaciens</i> C58 pTi plasmid	<input type="checkbox"/> <i>Agrobacterium tumefaciens</i> C58 pAT plasmid
<input type="checkbox"/> <i>Aquifex aeolicus</i>	<input type="checkbox"/> <i>Aquifex aeolicus</i> ec1 plasmid
<input type="checkbox"/> <i>Bacillus halodurans</i> C-125	<input type="checkbox"/> <i>Bacillus subtilis</i>
<input type="checkbox"/> <i>Barrelia burgdorferi</i>	<input type="checkbox"/> <i>Buchnera</i> sp. APS
<input type="checkbox"/> <i>Buchnera</i> sp. APS pTrp DNA plasmid	<input type="checkbox"/> <i>Buchnera</i> sp. APS pLeu DNA plasmid
<input type="checkbox"/> <i>Campylobacter jejuni</i>	<input type="checkbox"/> <i>Caulobacter crescentus</i>
<input type="checkbox"/> <i>Chlamydia pneumoniae</i>	<input type="checkbox"/> <i>Chlamydia trachomatis</i>
<input type="checkbox"/> <i>Chlorobaculum thalassium</i>	<input type="checkbox"/> <i>Chlorobaculum thalassium</i> pChB1 plasmid

Fig. 4. List of microbial proteome databases available in GWFasta server as represented in the submission page.

like FASTA. While the GWFasta server provides users the default values selected by the developers for giving the best results, users are provided options that can be manipulated. These include: word size, or ktup; expectation value; substitution matrix; gap initiation; and extension penalties. The ktup parameter determines how many consecutive identities are required in a match. For protein sequence comparison, ktup = 2 is frequently used. More sensitive searches can be done using ktup = 1. For DNA sequence comparisons, the ktup parameter can range from 1 to 6, while for short oligonucleotides or oligopeptides, ktup = 1 should be used. The E value is set to 10 as default, while the default matrix is BLOSUM50. The gap initiation penalty is -12 by default for proteins, -16 for DNA, and -15 for FAST[XY]/TFAST[XY]. The gap extension penalty is -2 by default for proteins, -4 for DNA, and -3 for FAST[XY]/TFAST[XY].

10.3. Submission of Sequences and Retrieval of Results

The input sequences can be either plain primary sequences (proteins or DNA) or can be in any one of the 18 formats allowed by the READSEQ program (developed by Don Gilbert, Bioinformatics Group, Biology Department and Center for Genomics and Bioinformatics, Indiana University, Bloomington, IN). The server handles quite a heavy

load through an automated queuing system for sequences that are submitted in the batch mode. CGI-PERL scripts handle the back-end programming. Users are also optionally notified through e-mail (provided by the user) on the completion of their queries. They are allowed to use the unique Job Identification number allotted on submission to retrieve their results within 5 d. They also have the option to go for further sequence analyses using integrated programs.

10.4. Searches Against Nonredundant Databases

The users get a typical FASTA report on completion of a run against the nonredundant databases (Fig. 5). The report is, however, provided with options for retrieving and analyzing the sequences. The analysis of the report includes multiple sequence alignment (MSA), compositional analysis (in the case of proteins), creation of a phylogenetic tree, and visualization of alignments of hits against database sequences. Appropriate buttons are provided on the top of the report page for each particular type of sequence analysis program. The FASTA alignment obtained from an initial search against the sequence databases can be visualized using the alignment viewer MView (Fig. 6). The difference from a command-line version of a FASTA report is that the GWFasta Web server does not provide a histogram in its output. While some visual graphics are lost due to this, the histograms do not provide any additional data other than what is in the report.

10.5. Searches Against Genome and Proteome Databases

Cross-species genome comparison draws attention as a method for reliable gene annotation. Sequence similarity search programs stand out as one of the quickest and readily available methods to characterize newly sequenced genes and assign function to them. Functionally important elements such as genes or regulatory elements can be highlighted when genomic sequences from related species are compared, because such functional elements are strongly conserved in evolutionary process, unlike “junk” DNA. Users are provided the option of searching their query sequences against genomic and proteome databases. Separate pages are provided for eukaryotic and microbial searches. The microbial organisms are further classified into archaeal and bacterial groups. The reports from these searches are parsed and presented in alphabetical order of organisms that were selected for similarity searching (Fig. 7). The server retrieves only the aligned portions of the nucleotide query and the database sequences in the case of DNA–DNA comparisons, while in the case of protein–protein comparisons or other DNA–protein comparisons, the server extracts whole protein sequences from the databases. The users also have the option of scanning the original FASTA report of the genomic searches, which is similar in presentation to the typical FASTA search report.

10.6. Postprocessing of FASTA Output

10.6.1. Multiple Sequence Alignment (MSA)

The simultaneous alignment of many nucleotide or amino acid sequences is now an essential tool in molecular biology. Multiple alignments are used to find diagnostic patterns to characterize protein families; to detect or demonstrate homology between new sequences and existing families of sequences; to help predict the secondary and tertiary structures of new sequences; to suggest oligonucleotide primers for polymerase chain reaction (PCR); and as an essential prelude to molecular evolutionary analysis.

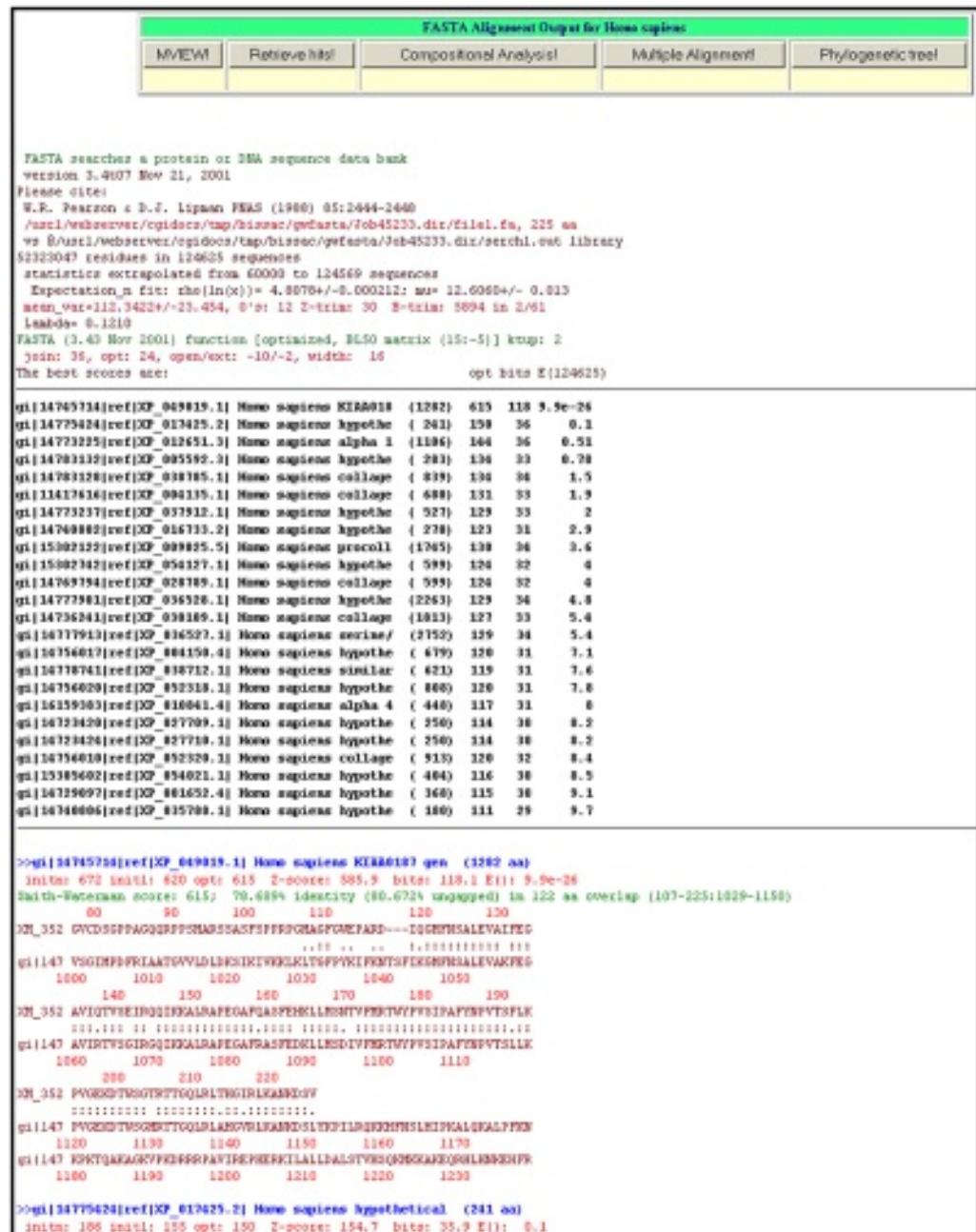


Fig. 5. Typical report generated after a GWFasta search against sequence databases with the histogram output suppressed

GWFASTA uses one of the most widely used bioinformatics tools for aligning multiple sequences. ClustalW is a general-purpose multiple sequence alignment program for DNA or proteins (24). It produces biologically meaningful multiple sequence alignments of divergent sequences. It calculates the best match for the selected sequences, and lines them up so that the identities, similarities, and differences can be seen. Evo-

PASTA Alignment with MView

denominator is required which requires for integer division to have 16 bits allocated to store the dividend by identifying a property

卷之三

卷之三

Fig. 6. Graphical presentation of FASTA alignment between query and database sequences using the alignment viewer MView.

RESULT PAGE FOR GWFASTA SERVER

Results of FASTA Search Against Eukaryotic Proteome [Only annotated Genomes] Database

Extract Top Hits for Alignment!

GENOME	Score	Expect Value	FASTA Alignment	Extract
<i>Arabidopsis thaliana</i>	294.9	1.6e-09	FASTA Output	Extract Sequence!
<i>Cenorhabditis elegans</i>	162.7	0.037	FASTA Output	Extract Sequence!
<i>Drosophila melanogaster</i>	364.6	2.1e-13	FASTA Output	Extract Sequence!
<i>Fugu rubripes</i>	234.2	3.8e-06	FASTA Output	Extract Sequence!
<i> Homo sapiens</i>	585.9	9.9e-26	FASTA Output	Extract Sequence!
<i>Plasmodium falciparum</i>	+	+	+	NO HITS!
<i>Saccharomyces cerevisiae</i>	362.6	2.7e-13	FASTA Output	Extract Sequence!

Fig. 7. Proteome search report generated by GWFasta server on searching XM_352357 human sequence against eukaryotic proteome databases.

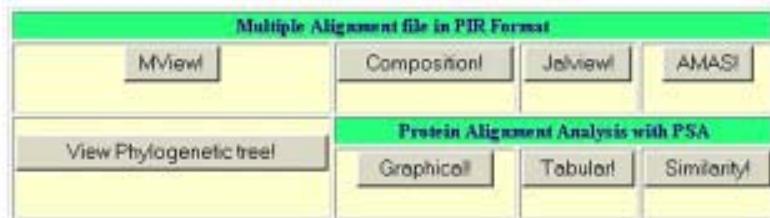
lutionary relationships can be seen by viewing cladograms or phyograms. ClustalW default output from the GWFasta server is in PIR format (Fig. 8).

10.6.2. Visualization of FASTA and ClustalW Alignments

MView is a tool for converting the results of a sequence database search (BLAST, FASTA, and so on) into the form of a colored multiple alignment of hits stacked against the query (25). Alternatively, an existing multiple alignment (MSF, PIR, CLUSTAL, and so on) can be processed. Colored alignment can help in detecting conserved regions and groups of residues having common properties with considerable ease. MView is integrated into the GWFasta server at two points—after FASTA searches (Fig. 6) and after MSA (Fig. 9). It thereby helps in providing an objective, beautiful, and useful report to the user.

10.6.3. Editing of Alignment

Jalview, an alignment viewer and editor program, is integrated with GWFasta to edit and manipulate MSA generated from hits reported by FASTA using ClustalW. Jalview is a Java-based tool that combines display speed and consensus color schemes with easy access to public databases using CORBA and CGI. The program colors residues by the physico-chemical properties of amino acids, similarity to consensus sequence, hydrophobicity, or secondary structure. It performs additional pairwise alignment using the Smith-Waterman algorithm, and can send colored postscripts of output by e-mail. Jalview is developed by M Clamp (Sanger Center, UK).



>P1;gi|7294027|gb|AAF49383.1|

```

MADDAGQDKKQHRRARQSGVKADIKKLLKAKQD-SMKEPELTARQRNPKAF
AINSAQRAKEPNFRKKEDLTAKKQMPVVDQTP-MVPPPVVIAVVGGPPVVG
KTTLIPDILIKSFTIRTMVTDIKGPIITIVTSKBRITLLECH-NDVNSMIDV
AKCADLVLLLCDASYGPMEIPEFNLNCQVHGMVKINGVLTMLDMIKKPK
QLRKRSQKLESHRFVTEVYDGAKLFLSGLLMGETLRLNEVHNLRGPFLSVHK
FRP1QWGRGAMSYLLVDRIEDVINTIDRVRED-PKCDREVVLVGTVRGVPLK
QE--HEVIVIAGLGDDARIDELNVIPDPCPLP-----
-----GTEKKRSLLERKLLYTAPHSVGGGIVYDKDAVYIELQGSHSH
KEQEQTAKA--AEQ----A--ELVNHKLIDHKATIDEQMEQGEFRLPSDA
KPIKSSDPRM-----DQDDEE

```

Fig. 8. Multiple sequence alignment (MSA) output in PIR format of top five similar sequences to XM_352357 human sequence after search against eukaryotic proteome databases.

CLUSTAL W Alignment With MView

Identities computed with respect to: (1) gi|7294027|gb|AAF49383.1|
 Maximum sequences to show: 10
 Colored by: identity + propensity

1	gi 7294027 gb AAF49383.1	100.0%
2	gi 7523707 gb AAF53146.1 AC011	37.0%
3	gi 14745714 ref XP_049019.1	40.6%
4	gi 13138699	26.7%
5	gi 6325839 ref NP_015107.1	35.0%
6	Query	16.8%

```

MADDAGQDKKQHRRARQSGVKADIKKLLKAKQD-SMKEPELTARQRNPKAF
AINSAQRAKEPNFRKKEDLTAKKQMPVVDQTP-MVPPPVVIAVVGGPPVVG
KTTLIPDILIKSFTIRTMVTDIKGPIITIVTSKBRITLLECH-NDVNSMIDV
AKCADLVLLLCDASYGPMEIPEFNLNCQVHGMVKINGVLTMLDMIKKPK
QLRKRSQKLESHRFVTEVYDGAKLFLSGLLMGETLRLNEVHNLRGPFLSVHK
FRP1QWGRGAMSYLLVDRIEDVINTIDRVRED-PKCDREVVLVGTVRGVPLK
QE--HEVIVIAGLGDDARIDELNVIPDPCPLP-----
-----GTEKKRSLLERKLLYTAPHSVGGGIVYDKDAVYIELQGSHSH
KEQEQTAKA--AEQ----A--ELVNHKLIDHKATIDEQMEQGEFRLPSDA
KPIKSSDPRM-----DQDDEE

```

MView 1.41, Copyright © Heid-P. Böker, 1993-2001.

Fig. 9. Visualization of multiple sequence alignment (MSA) using MView. The format chosen is the “new” format of the alignment viewer program.

10.6.4. Evolutionary and Phylogenetic Analysis

Any meaningful gene functional studies along the phylogenetic spectra from invertebrates to mammals require a set of orthologs that can be linked to one another. A genome-wide similarity search against genomic or proteomic sequence databases provides users a set of similar sequences that can be used to construct MSA. MSA of similar sequences provides an opportunity to study the phylogenetic divergence by

Phylogenetic tree

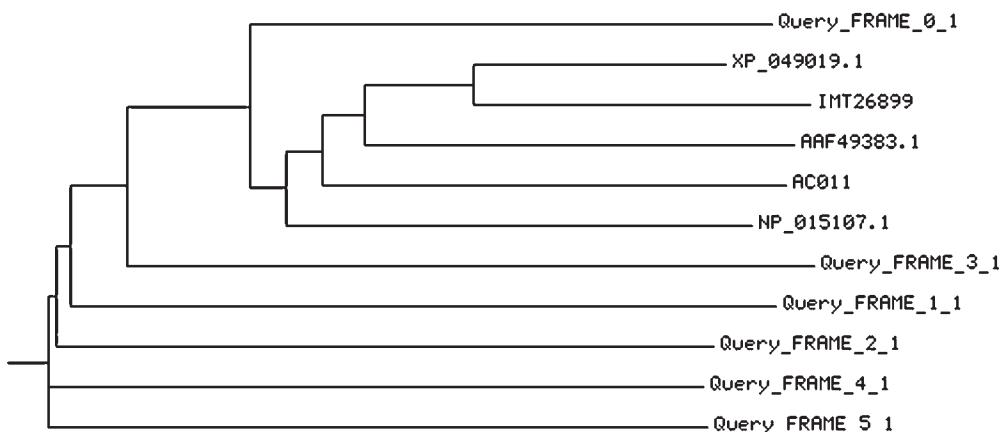


Fig. 10. Graphical presentation by Phydendron of a phylogenetic tree derived from an MSA of similar sequences from a proteome search of human sequence XM_352357.

observing the number or type of changes in the aligned residues. Phylogenetic analyses of aligned sequences allows for a better understanding of molecular relationships between sequences (26). A phylogenetic tree helps to visually appreciate the degree of divergence from each other. A Web server, Phydendron, is integrated with our server and allows the creation and visualization of a phylogenetic tree (Fig. 10). It provides several options, including the type of tree (e.g., cladogram, phenogram, swoopogram) and type of generated image (GIF, PostScript, or PDF). The orientation of the image can also be changed according to the user's choice.

10.6.5. Property Plots of Aligned or Single Sequences Using PSA

The presentation of physical properties of amino acids graphically along its primary structure (e.g., hydropathy plot, polarity plot, and so on) can help in understanding the function of a protein. PSAweb (27) analyzes the amino acid sequence and MSA of proteins generated by GWFasta to allow the users to plot amino acid properties along protein primary structure through these plots. Users can select up to 4 properties at a time, out of 36 available on the server, to plot in a single window or in multiple windows. Choices include (1) length of window used for averaging; (2) moment method to be applied before averaging; and (3) method used for averaging (mean over window, hat over window, median, data-sieve, and so on). It provides the following options on a single protein sequence: (a) computation of the physical properties of amino acids; (b) presentation of the properties along the primary structure of a protein by 2-D graph; (c) presentation of the amino acid properties along the primary structure as a table; and (d) highlight a residue or group of residues in the sequence that exhibits a specific function (e.g., hydrophobic residues, polar residues, and small residues).

The tool also allows plotting amino acid properties of protein sequences along sequence alignment, creates a separate graph corresponding to each sequence in the alignment, and presents all the user-selected properties in this graph (Fig. 11). Choices include (a) computing and presentation of the overall property of each position in the

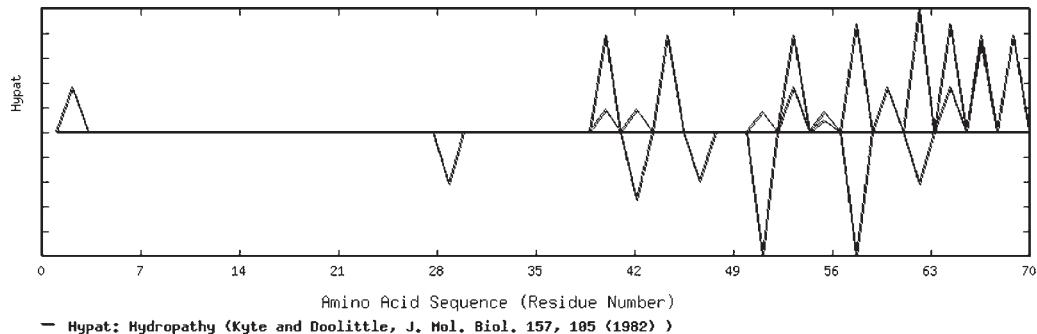


Fig. 11. Hydropathic plot (Kyte and Doolittle method, 1982) generated using PSAweb.

Position Specific Analysis

	%score	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Gaps
MM	67.3	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	9
AA	67.3	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9
DA	65.5	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9
DD	67.3	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9
M AE	50.9	1	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	8
E GLM	38.2	0	0	0	1	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	7
A QME	38.2	1	0	0	1	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	7
K DPQ	38.2	0	0	1	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	7
D KSS	40.0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	2	0	0	0	0	7
Q RHN	38.2	0	0	0	0	0	0	1	0	0	0	0	1	0	1	1	0	0	0	0	0	7
K KRK	43.6	0	0	0	0	0	0	0	0	3	0	0	0	0	0	1	0	0	0	0	0	7
K QSQ	40.0	0	0	0	0	0	0	0	0	1	0	0	0	0	2	0	1	0	0	0	0	7
H HHH	49.1	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	7
R RRR	49.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	7
K ATK	40.0	1	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	1	0	0	0	7
K RPA	38.2	1	0	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	7
N QKK	40.0	0	0	0	0	0	0	0	0	2	0	0	1	0	1	0	0	0	0	0	0	7
S SSE	43.6	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	7
G GGK	43.6	0	0	0	0	0	3	0	0	1	0	0	0	0	0	0	0	0	0	0	0	7
P VPN	40.0	0	0	0	0	0	0	0	0	0	0	0	1	2	0	0	0	0	1	0	0	7
K KTT	41.8	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	2	0	0	0	7
A AAA	49.1	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7
A DRK	38.2	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	7
K KKK	49.1	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	7
K KKK	49.1	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	7
K KSL	40.0	0	0	0	0	0	0	0	0	2	1	0	0	0	0	0	1	0	0	0	0	7
K LEH	38.2	0	0	0	1	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	7
R KLT	38.2	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1	0	1	0	0	7
L ADQ	38.2	1	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	7
L KKG	40.0	0	0	0	0	0	1	0	0	2	1	0	0	0	0	0	0	0	0	0	0	7
n vvh	40.0	n	n	n	n	n	n	n	1	n	2	n	n	n	n	1	n	n	n	n	n	7

Fig. 12. Position Specific Analysis of a multiple sequence alignment (MSA) of similar sequences to XM_352357 human sequence after search against eukaryotic proteome databases.

alignment; (b) highlight the conserved residues in the alignment; (c) highlight specific residue in the alignment that exhibits a specific function; (d) compute a position-specific score matrix (Fig. 12); and (e) similarity among the sequences present in the alignment. PSA also allows users to identify the protein sequence in MSA that has the highest similarity with most other sequences of a set, and categorize it as the representative protein of that set.

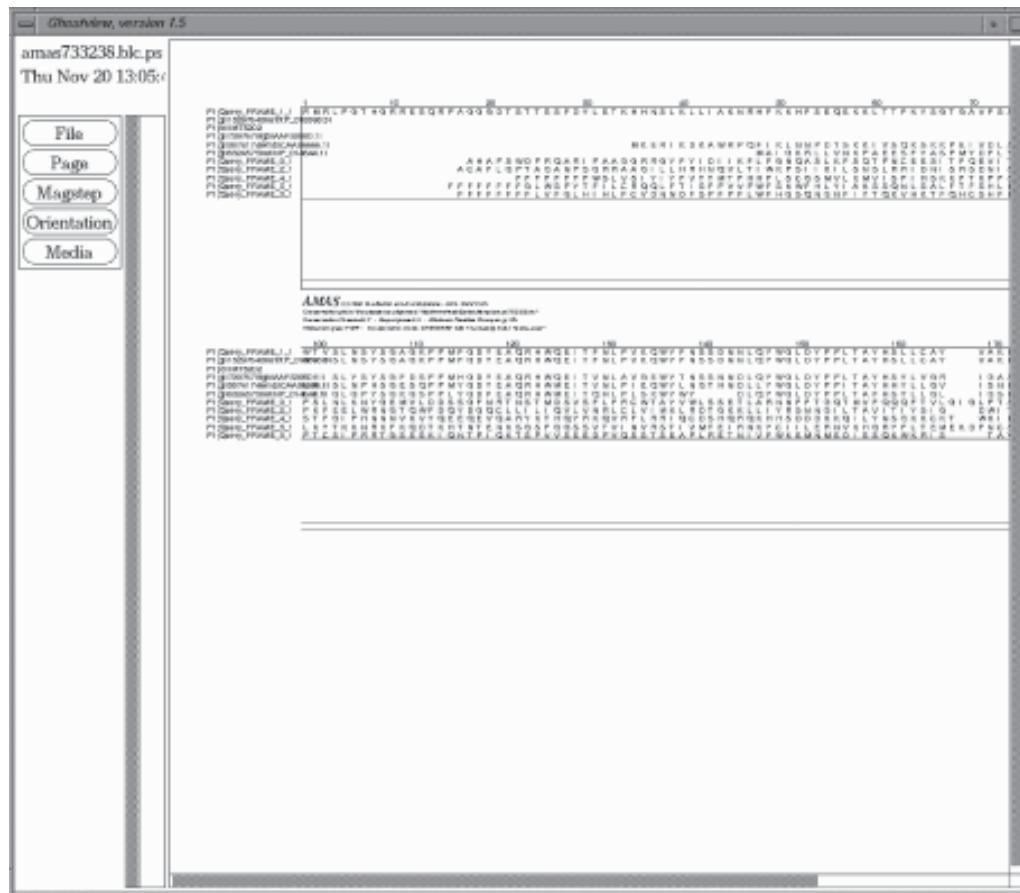


Fig. 13. A postscript image of analysis of a multiple sequence alignment (MSA) generated by AMAS using Alsprint program.

10.6.6. Analysis of Multiple Sequence Alignment Using AMAS

Analyzing an MSA can also identify substitutions of structural and functional importance. Another MSA analyzer program, AMAS (28), has a strategy based on a flexible set-based description of amino acid properties that defines the conservation between any groups of amino acids. It performs a systematic characterization of the physico-chemical properties seen at each position in a multiple protein sequence alignment. The sequences in the alignment are in subgroups based on sequence similarity, functional, evolutionary, or other criteria. The comparison of all pairs of subgroups highlights positions that confer the subgroup's unique features. It utilizes the Alsprint (29) program to present graphical interpretation of its results (Fig. 13). The server simplifies the analysis of multiple sequence data by condensing the mass of information present, and thus allows the rapid identification of substitutions of importance.

11. Implementation of GWFasta

The GWFasta Web-based server is available at <http://www.imtech.res.in/raghava/gwfasta/> (Fig. 14). It is installed on a SUN server (420E) with UNIX (SOLARIS) envi-



Biointeractive Centre
IMTECH Chandigarh, INDIA

[\[ABOUT GWFASTA\]](#) [\[HELP\]](#) [\[REFERENCES\]](#) [\[ACKNOWLEDGEMENT\]](#) [\[RETRIEVE RESULTS\]](#)

 [Merot State of India](#)

GWFASTA: Genome Wide FASTA Search

GWFASTA is a webserver intended for the objective of aiding researchers in their quest for sequence similarity search analysis on a genome wide scale.

Sequence alignments provide a powerful way to compare novel sequences with previously characterized genes. Both functional evolutionary information can be inferred from well designed queries and alignments. Pairs of protein and DNA sequences can be aligned using FASTA. It looks for matching sequence patterns or words, called k-tuples, and then attempts to build a local alignment based on these word matches. This is unlike a much slower but more sensitive Smith-Waterman algorithm developed for searching before FASTA. FASTA is comparable in reliability and in algorithm to BLAST but is more sensitive to searches for sequence families and should be preferred more for DNA searches instead of BLAST.

FASTA compares an input DNA or protein sequence to all sequences in a target sequence database, and then reports the best n sequences and local alignments of these matched sequence with the input sequence. The input sequence and database are usually in FASTA format.

If you use GWFASTA please cite the following paper:
Issac, B. and Raghava, G. P. S. (2002) GWFASTA: A server for FASTA search in Eukaryotic and Microbial genomes. *EvoTech* (3): 549-556.

SEARCH AGAINST

Standard Protein Databases	Standard Nucleotide Databases
Prokaryotic Proteome Databases	Prokaryotic Genomic Databases
Eukaryotic Proteome Databases	Eukaryotic Genomic Databases

FASTA AGAINST Proteome
[Only Annotated Genomes]

Prokaryote	Eukaryote
----------------------------	---------------------------

FASTA AGAINST Genomes

Prokaryote	Eukaryote
----------------------------	---------------------------

Fig. 14. Home Page of GWFASTA server available at <http://www.imtech.res.in/raghava/gwfasta/>.

ronment and has four shared 450-MHz UltraSparc II CPUs with 4-MB L2 Cache and 2-GB (8×256 MB) RAM. The >300-GB hard disk ensures that the server has enough memory and space to handle huge genomic sequences from various organisms. A redundant power supply and PERL v5.03-based CGI scripting ensures that the server is capable of handling a heavy load of queries from users. North American and European users can access a mirror site of GWFasta at <http://bioinformatics.uams.edu/raghava/gwfasta/>.

12. Conclusion

The FASTA packages provide a flexible set of sequence-comparison programs that are particularly valuable because of their accurate statistical estimates and high-quality alignments. Traditionally, sequence similarity searches have sought to ask one question: "Is my query sequence homologous to anything in the database?" Homology can be reliably inferred from statistically significant similarity. With the increasing number of genomic and proteomic databases available as a result of mega-genome sequencing projects running concurrently worldwide, similarity searching has arrived as a valuable tool for annotating new genomic DNA. Our aim by developing a tool like GWFasta is to increase the number of interfaces available to the users to analyze their sequences from a single platform. Multiintegrated servers such as GWFasta provide ample opportunity for serious researchers to go for subjective and intense sequence analysis.

Appendix I. Steps in FASTA Algorithm

- Step 1. Identify regions shared by the two sequences with the highest density of identities (ktup = 1) or pairs of identities (ktup = 2).
- Step 2. Rescan the 10 regions with the highest density of identities using the PAM250 matrix. Trim the ends of the region to include only those residues contributing to the highest score. Each region is a partial alignment without gaps.
- Step 3. If there are several initial regions with scores greater than the cutoff value, check whether the trimmed initial regions can be joined to form an approximate alignment with gaps. Calculate a similarity score that is the sum of the joined initial regions minus a penalty (usually 20) for each gap. This initial similarity score is used to rank the library sequences. The score of the single best initial region found in step 2 is reported.
- Step 4. Construct a Needleman-Wunsch-Sellers (NWS) optimal alignment of the query sequence and the library sequence, considering only those residues that lie in a band 32 residues wide centered on the best initial region found in step 2. FASTA reports this score as the optimized score.

References

1. Venter, J. C., Adams, M. D., Myers, E. W., et al. (2001) The sequence of the human genome. *Science* **291**, 1304–1351.
2. Lander, E. S., Linton, L. M., Birren, B., et al. (2001) Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.
3. Waterson, R. H., Lindblad-Toh, K., Birney, E., et al. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562.
4. Miyazaki, A., Sugawara, H., Gojobori, T., and Tateno, Y. (2003) DNA DataBank of Japan (DDBJ) in XML. *Nucleic Acids Res.* **31**, 13–16.

Appendix II

Web-Based Services Available on FASTA

Sl. No.	Description and link	Remarks
1.	EBI FASTA European Bioinformatics Institute http://www.ebi.ac.uk/fasta33	*****
2.	EBI Proteomes & Genomes FASTA3 European Bioinformatics Institute http://www.ebi.ac.uk/fasta33/genomes.html	*****
3.	FASTA Programs of U. of Virginia University of Virginia http://fasta.bioch.virginia.edu/	*****
4.	FASTA Sequence similarity search Genome Net http://fasta.genome.ad.jp/	*****
5.	FASTA: SEARCH AND ANALYSIS DDBJ Japan http://spiral.genes.nig.ac.jp/homology/fasta-e.shtml	*****
6.	FASTA Cambridge University http://www.bio.cam.ac.uk/cgi-bin/fasta3/fasta33.cgi.pl?conf=33	*****
7.	FASTA Cambridge University http://www.bio.cam.ac.uk/cgi-bin/fasta3/fasta3.pl	*****
8.	FASTA: Sequence database search (version 3) (W. Pearson) Pasteur Institute http://bioweb.pasteur.fr/seqanal/interfaces/fasta.html	*****
9.	FASTA/BLAST ANTHEPROT (ANalyse THE PROTeins) http://antheprot-pbil.ibcp.fr/fasta.html	*****
10.	MolbioWorkbench ANTHEPROT (ANalyse THE PROTeins) http://www.micro.bio.uni-giessen.de/LOCAL_SERVERS/fasta.html	*****

5. Manuel, A., Beaupain, D., Romeo, P. H., and Raich, N. (2000) Molecular characterization of a novel gene family (PHTF) conserved from *Drosophila* to mammals. *Genomics* **64**, 216–220.
6. Soliveri, J. A., Gomez, J., Bishai, W. R., and Chater, K. F. (2000) Multiple paralogous genes related to the *Streptomyces* coelicolor developmental regulatory gene *whiB* are present in *Streptomyces* and other actinomycetes. *Microbiology* **146**, 333–343.
7. Komeda, H. and Asano, Y. (2003) Genes for an alkaline D-stereospecific endopeptidase and its homolog are located in tandem on *Bacillus cereus* genome. *FEMS Microbiol Lett.* **228**, 1–9.
8. Gibbs, A. J. and McIntyre, G. A. (1970), The diagram, a method for comparing sequences. Its use with amino acid and nucleotide sequences. *Eur. J. Biochem.* **16**, 1–11.
9. Needleman, S. and Wunsch, C. (1970) A general method applicable to search for similarities in the amino acid sequences of two proteins. *J. Mol. Biol.* **48**, 444–453.

10. Smith, T. and Waterman, M. (1981) Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197.
11. Pearson, W. R. and Lipman, D. J. (1988) Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* **85**, 2444–2448.
12. Lipman, D. J. and Pearson, W. R. (1985) Rapid and sensitive protein similarity searches. *Science* **237**, 1435–1441.
13. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.
14. Altschul, S. F., Madden, T. L., Schaffer, A. A., et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402.
15. Pearson, W. (2000), Flexible sequence similarity searching with FASTA3 program package. “In *Bioinformatics Methods and Protocols*”, Misener, S., and Krawety, S. A. (eds.), Humana Press, Inc., Totowa, NJ, pp. 185–219.
16. Wilbur, W. J. and Lipman, D. J. (1983), Rapid similarity searches of nucleic acid and protein data banks. *Proc. Natl. Acad. Sci., USA* **80**, 726–730.
17. Pearson, W. R. (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol.* **183**, 63–98.
18. Anderson, I. and Brass, A. (1998), Searching DNA databases for similarities to DNA sequences: when is a match significant? *Bioinformatics* **14**, 349–356.
19. Pearson, W. R. (1995) Comparison of methods for searching protein sequence databases. *Protein Sci.* **4**, 1150–1160.
20. Pearson, W. R. (1996) Effective protein sequence comparison. *Methods Enzymol.* **266**, 227–258.
21. Pearson, W. R. (1998), Empirical statistical estimates for sequence similarity searches. *J. Mol. Biol.* **276**, 71–84.
22. Miller, W. (2000), Comparison of genomic DNA sequences: solved and unsolved problems. *Bioinformatics* **17**, 391–397.
23. Issac, B. and Raghava, G. P. S. (2002), GWFasta: a server for FASTA search in eukaryotic and microbial genomes. *BioTechniques* **33**, 548–556.
24. Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* **22**, 4673–4680.
25. Brown, N. P., Leroy, C., and Sander, C. (1998) MView: a Web-compatible database search or multiple alignment viewer. *Bioinformatics* **14**, 380–381.
26. Gogarten, J. P. and Olendzenski, L. (1999) Orthologs, paralogs and genome composition. *Curr. Opin. Genet. Dev.* **9**, 630–636.
27. Raghava, G. P. S. (2001), A graphical Web server for the analysis of protein sequences and alignment. *Biotech. Software and Internet Report.* **2**, 255–258.
28. Livingstone, C. D. and Barton, G. J. (1993) Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Comput. Appl. Biosci.* **9**, 745–756.
29. Barton, G. J. (1993) Alscript: a tool to format multiple sequence alignments. *Prot. Eng.* **6**, 37–40.

Protein Sequence Analysis and Domain Identification

Chris P. Ponting and Ewan Birney

1. Introduction

The fundamental unit of protein structure is the domain, defined as a region or regions of a polypeptide that folds independently and possesses a hydrophobic core with a hydrophilic exterior (see Note 1). Domains, particularly those with enzymatic activities, may possess functions independently of whether they are present in isolation or are part of a larger multidomain protein. Other domains confer regulatory and specificity properties to multidomain proteins, usually via the provision of binding sites. Because the majority of eukaryotic proteins, and a large number of eubacterial and archaeal proteins, are multidomain in character, the determination of the structures and functions of these proteins requires detailed consideration of their domain architectures.

Experience gathered from decades of structural and molecular biology shows that those domains that possess substantial similarities in sequence (>30% identity) also possess a common fold and usually possess similarities in function (1,2). These domains are thought to be homologous—that is, they are derived from a common ancestral gene following its duplication. The amino acid sequences of these domains, once identical at the time of gene duplication, have diverged increasingly during their evolution, yet have retained those amino acid properties that are essential for their proper domain folding, structure, and function. Similarities between homologous domains' sequences may have eroded sufficiently that their common evolutionary heritage is not apparent simply from their pairwise comparison. In these cases, the existence of a homology relationship may be inferred either from knowledge of their functions and tertiary structures, and/or application of sequence analysis methods that make use of multiple alignments. Many methods are now available that are used to predict homology relationships. In this chapter, we survey their use and provide indicators on how to realize the predictive potential of sequence analysis by database searching.

2. Materials

All that is needed to perform sequence analysis is a computer with access to the Internet. In **Table 1**, we list a variety of applications that can be accessed via the Worldwide Web (WWW). In general, the user cuts and pastes a query protein sequence into a form, sets a variety of parameters, and then initiates the application. Users with nucleotide sequences should first use gene-prediction algorithms to generate open reading

frame (ORF) predictions (a list of gene prediction tools is given at <http://www.ncbi.nlm.nih.gov/gene/> and these are reviewed in [ref. 3](#)).

Detailed analysis of sequences is best performed using tools compiled and running locally. Sequence databases, alignment programs, and database searching tools (*see Table 1*) are all available free via anonymous file transfer protocol (ftp) (log-in as username “anonymous” and fill in your email address as the password). Instructions on downloading and compiling files are usually provided at the ftp site. In general, programs are provided as “platform-specific” versions that can be installed on different computer systems.

The majority of the applications described here compare protein and not nucleotide sequences. Consequently, and also if computer disk space is limited, it is recommended to download protein sequence databases before nucleotide databases ([Note 2](#)). For historical reasons, different alignment programs or editors use different file formats, such as ClustalW (ALN), multiple sequence format (MSF), and Pearson (FASTA) formats. Format conversion is provided by, among others, the ClustalW, ClustalX, SeaView, and ReadSeq programs.

3. Methods

Later in this section we suggest a recipe for the analysis of a single protein sequence with respect to its domain architecture. This recipe makes use of various programs that provide different statistical indicators of the significance of predictions. Many of these indicators are derived from nontrivial analysis of score distributions. Anyone using such statistics to derive homology arguments is urged to understand their derivation and the limitations of their use.

3.1. Database Searching: Single or Multiple Sequence Queries

There are many methods that find domains in sequence databases (reviewed in [refs. 4,5](#)), of which some make use of distinctive motifs containing relatively few amino acids ([6](#)) and others make use of the degrees of amino acid conservation present throughout a domain structure, as represented in a multiple alignment. Many of the latter methods are based on similar algorithms. Differences among the methods are threefold and result from (1) assumptions made in the method; (2) the manner by which a domain is represented numerically in the method; and (3) parameters indicating the statistical significances of matches reported by the method.

An important consideration is whether a method uses a single sequence or multiple sequences as input: in general, multiple-sequence methods, especially when multiple alignments are constructed well, significantly outperform single-sequence methods. However, at the onset of an analysis, a researcher usually is aware of only a handful of homologs. As a result, single-sequence searches are essential to find sufficient examples of a domain to allow the use of more potent multiple-sequence methods.

3.2. Assumptions Made by the Methods

All methods assume that conservation patterns of different positions in a sequence or alignment are not correlated (principally this is for algorithmic reasons, because the consideration of such correlations would otherwise result in unreasonably long computation times). Whereas some methods allow gaps in matched regions of the domain

Table 1
Single Sequence Analysis Via the Worldwide Web

NCBI BLAST	(Sequence similarity)	http://www.ncbi.nlm.nih.gov/blast/index.shtml
WU-BLAST	(Sequence similarity)	http://www.ebi.ac.uk/blast2/
FASTA3	(Sequence similarity)	http://www.ebi.ac.uk/fasta33/
SignalP	(Signal peptides)	http://www.cbs.dtu.dk/services/SignalP-2.0/
COILS	(Coiled-coils)	http://www.ch.embnet.org/software/COILS_form.html
NetOGlyc	(O-glycosylation)	http://www.cbs.dtu.dk/services/NetOGlyc-3.0/
PSORT	(Protein localisation)	http://psort.nibb.ac.jp/form.html
JPred	(Secondary structure prediction)	http://www.compbio.dundee.ac.uk/~www-jpred/
PHD	(Secondary structure prediction)	http://www.ebi.ac.uk/~rost/predictprotein/
TMAP	(Transmembrane regions)	http://www.mbb.ki.se/tmap/index.html
TMpred	(Transmembrane regions)	http://www.ch.embnet.org/software/TMPRED_form.html
TMHMM	(Transmembrane regions)	http://www.cbs.dtu.dk/services/TMHMM/
Wise2	(Protein-Genome alignment)	http://www.ebi.ac.uk/Wise2/
Multiple sequence alignments via WWW		
CLUSTALW		http://www2.ebi.ac.uk/clustalw/
T-Coffee		http://www.ch.embnet.org/software/TCoffee.html
PCMA		ftp://iole.swmed.edu/pub/PCMA/
MultAlin		http://prodes.toulouse.inra.fr/multalin/multalin.html
SAM		http://www.cse.ucsc.edu/research/compbio/short_form.html
Sequence Databases via ftp		
Protein:		
NCBI's nr database		ftp://ftp.ncbi.nlm.nih.gov/blast/db/nr.tar.gz
SwissProt		ftp://ftp.expasy.ch/databases/swiss-prot/
TrEMBL		http://www.ebi.ac.uk/trembl/FTP/ftp.html
NRDB90		http://www.ebi.ac.uk/~holm/nrdb90/
Nucleotide:		
EMBL		ftp://ftp.ebi.ac.uk/pub/databases/emb/
DDBJ		http://www.ddbj.nig.ac.jp/anoftp-e.html
NCBI's nt database		ftp://ftp.ncbi.nlm.nih.gov/blast/db/nt.tar.gz

(continued)

Table 1 (Continued)
Single Sequence Analysis Via the Worldwide Web

Multiple Sequence Databases via WWW	Single Sequence Databases via WWW
BLOCKS	http://www.blocks.fhcrc.org/
InterPro	http://www.ebi.ac.uk/interpro/
Pfam	http://www.sanger.ac.uk/Software/Pfam/ or http://pfam.wustl.edu/
PRINTS	http://www.bioinf.man.ac.uk/dbrowser/PRINTS/
ProDom	http://protein.toulouse.inra.fr/prodom/current/html/home.php
ProfileScan	http://hits.isb-sib.ch/cgi-bin/PFSCAN
Prosite	http://us.expasy.org/prosite/
SMART	http://smart.embl.de or http://smart.ox.ac.uk
Database searching Programs via the Web or ftp	
BLAST	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/
HMMER	ftp://ftp.genetics.wustl.edu/pub/eddy/hmmr/
THOR	http://smart.ox.ac.uk/thor/
MoST	ftp://ftp.ncbi.nlm.nih.gov/pub/koonin/most/
Pfsearch	http://www.isrec.isb-sib.ch/ftp-server/pftools/
SAM	http://www.cse.ucsc.edu/research/compbio/sam.html
Search/FASTA	ftp://ftp.virginia.edu/pub/fasta/
Wise2	ftp://ftp.ebi.ac.uk/pub/software/unix/wise2/
Orthology/Paralogy Assignments via the Web	
Clusters of Orthologous Groups, COGs	http://www.ncbi.nlm.nih.gov/COG/
Eukaryotic gene orthologs	http://www.tigr.org/db/tgi/ego/
InParanoid Database	http://inparanoid.cgb.ki.se/
HomoloGene	http://www.ncbi.nlm.nih.gov/HomoloGene/
Putative orthology groups	http://www.cbil.upenn.edu/gene-family/
Other Programs available via ftp	
Belvu	(Alignment editing)
GDE	(Alignment editing)
SeaView	(Alignment editing)

Boxshade
CHRoma
MVIEW
CLUSTALW
CLUSTALX
Dotter
Prospero
MACAW
ReadSeq

(Alignment coloring)
(Alignment coloring)
(Alignment coloring)
(Multiple alignments)
(Multiple alignments)
(Dot-matrix program)
(Repeat detection)
(Multiple alignments)
(Format conversion)

http://www.ch.embnet.org/software/BOX_doc.html
<http://www.lg.ndirect.co.uk/chroma/>
<http://mathbio.nimr.mrc.ac.uk/~nbrown/mview/>
<ftp://ftp.ebi.ac.uk/pub/software/unix/clustalw/>
<ftp://ftp.ebi.ac.uk/pub/software/unix/clustalx/>
<ftp://ftp.cgb.ki.se/pub/esr/dotter/>
<http://www.well.ox.ac.uk/ariadne/>
<ftp://ftp.ncbi.nlm.nih.gov/pub/macaw/>
<ftp://ftp.bio.indiana.edu/molbio/readseq/>

alignment, others do not. Disallowing gaps simplifies the computation and also allows classical analytical statistics to be used to evaluate results (7). Gapped methods almost invariably rely on dynamic programming (8) methods or approximations to them. As might be expected, computation times increase significantly, and in addition the analytical statistical theory of Karlin and Altschul breaks down (**Note 3**). In these cases, statistical significance has to be estimated empirically or by using a Bayesian framework to evaluate the significance of the match (**Note 4**).

3.3. Derivation of the Parameters for the Method

Single-sequence search methods need to estimate the rates by which sequences have mutated during their evolution. These are usually provided by an amino acid substitution matrix (**Note 5**) and also, when using gapped alignments, a penalty for gaps (**Note 6**). Multiple-sequence methods need to represent numerically the patterns of conservation in a multiple alignment. These representations of amino acid distributions at each position in the alignment are often called “profiles.” A probabilistic interpretation of these profiles in the hidden Markov model (HMM) is called a profile-HMM. This does not change the algorithm used, but does change the statistical interpretation of the algorithm (**Note 7**). As with single-sequence searches, some multiple-sequence methods allow gaps in the alignment of the domain model with each sequence (**Note 8**).

3.4. Statistical Results

Database searching methods report scores for the comparison of a target sequence with a domain model. Usually methods provide two scores: one that is not an indicator of statistical significance and one that is. The statistical interpretation is an estimate of the likelihood that a sequence with that score is not related to the query sequence or model, and has matched “by chance.” It should be noted that such statistics are algorithm-specific, and are not necessarily related to the real biological significance of the hit: many algorithms detect only a small percentage of true homologs (true positives) with statistical significance. The scores of undetected homologs (false negatives) lie within the distribution of scores for unrelated proteins (true negatives). There is always a finite probability that scores for some unrelated proteins (false positives) may be greater than expected. Consequently, sequence analysts should always consider the biological contexts of possible false positives or true negatives, and also consider the results of complementary methods.

Statistical results derived from classical (frequentist) approaches provide some estimate of the probability that a sequence picked at random could have produced a score equal to, or greater than, the score of a real database sequence, i.e., $P(\text{score} > X)$. This probability estimate is of little use, since it takes no account of the size of the database searched. Statistics that do, and are commonly reported in database searches, are: (1) expect values (or E values), which represent the number of sequences with scores equal to X , or greater, expected absolutely by chance, which is simply $P(\text{score} > X) \times N$, where N is the number of sequences in the database; and (2) P values, which represent probabilities that, given a database of a particular size, random sequences score higher than X . This is not the same as $P(\text{score} > X)$!

An E value of less than 0.5 indicates possible significance of the hit, whereas an E value of less than 0.01 is likely to represent a homologous relationship. This does not

mean to say that all hits with E values equal to 0.01 represent real homologs, only that there is 99% confidence that the score has not been arrived at simply by chance. Do not, however, have 100% confidence in your algorithm. Be aware that some algorithms (including PSI-BLAST) appear to underestimate E values by at least one order of magnitude.

3.5. A Recipe for In-Depth Analysis of a Sequence

We suggest the following five-step protocol for sequence analysis. Because the requirements for one search often are different from those for another, this represents one of several possible approaches.

3.5.1. Step 1: Motifs, Patterns, and Profile Scans

Table 1 contains the addresses of several WWW sites that allow the prediction of a variety of molecular features from protein sequence information. These include coiled coils (9), transmembrane helices, protein localization, signal peptides, and glycosylation sites. Care should be taken in the interpretation of results from these methods: for example, predicted glycosylation sites in cytoplasmic proteins and kinase-mediated phosphorylation sites in extracellular proteins are both unlikely to be of biological relevance. In addition, these methods do not predict homology relationships.

Other Web sites provide easy comparisons of a user's sequence with large numbers of domain or motif (**Note 1**) alignments. If your sequence contains one or more domains that are represented among collections of multiple alignments, there is a fair chance that they shall be detected by either InterPro or Profilescan. These resources allow domains to be detected using several domain databases, such as Pfam, SMART, and Prosite profiles, at once. The SMART Web sites allow simultaneous prediction of domains, repeats, signal peptides, and transmembrane and compositionally-biased regions.

Use of these servers in combination is the most rapid way to arrive at functional prediction arguments. Successful domain, coiled-coil, or transmembrane helix prediction using these methods also serves to reduce the "searchspace": it is valid to concentrate subsequent searches only on those sequence regions that are not assigned with significant statistics by these methods.

3.5.2. Step 2: Pairwise Methods

A powerful and popular way to assign domains within a query sequence is via Basic Local Alignment Search Tool (BLAST) searches. Early versions of BLAST (10) provided significance estimates for ungapped pairwise alignments. More recent versions have provided two improvements: pairwise alignments that are gapped, and iterative searches derived from multiple alignments (11). BLAST searches may be initiated via Web servers (**Table 1**) or locally, using code contained in the National Center for Biotechnology Information (NCBI) toolbox (ftp://ftp.ncbi.nlm.nih.gov/toolbox/ncbi_tools). The BLAST family of programs includes: BLASTP, which compares a protein sequence with a protein database; BLASTN, which compares a nucleotide sequence with a nucleotide sequence; BLASTX, which compares a nucleotide sequence (in all reading frames) with a protein database; TBLASTN, which compares a protein sequence with a nucleotide sequence (in all reading frames); and TBLASTX, which compares a nucleotide sequence (in all reading frames) with a nucleotide database (in

all reading frames). Current versions of BLAST provide significance estimates in terms of E values (see **Note 3**). An alternative to BLAST is SSEARCH (12), which also ascribes E values to candidate homologs.

Several considerations must be borne in mind during such analyses. The first is that several nonoverlapping hits to different portions of your query sequence may indicate that it contains several domains or repeats. If this appears to be the case, then scan your BLAST output for hits from sequences contained in the protein data bank (PDB) (a database of known protein structures) or else initiate a BLAST search against the set of PDB sequences (try using the SUPERFAMILY resource: <http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/hmm.html>). A significant hit against a domain of known tertiary structure immediately provides an accurate prediction of domain boundaries for its homolog contained in your query sequence (see **Note 9**). Similarly, a significant hit against the whole of a polypeptide with bona fide N- and C-termini also provides accurate domain limits. As previously, successful prediction of a domain allows reduction of the searchspace in subsequent searches.

Second, the existence of compositionally biased regions in proteins, such as those in collagens, might distort the reported significance estimates of your findings. Ensure that such low-complexity regions are masked in your query sequence by preprocessing with the program SEG (13), or use the default composition-based statistics option when using NCBI BLASTP (11,14). Preprocessing your sequence with transmembrane helix and coiled-coil prediction algorithms (see **Subheading 3.5.1.**) is also highly recommended (see **Note 10**). However, such algorithms provide imperfect predictions, and regions that are misassigned as not coiled coil may yield low E values against known coiled coil regions in database proteins such as myosins and kinesins; such alignments are highly unlikely to be biologically relevant.

Third, be vigilant against errors. These can be present in all quarters. Your query sequence or a database sequence may contain one or more frameshift errors, or they may be artificially truncated or extended beyond the proteins' normal termini. Hypothetical protein sequences deduced from eukaryotic genome projects frequently are inaccurate as a result of errors in intron-exon boundary assignments. Such errors often are apparent upon construction of a multiple alignment of homologs: the error-ridden sequence usually demonstrates substantial nonconservation in regions that are relatively well conserved among its homologs. A strategy to combat frameshifts and unknown intron-exon boundaries is to compare a profile derived from either a similar sequence or a multiple alignment of homologs, with the relevant DNA sequence using GENEWISE (15,16). Finally, errors are not restricted to sequence. Functions of proteins assigned by some are found frequently to be at odds with those found by others. Unfortunately, this results not only in errors in the database annotation for the protein in question, but also propagation of the error to all homologs whose functions are predicted on the basis of the original misannotation.

3.5.3. Step 3: Multiple Sequence Analysis

At some stage in any sequence analysis project, a multiple alignment must be constructed. Constructing an optimal multiple alignment consists of two stages: calculation of a preliminary alignment using programs such as ClustalW and T-COFFEE, and its subsequent refinement using manual alignment editors such as ClustalX, GDE, and/or SeaView (see **Table 1**). Manual intervention in otherwise automated procedures is

an unfortunate consequence of the “alignment problem”: algorithms are currently unable to generate alignments with high accuracies, when compared with alignments generated from superimposed tertiary structures of homologs. This is due to the exponential increase in memory and computing power required to consider mathematically optimal alignments (see **Note 11**). However, manual editing can greatly improve upon alignments if the following guidelines (2) are followed:

- (1) Minimize the number of gaps in an alignment. Consider that insertions and deletions occur predominantly on the exterior surface of proteins and in loops between secondary structures. Loops are usually highly variable in structure and in sequence, and so need not be well aligned. Ensure that those residues in a loop region between two alignment blocks that are not aligned are shunted (without gaps) to the alignment block that is either N-terminal or C-terminal to it. Single-residue insertions can occur within β -strands as “ β -bulges”. α -helices are intolerant of insertions or deletions. In a small minority of cases, an insertion/deletion position in a loop may accommodate a whole domain structure as a “domain insertion” (17).
- (2) Maximize conservation of “core” hydrophobic residues. Hydrophobic residues comprise the majority of proteins’ interiors and therefore are subject to relatively strong evolutionary pressures. Ensure that hydrophobic residues in homologs are aligned within all predicted secondary structures. Note that the periodicity of hydrophobic residues in β -strands differs from that in α -helices, which provides information essential for secondary-structure prediction. Domains that are rich in cysteine residues, such as several small extracellular domains and zinc fingers, are exceptions in that the bulk of their hydrophobic core is provided by disulphide bridges and metal ions, respectively. These domains are often characterized by poor amino acid conservation and few secondary structures relative to their size.
- (3) Maximize conservation of residues known through experiment to be important for function. However, consider that homologs can possess nonidentical functions: there are many instances, for example, of enzyme homologs that are enzymatically inactive.
- (4) Minimize the presence of Pro and Gly in the middle of secondary structures, except for edge β -strands, and maximize the presence of strings of charged residues in insertion/deletion positions.
- (5) Ensure that all subfamilies in an alignment are represented equally by including only one of each pair of sequences that are greater than, for example, 80% identical (the program BELVU provides such an option).
- (6) Take care in your choice of domain boundaries. The best evidence for these boundaries are experimentally determined N- and C-terminal residues of the protein, or detection of a homolog of known tertiary structure. Other indicators to be considered are limits of domains contiguous to the sequence of interest, degrees of sequence conservation (or lack of conservation) between closely related homologs, and the presence of low-complexity regions (normally present within interdomain, rather than intradomain, regions). Tandem repeats, as indicated using algorithms such as Dotter or Prospero (**Table 1**), may also assist in determining boundaries.

3.5.4. Step 4: Multiple Sequence Analysis

Once a multiple alignment has been created to your satisfaction, then its representation—either a profile or a profile-HMM (see **Note 7**)—can be compared with databases of protein or nucleotide sequences. There are several tools that have been written to search databases (**Table 1**). Of these, the authors are most familiar with two (Wise2 and HMMer) that exploit profile-HMMs (see **Note 7**), and a third, the position-specific

iterated version of BLAST (PSI-BLAST) (11). All of these methods allow database searches for “local” similarity over only a portion of the alignment. Of the three, only PSI-BLAST is unable to perform searches for “global” similarity over the entire alignment.

Wise2 is best employed to compare profile-HMMs or protein sequences with DNA sequences. Although often computationally expensive, the genewise program of Wise2 exploits evolutionary information, such as splice site consensus sequences, base compositions, amino acid substitution matrices, and protein multiple-sequence alignments, to predict homologs and their gene structures, within genomic sequence. The HMMER package (v2) (18) allows construction of profile-HMMs from protein multiple-sequence alignments, and their comparisons with amino acid sequence databases. It is the primary search engine of Pfam and SMART domain databases, and its profile-HMMs may be used by genewise. PSI-BLAST allows a protein-sequence database search with a query sequence, against which all candidate homologs are aligned, and provides iterative comparisons with profiles derived from this and subsequent alignments, until convergence. In PSI-BLAST, these profiles are termed *position-specific score matrices* (PSSMs). These may be stored by the user using the -C (checkpoint) option, and re-used in PSI-BLAST using the -R (restart) option (14). PSI-BLAST has proven to be highly sensitive in revealing subtle homologies and, therefore, is a method of choice in detecting domain homologs.

HMMER and PSI-BLAST estimate E values for each alignment. By default in the download version, PSI-BLAST employs an E value threshold of 5×10^{-3} . Database sequences aligned with E values less than this threshold are used to generate an alignment and a profile for the subsequent search round. We suggest equivalent E value thresholds of 0.05 or 0.10 for successive rounds of database searching using HMMER.

Single ungapped motifs (defined in Note 1), such as those encompassing active or binding sites, may also be compared with databases using the Motif Searching Tool, MoST (19). This also is an iterative method, and allows choice of candidate homologs on the basis of E values (option: -e ; recommendation -e 0.05), and closely similar sequences to be discarded (option: -i; recommendation -i 80 [%]).

It is imperative that a hypothesis that two sequences represent domain homologs be justified statistically. This may be achieved using an E value threshold and programs such as HMMER, MoST, and PSI-BLAST. However, the user should be aware that the inclusion of a false-positive scoring lower than the supplied E value threshold following one iteration negates the identification of putative homologs detected in subsequent iterations. It is preferable that all candidate homologs be related by multiple instances of low E values from BLAST (or Ssearch) queries, and essential that their sequences display similarities in their patterns of conservation across their multiple alignment. The program THoR (20) can be used to ensure self-consistency among candidate domain homologs arising from both HMMER and PSI-BLAST predictions.

Nonstatistical evidence may provide information consistent with a homology hypothesis. Experimental and contextual information may be used to predict homology for sequences that score at levels similar to the top true negative in searches. For example, a predicted domain may possess limits exactly coincident with the boundaries of a region intervening between two properly annotated domains; or, the predicted domain may be known experimentally to possess a molecular function equivalent with

the functions of the proteins used to derive the query profile. In contrast, nonstatistical evidence may provide evidence that two sequences are *not* homologs. A sequence may score highly against a query profile, yet the pattern of conservation and/or domain limits apparent from the alignment of its close homologs may differ substantially from the conservation pattern of the query profile's alignment (e.g., refs. 21–23).

3.5.5. Step 5: Hypothesis Generation

A set of sequences that are predicted to be homologous on the basis of statistical and experimental, or contextual, information may then be used as the basis for experimentally testable hypotheses concerning function. The more ancient evolutionary relationships of homologs are, the more likely it is that they perform dissimilar molecular or cellular roles. Orthologs are genes that have arisen in divergent lineages as a result of speciation (24,25). These have sequences that are more similar to each other than they are to those of other homologs, and are likely to perform essentially identical molecular and cellular functions. They are distinct from paralogs, which are genes that have arisen via intragenome gene duplication (24,25) and often possess distinct, albeit related, functions. Distinguishing orthologs from paralogs therefore is crucial for inferring whether a protein's function is similar to that of a homolog that has been subjected to experimentation (Table 1). All is not lost if the functions of all homologs in a multiple alignment are unknown. Strict conservation of amino acid types, such as polar amino acids in the active sites of enzymes, or cysteine or histidine residues in Zn²⁺-binding sites, may yet provide clues to molecular function (26).

4. Notes

1. Here we employ a terminology discussed elsewhere (2). In brief: *motifs* are short, conserved regions that are localized stretches of domain sequences (for example, “binding-site motifs”); *patterns* are assemblies of one or more motifs; *alignment blocks* are ungapped alignments usually representing a single secondary structure; and *domains* are conserved structural entities with distinctive secondary structures and (often) a hydrophobic core. Thus, these terms are not mutually exclusive and, in particular instances, are interchangeable.
2. Databases are best maintained locally in a simple (FASTA) format of concatenated sequences separated by a single header line containing a “>” symbol, accession codes, and relevant species, gene, and molecular information. Sequence databases are often redundant: they contain several copies of identical sequences. Databases that are less redundant than others are SwissProt and NRDB90. The SwissProt database is recommended for its extensive annotation (see <http://www.expasy.ch/sprot/sp-docu.html>), whereas GenPept is recommended for its daily updates (see <ftp://ncbi.nlm.nih.gov/genbank/daily-nc/>). All database entries can be accessed via the highly informative Entrez system (see <http://www.ncbi.nlm.nih.gov/Entrez/>), which provides links between sequence, literature, structure, taxonomy, disease, and other information.
3. Ungapped BLAST algorithms (7) and the motif-searching tool MoST (19) use analytically derived statistics from the theory of Karlin and Altschul (7). This predicts that the distribution of scores of ungapped alignments should follow an extreme value distribution (EVD) with parameters that can be used to provide $P(\text{score} > X)$, and hence E and P values. Extending this theory to gapped searches is problematic: it has been suggested that scores from gapped alignments also vary according to an EVD (11,27–28). In this approximation, significance statistics may be estimated from the fit of parameters to pre-

sumed random sequences. These are either precalculated using a comparison to artificial random databases (as with gapped BLAST methods), or generated “on the fly” from the distribution of scores from presumed nonhomologs in the same database search (the “noise”) (as with HMMER, FASTA, and SSEARCH).

4. Inference on the probability of a match being a random sequence or not can be derived using Bayesian methods for profile-HMMs (18,29). The profile-HMM provides a likelihood that the sequence was an example of the domain. This likelihood is compared to the likelihood that the sequence was generated by a random model. Profile-HMMs use a simple random model of amino acids drawn randomly from a set of real sequences. The score reported by the program is a log likelihood ratio (the base of the logarithm is 2, hence the name “bit scores”). This implies that bit scores between different profile-HMMs are comparable, in contrast to other statistics. For the likelihood ratio to provide a probability that the sequence came from either the domain HMM or the random model, estimates of the probability of these outcomes before examining the sequence have to be made (18,29).
5. The amino acid substitution matrix represents what is known of the results of protein evolution: in trusted multiple alignments, pairs of amino acids that are often aligned against each are given higher scores relative to pairs of amino acids that are seldom paired in alignments. These scores are in qualitative agreement with known physical and chemical properties of amino acids. Numbers in matrices are given as log-odd ratios of the observed frequency of a pairing, relative to the frequency expected by chance. Commonly used score matrices are the BLOSUM (30) and Gonnet (31) series.
6. Gap penalties are a necessity since mathematically optimal alignments would otherwise be dominated by excessively gapped regions. There are two main forms of a gap penalty. Linear gap penalties are those used when individual gaps are penalized separately. Affine gap penalties, which are the more common, are those used when each gapped region is penalized both for the initiation of the gap and also for its extension when additional gap positions are needed. Unfortunately there is no analytical theory that reliably calculates “optimal” gap penalties. Thus many programs allow the user complete freedom in setting penalties. However, a series of studies using both random databases and databases of known structure have provided some empirically derived optima (28,32). The optimal gap penalty is linked to the comparison matrix used. For the common BLOSUM62 matrix, gap penalties of 12 for initiation of a gap and 2 for its extension are considered reasonable.
7. A trivial representation of amino acid conservation in a multiple alignment is the set of amino acid frequencies at each position. However, this is of little practical use because (1) multiple alignments are often strongly biased towards the detection of a particular subfamily of homologs, and (2) this provides insufficient information concerning amino acid substitution to identify more distant homologs. The problem of sequence bias is addressed by calculating weights for each sequence in the multiple alignment using a derived phylogenetic tree (33): similar sequences are weighted lower than dissimilar ones. A model of evolution is used to supplement the observed amino acid frequencies in the multiple alignment. In standard profile-based methods (33,34) the same amino acid substitution matrices that are used in single-sequence searches are used. In HMM-profiles, the model of evolution is represented as an under-sampling problem. This provides a consistent mathematical framework that combines frequencies observed in the alignment and an evolutionary model. Two models have been found to be useful. The first is the equivalent to the standard profile techniques recast to fit the Bayesian framework. The second is a model based around Dirichlet mixtures of expected amino acid distributions in multiple alignments (35).

8. Gaps generated in aligning a domain model with a target sequence are penalized to prevent an unrealistic, over-gapped alignment. In profile methods, gap penalties are set arbitrarily, in a similar manner to single-sequence searches. However, penalties are multiplied by an ad hoc position-specific ratio indicating the tolerance of insertions or deletions at this position. HMM-profiles have a stronger theoretical basis for setting gap penalties, being the best fit to an assumed distribution from the observed gap length at each position. As with the raw frequency of amino acids, the observed gap-length distribution is considered to be an under-sampling of real allowed gaps at each position, and the observed frequencies are modified by a model of gap evolution.
9. In rare cases, the order of secondary structures in pairs of homologous domains is known to be circularly permuted (36). Identifying such homologs and determination of their domain limits is not a trivial task using conventional methods (37).
10. In a few cases (e.g., ref. 38) conserved domains are known to contain coiled coils.
11. Methods used for multiple-alignment algorithms are either (1) progressive alignment procedures where the multiple alignment proceeds up an evolutionary tree, fixing the alignment at each node (as in CLUSTALW, PILEUP), or (2) iterative training procedures that derive a model of the resulting multiple alignment (as in HMMER, MoST).

References

1. Doolittle, R. F. (1995) The multiplicity of domains in proteins. *Annu. Rev. Biochem.* **64**, 287–314.
2. Ponting, C. P. and Russell, R. B. (2002) The natural history of protein domains. *Annu. Rev. Biophys. Biomol. Struct.* **31**, 45–71.
3. Mathe, C., Sagot, M. F., Schiex, T., and Rouze, P. (2002) Current methods of gene prediction, their strengths and weaknesses. *Nucl. Acids Res.* **30**, 4103–4117.
4. Bork, P. and Gibson, T. J. (1996) Applying motif and profile searches. *Methods Enzymol.* **266**, 162–184.
5. Ponting, C. P., Schultz, J., Copley, R. R., Andrade, M. A., and Bork, P. (2000) Evolution of domain families. *Adv. Prot. Chem.* **54**, 185–244.
6. Jonassen, I. (2000) Discovering patterns conserved in sets of unaligned protein sequences. *Methods Mol. Biol.* **143**, 33–52.
7. Karlin, S. and Altschul, S. F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. USA* **87**, 2264–2268.
8. Pearson, W. R. and Miller, W. (1992) Dynamic programming algorithms for biological sequence comparison. *Methods Enzymol.* **210**, 575–601.
9. Lupas, A. (1996) Coiled coils: new structures and new functions. *Trends Biochem. Sci.* **21**, 375–382.
10. Altschul, S. F., Boguski, M. S., Gish, W., and Wootton, J. C. (1994) Issues in searching molecular sequence databases. *Nat. Genet.* **6**, 119–129.
11. Altschul, S. F., Madden, T. L., Schäffer, A. A., et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402.
12. Pearson, W. R. (1991) Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. *Genomics* **11**, 635–650.
13. Wootton, J. C. and Federhen, S. (1996) Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.* **266**, 554–571.
14. Schäffer, A. A., Wolf, Y. I., Ponting, C. P., Koonin, E. V., Aravind, L., and Altschul, S. F. (1999) IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics* **15**, 1000–1011.

15. Birney, E., Thompson, J. D., and Gibson, T. J. (1996) PairWise and SearchWise: finding the optimal alignment in a simultaneous comparison of a protein profile against all DNA translation frames. *Nucleic Acids Res.* **24**, 2730–2739.
16. Birney, E. and Durbin, R. (2000) Using GeneWise in the *Drosophila* annotation experiment. *Genome Res.* **10**, 547–548.
17. Russell, R. B. (1994) Domain insertion. *Protein Eng.* **7**, 1407–1410.
18. Eddy, S. R. (1996) Hidden Markov models. *Curr. Opin. Struct. Biol.* **6**, 361–365.
19. Tatusov, R. L., Altschul, S. F., and Koonin, E. V. (1994) Detection of conserved segments in proteins: iterative scanning or sequence databases with alignment blocks. *Proc. Natl. Acad. Sci. USA* **91**, 12,091–12,095.
20. Dickens, N. J. and Ponting, C. P. (2003) THoR: a tool for domain discovery and curation of multiple alignments. *Genome Biol.* **4**, R52.
21. Ponting, C. P., Bork, P., Schultz, J., and Aravind, L. (1999) No Sec7-homology domain in guanine-nucleotide-exchange factors that act on Ras and Rho. *Trends Biochem. Sci.* **24**, 177–178.
22. Barnes, M. R., Russell, R. B., Copley, R. R., et al. (1999) A lipid-binding domain in Wnt: a case of mistaken identity? *Current Biol.* **9**, R717–R718.
23. Copley, R. R., Ponting, C. P., and Bork, P. (1999) Phospholipases A2 and Wnts are unlikely to share a common ancestor. *Current Biol.* **9**, R718.
24. Fitch, W. M. (1970) Distinguishing homologues from analogous proteins. *Syst. Zool.* **19**, 99–113.
25. Fitch, W. M. (1995) Uses for evolutionary trees. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **349**, 93–102.
26. Ponting, C. P. (2001) Issues in predicting protein function from sequence. *Brief. Bioinform.* **2**, 19–29.
27. Mott, R. (1992) Maximum-likelihood estimation of the statistical distribution of Smith-Waterman local sequence similarity scores. *Bull. Math. Biol.* **54**, 59–75.
28. Altschul, S. F. and Gish, W. (1996) Local alignments statistics. *Methods Enzymol.* **266**, 460–480.
29. Krogh, A., Brown, M., Mian, I. S., Sjolander, K., and Haussler, D. (1994) Hidden Markov models in computational biology. Applications to protein modeling. *J. Mol. Biol.* **235**, 1501–1531.
30. Henikoff, S. and Henikoff, J. G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* **89**, 10915–10919.
31. Benner, S. A., Cohen, M. A., and Gonnet, G. H. (1994) Amino acid substitution during functionally constrained divergent evolution of protein sequences. *Protein Eng.* **7**, 1323–1332.
32. Brenner, S. E., Chothia, C., and Hubbard, T. J. P. (1998) Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl. Acad. Sci. USA* **95**, 6073–6078.
33. Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680.
34. Gribskov, M. and Veretnik, S. (1996) Identification of sequence pattern with profile analysis. *Methods Enzymol.* **266**, 198–212.
35. Karplus, K. (1995) Evaluating regularizers of estimating distributions of amino acids. *ISMB* **3**, 188–196.
36. Lindqvist, Y. and Schneider, G. (1997) Circular permutations of natural protein sequences: structural evidence. *Curr. Opin. Struct. Biol.* **7**, 422–427.

37. Uliel, S., Fliess, A., Amir, A., and Unger, R. (1999) A simple algorithm for detecting circular permutations in proteins. *Bioinformatics* **15**, 930–936.
38. Weimbs, T., Low, S. H., Chapin, S. J., Mostov, K. E., Bucher, P., and Hofmann, K. (1997) A conserved domain is present in different families of vesicular fusion proteins: a new superfamily. *Proc. Natl. Acad. Sci. USA* **94**, 3046–3051.

Mammalian Genes and Evolutionary Genomics

Leo Goodstadt and Chris P. Ponting

1. Introduction

The availability of mammalian genome sequences is allowing differentiation between regions relatively disregarded by selection and those that have been subject to stronger selection and hence will probably be of the greatest biological interest. These selective pressures are most often purifying, so as to conserve genetic sequences whose maintenance is critical to the fitness of the individuals of a species. The identification of well-conserved regions, whether within protein-coding genes or inter-genic regulatory elements, such as promoters and enhancers, requires comparison of more distantly related vertebrates, such as human and fish or chicken, which display sufficient underlying sequence divergence for the conservation to be evident.

Conservation is not the only indicator of functional importance. On rare occasions, functional importance may be indicated by a surfeit of sequence change, rather than by a deficit. Genes, and other genomic regions, that have been subject to adaptive, Darwinian evolution have experienced greater changes to their sequences than have sequences not subjected to selection (Note 1). Adaptive evolution is best revealed by comparing closely related species where most nucleotides under neutral selection have not been substituted more than once since their common ancestor. The draft genome sequences of mouse (1) and rat (2) have proved to be the ideal yardsticks against which to measure the human genome. Genomes of mammals more closely related to humans, such as the chimpanzee, have too few sequence differences to allow informative comparisons. In these cases, too little evolutionary drift (Note 2) may have occurred to enable regions that have been subject to selection to be distinguished from those that have not.

Nucleotide substitution is not the only mediator of mutational change. Duplication of single genes (3), or multiple consecutive genes (4), in a genome provides an important substrate on which selection can act. Duplicate genes that become fixed in the population are the raw material for functional innovation. Often, these duplicates allow novel function to be generated while ancestral functions are preserved. Genes that have arisen via duplication within a genome are termed paralogs; corresponding genes that are found in different organisms due to speciation are termed orthologs (5,6).

For the initial analyses of the mouse and rat genome sequences, we employed two approaches to survey the effects of selection on mammalian genes. First, by comparing human genes and their rodent orthologs, we estimated the selective pressures borne by different gene categories since their common ancestor 70–110 million years ago (Mya).

From: *The Proteomics Protocols Handbook*
Edited by: J. M. Walker © Humana Press Inc., Totowa, NJ

Second, we performed a comparison of mouse or rat gene paralogs that have arisen, via intragenomic duplication, in the intervening years since this common ancestor. The former approach concentrates on nonduplicated genes, whereas the latter investigates duplicated genes that are specific to mouse and/or rat lineages. In this chapter, we outline the computational procedures that were required to perform these two analyses.

2. Materials

Gene comparisons require both a comprehensive set of sequences, and methods for their alignment and analysis. Unfortunately, the complete set of all amino acid and cDNA sequences for protein coding genes of mammalian genomes is an elusive quarry. This is because gene prediction algorithms cannot yet perfectly discriminate between the signal present in the approx 1.5% of the genome that contains coding sequence (1), and the noise present in the remaining 2.5–3.0 Gb sequence portion (7).

The gene sets used for the genome comparisons in Waterston et al. (1) and The Rat Genome Sequencing Consortium (2) can be obtained from the Ensembl database (<http://www.ensembl.org/>) (8). Data are also available from, for example, ftp://ftp.ensembl.org/pub/current_human/, containing FASTA-formatted cDNA, DNA, and amino acid sequences, and from the EnsMart data-mining tool (<http://www.ensembl.org/EnsMart/>). Ensembl gene predictions are available from the University of California at Santa Cruz (UCSC) genome browser database (<http://genome.cse.ucsc.edu/downloads.html>). This database contains a wealth of additional information, including known genes and gene predictions using Fgenesh++, SGP2, SLAM, and Twinscan methods. Data may also be downloaded from the National Center for Biotechnology Information (NCBI) (<ftp://ftp.ncbi.nih.gov/genomes/>).

Protein and transcript sequences of orthologs may be downloaded from the Compara database (8). Other orthology resources include HomoloGene (<http://www.ncbi.nlm.nih.gov/HomoloGene/>), Eukaryotic Gene Orthologs (<http://www.tigr.org/tdb/tgi/ego/ego.shtml>), and COGs (<http://www.ncbi.nlm.nih.gov/COG/>).

Protein multiple sequence alignment and the prediction of genes from genomic sequence using hidden Markov models are described elsewhere in this volume (*see* Chapter 49). To estimate the evolutionary rates K_A and K_S (see **Subheading 3.1.**) for a pair of orthologs, their pairwise amino acid sequence alignment needs to be constructed. Although pairwise sequence alignment algorithms such as ClustalW often provide excellent alignments, we prefer to use Basic Local Alignment Search Tool (BLAST)-2-Sequences (9) available from <ftp://ftp.ncbi.nlm.nih.gov/blast/executables/>. This is because the latter produces local alignments, which are less prone to inaccurate gene predictions than are global alignments. K_A and K_S can be calculated for large numbers of sequences with moderate computing demands with the efficient heuristic estimation algorithm in yn00 (10) from the widely used and freely available PAML package (<http://abacus.gene.ucl.ac.uk/software/paml.html>) (11). For smaller sequence sets and with greater computational resources, it may be preferable to have more accurate K_A and K_S calculated using algorithms incorporating maximum likelihood codon evolution models (e.g., codeml in the PAML package). These methods are, however, more than an order of magnitude slower.

3. Methods

3.1. Estimates of Evolutionary Rates

The impact of selection on protein-coding genes can be estimated because of the degeneracy of the triplet genetic code, where several trinucleotides can code for the same amino acid. For example, the codons for eight amino acids are entirely redundant at the third “wobble” position, which is thus known as a fourfold degenerate, or 4D, site (12): GCx (Ala), CCx (Pro), TCx (Ser), ACx (Thr), CGx (Arg), GGx (Gly), CTx (Leu), and GTx (Val). Nucleotide substitutions at the 4-D site do not produce changes in the translated protein sequence (i.e., are synonymous or “silent”), and their rates of substitution might therefore reflect the underlying neutral rate of change.

The number of synonymous changes at two-, three-, and fourfold degenerate sites can be counted along with the number of nonsynonymous changes to allow the calculation of K_A , the number of nonsynonymous substitutions per nonsynonymous site, and K_S , the number of synonymous substitutions per synonymous site (13) (Note 3 explains why this task is not trivial). K_A is a measure of mutations fixed in the presence of evolutionary selection, whereas K_S represents the background genetic drift. By comparing these two numbers (i.e., the K_A/K_S ratio, or ω) for the same protein-coding gene, we can reveal the consequences of natural selection. A K_A/K_S ratio of one implies that there has been no evolutionary discrimination between synonymous and nonsynonymous sites, and thus that these sequences have not been subject to selective pressure. Smaller K_A/K_S ratios, e.g., less than approx 0.15, imply that purifying selective pressure has predominated, whereas the rare cases where ω is much greater than one are the results of diversifying, adaptive evolution. Later (Subheading 3.5.), we review a method to identify within single gene families individual codons in each gene that have been subject to purifying selection and others that exhibit the effects of adaptive evolution.

3.2. Estimating K_A and K_S Values for Orthologs

Evolutionary rates can be estimated using either the yn00 or the codeml program in PAML. These programs take as input a file containing aligned coding DNA sequences (Note 4). It is recommended that the “F3x4” codon model be used in codeml (CodonFreq = 2), which allows for the transition/transversion rate and codon usage biases without over-parameterization. Otherwise, pairwise comparisons should include the settings in Note 4. The default parameters for yn00 are appropriate, with the possible use of unequal codon weighting for divergent sequences (weighting = 1).

3.3. Identification of Paralog Clusters

The predominant mechanism of gene duplication is thought to be unequal recombination (3,14). This gives rise to clusters of multiple, tandemly repeated paralogs with equivalent transcription strand orientations. To identify these clusters requires either an all-against-all gene sequence comparison, which is computationally expensive, or else a sliding-window comparison examining only the sequences of neighboring genes. For the Mouse Genome Sequencing Project (1), we used the sliding-window approach. Genes were ordered on the basis of their cytogenetic coordinates.

Then the protein sequence of each gene (see **Note 5**) was compared with those of the five preceding and five succeeding, genes, using BLAST-2-Sequences (9) and an expect (E) value upper threshold of 10^{-4} (see **Note 6**). In a subsequent study of the rat genome (2), the sliding window was extended to include 10 preceding and 10 succeeding genes, and coiled coil and compositionally biased sequences were masked (15,16) to reduce false-positive homolog identification.

In order to detect mouse-specific clusters of paralogous genes that are absent in the human genome, we first identified closely sequence-similar human homologs using BLASTP (17). Subsequently, human orthologs were assigned using the Ensembl ortholog pair set or by constructing dendrograms using Clustal-W (18) and neighbor-joining methods. These dendograms were inspected for occurrences where either a single or no human gene appeared to be an ortholog of at least four mouse genes.

3.4. Inferring Evolutionary Gene Histories Using K_S

K_S , the number of synonymous substitutions per synonymous site, is of use in predicting orthology and paralogy relationships. This is because the period of evolutionary time since the speciation event that gave rise to orthologs in separate species is directly proportional to K_S , provided that the underlying neutral rates for the orthologs have not diverged since speciation (**Note 7**). Similarly, K_S values can also be used to time gene duplication events that gave rise to paralog pairs. Thus, pairs of rat paralogs with K_S values less than the average K_S value between rat and mouse orthologs are likely to have arisen since the divergence of rat and mouse lineages (2).

To predict the evolutionary relationships among a group of closely related homologs, it is necessary to construct a tree (dendrogram) based on neutral rate (K_S) estimates. These estimates are generated using the PAML package (11), as in **Subheading 3.2**. A phylogenetic distance matrix can be constructed from the pairwise K_S calculations for all the genes in the group. The underlying phylogenetic tree can then be recovered using standard distance matrix tree-building algorithms, such as UPGMA, neighbor-joining, and the Fitch–Margoliash criterion (19). Phylip (<http://evolution.genetics.washington.edu/phylip.html>) is a free package of programs for inferring phylogenies that includes implementations of each of these algorithms.

Interpreting the resulting dendrogram is relatively straightforward and involves reconciling the gene tree with a known species tree. We will explain this procedure using a K_S -derived tree (**Fig. 1**) created from homologous genes from either species A or species B. The genomes of species A and B encode 9 and 4 homologs, respectively, but (pseudo)genes A7, A8, and B5 have accumulated stop codons and frame shifts due to nonfunctionality. We work down the gene tree from the leaves towards the root in a post-order tree traversal. Each node represents either a speciation (marked with “S”) or a gene duplication event. These are distinguished by considering whether the leaves that eventually sprout from the node are represented by one or by both species. Provided that no per-lineage gene losses have occurred, the first intermediate node that contains both species represents the last common ancestor of each lineage-specific gene family just prior to speciation (e.g., the node marked “S1” in **Fig. 1**). The two groups of genes from each species above the speciation node (i.e., A1–A6 and B1–B2) are orthologs to each other. When there are multiple genes for any species (e.g., A1–A6), these are paralogs arising from lineage-specific duplication.

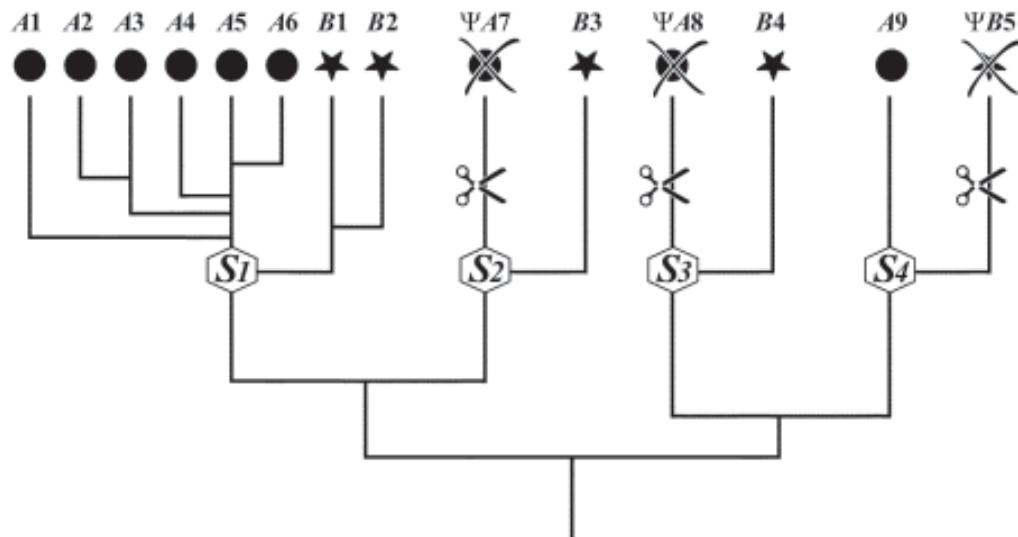


Fig. 1. Phylogenetic tree showing ortholog and paralogs from one gene family found in two species. The leaves of the tree at the top, representing genes from species A and B, are labeled accordingly (A1–A9 and B1–B5). The nodes of the tree are gene duplication events, with S1–S4 representing the speciation point that gave rise to A and B from their last common ancestor. Genes A7, A8, and B5 are nonfunctional pseudogenes (labeled with Ψ), which are generally either missing or excluded from sets of predicted genes in sequenced genomes.

In some cases, all the orthologs of one species will be missing in one branch of a tree, as a result of genomic deletion, disruption to the gene, or errors in gene prediction or genome assembly. Thus B3 is “orphaned” because its ortholog A7 is an unpredicted pseudogene. In this case, B3 will appear to join to the dendrogram below the speciation node S1 rather than at its real speciation node, S2. Lineage-specific gene losses thus can be identified in all such cases, as the branches will join below speciation nodes. In rare cases, there might be reciprocal lineage-specific gene losses in the two branches of the dendrogram that are the most closely related to each other, e.g., the loss of A8 and B5. In this case, the speciation nodes S3 and S4 will not be predicted, but instead, the branches for A8 and B5 will appear to encounter each other at the node above S3 and S4. It is usually easy to identify such examples, as the resulting mis-predicted speciation node will have a significantly greater phylogenetic distance to its genes than the norm.

3.5. Detecting Adaptive Evolution

Adaptive mutations that offer a selective advantage are fixed at a higher rate than the background neutral mutation rate, so that the K_A/K_S for residues under adaptation will be greater than one. However, most residues are conserved, and it is only the minority of amino acid sites that are responsible for molecular adaptation (20). Thus, even where there is diversifying selection, the K_A/K_S averaged over the entire coding sequence is usually less than one. It is necessary instead to calculate K_A/K_S values (ω values) for individual residues.

The codeml program in the PAML package implements maximum likelihood (ML) codon evolution models, which allow selection to vary among amino acid sites. Bayesian statistics is then used to predict with different degrees of confidence which individual residues belong to the diversifying group. The likelihood ratio test (LRT) can be used to compare nested evolutionary models that either do or do not allow for residues with $\omega > 1$. Adaptive evolution is predicted if (a) the LRT is statistically significant, (b) the ML model predicts a class of residues with $\omega > 1$, and (c) individual sites belonging to the high ω class can be identified with high posterior Bayesian probabilities.

Adaptive evolution increases the fitness of individuals of a species in a changing environment. It often occurs episodically over comparatively short periods of evolution. Even where there is sustained molecular innovation, e.g., in face of host/pathogen or sexual competition, the loci of adaptation may vary in different lineages of the phylogenetic tree. Over long periods of evolutionary time, such adaptive episodes will be lost in the general background of purifying selection or genetic drift. With large sequence divergences between branches of a phylogenetic tree, on average, each residue will have mutated more than once since the common ancestor, and it will be impossible to reliably estimate ancestral mutations giving rise to the modern sequences. Thus calculations for both K_A and K_S saturate for large values. Conversely, if the sequence divergence is small, too little evolutionary drift (**Note 2**) will have occurred to allow differentiation between regions that have been subject to selection and those that have not. Computer simulations show that on average, codeml gives conservative and reliable results, and has high resolving power (**21**). However, all this holds true only if there are adequate numbers of genes (>6) and sequence lengths (>50 – 100 codons), and where sequence divergence falls within a moderate range (e.g., S/N_{branches} approx 0.1–1). The latter quantity represents the average number of expected nucleotide changes per codon per branch, where:

$$N_{\text{branches}} = S/(2 \times N_{\text{genes}} - 3);$$

$$S = 3K_S p_S + 3K_A(1 - p_S);$$

S is a measure of tree length, the expected number of nucleotide substitutions per codon;

N_{genes} is the number of genes in the tree;

N_{branches} is the number of branches in the tree; and,

p_S is the proportion of synonymous sites. This should be calculated empirically for each family of genes, but typical values are 20–30%.

Phylogenetic trees of closely related paralogs and orthologs are thus useful starting points for identifying gene families or sub-families of the appropriate genetic divergence and biological interest, which can be analyzed further to detect adaptive diversifying evolution.

3.6. Estimating Per-Site Values (ω) of K_A/K_S

The codeml program can be used for estimating per-site values (ω) of K_A/K_S . This requires the multiply aligned sequences of all the requisite genes (**Note 4**) as well as the

reconstructed phylogenetic tree in New Hampshire format. Codeml includes three pairs of nested codon evolution models, each of which tests a null hypothesis using LRT to give statistical significance. Log likelihood values (l) are calculated for each model by maximum likelihood. The LRT consists of comparing the test statistic ($2\Delta l$) to critical values in the Chi square (χ^2) distribution calculated using the appropriate number of degrees of freedom (**Note 9**).

Models M0 and M3 assume respectively that sites fall into either one, or two or three, discrete classes of ω values. They are thus, strictly speaking, only a test of whether all residues in a sequence are under identical selection pressures. Model M1 assumes all sites are under either purifying or neutral selection ($\omega = 0$ or 1), and model M7 assumes a beta distribution of ω between 0 and 1. Models M2 and M8 extend M1 and M7 by adding a third class of residues with $\omega > 1$. Thus significant LRTs between both M1/M2 and M7/M8 are direct indications of positive selection. Each of the models M3, M2, and M8 also predict, with corresponding posterior Bayesian probabilities, lists of individual residues that may fall into any class of positively selected sites with $\omega > 1$. These predictions are of increasing stringency, from M3 to M2 to M8.

We suggest that evidence for the evolution of sites under positive selection is provided by multiple nested models. Determination of adaptive sites would require that (a) the LRT statistics are significant for each of the pairs of models M0/M3, M1/M2 and M7/M8; (b) some sites with $\omega > 1$ should be predicted by each of these three nested models; and (c) the residue in question should be predicted with a posterior Bayesian probability of at least 0.90 by M3, M2, or M8, and confirmed by a probability of > 0.5 in one other model.

3.7. Interpretation of ω Estimates

The identification of positively selected residues using the PAML package has been well validated by both computer simulations (21) and congruence with biological observations (22). Nevertheless, the interpretation of per site adaptive predictions requires due care. Sites subjected to positive selection are expected to have arisen as a result of interspecific or conspecific competition, mostly involving interfaces in protein–protein interactions. Thus, such sites are not expected in the buried interior of proteins, but are more likely to be spatially clustered at the solvent-accessible periphery of single molecules. It is important therefore to map these sites to homology models of known three-dimensional protein structures using a visualization tool, such as Swiss-PDBviewer (<http://ca.expasy.org/spdbv/>). Clusters of positively selected sites are excellent candidates for functional investigation using molecular biology techniques.

4. Notes

1. A small fraction of nucleotide changes between human and mouse genomes is likely to have been due to adaptive, or diversifying, selection. Adaptive evolution arises when a selectively favored allele appears within a population, sweeps through it, and becomes fixed. The minor allele achieves selective advantage in competitive situations, such as in mate selection, or by resisting debilitating infection. Thus, genes that have been subject to multiple bouts of adaptive evolution may be distinguished from genes constrained by purifying selection by characteristic differences in nucleotide substitution rates: they

contain sites whose inferred nucleotide substitution histories are not explicable by neutral evolution theory.

- Only approx 5% of bases in human or mouse genomes appear to have been subject to selection (1). These regions have been constrained by purifying selection, which has acted to prevent deleterious alleles from being fixed in their populations. Less than half of the approx 5% under selection can be accounted for by protein-coding genes. Thus, selection has acted extensively in the noncoding portions of the genomes. Sites in protein-coding genes that, when mutated, result in amino acid conservation, rather than substitution, are considered to be selectively neutral. It is estimated that approximately half of these sites have experienced a nucleotide substitution since the common ancestor of primates and rodents (1). It is not surprising, therefore, that much of the selectively neutral regions of the human and mouse genomes can be aligned, despite their divergence from a common ancestor over 70 Mya. The genomes of more closely related mammals, such as the chimpanzee, have too few sequence differences to allow differentiation between regions that have been subject to selection and those that have not.
- The apparent simplicity of the definitions of K_A and K_S appears to suggest that their calculation would be both trivial and obvious, requiring no more than counting the synonymous and nonsynonymous sites in, and difference between, two sequences. Over a dozen intuitive methods have been developed since the 1980s with varying degrees of sophistication, including attempts to correct for multiple substitutions. However, ignoring the biases between the rates for nucleotide transitions ($T \leftrightarrow C$ and $A \leftrightarrow G$) and transversions ($T/C \leftrightarrow A/G$), as well as codon usage affects the calculated values significantly. The former bias leads to underestimation of S , overestimation of K_S , and underestimation of ω ; the latter bias usually overestimates S and ω , depending on the GC content at the third codon position. Small errors in partitioning synonymous and nonsynonymous sites and substitutions lead to changes to the values of K_A and K_S in opposite directions. This results in large errors in the ratio K_A/K_S .
- codeml and yn00 in the PAML package can read sequence files in the format of the PHYLIP package. In sequential form, this includes the number of sequences and sequence length on a single line.

This is followed by the successive sequence names and nucleotide data (see **Table 1**).

The numerical methods for models M2 and M8 are particularly susceptible to the problem of multiple local minima, and it is therefore a good idea to run these with multiple starting ω values, for example, $\omega = 0.03, 0.3$, and 1.3 , and only use the results with the greatest log likelihood.

- An additional complication arises from alternative splicing of genes. Ideally, we need to identify transcripts from pairs of homologous genes that are homologous throughout, meaning that each pair of aligned nucleotides arose from a single nucleotide in the common ancestral gene. Alternative transcripts that use different exon combinations will not usually be homologous throughout their alignment. Identifying homologous transcripts is most simply achieved by comparing all transcripts of one gene with all those of another, and then choosing the pair of transcripts with the highest alignment (bit) score. More accurate but more complex methods involve mapping transcript alignments to genome–genome alignments (2).
- Choice of the E-value upper threshold should be determined by the gene count (N) in a genome. The number of different sequence comparisons using a sliding window containing n genes is approx $N(n - 1)/2$. If $N = 25,000$, and $n = 21$, this number is 2.5×10^5 . Thus, for an E-value upper threshold of 10^{-4} , the number of false-positive predictions of homology is $2.5 \times 10^5 \times 10^{-4}$, or 25. Since approx 3000 genes are found in paralog gene clusters in mammalian genes, this represents an error rate of less than 1%.

Table 1
Sequence Data and Codeml Parameters

2 12

sequence 1

ATGAAG---CAC

sequence 2

ATGAAGCTACAC

Each sequence should contain a whole number of codons.

codeml settings specified in the control file (usually codeml.ctl) for pairwise comparisons should include (comments follow asterisks):

seqfile = a.file	* specify your input file name here
outfile = output	* specify the name of the results file here
runmode = -2	* -2:pairwise
seqtype = 1	* 1: codons
CodonFreq = 2	* 2: F3X4
model = 0	* 0: one K_A/K_S ratio for branches
NSsites = 0	* 0: one
icode = 0	* 0: universal genetic code; * see PAML manual for more choices
fix_Kappa = 0	* 0: Kappa (transition/transversion ratio) * to be estimated
Kappa = 2	* initial or fixed Kappa
fix_omega = 0	* 0: K_A/K_S to be calculated
omega = .4	* initial or fixed omega
getSE = 0	* 0: discard standard errors, 1: keep S.E.s
Additional codeml settings for identifying positively selected sites should include:	
NSsites = 0 3 1 2 7 8	* ML models where * 0: M0 A single K_A/K_S throughout the sequence * 3: M3 Multiple discrete K_A/K_S classes * 1: M1 Neutral model: K_A/K_S = either 0 or 1 * 2: M2 As M1 but with extra sites of $K_A/K_S > 1$ * 7: M7 K_A/K_S follows beta distribution * 8: M8 as M7 but with extra sites of $K_A/K_S > 1$ * specify multiple models for batch runs
omega = .03	* initial or fixed omega, save the best * results where omega = 0.03,0.3,1.3 for * NSsites = 2 and 8
ncatG = 3	* # of discrete classes in M3

7. Neutral rates vary greatly among different mammalian chromosomes and genomic regions (1,12,23,24). This variation might be thought to confound the prediction of orthology and paralogy using K_S values from yn00. However, many of the gene duplication events that produce paralogs are local syntenic duplications within regions that may share the same neutral rate. In general, this problem is least worrying when the average neutral rates between pairs of species are greater than the rate variations within their genomes, which is the case for comparisons among human, mouse, and rat sequences. The rate variations certainly do confound comparisons between more closely related species, such as chimpanzee and human.

8. The M0/M3 LRT is compared to the χ^2 distribution with 4 degrees of freedom (d.f.), whereas both the M1/M2 and M7/M8 LRTs have 2 d.f. The χ^2 distribution for the appropriate d.f. can be obtained by running the chi2 program in the PAML package.

Acknowledgments

We thank Scott Beatson, Richard Copley, Richard Emes, and Ewan Birney for helpful discussions, and the Medical Research Council UK for funding.

References

1. Waterston, R. H., Lindblad-Toh, K., Birney, E., et al. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562.
2. Gibbs RA, Weinstock GM, Metzker ML, et al. (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493–521.
3. Ohno, S. (1970) *Evolution by Gene Duplication*. Springer-Verlag, New York.
4. Bailey, J. A., Gu, Z., Clark, R. A., et al. (2002) Recent segmental duplications in the human genome. *Science* **297**, 1003–1007.
5. Fitch, W. M. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool.* **19**, 99–113.
6. Fitch, W. M. (1995) Uses for evolutionary trees. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **349**, 93–102.
7. Ureta Vidal, A., Ettwiller, L., and Birney, E. (2003) Comparative genomics: genome-wide analysis in metazoan eukaryotes. *Nat. Rev. Genet.* **4**, 251–262.
8. Clamp, M., Andrews, D., Barker, D., et al. (2003) Ensembl 2002: accommodating comparative genomics. *Nucleic Acids Res.* **31**, 38–42.
9. Tatusova, T. A. and Madden, T. L. (1999) BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol. Lett.* **174**, 247–250.
10. Yang, Z. and Nielsen, R. (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* **17**, 32–43.
11. Yang, Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**, 555–556.
12. Hardison, R. C., Roskin, K. M., Yang, S., et al. (2003) Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Res.* **13**, 13–26.
13. Hurst, L. D. (2002) The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends Genet.* **18**, 486.
14. Smith, G. P. (1976) Evolution of repeated DNA sequences by unequal crossover. *Science* **191**, 528–535.
15. Lupas, A., Van Dyke, M., and Stock, J. (1991) Predicting coiled coils from protein sequences. *Science* **252**, 1162–1164.
16. Wootton, J. C. and Federhen, S. (1993) Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem.* **17**, 149–163.
17. Altschul, S. F., Madden, T. L., Schaffer, A. A., et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402.
18. Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680.
19. Fitch, W. M. and Margoliash, E. (1967) Construction of phylogenetic trees. *Science* **155**, 279–284.

20. Golding, G. B. and Dean, A. M. (1998) The structural basis of molecular adaptation. *Mol. Biol. Evol.* **15**, 355–369.
21. Anisimova, M., Bielawski, J. P., and Yang, Z. (2002) Evaluation of the Bayesian approach to detecting codon sites under positive Darwinian selection. *Mol. Biol. Evol.* **19**, 950–958.
22. Emes, R. D., Beatson, S. A., Ponting, C.P., and Goodstadt, L. (2004) Evolution and comparative genomics of odorant- and pheromone-associated genes in rodents. *Genome Res.* **14**, 591–602.
23. Makalowski, W. and Boguski, M. S. (1998) Synonymous and nonsynonymous substitution distances are correlated in mouse and rat genes. *J. Mol. Evol.* **47**, 119–121.
24. Casane, D., Boissinot, S., Chang, B. H., Shimmin, L. C., and Li, W. (1997) Mutation pattern variation among regions of the primate genome. *J. Mol. Evol.* **45**, 216–226.

Computational Identification of Related Proteins

BLAST, PSI-BLAST, and Other Tools

Qunfeng Dong and Volker Brendel

1. Introduction

Molecular sequences that share a high degree of similarity often are thought to have evolved from common ancestral genes. Closely related protein sequences will presumably correspond to similar three-dimensional structures and conserved biological functions (although the reverse is not necessarily true: similar structures and conserved functions do not imply that the corresponding protein sequences will be similar; reviewed in **ref. 1**). These assumptions provide the basis for computational gene annotation. Typically, the first step in characterizing a novel gene is to compare its sequence against known sequences in available databases and to predict its origin and function by copying the annotation of those previously characterized sequences. This approach has been highly successful and is probably the only practical method applicable to large-scale annotation efforts at present. It should be pointed out, however, that this practice is not without its limitations (and is also unsatisfactory from the more theoretical perspective of those who wish to determine structure and function from primary sequence; for a provocative editorial on this subject, *see ref. 2*). The intrinsic problems of transitive propagation of historical annotation errors have been discussed elsewhere (3) and are all too familiar to any biologist who has looked into the databases only to find puzzling annotations that make no sense with current knowledge.

Basic Local Alignment Search Tool (BLAST [4,5]) is the most popular computer program used to perform database similarity searches. The program allows rapid identification of sequences from large databases that are similar to the query and provides sound statistical evaluation of the significance of the findings. This chapter reviews briefly the standard use of this tool and, in greater detail, the use of lesser known applications, along with additional references to other useful tools to help with nonstandard queries. The scope of the chapter is limited to practical aspects of these tools addressing biological research problems, and only marginally deals with computational and statistical issues.

2. Web Resources and Source Code

1. NCBI BLAST, <http://www.ncbi.nlm.nih.gov/BLAST/>.
2. WU-BLAST, <http://blast.wustl.edu/>.
3. MuSeqBox, <http://bioinformatics.iastate.edu/bioinformatics2go/mb/MuSeqBox.html>.
4. PhyloBLAST, <http://www.pathogenomics.bc.ca/phyloBLAST/>.

From: *The Proteomics Protocols Handbook*
Edited by: J. M. Walker © Humana Press Inc., Totowa, NJ

5. FASTA, <http://fasta.bioch.virginia.edu/fasta/>.
6. GeneSeqr, <http://bioinformatics.iastate.edu/cgi-bin/gs.cgi>.
7. BLAT, <http://genome.ucsc.edu/cgi-bin/hgBlat/>.
8. Vmatch, <http://www.vmatch.de/>.

3. Methods

3.1. The NCBI BLAST Web Server

The huge popularity of the BLAST program is a tribute to the excellent service and continuing development efforts of the US National Center for Biotechnology Information (NCBI). The NCBI BLAST Web server (<http://www.ncbi.nlm.nih.gov/BLAST/>) is accessed tens of thousands times per day, and provides the easiest access to the program for most biologists, encompassing a variety of sequence comparison tools. Users can match their sequence of interest against different protein or nucleotide sequence databases. For general purpose searches, the “nonredundant” protein or nucleotide entries stored in GenBank (abbreviated *nr*; [6]) are commonly used. However, users need to be aware of the definition of redundancy in this context. For example, a laboratory could have submitted a partially determined gene sequence first and a full-length sequence later on, or two laboratories could have submitted sequences for the same gene independently. In each case, the two entries would be considered independent, “nonredundant” entries in the *nr* database. Thus, redundancy is defined on the level of the data record, not on the sequence level. In addition to the *nr* databases, NCBI also provides specialized databases that are not included in *nr*. For example, users can search their protein sequence against *pdb*, which comprises protein sequences whose structures are available at the Protein Data Bank (PDB; [7]). Users can also compare their nucleotide sequence against the *est* database, consisting of expressed sequence tags (ESTs), a query useful for gene finding or gene structure annotation. In addition, the NCBI server allows users to narrow their searches within subsets of the chosen databases to specific molecule types (e.g., kinase) or organisms, with the “Limit by entrez query” option available on all the regular BLAST search forms.

NCBI also provides a set of specialized database search tools. For example, short query sequences can be searched at a dedicated page with BLAST parameters optimized to suit such searches (e.g., this function could be used to check the uniqueness of reverse transcriptase-polymerase chain reaction (RT-PCR) primers against the EST database). Protein sequences can be queried for domain structures (RPS-BLAST, <http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi>; CDART, <http://www.ncbi.nlm.nih.gov/Structure/lexington/lexington.cgi?cmd=rps>). In fact, for protein queries, the Web server will first try to identify any conserved domains by a BLAST search of the query against the “Conserved Domain Database” (<http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml>) by default. Note, however, that complete protein domain searches usually involve methods other than BLAST-based algorithms. It is advised that users should also search other specialized protein domain databases/tool combinations such as BLOCKS (<http://blocks.fhcrc.org/>), Pfam (<http://www.sanger.ac.uk/Pfam1>), and SMART (<http://smart.embl-heidelberg.de/>).

NCBI also provides dedicated Genomic-BLAST search pages for those organisms with complete or nearly completed genome sequences that are in the process of being finished by large-scale sequencing projects. The results of the Genomic-BLAST can be

visualized in the context of the genome (Genome View). Another commonly used tool at the Web site is vector screening, which screens a query sequence for matches against common cloning vectors that are compiled in a dedicated vector sequence database (UniVec; <http://www.ncbi.nlm.nih.gov/VecScreen/VecScreen.html>). To screen large numbers of sequences, the UniVec database can be downloaded and searched with other programs such as Vmatch (<http://www.vmatch.de/>) or cross_match (P. Green, unpublished; <http://www.phrap.org/phredphrapconsed.html>).

Besides database searches, the NCBI BLAST Web server also allows pairwise comparison of two sequences (<http://www.ncbi.nlm.nih.gov/blast/bl2seq/bl2.html>). Although in theory the algorithm is not guaranteed to return an optimal alignment, in practice it provides a reliable and convenient way to perform pairwise alignments. It is beyond the limits of this short chapter to review all the excellent functions linked to the NCBI BLAST Web service. Readers are strongly encouraged to visit NCBI's comprehensive on-line tutorials and documentation (<http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/information3.html>). In the remainder of this chapter, we describe the essential BLAST capabilities beyond what is available at the NCBI Web server, in particular focusing on the standalone BLAST and PSI-BLAST programs.

3.2. The Standalone BLAST Program

Although the NCBI BLAST Web server is an excellent and versatile tool, installation of BLAST to run locally on your own computer has some advantages relative to completely relying on this server. For example, the NCBI standard BLAST site currently does not allow multiple queries to be run simultaneously (except for the MEGABLAST server, which is optimized for matching nucleotide query sequences with near-identical nucleotide database sequences; <http://www.ncbi.nlm.nih.gov/blast/megablast.shtml>). Thus, if a researcher wishes to determine the set of sequences related to a query set of, say, 100 sequences, 100 separate BLAST submissions to the server would be required, and subsequent results would lack collation. By contrast, such tedious submission can be entirely avoided if the BLAST search were run locally (an alternative approach, which does permit multiple query input, is the use of network-client BLAST; *see Note 1*).

To distinguish these options, it may be helpful to dissect the NCBI BLAST Web server a bit (the same dissection would apply to most Web services). The server consists of three major components. One component comprises the Web forms used to input a query sequence and select search parameters. The second component is the actual BLAST program, installed on a dedicated powerful computer to perform the search tasks. This program compares a user-submitted query sequence against all selected sequences stored at NCBI. The third component consists of a suite of tools that graphically and textually present the search results and alignments. Although the entire NCBI BLAST interface cannot be duplicated on your local computer, the second component (the BLAST program itself) is freely available in both source code and executable version for anyone to install locally. Of course, even skilled bioinformaticians regularly use the NCBI BLAST Web server, because of its connection to large, up-to-date databases, excellent presentation of results, and linkages to other useful data. Running BLAST locally should be considered to be an approach complementary to using the Web service, allowing for more versatile functions not offered by the Web service.

Most biologists are not unaccustomed to the idea of installing computer programs. For example, the CLUSTAL multiple sequence alignment program (8) has been widely distributed and is routinely used locally despite the availability of excellent Web servers (e.g., the CLUSTAL server at European Bioinformatics Institute, <http://www.ebi.ac.uk/clustalw/>). Similarly, biologists commonly install and use programs for phylogenetic analysis on their personal computers (e.g., MEGA [9]; PHYLIP [10]). However, the installation of the BLAST program may seem more difficult to many biologists, because installation is not quite as simple (although sufficient instructions are supplied in the documentation files in the BLAST package). In this section, we will go over the installation process in detail, covering only NCBI-BLAST. However, it should be noted that another popular version of BLAST is available from Washington University (WU-BLAST; <http://blast.wustl.edu/>). The two versions perform database searches based on the same principles, but differ with respect to implementation details and the statistical methods used to evaluate results.

The NCBI BLAST package can be downloaded from the NCBI ftp site (<ftp://ftp.ncbi.nlm.nih.gov/blast/executables/LATEST>). At the site is a list of files with names indicating program, version number, and suitable operating system for each available download. Currently, BLAST is available for a variety of operating systems, including various versions of Unix, Linux, Microsoft Windows, and MacOS X. As an example, *see Note 1* for step-by-step instructions for installation of BLAST on Microsoft Windows computers.

3.2.1. Setting Up the Database

Once you have installed the BLAST program on your own computer, you will need to build a BLAST database (the set of sequences you wish to query with your sequence(s) of interest). A BLAST database file is simply a text file of nucleotide or protein sequences in what is called “FASTA format” (**Fig. 1**). The database can be large, such as the GenBank *nr* protein database (more than 1,600,000 sequences at present), or small (consisting of at least one sequence). Note that nucleotide and protein sequences cannot be mixed in the same database. Typically, a researcher might create a database of all proteins from a given organism, or all genes of a particular functional group. Once a suitable set of sequences has been compiled in a FASTA-formatted file, the next step is to build indices for your database sequences. “Index” is a concept of computer science, a mechanism to locate data points in a file faster than by simple scanning. Specifically for BLAST, the indices are used to quickly identify exact matching regions between your query sequence and database sequences. The indices are created using the *formatdb* program included in the BLAST package. Unfortunately, the *formatdb* and all the other programs in the package do not have a graphical user interface (GUI). This means that you cannot simply select files and parameters, then run the program using the computer mouse in the way you are accustomed to for most other Windows programs. Instead, you will have to run the program from the command line (meaning you have to type some simple computer instructions at a prompt). *See Note 2* for an example of using the *formatdb* program to build a BLAST index. You need to build the index only once; then you are ready to perform your local BLAST searches.

```

>gi|4034483|emb|CAA10169.1| (AJ012753) RAGE protein
GAVVGAQNITARIGEPLVLSVRGPPRNHPSGWNGNW
>gi|28680|emb|CAA47406.1| (X67016) amphiglycan
MAPARILFALLLFLVGGVAESIRETEVIDPQDLLEGRYFSGALPDEDVVPGQESDDFELSGSGDLDDLEDSMIGPEVVHPLVPLDNHI
PERAGGSQVPTEPKLEENEVPIKRRISPVEESEDVSNKVSMSSTVQGSNIFERTEVLAALIVGGIVGILFAVFLILLMYRMKKKDEG
SYDLGKPKIYKKAPTENEYA
>gi|4225951|emb|CAA51170.1| (X72579) collagen X
MLPQIPFLVLSNLVHGVFYAERYQMPGKIKGPFLPNTKTQFFIPTYIKSKGIAVRGEQGTPGPPGPAGPRGHPGPSGPPGKPGYGS
LQGEPGLPGPPGSAVKGPGVPGLPGKPGERGPYGPKGVDVGPAGLPGPRGP

```

Fig. 1. A sample Basic Local Alignment Search Tool (BLAST) database file. A BLAST database file is a text file composed of sequences in FASTA format. A sequence in FASTA format begins with a > symbol and one-line description, followed by lines of sequence. The database can contain either nucleotide or protein sequences, but not both. In order to make the database usable by the BLAST program, it needs to be indexed first (see text and **Note 2**).

3.2.2. Running BLASTALL

The general BLAST search is carried out by the *blastall* program. The program can invoke several sub-programs (**Table 1**), each distinguished by the type of query and database sequences it utilizes. See **Note 3** for an example of invoking the sub-program BLASTP by the -p option of *blastall*. The BLASTN and BLASTP sub-programs are straightforward, comparing a nucleotide or protein query sequence against a nucleotide or protein database, respectively. The other programs all involve automatic six-frame translation of nucleotide sequences in the query, or database, or both. By doing so, they compare sequences at the protein level, thus increasing the chance of finding a match that may otherwise be masked by differential codon usage. For example, the BLASTX program translates a query nucleotide sequence in all reading frames and compares it to a protein database. This is useful when assigning tentative function to a novel nucleotide sequence (such as an EST) without prior knowledge of the position of the correct reading frame. The TBLASTN program compares a protein query sequence against a translated nucleotide database. For example, if a user has a protein query sequence from a particular species, a TBLASTN query against EST sequences isolated from other species could be used to determine whether homologous genes are expressed in those species. The TBLASTX program requires the most computational resources, because it translates both the nucleotide query and the nucleotide database in all six frames before comparisons are made on the level of translation. For example, one could use this program to compare human ESTs against mouse ESTs to look for homologous genes.

3.2.3. Alignment Display Options

Figure 2 shows a typical alignment returned from a BLAST search. Formally, an alignment between two sequences (here, the query sequence and a database sequence) is a series of columns representing matches, mismatches, or gaps. A match is a pair of identical symbols (nucleotide or amino acid). A mismatch is a pair of different symbols. A gap is a special case of mismatch with one symbol being represented as a dash character (-).

The default display of BLAST alignments is in pairwise format, in which the query and each database sequence are aligned to one another. Sometimes, it is more conve-

Table 1
BLAST Subprograms

Programs	Description
BLASTN	Search DNA database with DNA sequence as query
BLASTP	Search protein database with protein sequence as query
BLASTX	Search protein database with DNA sequence as query
TBLASTN	Search DNA database with protein sequence as query. The DNA database is automatically translated in all reading frames.
TBLASTX	Search DNA database with DNA as query. Both the query and database are automatically translated in all reading frames.

A

```
>gb|AAC73282.1| (AE000126) uridylylate kinase [Escherichia coli]
Length = 241

Score = 41.6 bits (96), Expect = 3e-004
Identities = 28/97 (28%), Positives = 44/97 (45%), Gaps = 5/97 (5%)
Frame = +1

Query: 931  LVVLGRNGSDYSAAVLAACLR----ADCCEIWTVDGVYTCPRQVPDARLLKSMSYQE 1095
      +++      G+ +      AACLR      AD      T VDG+T DP + P A + + +Y E
Sbjct: 132  VILSAGTGNPFTTDSAACLRGIEIEADVVVLKATKVDGVFTADPAKDPATMYEQLTYSE 191

Query: 1096 AMELSYFGAKVLHPRTITPIAQFQIPCLIKNTGNPQA 1206
      +E          KV+      T      ++P + N      P A
Sbjct: 192  VLEKE---LKVMIDLAAFTLARDHKLPPIRVFNMNPKGA 225
```

B

QUERY	239	tctgcacgg-acc-gaagcaaccaaatcagccagaat--ccggaaagcggcggtggcag	294
16838789	383	tctgcacgg-acc-gaagcaaccaaatcagccagaat--ccggaaagcggcggtggcag	328
16839156	402	tctgcacgggacctgaagcaccaaatcagccagaatccggaaagcggcggtggcag	343
16835449	410	tctgcacgg-acc-gaagcaaccaaatcagccagaat--ccggaaagcggcggtggcag	355
13392607	407	tctgcacga-gcc-gaagaa 389	
16826006	200	tctgcacga-gcc-gaagaa 182	

Fig. 2. Basic Local Alignment Search Tool provides different alignment views for its output. (A) The default “pairwise alignment” view shows pairwise alignments between the query sequence (Query lines) and each database hit (Sbjct lines). (B) The optional “Query-anchored” multiple alignment view aligns (anchors) each database hit to the query sequence.

nient to have a multiple-sequence alignment display to show the conserved sites between the query and all its database hits, highlighting conserved motifs. Although this is not strictly a multiple sequence alignment (because only query to database sequence comparisons were performed), BLAST provides several “Query-Anchored” formats to view an implied multiple sequence alignment based on the pairwise comparisons (Fig. 2B). These formats can be selected by parameter choices at the Web site or with the -m option of the *blastall* program.

3.2.4. Choice of Score Matrix

Critical in a database search is the assignment of a score to each alignment that reflects a ranking of the alignments. Different scoring schemes are appropriate for dif-

ferent search goals (11). Here, we discuss only the overarching goal of distinguishing statistically significant alignments from alignments of a quality that might be expected to arise by chance. Obviously, the more similar two sequences are, the more identical matches are expected in the alignment. Therefore, the basic logic of a scoring scheme is to award each match column a positive score and penalize each column of mismatch or gap. Then the total score of the alignment is the sum of scores from all columns. For nucleotide sequences, the score scheme is usually uniform. Each match column is given a positive score (e.g., +2), and mismatch and gap columns are given a negative score (e.g., -1 and -5, respectively). For protein sequences, the scoring schemes are more complicated, based on the observation that in many instances amino acids that share similar physiochemical properties can substitute for each other without disrupting protein structure and function. A reasonable strategy to account for this is to distinguish different pairs of mismatches (substitution events) in a scoring matrix. The functional, “conservative” substitutions should be less penalized than “nonconservative” changes. For example, an alignment between amino acids Val and Leu (both hydrophobic, differing by only one CH_2 moiety) will generally be given a higher score than the alignment between Val and Lys (because Lys is a basic hydrophilic residue). Therefore, the challenge of scoring protein alignments is to assign different scores for each individual amino acid pair (amounting to 210 distinct parameters for the alignment algorithm!). To meet this challenge, many different amino acids score matrices (also called amino acid substitution matrices) have been developed.

An amino acid substitution matrix is a table with the 20 naturally occurring amino acids labeling the rows and columns and table cells containing the score values for the respective substitutions. The table may also contain scores for alignments with degenerate amino acid symbols and gap columns. A commonly used amino acid scoring matrix for general protein database searches is the BLOSUM62 matrix (Fig. 3). BLOSUM refers to a set of score matrices (12). The score values reflect observed substitution frequencies of amino acid pairs in related proteins. In particular, to derive the BLOSUM matrices, a large data set of local multiple sequence alignments were extracted from the BLOCKS database (<http://blocks.fhcrc.org/>). Substitution frequencies were deduced, and BLOSUM matrix scores were derived as scaled log-odds ratios comparing the observed substitution frequencies to the expected frequencies obtained by multiplying the respective overall amino acid frequencies. Because within each protein family (block) some member sequences will be globally more similar to each other than to other sequences (for example, as a result of a sampling bias for closely related species), closely related sequences must be merged into one representative in each alignment block before counting the substitutions, to avoid artifactual over-representation of some amino acid substitutions. Different cutoff criteria that define closely related sequences were used to produce a set of BLOSUM matrices. For example, when sequences that were 62% identical were defined as closely related, the resulting matrix was called BLOSUM62. BLOSUM62 is the default choice in both the standalone *blastall* program and the NCBI BLAST Web site; however, users can also choose other types of scoring matrices. For the *blastall* program, the choice can be invoked by the -M option. Other score matrices include the PAM series, which seek to model amino acid substitutions occurring during protein evolution (13).

```

#  Matrix made by matblas from blosum62.ijj
#  * column uses minimum score
#  BLOSUM Clustered Scoring Matrix in 1/2 Bit Units
#  Blocks Database = /data/blocks_5.0/blocks.dat
#  Cluster Percentage: >= 62
#  Entropy = 0.6979, Expected = -0.5209
  A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V  B  Z  X  *
A  4  -1  -2  -2  0  -1  -1  0  -2  -1  -1  -1  -1  -2  -1  1  0  -3  -2  0  -2  -1  0  -4
R  -1  5  0  -2  -3  1  0  -2  0  -3  -2  2  -1  -3  -2  -1  -1  -3  -2  -3  -1  0  -1  -4
N  -2  0  6  1  -3  0  0  0  1  -3  -3  0  -2  -3  -2  1  0  -4  -2  -3  3  0  -1  -4
D  -2  -2  1  6  -3  0  2  -1  -1  -3  -4  -1  -3  -3  -1  0  -1  -4  -3  -3  4  1  -1  -4
C  0  -3  -3  -3  9  -3  -4  -3  -3  -1  -1  -3  -1  -2  -3  -1  -1  -2  -2  -1  -3  -3  -2  -4
Q  -1  1  0  0  -3  5  2  -2  0  -3  -2  1  0  -3  -1  0  -1  -2  -1  -2  0  3  -1  -4
E  -1  0  0  2  -4  2  5  -2  0  -3  -3  1  -2  -3  -1  0  -1  -3  -2  -2  1  4  -1  -4
G  0  -2  0  -1  -3  -2  -2  6  -2  -4  -4  -2  -3  -3  -2  0  -2  -2  -3  -3  -1  -2  -1  -4
H  -2  0  1  -1  -3  0  0  -2  8  -3  -3  -1  -2  -1  -2  -1  -2  -2  2  -3  0  0  -1  -4
I  -1  -3  -3  -3  -1  -3  -3  -4  -3  4  2  -3  1  0  -3  -2  -1  -3  -1  3  -3  -3  -1  -4
L  -1  -2  -3  -4  -1  -2  -3  -4  -3  2  4  -2  2  0  -3  -2  -1  -2  -1  1  -4  -3  -1  -4
K  -1  2  0  -1  -3  1  1  -2  -1  -3  -2  -5  -1  -3  -1  0  -1  -3  -2  -2  0  1  -1  -4
M  -1  -1  -2  -3  -1  0  -2  -3  -2  1  2  -1  5  0  -2  -1  -1  -1  1  -3  -1  -1  -4
F  -2  -3  -3  -3  -2  -3  -3  -3  -1  0  0  -3  0  6  -4  -2  -2  1  3  -1  -3  -3  -1  -4
P  -1  -2  -2  -1  -3  -1  -1  -2  -2  -3  -3  -1  -2  -4  7  -1  -1  -4  -3  -2  -2  -1  -2  -4
S  1  -1  1  0  -1  0  0  0  -1  -2  -2  0  -1  -2  -1  4  1  -3  -2  -2  0  0  0  -4
T  0  -1  0  -1  -1  -1  -1  -2  -2  -1  -1  -1  -1  -2  -1  1  5  -2  -2  0  -1  -1  0  -4
W  -3  -3  -4  -4  -2  -2  -3  -2  -2  -3  -2  -3  -1  1  -4  -3  -2  11  2  -3  -4  -3  -2  -4
Y  -2  -2  -2  -3  -2  -1  -2  -3  2  -1  -1  -2  -1  3  -3  -2  -2  2  7  -1  -3  -2  -1  -4
V  0  -3  -3  -3  -1  -2  -2  -3  -3  3  1  -2  1  -1  -2  -2  0  -3  -1  4  -3  -2  -1  -4
B  -2  -1  3  4  -3  0  1  -1  0  -3  -4  0  -3  -3  -2  0  -1  -4  -3  -3  4  1  -1  -4
Z  -1  0  0  1  -3  3  4  -2  0  -3  -3  1  -1  -3  -1  0  -1  -3  -2  -2  1  4  -1  -4
X  0  -1  -1  -1  -2  -1  -1  -1  -1  -1  -1  -1  -1  -2  0  0  -2  -1  -1  -1  -1  -1  -4
*  -4  -4  -4  -4  -4  -4  -4  -4  -4  -4  -4  -4  -4  -4  -4  -4  -4  -4  -4  -4  -4  -4  -4  1

```

Fig. 3. The BLOSUM62 Score Matrix. Positive scores indicate similar amino acids that are frequently seen to substitute for each other in alignments. Negative scores indicate rare substitutions that typically change protein structure and function.

3.2.5. Interpreting the E Value

When comparing a query sequence against a large database, there are usually many database sequences that can be reasonably aligned to the query. But how does one determine the likelihood that an alignment represents a true evolutionary relationship or structural/functional conservation instead of being the result of random matching? This question requires a statistical basis for deciding whether the scores reflect matching of truly (biologically) similar sequences or chance events. Statistically, the theoretical distribution of random (local) alignment scores follows an extreme value distribution, which is characterized by the score matrix and the scale of the search space (reviewed in ref. 14). Based on this theory, the statistical significance of any alignment score can be evaluated by its expectation (E) value (see Note 4). High alignment scores lead to low E values. The E value provides an estimate of the number of alignments one would expect to find by chance having the same or greater alignment score. Thus, a high E value casts doubts on the biological relevance of the alignment, because an alignment of equal quality could also easily result by chance. For example, if a particular database sequence retrieved from a BLAST search against NCBI's *nr* database has an E value of 0.1, this means that 10% of an equal-sized database of randomized sequences would be expected to yield alignments of at least equal quality. The cutoff for the E value can be set by the user both at the NCBI BLAST Web server

and in the standalone *blastall* program using -e option, resulting in only alignments with this E value or lower being displayed. Users should note that reported E values are adjusted for the multiple comparisons made by searching against the many sequences in the database. Thus, the same alignment would have a higher E value when derived from querying a large database compared to the value calculated for querying a small database, resulting from the expectation that a large database may produce more random alignments by chance. Therefore, when reporting or comparing database searches, users should state both the E value and the size of the database searched (in addition to its general composition). In this context, it is also important to recall that BLAST is a local alignment tool. If two sequences share overall similarity, the “local” alignment will extend over the entire sequences. However, many proteins are modular in nature, consisting of multiple independent domains. Alignments between a query and a database sequence may be the result of a single shared domain. Therefore, search results must always be carefully analyzed and interpreted (for example, quoting a BLAST-derived similarity score is meaningless unless accompanied by a statement as to the matching segment length). Similarly, choices of query sequence length and database may critically affect the ability to identify more divergent relatives of the query (for example, a shared motif in the N-terminal half of two protein sequences may be statistically insignificant in a search with the entire query sequence, but could show up as a hit when searching with the motif segment only). In addition, BLAST (as well as many other bioinformatics tools) is only able to give the best hints to guide the direction of study, and predicted functions by BLAST will ultimately have to be tested experimentally rather than taken for granted. It has been shown that functional annotation with low BLAST E values (below 1e-50 against large protein databases) still produces errors (15).

3.2.6. Sequence Masking

If you perform a BLAST search against NCBI’s *nr* database with a poly-A DNA sequence using the Web server’s default parameters, you will not be able to retrieve any matching database sequences. However, if you uncheck the “Low complexity” filter and re-run the same poly-A query, you will see a lot of matched database sequences. The poly-A sequence is an extreme example of so-called low-complexity sequences, which have highly biased base frequencies. Similarly, some protein sequences also contain low-complexity sequences that are highly enriched in certain amino acids. Sequences with such low-complexity segments are likely to produce many false-positive database matches (14). Therefore, by default, both the BLAST Web server and the standalone *blastall* program mask low-complexity regions, replacing the original letters with the special symbol X. If your query sequence is completely masked, you will get an error message showing “ERROR: BLASTSetUpSearch: Unable to calculate Karlin-Altschul params, check query sequence.” To turn off the masking, you can set the -F option of *blastall* to *F*. This feature can become critical in certain applications (for example, if one wanted to detect homologs of proteins with glycine-rich or serine/arginine-rich domains) and illustrates the necessity for users to carefully consider whether the default options supplied by the Web server or the standalone program are appropriate for the given task. Although this is an apparently trivial remark, in practice many initially puzzling results derive from inappropriate parameter choices.

3.3. Postprocessing BLAST Output

In many applications, particularly those involving multiple query sequences, the BLAST result file can be very large. In this case, postprocessing programs are needed to help examine the output. For example, MuSeqBox (16) was designed for multiquery-sequence BLAST output examination arising from a large-scale EST annotation project. The program parses the BLAST output, extracts the informative parameters of BLAST hits, and saves this information in tabular form in either text or HTML format. The hit tables are optionally further analyzed with the program to produce subsets of BLAST hits according to user-specified criteria. For example, MuSeqBox screens the BLASTX output from EST vs EST and protein database queries, to indicate sequences that potentially represent alternatively spliced transcripts, full-length coding sequences, or contain repetitive structures. The MuSeqBox program can be downloaded freely for academic use and is also accessible online with extensive tutorials at <http://bioinformatics.iastate.edu/bioinformatics2go/mb/MuSeqBox.html>. A number of other BLAST output processing programs are available, including BEAUTY (17), PhyloBLAST (18), and Zerg (19), to facilitate alignment display, domain searches, phylogenetic analyses, and other applications.

3.4. PSI-BLAST

Regular BLAST compares the query sequence with each database sequence, and the search stops when all sequences have been compared. Therefore, the result is the set of all sequences that show direct similarity to the single query sequence. However, sometimes this is not sufficient to produce the desired result. For example, if one wishes to retrieve all the members of a certain gene family from the database, then limiting the search to any single member as query may retrieve only closely related family members. PSI-BLAST was designed to identify distant homologs that share only weak sequence similarity with the starting query sequence. PSI stands for “Position-Specific Iterated.” The strength of PSI-BLAST is its ability to derive a “profile” for a protein family. The profile characterizes the extent of conservation in certain parts of the protein sequences. This profile, instead of any individual family member, is then used to search the database, which increases the chances of retrieving all the members of the gene family by focusing on the conservation pattern. PSI-BLAST starts the search as a regular BLAST search comparing a single query sequence to the database. Then the resulting hits and the query are aligned to derive a representative profile for this presumed protein family. The profile determines position-specific amino acid substitution scores, which modulates the scores in the general scoring matrix being used (such as a BLOSUM matrix, as discussed in **Subheading 3.2.4.**). The length of the profile corresponds to the length of the initial query sequence. Each profile column corresponds to a position in the query, and each profile row records the scores of all possible residues in that position. High scores are assigned to conserved residues, whereas negative scores are assigned to weakly conserved or nonconserved residues. This profile is then used to search the database again to identify more sequences matching the newly created profile. If more sequences are detected, they are used to refine the current profile. Then the refined profile is used to search the database again. These steps are iterated many times at the user’s discretion or until no more similar sequences can be identified. There are many successful examples of using PSI-BLAST to identify distant

homologs (e.g., 20,21). NCBI also provides excellent tutorials for PSI-BLAST (<http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/psi1.html>).

Here we want to emphasize the power of the PSI-BLAST standalone version. When you install the BLAST package, you also install standalone PSI-BLAST, called *blastpgp*. The greatest advantage of using the standalone *blastpgp* is that it allows you to build a profile with one database and then search the profile against another database (see Note 5). For example, you may want to find out whether the sequence of your interest has any homologs whose three-dimensional structures are available in the protein data bank (PDB). If you simply use BLAST (or even PSI-BLAST) to query your sequence against the PDB database, it is quite possible that you will not find any significant matches. This is a result of the limited number of structures available currently relative to the number of sequences available from a given protein database. However, you can increase your chances of detecting a more distant homolog by performing a PSI-BLAST search against a large database (e.g., the NCBI *nr* protein database, with more than 1,600,000 entries now) first, then saving the profile, and subsequently using this profile to search PDB (unfortunately, you cannot perform such a strategy at the NCBI's PSI-BLAST server, which allows you to build and search the profile only against the same database). **Figure 4** shows a successful example of this strategy. The task in this example was to annotate the potential structure/function information of APPL protein that was originally identified as an adaptor molecule interacting with the oncogene serine/threonine kinase AKT2 (22). Two domains—PH (pleckstrin homology) and PTB (phosphotyrosine binding)—had been identified in the middle and C-terminus of this protein, indicating its role in signal transduction. However, its N-terminal sequences showed no obvious similarity to any known domains. Using the outlined PSI-BLAST strategy, we were able to show that the N-terminal region is significantly similar to a BAR domain region in a *Drosophila* amphiphysin protein (dAmph, residues 1 to 245), which has an available three-dimensional structure (PDB accession 1URU). Our prediction of a BAR domain is in agreement with an independent analysis by Miaczynska et al. (23) using a different motif search algorithm. In addition, a class of proteins has already been shown to have a BAR domain adjacent to a PH domain to function in enhancing lipid interactions (24). The APPL protein is possibly a new member of this protein family.

3.5. Other Programs

Before the BLAST program was created, FASTA (25,26) was the only fast database search program. The FASTA program still is in popular use today (Web server at <http://fasta.bioch.virginia.edu/fasta/>). Both FASTA and BLAST are based on heuristic search algorithms (see Note 6), directly implying that neither is guaranteed to find all optimal alignments. The algorithms sacrifice some sensitivity to gain fast search speeds, which is critical for searching against large sequence databases. For more sensitive searches, a slower program called SSEARCH (<ftp://ftp.virginia.edu/pub/fasta>), which is based on the Smith-Waterman dynamic programming algorithm (27), can be used. Currently, a new generation of database similarity search programs based on suffix tree or suffix array data structures (e.g., Vmatch, <http://www.vmatch.de/>) are showing great promise in both speed and sensitivity when solving large-scale sequencing matching tasks.

Every computer program has its scope, and so does the BLAST program (although sometimes users tend to expect everything from this single program!). Certain data-

```

>pdb|1URU|A Chain A, Amphiphysin BAR Domain From Drosophila
Length = 244

Score = 83.9 bits (206), Expect = 2e-17
Identities = 33/234 (14%), Positives = 68/234 (29%), Gaps = 20/234 (8%)

Query: 4   IDKLPIEETLEDSPQTRSLLGVFEEDATAISNYMNQLYQAMHRIYDAQNELSAATHLTSK 63
        + L + D L F N+L + + AA+
Sbjct: 25  LQNLGKVDRTADEI-FDDHLNNFNRQ---QASANRLQKEFNNYIRCVRRAAQAAASKTLXD 79

Query: 64  LLKEYEKQRFPGLGGDEVMSSTLQQFSKVIDELSSCHAVALSTQLADAMMFPISQFKERDL 123
        + E + ++ + Q + L Q+ + QF E
Sbjct: 80   SVCEIYEPQWSG-----YDALQAQGTGASESLWADFAHKLGDQVLIPLNNTYTGQFPEXKK 133

Query: 124  KEITLTLKEVQFQIASNDHAAINRYSRLSK---KRENNDKVKYEVTEDVYTSRKQHQQTMMH 180
        K +++ D+D + + L KR+D + E + +R+
Sbjct: 134  KVEKRNKRKL----IDYDGQRHSFQNLQANANRKDDVKLTKGREQLEEARRTYEILNTE 188

Query: 181  YFCALNTLQYKKKIALLEPLLGYMQAQISFFKMGSENLngQLEEFLANIGTSVQ 234
        L L + + L+ L + F + ++ + LE + + T Q
Sbjct: 189  LHDELPALYDSRILFLVTNLQTLFATEQVFHNETAKIYSE-LEAIVDKLATESQ 241

```

Fig. 4. Identification of a putative BAR domain in the mouse APPL protein by PSI- Basic Local Alignment Search Tool (BLAST). The N-terminus of the mouse APPL protein (GenBank Accession NP_060245; residues 1 to 280) was searched against the National Center for Biotechnology Information *nr* protein database with the PSI-BLAST program using default parameters and a maximum of ten iterations. The profile was saved and used to search the *pdb* protein database with the PSI-BLAST program using default parameters and a maximum of two iterations. The displayed alignment was taken from the top hit, a known BAR domain from the *Drosophila* amphiphysin protein. Positions Lys¹⁶¹, Lys¹⁶³, and Arg¹⁴⁰ of the amphiphysin BAR domain have been shown to be critical for its binding to lipid membranes by mutation studies (24). These critical amino acids are also conserved in the APPL protein.

base similarity search tasks are better handled by other specialized programs. For example, when aligning genomic DNA to EST or cDNA sequences, the ESTs or cDNAs have to be correctly aligned to display the exons while indicating the correct positions of the introns. However, BLAST can only report a list of potential exons separately, without indicating the precise borders and complete alignment. However, a number of specialized spliced-alignment programs such as GeneSequer (28; <http://bioinformatics.iastate.edu/cgi-bin/gs.cgi>) and BLAT (29; <http://genome.ucsc.edu/cgi-bin/hgBlat/>) are designed to provide such alignments.

4. Notes

1. The BLAST package is freely available at NCBI (<ftp://ftp.ncbi.nlm.nih.gov/blast/executables/LATEST>). For the Windows version, download the installation file with a name like *blast-x.x.x-ia32-win32.exe*, where “x” indicates the version number and “win” indicates that it is for Windows. Save the installation file into a folder, e.g., C:\BLAST, and then double-click it. The program will self-extract into the following contents:
 - a. Documentation files—plain text files describing how to install and execute all the programs (for example, *blast.txt* describes how to install standalone BLAST programs).
 - b. Executables—the binary files to run the programs. The *blastall* program is the core program that people refer to as the standalone BLAST program. The other programs included are *formatdb*, *fastacmd*, *copymat*, *makemat*, *rpsblast*, *seedtop*, *blastclust*,

- blastpgp, impala*. They are either BLAST-based database search programs or utility programs. This chapter covers the basic usage of the *formatdb*, *blastall*, and *blastpgp* programs. For all other usage, see the program documentation.
- c. Data—a folder containing data files used by BLAST programs. For example, amino acid score matrices used for aligning protein sequences are stored in the data folder. Now create a text file with filename *ncbi.ini* (using any editor, e.g., WordPad; the filename must be *ncbi.ini*) and type into the file on two separate lines [NCBI] and Data=C:\BLAST\data, respectively. Then put the file inside Windows' system folder C:\Windows or C:\WINNT. This file tells the BLAST programs where to locate the data folder (thus, if you change the root C:\BLAST, you will need to edit the *ncbi.ini* file accordingly). Now you have finished the installation of the standalone BLAST program. The above procedure is also described in the *blast.txt* documentation file. In addition to the *blastall* program, NCBI also provides a network-client version of the BLAST program, *Blastcl3*, which has similar uses to *blastall*, except that *Blastcl3* searches remote databases at NCBI while running on your local computer. Therefore, if you do not want to download NCBI's databases for local execution of the *blastall* program (and you won't unless you have lots of disk space available!), you can still submit (multiple) query sequences from a local file to search NCBI's databases using *Blastcl3*. Because *Blastcl3* relies on the network, sometimes it can be very slow, especially when submitting large numbers of query sequences. *Blastcl3* is part of the *netblast* package, available from <ftp://ftp.ncbi.nlm.nih.gov/blast/executables/LATEST/>. For Windows, users should download the installation file *netblast-date-ia32-win32.exe*. Double-click it for self-extraction, and you will see the executable program *Blastcl3*. The installation is done. Should any documentation files be missing from the package or get lost, you can always retrieve them from <ftp://ftp.ncbi.nlm.nih.gov/blast/documents/>.
 2. Suppose that in your C:\Database folder (or any folder of your choice) you have a protein database file named *database.fsa* (or any filename of your choice). The file could be created by you or downloaded from other places, as long as it is in FASTA format (see **Fig. 1**). For example, you can download *pdb* protein sequences from <ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/pdbaa.gz>. To build the indices of the *database.fsa* file, you need to first invoke the Windows command prompt by clicking the Start menu and select Run to pop up a small window with a text field. Type in cmd and click OK. Now you should see a black window with cursor blinking after the > sign. The > sign is called a prompt, and the blinking cursor is where you type in your computer commands (instructions to tell your computer programs what to do) to execute computer programs. Now type C:\BLAST\formatdb -i C:\Database\database.fsa -p T and then hit the Return (or Enter) key to execute the command. Notice that C:\BLAST\formatdb specifies the path to the *formatdb* program, and similarly C:\Database\database.fsa specifies the path to the *database.fsa* file. Otherwise, the computer will not be able to find the files *formatdb* and *database.fsa*. The *formatdb* program has two options (-i and -p). The -i option specifies the name of the input file. The -p option tells the *formatdb* whether the input file is a protein sequence file or a nucleotide sequence file. If protein, the -p option should be set to T; otherwise it should be set to F. Now you have finished building the indices. Notice that three index files were created. The use of *formatdb* is also described in the *formatdb.txt* documentation file.
 3. After you have installed the BLAST program (see **Note 1**) and built the indices for your database (see **Note 2**), you are ready to run a BLAST search with your query sequences against the database. Suppose you have a protein query file named *query.fsa*. You can have one or multiple protein sequences in the query file, all in FASTA format. To perform the search, you need to execute the *blastall* program by typing at the command prompt C:\BLAST\blastall -p blastp -d C:\Database\database.fsa -i query.fsa -e 1e-4 -o Blast-

- Output. The *-p* option specifies the flavor of BLAST program (see **Table 1**). The *-d* option specifies the location and name of the database file. The *-i* option specifies the location and name of the query file. The *-e* option specifies the threshold E value. Only hits with E value below the threshold will be reported. The *-o* option specifies the location and name of the BLAST output file. The output file is a text file and can be viewed using any text editor, such as *WordPad*. The use of *blastall* is also described in the *blast.txt* documentation file.
4. The score S of an alignment, also called *raw score*, is the sum of scores from all pairs of aligned residues. The score of each pair of residues is specified in the score matrix. Obviously, raw scores depend on the specific score matrix being used, and therefore are of limited value for comparisons with other alignments using different score matrices. Thus, raw scores are normalized into *bit scores* S' by the equation $S' = (\lambda S - \ln K) / \ln 2$, where λ and K are statistical parameters reported by the program. Then the expected number (E value) of alignments with normalized scores no less than S' can be approximated by $E = mn/2^{S'}$, where m is the length of the query sequence and n is the total length of the database sequences. As shown in the formula, high S' scores produce low E values, and a large database will give a higher E value for the same raw score.
 5. The standalone PSI-BLAST *blastpgp* is part of the BLAST package you have already installed (see **Note 1**). The program can save a profile that is built from a protein query search against a protein database, and then search this profile against another protein database. For example, to save a profile from your query file searched against NCBI's *nr* protein database, you need to execute the *blastpgp* program by *C:\BLAST\blastpgp -d C:\Database\nr -i query.fsa -j 10 -h 1e-4 -C query.profile -o nr.output*. The *-j* option specifies the maximum number of iterations of the PSI-BLAST search against the database. The *-h* option sets the E value threshold for database hits to be included in the profile. Only hits with E value below the threshold will be used to build the profile. The *-C* option specifies the name of the profile. Having built the profile, you can search another database with it. For example, you can search against the *pdb* database by *C:\BLAST\blastpgp -d C:\Database\pdb -i query.fsa -j 2 -R query.profile -o pdb.output*. The *-R* option reads the profile file *query.profile* (note that this is a binary file, which cannot be viewed using a text editor). The search result is located in the text file *pdb.output*. In the above example, the search is conducted against a protein database. You can also use the profile to search against a nucleotide database. In that case, you will need to use the *blastall* program to invoke another subprogram, PSIBLASTN, by *C:\BLAST\blastall -p psiblastn -d C:\Database\nucleotideDB -i query.fsa -j 2 -R query.profile -o nucleotideDB.output*. The use of PSI-BLAST is also described in the *blast.txt* documentation file.
 6. A detailed description of the BLAST algorithm is beyond the scope of this chapter. Interested readers are referred to the listed references (5,30). The central idea of the BLAST algorithm is to rapidly locate the exact-matching regions between query and database sequences and then extend those exact-matching regions only to longer alignments. Basically, BLAST generates all short subsequences from the query. These subsequences are called *words*. The length of each *word* is typically set to 3 for proteins and 11 for nucleotide sequences. The list of *words* is enlarged by including all the *neighborhood words*, defined as words that align with the original words above a certain threshold. Then BLAST scans each preindexed database sequence for exact matches to the list of words. Each exactly matched word pair is called a *hit*. If two nonoverlapping hits, each with a score of at least *T*, are found within a certain distance *A*, an ungapped extension is invoked until the alignment score cannot be improved by further extension. The resulting alignment is called a “high-scoring segment pair,” or HSP. Then if the score of an HSP is higher than a certain threshold *S*, a gapped extension is triggered to further extend the alignment in both direc-

tions until the score drops no more than some defined value X below the best score discovered so far.

Acknowledgments

This work was supported in part by NSF Plant Genome Research Projects grants DBI-9872657 and DBI-0110254. We wish to thank our colleagues Carolyn Lawrence and Trent Seigfried for critical reading and editing of earlier versions of the manuscript.

References

1. Weir, M., Swindells, M., and Overington, J. (2001) Insights into protein function through large-scale computational analysis of sequence and structure. *Trends Biotechnol.* **19**, S61–S6.
2. Konopka, A. K. (2003) Selected dreams and nightmares about computational biology. *Comp. Biol. & Chem.* **27**, 91–92.
3. Brendel, V. (2002) Integration of data management and analysis for genome research. In Schubert, S., Reusch, B., and Jesse, N. (eds.), “Informatik bewegt”. *Lecture Notes in Informatics (LNI)—Proceedings P-20*, 10–21.
4. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.
5. Altschul, S. F., Madden, T. L., Schäffer, A. A., et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402.
6. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Wheeler, D. L. (2003) GenBank. *Nucleic Acids Res.* **31**, 23–27.
7. Westbrook, J., Feng, Z., Chen, L., Yang, H., and Berman, H. M. (2003) The Protein Data Bank and structural genomics. *Nucleic Acids Res.* **31**, 489–491.
8. Higgins, D. G., Thompson, J. D., and Gibson, T. J. (1996) Using CLUSTAL for multiple sequence alignments. *Methods Enzymol.* **266**, 383–402.
9. Kumar, S., Tamura, K., and Nei, M. (1994) MEGA: Molecular Evolutionary Genetics Analysis software for microcomputers. *Comput. Appl. Biosci.* **10**, 189–191.
10. Felsenstein, J. (1989) PHYLIP—Phylogeny Inference Package (Version 3.2). *Cladistics* **5**, 164–166.
11. Vogt, G., Etzold, T., and Argos, P. (1995) An assessment of amino acid exchange matrices in aligning protein sequences: the twilight zone revisited. *J. Mol. Biol.* **249**, 816–831.
12. Henikoff, S. and Henikoff, J. G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* **89**, 10,915–10,919.
13. Dayhoff, M. O., Schwartz, R. M., and Orcutt, B. C. (1978). A model of evolutionary change in proteins. In: (Dayhoff, M. O., ed.) *Atlas of Protein Sequence and Structure*, National Biomedical Research Foundation, Washington, DC: pp. 345–362.
14. Altschul, S. F., Boguski, M. S., Gish, W., and Wootton, J. C. (1994) Issues in searching molecular sequence databases. *Nat. Genet.* **6**, 119–129.
15. Rost, B. (2002) Enzyme function less conserved than anticipated. *J. Mol. Biol.* **318**, 595–608.
16. Xing, L. and Brendel, V. (2001) Multi-query sequence BLAST output examination with MuSeq Box. *Bioinformatics* **17**, 744–745.
17. Worley, K. C., Wiese, B. A., and Smith, R. F. (1995) BEAUTY: an enhanced BLAST-based search tool that integrates multiple biological information resources into sequence similarity search results. *Genome Res.* **5**, 173–184.

18. Brinkman, F. S., Wan, I., Hancock, R. E., Rose, A. M., and Jones, S. J. (2001) PhyloBLAST: facilitating phylogenetic analysis of BLAST results. *Bioinformatics* **17**, 385–387.
19. Paquola, A. C., Machado, A. A., Reis, E. M., Da Silva, A. M., and Verjovski-Almeida S. (2003) Zerg: a very fast BLAST parser library. *Bioinformatics* **22**, 1035–1036.
20. Altschul, S. F. and Koonin, E. V. (1998) Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases. *Trends Biochem. Sci.* **23**, 444–447.
21. Jones, D. T. and Swindells, M. B. (2002) Getting the most from PSI-BLAST. *Trends Biochem. Sci.* **27**, 161–164.
22. Mitsuuchi, Y., Johnson, S. W., Sonoda, G., Tanno, S., Golemis, E. A., and Testa, J. R. (1999) Identification of a chromosome 3p14.3-21.1 gene, APPL, encoding an adaptor molecule that interacts with the oncoprotein-serine/threonine kinase AKT2. *Oncogene* **18**, 4891–4898.
23. Miaczynska, M., Christoforidis, S., Giner, A., et al. (2004) APPL proteins link Rab5 to nuclear signal transduction via an endosomal compartment. *Cell* **116**, 445–456.
24. Peter, B. J., Kent, H. M., Mills, I. G., et al. (2004) BAR domains as sensors of membrane curvature: the amphiphysin BAR structure. *Science* **303**, 495–499.
25. Lipman, D. J. and Pearson, W. R. (1985) Rapid and sensitive protein similarity searches. *Science* **227**, 1435–1441.
26. Pearson, W. R. and Lipman, D. J. (1998) Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* **85**, 2444–2448.
27. Smith, T. and Waterman, M. (1981) Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197.
28. Usuka, J., Zhu, W., and Brendel, V. (2000) Optimal spliced alignment of homologous cDNA to a genomic DNA template. *Bioinformatics* **16**, 203–211.
29. Kent, W. J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664.
30. Pertsemlidis, A. and Fondon, J. W. 3rd. (2001) Having a BLAST with bioinformatics (and avoiding BLASTphemy). *Genome Biol.* **2**, reviews 2002.1–2002.10.

Protein Identification and Analysis Tools on the ExPASy Server

Elisabeth Gasteiger, Christine Hoogland, Alexandre Gattiker, Séverine Duvaud, Marc R. Wilkins, Ron D. Appel, and Amos Bairoch

1. Introduction

Protein identification and analysis software performs a central role in the investigation of proteins from two-dimensional (2-D) gels and mass spectrometry. For protein identification, the user matches certain empirically acquired information against a protein database to define a protein as already known or as novel. For protein analysis, information in protein databases can be used to predict certain properties about a protein, which can be useful for its empirical investigation. The two processes are thus complementary. Although there are numerous programs available for those applications, we have developed a set of original tools with a few main goals in mind. Specifically, these are:

1. To utilize the extensive annotation available in the Swiss-Prot database (1) wherever possible, in particular the position-specific annotation in the Swiss-Prot feature tables to take into account posttranslational modifications and protein processing.
2. To develop tools specifically, but not exclusively, applicable to proteins prepared by two-dimensional gel electrophoresis and peptide mass fingerprinting experiments.
3. To make all tools available on the World-Wide Web (WWW), and freely usable by the scientific community.

In this chapter we give details about protein identification and analysis software that is available through the ExPASy World Wide Web server (2).

Analysis tools include Compute pI/Mw, a tool for predicting protein isoelectric point (pI) and molecular weight (Mw); ProtParam, to calculate various physicochemical parameters; PeptideMass, a tool for theoretically cleaving proteins and calculating the masses of their peptides and any known cellular or artifactual posttranslational modifications; PeptideCutter, to predict cleavage sites of proteases or chemicals in protein sequences; ProtScale, for amino acid scale representation, such as hydrophobicity plots.

Protein identification tools include TagIdent, a tool that lists proteins within a user-specified pI and Mw region, and allows proteins to be identified through the use of short “sequence tags” up to six amino acids long; AACompIdent, a program that identifies proteins by virtue of their amino acid (AA) compositions, sequence tags, pI, and Mw; AACompSim, a program that matches the theoretical AA composition of proteins against the Swiss-Prot database to find similar proteins; MultiIdent, a combination of

other tools mentioned above that accepts multiple data types to achieve identification, including protein pI, Mw, species of interest, AA composition, sequence tag, and peptide masses; and Aldente, a powerful peptide mass fingerprinting identification (PMF) tool.

Protein characterization tools in the context of PMF experiments include FindMod, to predict posttranslational modifications and single-amino acid substitutions; GlycoMod, a tool to predict the possible compositions for glycan structures, or compositions of glycans attached to glycoproteins; FindPept, to predict peptides resulting from unspecific proteolytic cleavage, protease autolysis, and keratin contaminants; and BioGraph to visualize the results of the ExPASy identification and characterization tools.

The tools described here are accessible through the ExPASy WWW server, from the tools page, <http://www.expasy.org/tools/> (see **Fig. 1**). In addition to the tools maintained by the ExPASy team, this page contains links to many analysis and prediction programs provided on Web sites all over the world. The “local” ExPASy tools can be distinguished by the small ExPASy logo preceding their name. They are continually under development and thus may change with time. We document new features of tools in the “What’s new on ExPASy” Web page at <http://www.expasy.org/history.html>. Feedback and suggestions from users of the tools is very much appreciated and can be sent by e-mail to tools@expasy.org. Detailed documentation for each of the programs is available from the Web site.

2. The Swiss-Prot Database

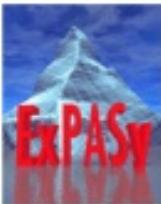
The identification tools described below all work directly and exclusively with the Swiss-Prot protein knowledgebase and its automatically annotated supplement TrEMBL (**1**). Since the maintainers of Swiss-Prot and TrEMBL (the Swiss Institute of Bioinformatics and the European Bioinformatics Institute) joined forces with the PIR group at Georgetown University to form the UniProt consortium (**3**), Swiss-Prot and TrEMBL are also known as the “UniProt Knowledgebase.”

In order to make the most of the tools, it is helpful to understand a number of concepts applied in Swiss-Prot and TrEMBL. The Swiss-Prot user manual (<http://www.expasy.org/sprot/userman.html>) provides a detailed description of the database format and scope, and complements the information in this section.

2.1. Annotation Quality

Swiss-Prot is known for its extensive manual annotation, whereas the vast majority of TrEMBL entries are unannotated or automatically annotated. This has a number of implications for the user of proteomics tools.

Identification results usually show the description (DE) line of protein entries matching the experimental data (sometimes, this description may be truncated if it is longer than the space available in the output tables). Whereas all Swiss-Prot description lines are manually created and verified to list the most common name and synonyms used for a protein, enforcing standardized nomenclature, TrEMBL DE lines usually consist of the phrase typed in by the submitter of the underlying nucleotide coding sequence, or of a protein name inferred by automatic annotation procedures. As far as keywords (KW lines) and feature tables (FT lines) are concerned, the situation is similar: all



[ExPASy Home page](#) [Site Map](#) [Search ExPASy](#) [Contact us](#) [Swiss-Prot](#) [PROSITE](#) [SWISS-2DPAGE](#)

Hosted by SIB Switzerland Mirror sites: [Australia](#) [Bolivia](#) [Canada](#) [China](#) [Korea](#) [Taiwan](#) [USA](#)

Search for

ExPASy Proteomics tools

The tools marked by  are local to the ExPASy server. The remaining tools are developed and hosted on other servers.

[\[Protein identification and characterization\]](#) [\[DNA > Protein\]](#) [\[Similarity searches\]](#) [\[Pattern and profile searches\]](#)
[\[Post-translational modification prediction\]](#) [\[Topology prediction\]](#) [\[Primary structure analysis\]](#) [\[Secondary structure prediction\]](#)
[\[Tertiary structure\]](#) [\[Sequence alignment\]](#) [\[Biological text analysis\]](#)

Protein identification and characterization

- [AACompIdent](#)  - Identify a protein by its amino acid composition
- [AACompSim](#)  - Compare the amino acid composition of a Swiss-Prot entry with all other entries
- [MultIdent](#)  - Identify proteins with *pI*, *Mw*, amino acid composition, sequence tag and peptide mass fingerprinting data
- [PeptIdent](#)  - Identify proteins with peptide mass fingerprinting data, *pI* and *Mw*. Experimentally measured, user-specified peptide masses are compared with the theoretical peptides calculated for all proteins in Swiss-Prot, making extensive use of database annotations
- [TagIdent](#)  - Identify proteins with *pI*, *Mw* and sequence tag, or generate a list of proteins close to a given *pI* and *Mw*
- [FindMod](#)  - Predict potential protein post-translational modifications and potential single amino acid substitutions in peptides. Experimentally measured peptide masses are compared with the theoretical peptides calculated from a specified Swiss-Prot entry or from a user-entered sequence, and mass differences are used to better characterize the protein of interest.
- [GlycoMod](#)  - Predict possible oligosaccharide structures that occur on proteins from their experimentally determined masses (can be used for free or derivatized oligosaccharides and for glycopeptides)
- [GlycanMass](#)  - Calculate the mass of an oligosaccharide structure
- [FindPept](#)  - Identify peptides that result from unspecific cleavage of proteins from their experimental masses, taking into account artefactual chemical modifications, post-translational modifications (PTM) and protease auto-lytic cleavage
- [PeptideMass](#)  - Calculate masses of peptides and their post-translational modifications for a Swiss-Prot or TrEMBL entry or for a user sequence
- [PeptideCutter](#)  - Predicts potential protease and cleavage sites and sites cleaved by chemicals in a given protein sequence
- [PepMAPPER](#) - Peptide mass fingerprinting tool from UMIST, UK
- [Mascot](#) - Peptide mass fingerprint, sequence query and MS/MS ion search from Matrix Science Ltd., London
- [PepSea](#) - Protein identification by peptide mapping or peptide sequencing from Protana, Denmark
- [PeptideSearch](#) - Peptide mass fingerprint tool from EMBL Heidelberg
- [ProteinProspector](#) - A variety of tools from UCSF (MS-Fit, MS-Tag, MS-Digest, etc.) for mining sequence databases in conjunction with mass spectrometry experiments [Mirrors at [UCL-Ludwig](#), UK / [Ludwig Institute Melbourne](#) (Australia)]
- [PROWL](#) - Protein chemistry and mass spectrometry resource from Rockefeller and NY Universities [or from [Genomic Solutions](#)]
- [PFMUTS](#) - Shows the possible single and double mutations of a peptide fragment from MALDI peptide mass fingerprinting

Fig. 1. The ExPASy tools page, <http://www.expasy.org/tools/>. All underlined text represents hypertext links, which, when selected with a computer mouse, take the user to the corresponding page for the chosen tool. The tools whose names are preceded by a small ExPASy logo are maintained by the ExPASy team; all other links lead to external servers.

Swiss-Prot entries are assigned a comprehensive list of keywords as part of the manual annotation process; TrEMBL, however, has very few, but automatically assigned keywords. Even more importantly for identification tools, feature tables, which contain information about known position-specific events in the sequence, such as posttranslational modifications and processing, or sequence variants, are very complete in Swiss-Prot and scarce in TrEMBL. Finally, the sequences themselves are carefully checked in Swiss-Prot and much less likely to contain errors (e.g., frameshifts) than in TrEMBL.

2.2. Alternative Splicing

Many proteins exist in more than one isoform, one cause of which is alternative (differential) splicing. Splice isoforms may differ considerably from one another, with potentially less than 50% sequence similarity between isoforms. In the Swiss-Prot database, one sequence (usually that of the longest isoform) is displayed for each protein. Known variations of this sequence are recorded in the feature table (using the VARSPLIC key), together with the name(s) of the isoform(s) in which each variant occurs. Unique and stable identifiers have been assigned to all alternative splice isoforms, and the sequences of these isoforms are distributed with Swiss-Prot. The unique splice isoform identifiers (of the form P19491-2, where P19491 is the accession number of the “original” Swiss-Prot entry, and “-2” denotes the second annotated splice isoform in that entry) can be submitted to the ExPASy analysis tools. For identification tools, the databases that constitute the search space include the alternative splice isoform sequences annotated in Swiss-Prot and TrEMBL in addition to the canonical sequences contained in those databases. For each isoform, the ExPASy server provides a page displaying the complete sequence of that isoform, with direct links to submission forms of the analysis tools described in this chapter.

2.3. Posttranslational Modifications

Posttranslational modification annotation (4,5) in Swiss-Prot, particularly in the feature table, is currently undergoing a major overhaul and standardization process. Controlled vocabularies are introduced for the feature descriptions corresponding to the feature keys MOD_RES (used for processes like phosphorylation, acetylation, sulfation, and so on), LIPID (for palmitoylation, farnesylation, geranyl-geranylation, and so on), CROSSLNK (for thioether, thioester, and other bonds) and DISULFID. This facilitates the task of reliably parsing out information about posttranslational modification events and applying the corresponding mass corrections to affected peptides. A database of modifications, containing the biological mechanism, and the conditions for occurrence (taxonomy, type of amino acid, position within the sequence) for each stored modification, is being built and will be made available via ExPASy, extensively linked to Swiss-Prot entries and proteomics tools.

It should be noted that while mass calculations can take into account known posttranslational modifications if they consist in the addition of simple groups (e.g., phosphorylation, acetylation), the algorithm used for the calculation of isoelectric points (and used by many of the tools described later) does not.

2.4. Swiss-Prot-Related Conventions for the ExPASy Tools

Unless otherwise stated, the ExPASy tools use Swiss-Prot annotations to process polypeptides to their mature forms before using them for calculations or protein identification procedures. Thus, protein signal sequences and propeptides are removed where found, and precursor molecules processed into their resulting chains.

The characterization and analysis tools described in this chapter all accept Swiss-Prot/TrEMBL identifiers (including splice isoform identifiers) as well as raw sequences as input.

When entering sequence data into text boxes for the tools, note that any spaces, newline (return) characters, and numbers will be ignored. This allows sequences in other formats, for example GCG format, to be used directly in the programs without first removing any numbering or other formatting. When using FASTA format, the first (header) line should be removed before submitting to the server.

The numbering used by the tools for amino acids in protein sequences refers to the Swiss-Prot entry. If proteins are processed to mature forms, the number of the N-terminal amino acid will remain the same as it was in the unprocessed protein sequence.

2.5. Stability of Swiss-Prot Entry Names Is Not Guaranteed

In 2004, the format of Swiss-Prot entry names (ID) will be extended from 4letters/underscore/5letters to at most 5letters/underscore/5letters. We have never claimed that Swiss-Prot IDs are stable, and have always strongly recommended the use of primary accession numbers instead. The months following the publication of this chapter will see a particularly large number of ID changes, as a result of this format change. Here, we identify all Swiss-Prot entries by their ID and AC, but would like to insist that the only identifiers whose stability we can guarantee are the accession numbers.

3. Single-Protein Analysis Tools on the ExPASy Server

3.1. Compute pI/Mw Tool

This tool (http://www.expasy.org/tools/pi_tool.html) calculates the estimated pI and Mw of a specified Swiss-Prot/TrEMBL entry or a user-entered AA sequence (see **Notes 1, 2**). These parameters are useful if you want to know the approximate region of a 2-D gel where a protein may be found.

To use the program, enter one or more Swiss-Prot/TrEMBL identification names (e.g., LACB_BOVIN) or accession numbers (e.g., P02754) into the text field, and select the “click here to compute pI/Mw” button. If one entry is specified, you will be asked to specify the protein’s domain of interest for which the pI and mass should be computed. The domain can be selected from the hypertext list of features shown, if any, or by numerically specifying the domain start and end points.

If more than one Swiss-Prot/TrEMBL identification name is entered, all proteins will automatically be processed to their mature forms, and pI and Mw values calculated for the resulting chains or peptides. If only fragments of the protein of interest are available in the database, no result will be given and an error message will be shown to highlight that the pI and mass cannot be returned accurately. Some database entries

have signal sequences or transit peptides of unknown length (e.g., Q00825; ATPI_ODOSI). In those cases, an average-length signal sequence or transit peptide is removed before the pI and mass computation is done (*see Note 3*). In Swiss-Prot release 42.6 of 28-Nov-2003, the average signal sequence length is 22 amino acids for eukaryotes and viruses, 26 amino acids for prokaryotes and bacteriophages, and 31 for archaeabacteria. Transit peptides have an average length of 57 amino acids in chloroplasts, 34 for mitochondria, 34 for microbodies, and 65 for cyanelles.

If your protein of interest is not in the Swiss-Prot database, you can enter an AA sequence in standard single letter AA code into the text field, and select the “click here to compute pI/Mw” button. The predicted pI and Mw of your sequence will then be displayed. A typical output from the program is shown in **Fig. 2A**.

Alternatively to the verbose html output, the result for a list of Swiss-Prot/TrEMBL entries can also be retrieved in a numerical format, with minimal documentation. A file containing four columns—ID, AC, pI, and Mw—is generated and can be loaded into an external application, such as a spreadsheet program. A typical file output is shown in **Fig. 2B**.

3.2. *ProtParam* Tool

3.2.1. Using *ProtParam*

ProtParam (<http://www.expasy.org/tools/protparam.html>) computes various physico-chemical properties that can be deduced from a protein sequence. No additional information is required about the protein under consideration. The protein can either be specified as a Swiss-Prot/TrEMBL accession number or ID, or in the form of a raw sequence. White space and numbers are ignored. If you provide the accession number of a Swiss-Prot/TrEMBL entry, you will be prompted with an intermediary page that allows you to select the portion of the sequence on which you would like to perform the analysis. The choice includes a selection of mature chains or peptides and domains from the Swiss-Prot feature table (which can be chosen by clicking on the positions), as well as the possibility to enter start and end position in two boxes. By default (i.e., if you leave the two boxes empty) the complete sequence will be analyzed (*see Note 4*).

3.2.2. The Calculated Parameters

The parameters computed by *ProtParam* include the molecular weight, theoretical pI, amino acid composition, atomic composition, extinction coefficient, estimated half-life, instability index, aliphatic index, and grand average of hydropathicity (GRAVY). Molecular weight and theoretical pI are calculated as in Compute pI/Mw. The amino acid and atomic compositions are self-explanatory. All the other parameters will be explained below.

3.2.2.1. EXTINCTION COEFFICIENTS

The extinction coefficient indicates how much light a protein absorbs at a certain wavelength. It is useful to have an estimation of this coefficient for following a protein which a spectrophotometer when purifying it (*see Note 5*)

It has been shown (6) that it is possible to estimate the molar extinction coefficient of a protein from knowledge of its amino acid composition. From the molar extinction coefficient of tyrosine, tryptophan, and cystine (cysteine does not absorb appreciably at wavelengths >260 nm, while cystine does) at a given wavelength, the extinction

ALACB_BOVIN (P02754)

DE Beta-lactoglobulin precursor (Beta-LG) (Allergen Bos d 5).
 OS Bos taurus (Bovine).

The parameters have been computed for the following feature:
 FT CHAIN 17 178 BETA-LACTOGLOBULIN.

Considered sequence fragment:

1	11	21	31	41	51		
1	1	1	1	1	1		
61	LIVT	QTMKGLDIQQ	VAGTNWYSLAM	AASDISLLDA	QSAPLRLVYVE	60	
61	ELKPTPEGDL	EILLQKWWENG	ECAQKIIIAE	KTKIPAVPKI	DALNENKVLV	LDTDYKKYLL	120
121	FCMENSAEPE	QSLACQCLVR	TPEVDDEALE	KFDKALKALP	MHIRLSFNPT	QLEBQCHI	

Molecular weight: 18281.21

Theoretical pI: 4.83

B

ARS1_MOUSE	054984	4.80	39065.08
ARSA_MOUSE_1	P50428	5.50	52173.26
ARSB_MOUSE	P50429	FRAGMENT	0.00
ARX_MOUSE	O35085	5.19	58504.37
ARY1_MOUSE	P50294	5.10	33713.36
ARY2_MOUSE	P50295	5.63	33701.41
ARY3_MOUSE	P50296	6.07	33685.69
ASAH_MOUSE_1	Q9WV54	6.11	13797.05
ASAH_MOUSE_2	Q9WV54	8.87	29017.27

Fig. 2. (A) Sample output from the Compute pI/Mw tool, where the program was requested to calculate the theoretical pI and Mw for the Swiss-Prot entry LACB_BOVIN (P02754). Note that the Compute pI/Mw tool shows the sequence of the region of the protein that is under consideration. In this case, the sequence of the mature beta-lactoglobulin is shown, which results when the secretion signal sequence is removed from the precursor polypeptide. (B) Output file sample retrieved from the Compute pI/Mw tool, where the program was requested to calculate the theoretical pI and Mw for a list of Swiss-Prot/TrEMBL entries. Note that the numerical format is minimal, to be exported into an external application. If pI and Mw cannot be computed, a value of "0.00" appears in the Mw column, and the reason for this is displayed in the pI column in the form of a code, the meaning of which is as follows:

FRAGMENT Incomplete CHAIN/PEPTIDE: pI/Mw cannot be computed

UNDEFINED Unknown start- or endpoints: pI/Mw cannot be computed

XXX Sequence contains several consecutive undefined AA: pI/Mw cannot be computed

If a Swiss-Prot/TrEMBL entry has one or more mature chains/peptides documented, this is indicated by "_1", "_2", etc. appended to the ID. An appended "_1," "_2," and so on, indicates that the considered sequence is that corresponding to the first, second, and so on, CHAIN or PEPTIDE documented in the feature table.

coefficient of a denatured protein can be computed (see Note 6). Two tables are produced by ProtParam, the first one showing the computed values based on the assumption that all cysteine residues appear as half cystines, and the second one assuming that no cysteine appears as half cystine.

3.2.2.2. IN VIVO HALF-LIFE

The half-life is a prediction of the time it takes for half of the amount of protein in a cell to disappear after its synthesis in the cell. The prediction is given for three organisms (human, yeast, and *E. coli*), but it is possible to extrapolate the result to similar organisms. ProtParam estimates the half-life by looking at the N-terminal amino acid of the sequence under investigation (see Note 7).

3.2.2.3. INSTABILITY INDEX (II)

The instability index provides an estimate of the stability of your protein in a test tube. It can be predicted as described in Note 8. A protein whose instability index is smaller than 40 is predicted as stable; a value above 40 predicts that the protein may be unstable.

3.2.2.4. ALIPHATIC INDEX

The aliphatic index of a protein is defined as the relative volume occupied by aliphatic side chains (alanine, valine, isoleucine, and leucine). It may be regarded as a positive factor for the increase of thermostability of globular proteins. Note 9 details how the aliphatic index is computed.

3.2.2.5. GRAND AVERAGE OF HYDROPATHY

The grand average of hydropathy (GRAVY) value for a peptide or protein is calculated as the sum of hydropathy (7) values of all the amino acids, divided by the number of residues in the sequence.

3.3. PeptideMass

The PeptideMass tool (<http://www.expasy.org/tools/peptide-mass.html>) is designed to assist in peptide-mapping experiments, and in the interpretation of peptide-mass fingerprinting (PMF) results and other mass-spectrometry data (8) (see Note 10). It cleaves *in silico* a user-specified protein sequence or a mature protein in the Swiss-Prot/TrEMBL databases with an enzyme or reagent of choice, to generate peptides. Masses of the peptides are then calculated and displayed. If a protein from Swiss-Prot has annotations that describe discrete posttranslational modifications (specifically acetylation, amidation, biotinylation, C-mannosylation, formylation, farnesylation, γ -carboxy glutamic acid, geranyl-geranylation, lipoyl groups, *N*-acyl glycerides, methylation, myristoylation, NAD, *O*-GlcNAc, palmitoylation, phosphorylation, pyridoxyl phosphate, pyrrolidone carboxylic acid, or sulfation), the masses of these modifications will be considered in peptide mass calculations (see Note 11). Post-translational modifications can also be specified along with a user-entered sequence that is not in Swiss-Prot or TrEMBL. Guidelines for the input format of posttranslational modifications (PTMs) are accessible directly from the PeptideMass input form (see Note 12). The mass effects of artifactual protein modifications such as the oxidation of methionine or acrylamide adducts on cysteine residues can also be considered. The program can supply warnings where peptide masses may be subject to change from protein isoforms, database conflicts, or mRNA splicing variation.

To use the program, enter one or more Swiss-Prot identification names (e.g., TKN1_HUMAN) or any Swiss-Prot/TrEMBL accession number (e.g., P20366) into the text field, or enter a protein sequence of interest using the standard one-letter AA code. User-specified sequences should not contain the character X, but can contain the

character *J*, to represent either Ile or Leu, which are of the same mass. The enzyme or reagent to use to theoretically cleave the protein sequence should then be specified, and whether any missed cleavages should be allowed. You can select to exclude masses below a certain threshold (e.g., 500 Daltons) which might be too small to be visible in a mass spectrum. The PeptideMass output will include the portions of the sequence covered by only the fragments that are above that threshold. Special treatment (if any) of cysteine residues or oxidation of methionine should be selected, and whether results are desired as monoisotopic or average masses. Finally, click on the “Perform” button to send data to the program. **Figure 3** shows a typical output of the program PeptideMass, illustrating some of its features.

3.4. PeptideCutter

PeptideCutter (<http://www.expasy.org/tools/peptidecutter/>) predicts cleavage sites of proteases or chemicals in a protein sequence. Protease digestion can be useful if one wants to carry out experiments on a portion of a protein, separate the domains in a protein, remove a tag protein when expressing a fusion protein, or make sure that the protein under investigation is not sensitive to endogenous proteases. One or several reagents can be selected from a list of (currently) 33 proteases and chemicals.

The protein sequence can be entered in the form of a Swiss-Prot/TrEMBL accession number, a raw sequence, or a sequence in FASTA format, in one-letter amino acid code. Letters that do not correspond to an amino acid code (*B*, *J*, *O*, *U*, *X*, or *Z*) will cause an error message, and the user is required to correct the input. Please note that only one sequence can be entered at a time.

You have the possibility to select one or a group of enzymes and chemicals. Most of the cleavage rules for individual enzymes were deduced from specificity data summed up by Keil (9), and the rules are listed as part of the PeptideCutter documentation. You can also ask the program to consider only enzymes that cut the sequence a chosen number of times, which may be of particular interest if you have selected a large number of cleavage agents.

For the enzymes trypsin and chymotrypsin, enough experimental data were available to study cleavage by these enzymes at sites different from the otherwise widely used motifs “after K/R but not before P” for trypsin, and “after F/Y/W but not before P” for chymotrypsin. Keil (9) created probability tables for cleavage between all pairs of amino acids in the positions N- and C-terminal to the cleavage site. This more “sophisticated” model is available for use with PeptideCutter, and with this option, the output includes the cleavage probability for any potential site.

For the display of your results, there are three different output options:

1. For every selected enzyme, the number of cleavage sites and their positions are enumerated in a first table, in alphabetical order by enzyme name.
2. The second table (not displayed by default) displays all cleavage sites sequentially in the sequence, one site per line. For each position, the cleaving enzyme, the resulting peptide sequence, the peptide length, and its mass are listed (see Note 13). The peptides displayed are calculated based on the assumption that all chosen enzymes are present during digestion. If you want to have a list of peptides resulting from cleavage by only one enzyme or chemical, select only this enzyme and deselect all others.
3. Finally, there is also the possibility to show all results using a map. The entered protein sequence is marked with a “|” above an amino acid when there is a cleavage site between

You have selected TKN1_HUMAN (P20366) from Swiss-Prot:
 Protachykinin 1 precursor (PPT) [Contains: Substance P; Neurokinin A (NKA) (Substance K) (Neuromedin L); Neuropeptide K (NPK); Neuropeptide gamma; C-terminal flanking peptide].
 Signal and propep in positions 1-56 have been removed.

- Peptide SUBSTANCE P at positions 58 - 68 [Theoretical pl: 11.00 / Mw (average mass): 1348.63]

mass	position	#MC	modifications	peptide sequence
1349.6362	58-68	0	AMID: 68	1348.6515 RPKPQQFFGLM

100.0% of sequence covered (you may modify the input parameters to display also peptides < 500 Da):

1 11 21 31 41 51
 1 | | | | | |
 61 PQQFFGLM RPK 60

- Peptide NEUROPEPTIDE K at positions 72 - 107 [Theoretical pl: 8.40 / Mw (average mass): 3981.54]

mass	position	#MC	modifications	peptide sequence
1211.3653	86-96	0		ALYGHGQQISHK
870.0077	100-107	0	AMID: 107	869.0230 TDSFVGLM
864.8841	72-79	0		DADSSIEK
671.8577	80-85	0		QVALLK

91.7% of sequence covered (you may modify the input parameters to display also peptides < 500 Da):

61 71 81 91 101 111
 1 | | | | | |
 61 DADSSIEKQ VALLKALYGH GOISHKrhkT DSFVGLM

- NEUROPEPTIDE GAMMA 2ND PART at positions 89 - 107 [Theoretical pl: 9.99 / Mw (average mass): 2135.43]

mass	position	#MC	modifications	peptide sequence
870.0077	100-107	0	AMID: 107	869.0230 TDSFVGLM
863.9511	89-96	0		GHGQISHK

84.2% of sequence covered (you may modify the input parameters to display also peptides < 500 Da):

61 71 81 91 101 111
 1 | | | | | |
 61 GH GOISHKrhkT DSFVGLM

Fig. 3. Sample output from the PeptideMass tool. The protein selected was TKN1_HUMAN (P20366), and the program was requested to cleave with trypsin, show all peptides, sort peptides by mass, show all known modifications, use average masses, and display masses as $(M + H)^+$. The figures in the "modified mass" column in this case show the predicted masses of peptides known to be amidated. Note that there are three lists of peptides, which correspond to the cleavage of different products known to be created from the same initial polypeptide. Underlined text and numbers represent hypertext links in the output that, if selected, show either the Swiss-Prot entry for the protein (e.g., TKN1_HUMAN), or the sequence of any portion of a protein specified with numbers (e.g., 58-68), or a relevant section of the online documentation (e.g., modifications). The feature table of the entry P20366 describes three more mature peptides, for which the program output is not shown here.

this amino acid and the neighboring amino acid in the C-terminal direction (i.e., directly on the “right side” of the marked amino acid). The sequence map is displayed in portions of 10 to 60 amino acids. The number of amino acids displayed per line can be modified, which may be particularly useful when printing out the map. If you have selected several enzymes and find the map too overloaded, it is possible to reduce the information and display only the cleavage sites of one enzyme by clicking on its name in the map.

3.5. ProtScale

ProtScale (<http://www.expasy.org/tools/protscale.html>) allows computation and representation (in the form of a 2-D plot) of the profile produced by any amino acid scale on a selected protein (see **Note 14**). ProtScale can be used with 50 predefined scales entered from the literature. The scale values for the 20 amino acids, as well as a literature reference, are provided on ExPASy for each of these scales. To generate data for a plot, the protein sequence is scanned with a sliding window of a given size. At each position, the mean scale value of the amino acids within the window is calculated, and that value is plotted for the midpoint of the window.

You can set several parameters that control the computation of a scale profile, such as the window size, the weight variation model, the window edge relative weight value, and scale normalization.

3.5.1. Window Size

The window size is the length of the interval to use for the profile computation, i.e., the number of amino acids examined at a time to determine a point of hydrophobic character. When computing the score for a given residue i , the amino acids in an interval of the chosen length, centered around residue i , are considered. In other words, for a window size n , we use the $i-(n-1)/2$ neighboring residues on each side of residue i to compute the score for residue i . The score for residue i is the sum of the scale values for these amino acids, optionally weighted according to their position in the window.

One should choose a window that corresponds to the expected size of the structural motif under investigation: A window size of 5 to 7 is appropriate for finding hydrophilic regions that are likely to be exposed on the surface and may potentially be antigenic. Window sizes of 19 or 21 will make hydrophobic, membrane-spanning domains stand out rather clearly (typically >1.6 on the Kyte-Doolittle scale [7]).

3.5.2. Relative Weight of the Window Edges

The central amino acid of the window always has a weight of 100%. By default, the amino acids at the remaining window positions have the same weight, but you can attribute a larger weight (in comparison with the other residues) to the residue at the center of the window by setting the weight value for the residues at the extremities of the interval to a value between 0 and 100%. The decrease in weight between the center and the edges will either be linear or exponential, depending on the setting of the weight variation model option. The ProtScale documentation includes graphic illustrations of the two available models.

3.5.3. Scale Normalization

You can choose whether to use the unmodified selected scale values from the literature or to normalize the values so that they all fit into the range from 0 to 1. Normalization is useful if you want to compare the results of profiles obtained with different scales, and makes plots with a more uniform appearance.

3.5.4. Interpreting Results

The method of sliding windows, and hence ProtScale, only provides a raw signal and does not include interpretation of the results in terms of a score. When interpreting the results, one should consider only strong signals. In order to confirm a possible interpretation, one could slightly change the window size, or replace the scale by another similar one (e.g., two different hydrophobicity scales), and ensure that the strong signal is still present.

4. Protein Identification and Characterization Tools on ExPASy

4.1. TagIdent Tool

The TagIdent tool (<http://www.expasy.org/tools/tagident.html> [10,11]) serves two main purposes. Firstly, it can create lists of proteins from one or more organisms that are within a user-specified pI or Mw range (see Note 15). This is useful to find proteins from the database that may be in a region of interest on a 2-D gel. Secondly, the program can identify proteins from 2-D gels by virtue of their estimated pI and Mw, and a short protein “sequence tag” of up to six amino acids. The sequence tag can be derived from protein N-termini, C-termini, or internally, and generated by chemical- or mass-spectrometric sequencing techniques. As sequence tags are highly specific (e.g., there are 160,000 different combinations of four amino acid sequence tags) they represent a form of protein identification that is useful for organisms that are molecularly well defined and have a relatively small number of proteins (e.g., *Escherichia coli* or *Saccharomyces cerevisiae*). Interestingly, we have shown that C-terminal sequence tags are more specific than N-terminal tags (11); however, it remains technically more difficult to generate high-quality C-terminal protein sequence data. Thirdly, the sequence tag can be used, together with a very precise protein mass obtained from mass spectrometry, to identify a protein after peptide fragmentation. One can also specify terms to limit the search to a range of organisms or to a specific organism (species). Additionally, Swiss-Prot keywords can also be used in the identification procedure.

4.1.1. Use of TagIdent to List Proteins in a Defined pI and/or Mw Region

TagIdent can generate a list of proteins in a pI and Mw range of interest, which can be sent to the user by e-mail, if a valid e-mail address is specified, or displayed in the browser window. Queries can usually be dealt with within a few seconds in your browser window. However, if you wish to submit many different queries, you may prefer to receive the results by e-mail, for easier archiving, and also in order not to have to wait for the result of one query to come back before submitting the next one. When many queries are submitted and the results are requested to be sent by e-mail, the server schedules the execution of the different identification tasks in such a way that the server CPU remains easily accessible to other users. If desired, a name can be given to your query, which will appear as the subject of the e-mail message, or at the top of the result page. This is useful for archiving purposes or if many different queries are to be submitted to the program at the same time. You should then specify the pI and Mw regions within which you would like to search (e.g., pI of 5.5 ± 0.5 units and Mw $20,000 \pm 10\%$). If you would like to search using only one of the pI or Mw parameters, you can specify an unrestricted window to cover all possibilities for the other parameter (see Note 16). For example, a search where pI is set to 7.0 ± 5 units but where a

Mw window of $20,000 \pm 10\%$ is used, will return all proteins of sizes 18,000 to 22,000 Da, regardless of their pI. In the search, you can specify one or more terms matching those in the Swiss-Prot OS (species) or OC (classification) lines to limit the search to one organism, or a range of organisms. A document containing a full list of all Swiss-Prot species can be found at <http://www.expasy.org/cgi-bin/specist>. Thus if you want to investigate proteins exclusively from *S. cerevisiae*, you can specify “cerevisiae.” This is better than specifying “yeast,” a word common to the classification of many yeasts, which includes not only proteins from *Saccharomyces cerevisiae*, but also those from *Candida albicans* and *Schizosaccharomyces pombe*. The same applies for *Homo sapiens*, where “sapiens” will search only for human proteins while “human” will include proteins from human viruses. If you would like to investigate proteins from a broader range of species, it is possible to specify a classification like “mammalia,” which will return all mammalian proteins within the specified pI and Mw region. Use of the word *ALL* will search all species in the database; however, this is not recommended, given the size of protein databases, unless all other input data are extremely specific. If desired, searches can also be restricted through use of a Swiss-Prot keyword, such as *Plasmid* or *Alzheimer’s disease*. A document containing a full list of all Swiss-Prot keywords can be found at <http://www.expasy.org/cgi-bin/keylist.pl>. By clicking on one of the keywords in this list, one obtains the definition of the keyword’s usage in Swiss-Prot, its mapping to GeneOntology terms (GO) (12) (if any), the keyword hierarchy and category, as well as a list of all Swiss-Prot entries annotated with that keyword. Keywords will be used only to restrict searches in Swiss-Prot. Any specified keyword will be ignored for TrEMBL, whose keyword annotation is only partial, and largely created by automated procedures without any manual intervention. Finally, select the “Start TagIdent” button to submit the request to ExPASy.

4.1.2. Use of TagIdent to Identify Proteins From a 2-D Gel

TagIdent can identify proteins by matching sequence tags against proteins in Swiss-Prot from one or more species within a specified pI and Mw range (see Note 17). To use TagIdent for identification purposes, first specify the pI and Mw of the protein of interest as estimated from the 2-D gel. Then specify error margins that reflect the known accuracy of these estimates. (See ref. 11 for an example of how pI and Mw ranges can be defined.) The species and keyword in the database to match against should then be specified (see Subheading 4.1.1.), the “Tagging” option selected by clicking in the small box, and the sequence tag entered in single-amino acid code in the “Tag” text box. Note that the sequence tag can contain one or more X to represent any unknown amino acid. Finally, you should specify the source of your protein sequence (N-, C-terminal, or internal), such that the program can show the protein area of interest in the search results. Thus, for example, if you have generated an N-terminal protein sequence tag by Edman degradation, you should request the program to show predicted protein N-termini. Finally, submit the search to the ExPASy server by selecting the “Start TagIdent” button. A typical output is shown in Fig. 4.

4.1.3. Interpretation of TagIdent Results for Protein Identification

Accurate identification of proteins with sequence tags relies on all proteins from an organism being in sequence databases. In this manner, if only one protein within a given pI and Mw range is found to contain a certain N-, C-terminal, or internal sequence

```

Search performed in Swiss-Prot with following values:
  pI =      5.97
  delta-pI =  0.50
  Mw =      45098
  delta-Mw = 9019
  OS or OC =  ESCHERICHIA COLI
  KW keyword =  ALL
  Display the N-terminal sequence.
  Tag = MDQT
-----
Scan done on 11-Dec-2003.
Swiss-Prot Release 42.6 of 28-Nov-2003: 139947 entries
-----
462 proteins found in the specified pI/Mw ranges
Results with tagging: 1 found
-----
The number before the sequence indicates the position in the
mature protein where your tag MDQT has been found (first occurrence).
If the protein displayed results from the processing of a
precursor, the position of the tag in the precursor polypeptide
will be given in brackets.
The sequence tag itself is printed in lowercase.
---
DHE4_ECOLI (P00370)
  NADP-specific glutamate dehydrogenase (EC 1.4.1.4) (NADP-GDH).
  pI: 5.98, MW: 48581.37
1  mdqtySLESFLNHVQKRDPNQTEFAQAVREVMTLWPFL...
---
Results without tagging: 461 found
(Printing the N-terminal sequence)
---
AAT_ECOLI (P00509)
  Aspartate aminotransferase (EC 2.6.1.1) (Transaminase A) (ASPAT).
  pI: 5.54, MW: 43573.36
  MFENITAAPADPII1GLADLFRADEPGKINLGIGVYKDET...
ABGA_ECOLI (P77357)
  Aminobenzoyl-glutamate utilization protein A.
  pI: 5.51, MW: 46588.22
  MESLNQFVNSSLAPKLSHWRRDFHHYAESGWVEFRTATLVA...
...

```

Fig. 4. Output of the TagIdent tool where it was used for protein identification. A protein from an *Escherichia coli* 2-D gel was uniquely identified by virtue of its N-terminal sequence tag, estimated pI, and mass. Although the program was requested here to display protein N-termini, it will show any protein that carries a specified tag in the “results with tagging” list, be it found at a protein N-terminus, C-terminus, or internally. Here the identification of the protein as DHE4_ECOLI (P00370) is convincing not only because the tag is at the amino terminus, but because the tag was not found anywhere in the sequence of the other 462 proteins also within the specified pI and Mw window. The TagIdent output has been shortened for this figure. Note that this approach can also be used where the mass of an entire protein has been accurately determined by mass spectrometry. In such a case, the mass window used for searching can be quite small (e.g., mass \pm 0.5%).

tag, one can be confident that there is no other, as yet undescribed protein that could otherwise match the tag (see **Note 18**). In fully sequenced organisms, the procedure is thus self-checking. Because of this, the TagIdent approach is very useful for organisms whose genomes are known, such as *Haemophilus influenzae*, *Mycoplasma genitalium*,

Methanococcus jannaschii, *Escherichia coli*, and even the eukaryote *Saccharomyces cerevisiae*. A TagIdent output for a protein from *E. coli* is shown in **Fig. 4**, and illustrates the specificity of the approach. Caution is advised when using TagIdent for the identification of proteins from poorly molecularly defined organisms, or organisms that contain large numbers of proteins (e.g., human) (see **Note 19**). A four-amino acid sequence tag (of which there are 160,000 different combinations) can be unique in microorganisms that have a total protein count of 500 to 6000, but less useful in human, for example, which has about 25,000 different known genes, and many more (possibly up to 100,000) different proteins resulting from alternative splicing. If protein identification results with TagIdent show more than one protein carrying the sequence tag in the expected region, the same sequence tag, pI, and Mw data can be used in conjunction with protein AA composition for identification with the AACompIdent tool (see **Subheading 4.2.**).

4.2. AACompIdent Tool

The AACompIdent tool (<http://www.expasy.org/tools/aacomp/> [13]) can identify proteins by their amino acid (AA) composition. The program matches the percent empirically measured AA composition of an unknown protein against the theoretical percent AA compositions of proteins in the Swiss-Prot/TrEMBL database. A score, which represents the degree of difference between the composition of the unknown protein and a protein in the database, is calculated for each database entry by the sum of the squared difference between the percent AA composition for all amino acids of the unknown protein and the database entry. All proteins in the database are then ranked according to their score, from lowest (best match) to highest (worst match). Estimated protein pI and Mw, as well as species of interest and keyword, can also be used in the identification procedure.

4.2.1. Basic Use of the AACompIdent Tool

After selecting the AACompIdent tool from the ExPASy Tools page, you must first choose the relevant AA constellation to use in matching. For AA compositions determined by standard methods, use Constellation 2. This constellation is for 16 AAs (Asx, Glx, Ser, His, Gly, Thr, Ala, Pro, Tyr, Arg, Val, Met, Ile, Leu, Phe, Lys), does not consider Cys or Trp, and calculates Asn and Asp together as Asx, and Glu and Gln together as Glx (see **Note 20**). You should specify the e-mail address to which the results should be sent, then scroll down to the “Unknown Protein” field. Here you should specify a name for the search that will appear as the subject of the e-mail message, the protein pI and Mw estimated from the 2-D gel, as well as error ranges that reflect the accuracy of these estimates. You should also specify a keyword if appropriate (see **Subheading 4.1.1.** and the Swiss-Prot list of keywords, <http://www.expasy.org/cgi-bin/keylist.pl>), and one or more terms matching those in the Swiss-Prot OS (species) or OC (classification) lines to limit the search to one organism, or a range of organisms (see **Subheading 4.1.1.** and the Swiss-Prot list of species abbreviations, <http://www.expasy.org/cgi-bin/speclist>). Matching can also be done against all species in the database by specifying “ALL.” Finally, specify the experimentally determined AA composition of the protein, with compositional data expressed as molar percent. If you have analyzed a calibration protein in parallel with unknowns as part of your AA analysis procedure, the composition of this protein can be used to compensate for error

inherent to the AA analysis procedure (*see Note 21*). To do this, go to the “Calibration Protein” field, specify the Swiss-Prot ID name for the protein (e.g., ALBU_BOVIN for bovine serum albumin) and enter the experimentally determined AA composition of the protein, with data expressed as molar percent. Finally, select the “Run AACompIdent” button to submit the data to the ExPASy server. Results will be sent to your e-mail address.

4.2.2. Use of the AACompIdent Tool With Sequence Tags

Protein samples from 2-D gels can be submitted to Edman protein sequencing to create a sequence tag of three or four amino acids, after which the same protein sample can be used for AA composition analysis (14). This approach provides protein identification of higher confidence than identification by amino acid composition analysis alone. To use AA composition and sequence tag data together for protein identification, fill in the AACompIdent form as for the basic use described above but do not immediately submit it to ExPASy. Go to the bottom of the form, select the tagging option by clicking in the small box, and enter a protein sequence tag of up to six amino acids in single-AA code into the “Tag” text field. Finally, specify whether the sequence tag is N- or C-terminal, and select the “Run AACompIdent” button to submit the data to the ExPASy server. Results will be sent to your e-mail address.

4.2.3. Interpretation of AACompIdent Results

The output of AACompIdent contains three lists of proteins ranked according to their AA score (Fig. 5). The first list is the result of matching the AA composition of the query protein against all proteins from the species of interest that have the specified keyword (if any), but without considering the specified pI and Mw. The second list shows the result of matching the AA composition of the query protein against all proteins from all species in Swiss-Prot that have the specified keyword, again without considering pI and Mw. The third list contains the results of matching the AA composition of the query protein only against the proteins from the species of interest that lie within the specified pI and Mw range (*see Note 15*) and that also have the appropriate keyword. The third list is the most powerful search. In all lists, a score of 0 indicates a perfect match between the query protein and a protein in the database, with larger scores indicating increasing difference.

We have found that a top-ranked protein is likely to represent a correct identification if it meets three conditions (Fig. 5). Firstly, the same protein, or type of protein, should appear at the top of the three lists. Secondly, the top-ranked protein in the third list should have a score lower than 30 (indicating a “good fit” of the query protein with that database entry). Finally, the third list should show a large score difference between the top-ranked protein and the second ranked protein (indicating a unique matching of the query protein with the top-ranked database entry). For proteins from *E. coli*, we have shown that a score difference greater than a factor of 2 gives high confidence that the top-ranked protein represents the correct identity (13). If the top-ranking protein in the results does not meet these three conditions, the correct identity is often within the list of best-matching proteins (*see Note 22*). In such cases, the use of AACompIdent with a protein sequence tag can provide unambiguous identification due to the high specificity of sequence tag data (14). **Figure 5** shows the result of protein identification by AA composition, pI, Mw, species, and sequence tag. Note that when the sequence tag

SEARCH VALUES:
 Constellation 2
 Species searched: ESCHERICHIA COLI
 Keyword searched: ALL
 Name given to unknown protein: coli147
 pI: 5.70 Range: (5.20, 6.20)
 Mw: 34894 Range: (27916, 41872)
 Calibration protein: OVAL_CHICK (P01012)
 Tag= MKVA
 An asterisk (*) is printed to the left of a protein's rank if it carries the sequence tag.

 Scan the Swiss-Prot database (139947 entries)

The closest Swiss-Prot entries (in terms of AA composition) for the species ESCHERICHIA COLI:

Rank	Score	Protein	(pI	Mw)	Description
*	1	MDH_ECOLI	5.61	32337	Malate dehydrogenase (EC 1.1.1.37).
	2	YHDH_ECOLI	5.63	34724	Protein yhdH.
	3	K6P2_ECOLI	5.75	32388	6-phosphofructokinase isozyme 2
	4	ALKH_ECOLI	5.57	22284	KHG/KDPG aldolase
	5	YEIN_ECOLI	5.37	32910	Hypothetical protein yein.

The closest Swiss-Prot entries (in terms of AA composition) for any species:

Rank	Score	Protein	(pI	Mw)	Description
*	1	MDH_SALTY	6.02	32451	Malate dehydrogenase (EC 1.1.1.37).
*	2	MDH_ECOLI	5.61	32542	Malate dehydrogenase (EC 1.1.1.37).
*	3	MDH_PHOPR	5.15	32391	Malate dehydrogenase (EC 1.1.1.37).
*	4	MDH_HAEIN	5.86	32542	Malate dehydrogenase (EC 1.1.1.37).
	5	PUR2_CHICK	7.51	106543	Trifunctional purine biosynthetic

The closest Swiss-Prot entries (in terms of AA composition) and having pI and Mw values in the specified range for the species ESCHERICHIA COLI:

Rank	Score	Protein	(pI	Mw)	Description
*	1	MDH_ECOLI	5.61	32337	Malate dehydrogenase (EC 1.1.1.37).
	2	YHDH_ECOLI	5.63	34724	Protein yhdH.
	3	K6P2_ECOLI	5.75	32388	6-phosphofructokinase isozyme 2
	4	YEIN_ECOLI	5.37	32910	Hypothetical protein yein.
	5	SUCC_ECOLI	5.37	41392	Succinyl-CoA synthetase beta chain

Fig. 5. Output from the AACompIdent tool where a protein from an *Escherichia coli* 2-D gel has been correctly identified by matching its amino acid composition, sequence tag, estimated pI, and Mw against database entries for *E. coli*. The correct protein identity, malate dehydrogenase, was the top-ranked protein in the three lists, and showed a large score difference between the top- and second-ranked proteins in the third list (where pI and Mw windows are applied), and also in the first list. See text for more details about the significance of score patterns for identification confidence. In addition, the sequence tag MKVA has been found only for that protein (as shown by the asterisk). In the second list, where the amino acid composition of the query protein was matched against all entries in the Swiss-Prot database without considering protein pI and Mw, malate dehydrogenase from four different species was ranked in the top four positions. This illustrates that protein amino acid composition is well-conserved across species boundaries. The AACompIdent output has been shortened for this figure.

option is selected, the AACompIdent output will show either 40 amino acids of each protein's predicted N- or C-terminal sequence or its description, and show an asterisk to the left of a protein's rank if the protein carries the sequence tag anywhere in its sequence. If the tag is found in the displayed N- or C-terminal sequence, it will be shown in lowercase letters. We are confident that a protein from Swiss-Prot represents a correct identification if the query protein's empirically determined sequence tag of three amino acids or more is present at the expected N- or C-terminal position, and that this protein is ranked within the first 10 or so closest entries by amino acid composition.

4.3. AACompSim Tool

The AACompSim tool (<http://www.expasy.org/tools/aacsim/>) allows the theoretical AA composition of one protein in the Swiss-Prot database to be compared to proteins from one or all species in the database (see Note 23). This serves two main purposes: first, to allow the simulation of matching undertaken for identification purposes with AACompIdent (see Subheading 4.2.); second, to allow the detection of weak similarities between proteins by comparison of their compositions rather than sequences, as explored by Hobohm and Sander (15).

To use AACompSim, first select the constellation of amino acids you wish to work with (see Subheading 4.2.1.). If you wish to simulate matching undertaken with empirical data, you should specify constellation 2. To match against the database for detecting protein similarities, you should use all 20 amino acids in constellation 0. Then specify an e-mail address where the results can be sent, the Swiss-Prot identification name (e.g., IPIA_TOBAC) or accession number (e.g., Q03198) of the protein you would like to compare against the database, and the Swiss-Prot abbreviation for the species to match against (e.g., SALTY for *Salmonella typhimurium*). A document containing a full list of all Swiss-Prot species and their organism codes can be found at <http://www.expasy.org/cgi-bin/speclist>. If desired, matching can be done against all species in the database by specifying "ALL." Finally, select the "Search" button to submit the query to the ExPASy server. Results will be sent to your e-mail address. AACompSim will return three lists of proteins, similar to those from AACompIdent (see Notes 23, 24).

4.4. Multident Tool

Proteins can be identified by virtue of their peptide masses alone, but frequently other data are needed to provide high-confidence identification. The same is true for protein identification with AA composition. Following our earlier observations that high-confidence protein identification can be achieved with a combination of peptide mass and AA composition data (16,17), we have developed the protein identification tool MultiIdent (<http://www.expasy.org/tools/multiident/>). This tool uses parameters of protein species, estimated pI and Mw, keyword, AA composition, sequence tag, and PMF data to achieve protein identification (18). Currently, the program works by first generating a set of proteins in the database with AA compositions close to the unknown protein, as for AACompIdent (see Subheading 4.2.). Theoretical peptide masses from the proteins in this set are then matched with the peptide masses of the unknown protein to find the number of peptides in common (number of "hits"). Three types of lists are produced in the results: first, a list where proteins from the database are ranked

according to their AA composition score (see **Subheading 4.2.**); second, a list where proteins are ranked according to the number of peptide hits they showed with the unknown protein; and thirdly, a list that shows only proteins that were present in the both the above lists, where these proteins are ranked according to an integrated AA and peptide hit score. In all these lists, protein pI, Mw, species of origin, and Swiss-Prot keyword can be used as in AACompIdent to increase the specificity of searches.

4.4.1. Use of the MultIdent Tool

After selecting MultiIdent from the Tools page, you must first choose the constellation of amino acids you wish to work with. Then provide information including your e-mail address, details about the unknown protein (name, pI and Mw estimations, amino acid composition, sequence tag data if available), species of interest for matching (see **Notes 16, 22**), and Swiss-Prot keyword (if any). This should be done in essentially the same manner as for AACompIdent (see **Subheading 4.2.**). To include peptide masses for protein identification, first specify the size of the list to be created with the query protein's AA composition (e.g., 500). Then click in the checkbox next to the "Peptide Mass Fingerprinting" title in the program to enable this option, enter the list of peptide masses into the text box with an accuracy of at least one decimal place, and specify whether masses are monoisotopic or average. Then specify the enzyme used to create the peptides (e.g., trypsin), whether the protein was reduced and alkylated with any reagent (e.g., iodoacetamide, iodoacetic acid, 4-vinylpyridine) before cleavage, whether artificial protein modifications such as oxidation of methionine or acrylamide adducts to cysteine are expected, and the mass tolerance to be used in matching with peptides. The mass tolerance should reflect the known accuracy of your mass spectrometer. Finally, specify which results lists you would like to see in the MultiIdent output, and select the "Perform" button to submit the match to ExPASy. The results will be sent to you by e-mail.

4.4.2. Interpretation of Results

The list of closest Swiss-Prot entries in terms of protein AA composition is the same as for the AACompIdent output, and thus results and scores for proteins in this list can be interpreted as in **Subheading 4.2.3.** If sequence tags are used as part of a search strategy, the list of closest proteins in terms of protein AA composition will show the predicted protein N- or C-terminal sequence, and any sequence tags present will be highlighted in lowercase letters. Asterisks are also shown to the left of protein rank numbers to indicate that the sequence tag is present in the corresponding protein. These asterisks are used in the list of best matches by AA composition, as well as the lists of proteins generated by PMF and the integrated score.

The list of closest Swiss-Prot entries in terms of peptide hits is simply the list of proteins that have the most peptides in common with the query protein. The "hits" are the number of peptides that match with a database entry, and the peptide masses shown in the output are those from the database entry that match with those from the query protein. The top-ranked protein in this list will be the most likely identification of the protein; however, this may not be so if matching has been done with very large Mw windows. In any case, use of sequence tags of even three or four amino acids with peptide mass data can greatly increase the confidence of a database entry representing a correct identification. Note that for this purpose, sequence tags generated by tandem

mass spectrometry (MS/MS) or by postsource decay matrix-assisted laser desorption/ionization (MALDI)-time-of-flight (TOF) techniques can be used in MultiIdent as well as tag data generated at protein N- or C-termini.

The list of proteins with best integrated scores represents the most powerful form of matching (*see Note 25*). It can simultaneously consider the protein parameters of pI, Mw, AA composition, sequence tag, and peptide masses in order to rank proteins from the database for the species of interest (*see Note 22*) with a given keyword. The integrated score is a measurement of difference between the query protein and a database entry, and is derived by dividing the AA analysis score by the number of peptide hits that were found for that protein. Accordingly, an integrated score of 0 represents a perfect match for a query protein, with larger scores representing increasing differences. We find that the integrated score is useful for defining confidence limits if it is not immediately apparent whether a protein has been correctly identified.

4.5. *Aldente* Tool

4.5.1. Description

Aldente (Advanced Large-scale iDENTification Engine, <http://www.expasy.org/tools/aldente/>) is a tool that allows the identification of proteins using peptide-mass fingerprinting data.

Experimentally measured, user-specified peptide masses are compared with the theoretical peptides calculated for all proteins in the Swiss-Prot/TrEMBL databases. Isoelectric point, molecular weight, and a species (or group of species) can be specified in order to restrict the number of candidate proteins and reduce false-positive matches.

The main features of Aldente are:

- Use of a robust method (the Hough transform) to determine the deviation function of the mass spectrometer and to resolve peptide match ambiguities. In particular, the method is relatively insensitive to noise (*see Note 26*).
- Tuneable score parametrization: the user can choose the parameters he or she wants to take into consideration in the score and in which proportion.
- Extensive use of the annotations (protein mature form, posttranslational modifications, alternative splicing) in Swiss-Prot/TrEMBL, offering a degree of protein characterization as part of the identification procedure (*see Note 27*).
- Consideration of user-defined chemical amino acid modifications (oxidation of methionine, acrylamide adducts on cysteine residues, alkylation products on cysteine residues), and the possibility to define their contribution to the score.

4.5.2. Use of Aldente

After selecting the Aldente tool from the ExPASy Tools page, you should enter the sample information, with a query name, pI, and Mw estimations if known (*see Note 28*). A list of experimental peaks (*see Notes 10, 38*) can be entered directly into the text area (peptide masses with or without peak intensities), or by uploading a text file (*see Note 29*). You should then select the database to search (Swiss-Prot and/or TrEMBL), and the species (or group of species) if you want to restrict your query (*see Note 30*).

You have to specify the enzyme that was used to generate the peptides. In order to take into account partial cleavages, you can specify 0 or 1 missed cleavage sites to be allowed. You must define a minimum number of peptide-mass hits required for a matching protein to be included in the result list. The default value is 4.

Mass tolerance has to be set on protein and peptide levels. On protein level, upper and lower mass limits can be specified, which serve to filter the results but are not taken into account in the scoring (in contrast to the Mw estimation mentioned above). Peptide mass tolerance corresponds to the estimated internal precision of the mass spectrometer: the instrument's accuracy can be specified, either with an absolute value in Daltons or with a relative value in ppm (parts per million), or with both (see **Note 31**). Less accurate peptide mass data will require a larger mass tolerance and will result in a lower accuracy of your search (see **Note 32**).

Then specify the chemical modifications occurring on the unknown protein before cleavage, and the way to take them into account in the score (see **Note 33**).

You may also choose which types of PTMs annotated in Swiss-Prot you want to take into account (e.g., only experimentally proven or also computationally predicted ones) and the way to consider them in the score (see **Notes 34** and **35**).

Finally specify the maximum number of proteins you want to be displayed in the Aldente output, and select the “Submit” button to send your query to ExPASy. Depending whether you provide your e-mail address or not, the results will be sent to you by e-mail or displayed directly in your browser.

4.5.3. Aldente Output

The top part of the output result provides the date of the query, the database release number and current number of entries, and some statistics about the search. The protein statistics give the total number of proteins in the selected database and taxa, the number of proteins in the protein mass range, the number of proteins with enough peptides in the peptide mass tolerance, the number of proteins with the minimum number of hits after alignment, and the number of displayed proteins (see **Note 36**). Then the peptide statistics give the number of peptides generated in the mass range of your sample, the number of peptides matching a peak of the sample, and the average number of theoretical peptides per protein in the mass range.

Then follows a summary of the best-matching proteins from the database (**Fig. 6**), with a “quick jump” link to detailed peptide information provided further down in the same page.

After that, for each matching protein, detailed information concerning matching peptides is given (**Fig. 7**), with individual score, difference between the experimental and calculated masses, information regarding PTM or chemical modification if any, peptide position, and sequence. Finally the protein sequence is visualized with identified peptides in blue and uppercase, where trypsin loci (K, R) are shown in red.

Aldente results are displayed online or sent by e-mail, in the form of an html table, or in XML or text format for easier parsing. The html result contains direct links to FindMod (see **Subheading 4.6.**), GlycoMod (see **Subheading 4.7.**), and FindPept (see **Subheading 4.8.**) to further characterize matching proteins by predicting potential protein post-translational modifications, finding potential single-amino acid substitutions and potential unspecific cleavage, to PeptideMass (see **Subheading 3.3.**), and to BioGraph (see **Subheading 4.9.**), for the graphical representation of the theoretical spectrum. Relevant input data and/or information about the matching database entry are automatically transferred to those programs.

A new Aldente search can be launched directly from the result output. This allows the user to submit a second search with slightly modified parameters, i.e., with modi-

Aldente version Beta 15/01/2004 feedback is welcome [Documentation](#) [Input summary](#) [Printable page](#)

Date 22/01/2004 10:34:04 UTC

Release Swiss-Prot Release 42.7 of 15-Dec-2003: 141681 entries

Proteins Scanned 163999 / In mass range 155784 / Enough hits before alignment 53888 / Enough hits after alignment 33067 / Displayed 20

Peptides Generated 8507732 / Matching 473381 / Average of 54 peptides per protein

Rank	Z-score	Hits	Taxon	AC	ID	Mass	pI	Shift	Slope	DE
1	25.61	20	Homo sapiens	P17844	DDX5_HUMAN	69147	9.06	-0.016	37	Probable RNA-dependent helicase p68 (DEAD-box protein p68) (DEAD-box)
2	20.62	13	Homo sapiens	Q8Y2X3	NOP5_HUMAN	59577	9.03	-0.027	47	Nucleolar protein NOP5 (Nucleolar protein 5) (NOP5B) (HSPC120).
3	17.90	17	Mus musculus	Q81656	DDX5_MOUSE	69319	9.06	-0.035	50	Probable RNA-dependent helicase p68 (DEAD-box protein p68) (DEAD-box)
4	10.96	8	Schizosaccharomyces pombe	Q12524	IDHP_SCHPO	47293	8.86	0.104	-37	Probable isocitrate dehydrogenase [NADP], mitochondrial precursor
5	10.88	9	Rattus norvegicus	Q9Q288	NOP5_RAT	60070	8.7	-0.008	27	Nucleolar protein NOP5 (Nucleolar protein 5) (Nopp140 associated)
6	10.75	4	Escherichia coli	P09168	OGT_ECOLI	19179	8.48	0.152	-50	Methylated-DNA-protein-cysteine methyltransferase (EC 2.1.1.63) (6-O-Methylated-DNA-protein-cysteine methyltransferase (EC 2.1.1.63) (6-O-Methyltransferase))
7	10.75	4	other Bacteria	P09168	OGT_ECOLI	19179	8.48	0.152	-50	Methylated-DNA-protein-cysteine methyltransferase (EC 2.1.1.63) (6-O-Methyltransferase (EC 2.1.1.63) (6-O-Methyltransferase))
8	10.61	3	Escherichia coli	P10396	REP1_ECOLI	8941	10.92	0.095	-10	Replication initiation protein (Fragment).
9	10.58	8	other Bacteria	P51952	RIBB_PHOPO	40064	5.3	-0.069	70	3,4-dihydroxy-2-butanoate 4-phosphate synthase (DHBPS synthase).
10	10.04	6	Viruses	Q05163	9GL_ASFB7	14378	8.52	-0.013	40	late protein 9GL.
11	9.91	6	Viruses	P03346	GAG_HTLV2_C3	24386	8.4	-0.069	80	CORE PROTEIN P12 (P15).
12	9.80	5	Chlorophyta	Q9TK2B	R817_NEPOL	17812	10.78	-0.080	63	Chloroplast 30S ribosomal protein S7.
13	9.28	5	Archaea	Q8PTU1	PSMA_METMA	27353	5.14	0.264	-87	Proteasome alpha subunit (EC 3.4.25.1) (Multicatalytic endopeptidase).
14	9.18	4	other Bacteria	Q8DC31	RL19_VIBVU	13187	10.43	0.328	-233	50S ribosomal protein L19.
15	8.94	7	Chlorophyta	P54214	SFAS_DUNBI	31412	5.42	0.295	-173	SF-assassin.
16	8.84	5	other Bacteria	Q87C45	PRMA_XYLFT	33029	4.43	0.144	-33	Ribosomal protein L11 methyltransferase (EC 2.1.1.1-) (L11 Ntase).
17	8.62	5	Homo sapiens	Q16649	TR13_HUMAN	17318	5.52	0.203	-120	Thyroid receptor interacting protein 3 (TRIP-3) (Fragment).
18	8.20	4	other Bacteria	Q8KYP1	PAAD_STRCO	22614	8.89	0.155	-103	Probable aromatic acid decarboxylase (EC 4.1.1.-).
19	8.11	6	other Streptophyta	Q30358	M125_ORYSA	22087	9.5	0.245	-210	Mitochondrial 22 kDa protein (ORF 25).
20	7.98	5	other Bacteria	Q9PJL6	RL23_CHLMU	12134	9.98	-0.011	7	50S ribosomal protein L23.

[Resubmit](#)

Graphical visualisation of the results : [BioGraph](#)

Fig. 6. First part of the Aldente output, showing the summary of the best-matching proteins from the database. The top of the page summarizes some statistics about processed proteins and peptides, followed by the list of best-matching proteins, with related information. Access to documentation, submitted parameters, printable page, graphical visualization of the results with BioGraph (see Subheading 4.9. and Fig. 8), and “resubmit” function are provided.

fied molecular weight or pI ranges, number of missed cleavages, taxonomic range; or to resubmit an archived query at a later stage, for later database releases. This is particularly useful if the initial identification was unsuccessful or ambiguous.

4.6. FindMod Tool

The FindMod, GlycoMod (see Subheading 4.7.), and FindPept (see Subheading 4.8.) tools are used to identify the origin of peptide masses obtained by PMF that are not matched by protein identification tools such as Aldente. They also take into account posttranslational modifications annotated in Swiss-Prot or supplied by the user, and chemical modifications of peptides. It is quite common for PMF tools not to be able to find matching theoretical peptides for a few of the less intense peaks that were detected and submitted to the identification process.

FindMod (<http://www.expasy.org/tools/findmod/> [19]) is a program for *de novo* discovery of protein PTM or single-amino acid substitutions. It examines PMF results of known proteins for the presence of more than 20 types of PTMs of discrete mass, such

2) Q9Y2X3 NOP5_HUMAN Swiss-Prot : Homo sapiens Nucleolar protein NOP5 (Nucleolar protein 5) (NOP58) (HSPC120). [Up](#)

Z-score : 20.62 Hits : 13 Mw : 59677 pi : 9.03 Coverage : 28% Shift : -0.026667 dalton Slope : 47 ppm

Exp	Theo	Intensity			Delta		Dev	Cont	MC	CAM	MSO	PTM	Position	Sequence	
		Da	Da	%	rank	Da	ppm	ppm					start	end	
975.502	975.504004	1795	13	51	-0.00	-2	-21	-	1	1/1	-		207	- 214	CLQKVQDR
*	1327.712	1327.664064	13417	100	1	0.05	36	9	-	-	-		269	- 278	TQLYEYQLQNR
1398.754	1398.704546	8105	60	3	0.05	35	7	-	-	-	0/1		121	- 133	SQMDGLIPGVGPR
1414.74	1414.69948	2545	18	28	0.04	28	0	-	-	-	1/1		121	- 133	SQMDGLIPGVGPR
1584.875	1584.836684	2450	18	33	0.04	24	-5	-	-	-	-		222	- 235	LSSELLPVEEAEVK
*	1733.942	1733.922062	1453	10	67	0.02	11	-20	-	-	-		338	- 353	YGLIYHASLVLQQTSPK
1834.818	1834.754782	4564	34	8	0.06	34	2	-	-	-	0/1		372	- 388	YDAFGEDSSSAIGVNR
1850.814	1850.749706	1481	11	65	0.06	34	2	-	-	-	1/1		372	- 388	YDAFGEDSSSAIGVNR
1882.02	1881.967308	4014	29	12	0.05	27	-4	-	-	1/1	0/1		102	- 117	LNLSCHNSPVVNELMVR
1959.07	1959.022330	2133	15	40	0.05	24	-9	-	-	-	0/3		279	- 297	MMAIAPNVTVMVGLVGAR
1978.001	1977.98036	1220	9	83	0.02	10	-23	-	1	-	-		22	- 37	LQEVDLSLWNEFETPEK
2122.043	2121.855714	842	6	100	0.09	41	6	-	1	-	-		468	- 485	VEEEEEEKVAAEEEETSVK
2139.182	2139.104858	896	6	97	0.08	36	1	-	1	1/1	0/1		100	- 117	EKLNLSCHNSPVVNELMR

Details of the alignment

```

1 mlvlfetsvg ysifkvlnek RLQEVDSLAKK EFETPEKank iVklkhfkfkf qdtasalaaF talmeqkink qlkkvkkiv
81 kseahplava ddklggvkikl KLNLSCHNSPVVNELMRqir SQMDGLIPGVGPR EPHemaacl qIahsllsryr lkksadkvd
161 mvgaisllld ddkhlnnyi mscrwvghf fpeIgkiiad nlyckCLQK VGDmknryasa kLSSELLFEEV EAEVWaaesi
241 smygttldc inchnlhtq vleiseyrTQ LYETLQHNMNM AIAFHNTVHV GDLVGSARlia hagsllnlak haastvqllg
321 aekhfralk srdtptvYGL IYHASLWQGT EPrhkgkisr mlaaktvlaI rTDAFGEDSS SAMGVENNAk learltitled
401 rgirkllsgtq ksalaktekye hksvektlyp sgdstlptcs kkrkklieqvdk edeitekkak kakikwvKEE EEEKEVAAEE
481 ETGVKKKKkr gkkhkhipee leeeepctst aiaspekkkk kkkkrened

```

[GlycoMod](#) [FindMod](#) [FindPept](#) [ProteinTools](#)

Fig. 7. Second part of the Aldente output, showing details for one of the matching proteins. The output shows information on the matched peptides, including individual score, difference between the experimental and calculated masses, information regarding posttranslational modification (PTM) or chemical modification if any, peptide position, and sequence. The protein sequence is displayed with identified peptides in blue and upper case, and trypsin loci (K, R) (not visible in this figure). Links to ExPASy characterization tools are also provided.

as acetylation, amidation, biotin, C-mannosylation, deamidation, *N*-acyl diglyceride cysteine (tripalmitate), FAD, farnesylation, formylation, geranyl-geranyl, γ -carboxyglutamic acid, *O*-GlcNAc, hydroxylation, lipoyl, methylation, myristylation, palmitoylation, phosphorylation, pyridoxal phosphate, pyrrolidone carboxylic acid, and sulfation.

This is done by looking at mass differences between experimentally determined peptide masses and theoretical peptide masses calculated from a specified protein sequence. If a mass difference corresponds to a known PTM not already annotated in Swiss-Prot, rules are applied that examine the sequence of the peptide of interest and make predictions as to what amino acid in the peptide is likely to carry the modification. The same method is applied when predicting potential amino acid substitutions.

4.6.1. Input Parameters

FindMod is usually launched after a PMF identification run, for the most likely protein suggested by an identification program such as Aldente. The output of Aldente contains a link to the FindMod submission form with most parameters already filled in. If you wish to launch FindMod directly, you should specify the sequence of the protein you would like to characterize and for which you have determined a set of peptide masses. If this protein is known in Swiss-Prot/TrEMBL, enter the Swiss-Prot ID code or the protein accession number. Otherwise, you can enter the sequence of your protein of interest, in single-letter amino acid code, in either upper or lower case (see Note 37).

Protein sequences from other sources (e.g., word-processor programs or other Web pages) can be copied and pasted directly into this field. If there are spaces in your sequence, these will be ignored.

The characters *B*, *X*, or *Z* are accepted, but no masses are computed for peptides containing one or more of these characters.

A set of experimental masses must also be provided (see **Notes 29, 38**). The experimental peptide masses will first be compared to theoretical unmodified peptides and to peptides modified as documented in Swiss-Prot or by chemical modifications. The user can choose whether all peptide masses or only those that have not been attributed a theoretical peptide in this process should be examined for potential PTMs and/or single-amino acid substitutions.

If you wish to take into account other posttranslational modifications than those already predictable by FindMod, you can enter, for each of these modifications, its name, its atomic composition, and the amino acids on which this modification can be observed.

Further parameters are isotopic resolution (average or monoisotopic masses), chemical treatment of cysteine (see **Note 39**), oxidation state of methionine (see **Note 40**), mass tolerance (in ppm or in Daltons), digestion agent, and number of missed cleavages (up to three). You can enter the masses of your peptides as [M] or as [M+H]⁺ (see **Note 10**).

4.6.2. FindMod Output

The results from FindMod are divided into a header and up to three tables.

The header contains information about the submitted protein: a link to the Swiss-Prot/TrEMBL entry and the description line (if the protein is in Swiss-Prot/TrEMBL), pI, and molecular weight. Then the input parameters are listed, followed by an active link to PeptideMass. This allows the user to perform a theoretical cleavage of the protein of interest.

The tables report the peptides whose experimental masses match unmodified or modified theoretical digest products of the protein of interest.

The first table reports matches to theoretical digest products as unmodified, modified with the annotations in Swiss-Prot, and chemically modified as specified in the input form.

The second table reports those user masses that differ from a theoretical database mass by a mass value corresponding to one of the considered PTMs. These peptides are further examined, and FindMod checks whether the peptide sequences contain amino acids likely to carry the modification in question. This is done by applying a set of prediction rules that have been defined using information in the PROSITE database (20), examining all the PTM annotations in Swiss-Prot and information in the literature. The program first lists the matches conforming to these rules, highlighting potentially modified residues in color. Potential PTMs detected by mass difference but not confirmed by the rules are included in a second list.

The third table shows potential single-AA substitutions detected by mass difference. The following particularities are worth pointing out:

1. A BLOSUM62 score (21) is given for each suggested single-AA substitution. This provides information about the probability of substitution: Lowest score: -4 (low probability of substitution); highest score: 11 (high probability of substitution).

2. Potential single-amino acid substitutions are not displayed if they occur on the cleavage site and substitute the AA for an AA after which the digestion enzyme does not cleave.
3. If the suggested AA substitution corresponds to a sequence variant or conflict as annotated in the Swiss-Prot feature table, this substitution is highlighted in color, and a hypertext link is provided to the corresponding annotated variant or conflict.

At the end of the output page, the user will find a list of those entered matches that did not match in any of the previous tables (if any).

4.7. GlycoMod and GlycanMass Tools

Protein glycosylation is one of the most common and most complex post-translational modifications. Although the problem of predicting glycosylation from peptide-mass fingerprinting data is in principle the same as the one addressed by FindMod, the complexity and heterogeneity (the high number of possible combinations of monosaccharides forming glycan structures) made it necessary to conceive a separate tool, specializing in glycan structures and glycopeptides, GlycoMod.

GlycoMod (<http://www.expasy.org/tools/glycomod/> [22, 23]) finds all possible compositions of a glycan structure from its experimentally determined mass. It may be used to calculate the possible compositions of free or derivatized glycan structures, or compositions of glycans attached to glycoproteins and glycopeptides. The motivation and use of the tool are quite similar to FindMod. As there has been a recent book chapter devoted entirely to the use of GlycoMod (23), we will not detail its use here.

GlycanMass (<http://www.expasy.org/tools/glycomod/glycanmass.html>) allows the user to calculate the mass of a glycan from its monosaccharide composition. Available elements to build the oligosaccharide are hexose, HexNAc, deoxyhexose, NeuAc, NeuGc, pentose, sulfate, phosphate, KDN, and HexA. The user has the possibility to specify whether the monosaccharide residues are underivatized, permethylated, or peracetylated, and whether to use average or monoisotopic mass values.

4.8. FindPept Tool

4.8.1. Description

FindPept (<http://www.expasy.org/tools/findpept/>, [24]) is designed to predict peptides resulting from the following causes: unspecific proteolytic cleavage, missed cleavage, protease autolysis, and keratin contaminants (see **Notes 41, 42**).

Unspecific cleavage is the process by which peptides whose termini do not correspond to the cleavage specificity rules implemented in computer programs are produced by proteolysis. These rules are often simplistic and reflect our incomplete understanding of the specificity of certain enzymes (see **Note 43**). Other causes include a contamination with other proteases (e.g., trypsin usually contains traces of chymotrypsin), biological processes such as protein degradation, or a change in enzyme specificity over time (see **Note 44**).

4.8.2. Using FindPept

FindPept is not a part of the identification procedure, as it requires a protein as input. Therefore, a search with a tool like Aldente should be carried out first to match peptides resulting from specific cleavage and report a candidate protein. FindPept can be directly launched from the Aldente output, or can be accessed via its submission form, but also from the results of the FindMod or GlycoMod programs.

A protein sequence should be provided. If it is a Swiss-Prot/TrEMBL accession number, the program will read the annotation relative to posttranslational modifications in Swiss-Prot feature tables, and use these to generate a table of posttranslational modifications (*see Notes 45, 46*). If it is a user-entered sequence, expected post-translational modifications may be specified by entering their abbreviations within brackets. User-defined modifications can also be applied to any position or residue of choice. They can be entered by specifying the name and atomic composition of the PTM and the position(s) to which the PTM applies. A position can be supplied as a number (“18” = residue number 18 relative to the Swiss-Prot or user-entered sequence), one or more amino acids (“E D” = all glutamate and aspartate residues), or an anchor (“<” = N-terminus of each peptide). This functionality is especially useful if an atypical chemical reactant has been added during the experiment.

A set of experimental masses must also be provided, in the same format as specified for FindMod (*see Subheading 4.6*). Expected chemical modifications should be supplied (*see Notes 39, 47*).

The enzyme or chemical reagent used to generate the peptides can optionally be indicated. In this case, cleavage sites that obey cleavage rules at either end are highlighted in the results with a red slash, and peptides that obey them at both ends are displayed in a separate table. Additionally, a set of human keratins is theoretically digested, and matching masses are reported. A drop-down list adjacent to that used to select the enzyme can be used to specify the source of the enzyme from a list of the most common sources and variants, when the sequence of these enzymes is known and present in Swiss-Prot/TrEMBL. If one is selected, its sequence is submitted to a theoretical self-digestion, and the user masses are checked for autolysis fragments, accepting missed cleavages. PTMs present in the feature table of the Swiss-Prot entry for the enzyme are also taken into account (e.g., phosphorylation of pig pepsin, accession number P00791).

The output page (*see Note 48*) is divided into a header and up to seven tables. Each table is displayed only if matching peptides/PTMs have been found in the given category. The tables “Post-Translational/Artefactual Modifications for the protease/ the protein” list the PTMs applied to the protease, and to the studied protein. The table “Masses resulting from possible contaminants” lists the masses that correspond to the specific cleavage of a number of human keratins (*see Note 49*). The table “Peptides resulting from protease autolysis” lists the peptides obtained by specific self-digestion of the protease that match the user masses. The table “Matching peptides for specific cleavage” lists the peptides obtained by digestion of the studied protein for which both ends are either sites of specific cleavage by the protease, or the extremities of the original peptide. The table “Matching peptides for unspecific cleavage” lists all peptides obtained by allowing cleavage at any position in the sequence of the studied protein that match the user masses, for which at least one cleaved extremity does not match the enzyme cleavage rules. The table “Unmatched masses” lists the masses that could not be identified by the program, i.e., not assigned to one of the tables above. A set of buttons allows you to directly submit those masses to the FindMod and GlycoMod tools, and possibly identify PTMs or glycosylated sites on the protein.

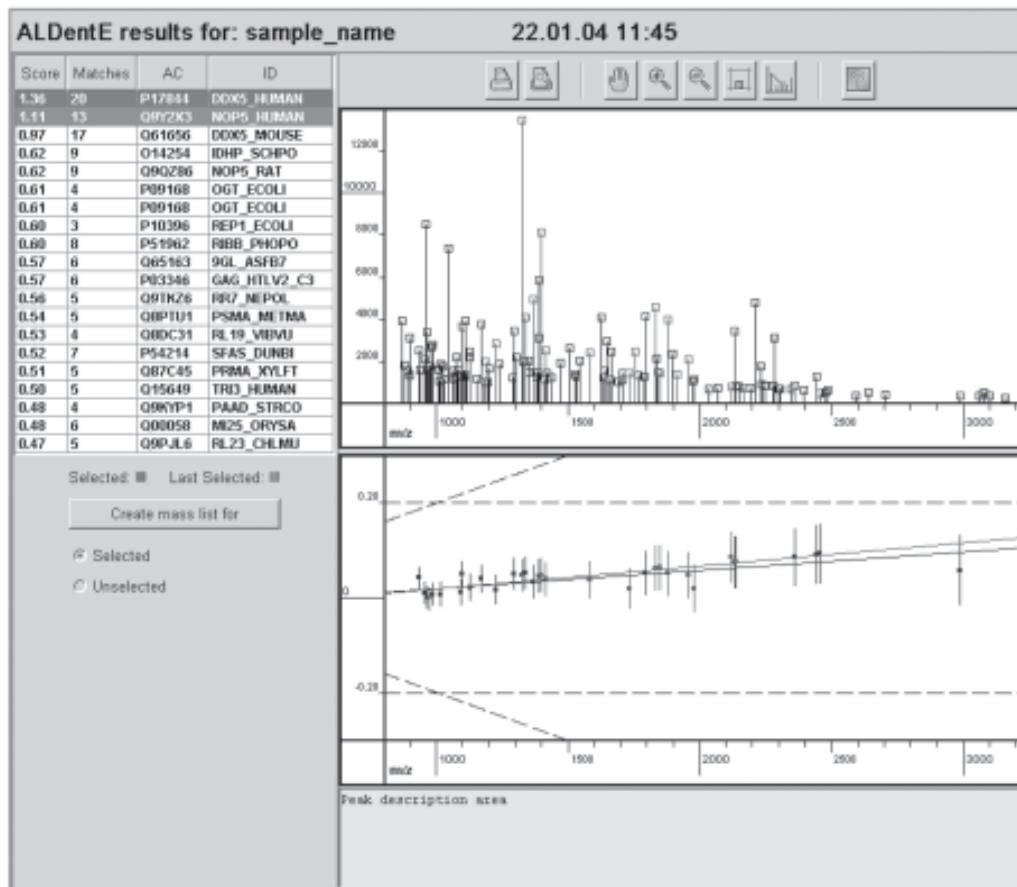


Fig. 8. The BioGraph applet, used to visualize an Aldente identification result. The lower part of the spectrum analysis panel shows the m/Z ratio on the horizontal axis and the arithmetic difference between theoretical and experimental mass (in a user-defined range) on the vertical axis. The two best-matching proteins, P17844 and Q9Y2X3, are selected in the score list, which causes the straight lines corresponding to their spectra to appear in this panel. As both lines have very similar slopes, this validates the assumption of their co-occurrence in the sample.

4.9. BioGraph Tool

4.9.1. Description

BioGraph (<http://www.expasy.org/tools/BiographApplet/>) is a Java applet that aims at providing ExPASy users with an interactive interface to visualize results of some proteomics tools. BioGraph is therefore accessible from Aldente, FindMod, or FindPept results by clicking on the “BioGraph” button.

This viewer is composed of three main components, or panels (see Fig. 8, which shows an Aldente identification result, visualized with BioGraph): first, the “Title panel,” intended to give general information about the source program; then the “Tool

results panel,” to summarize the source program results and interact with the spectrum; and lastly, the “Spectrum manipulation panel,” to interactively visualize the user-entered spectrum.

4.9.2. General Features

The “Title panel” provides three information items:

- The source tool name (either Aldente, FindMod or FindPept),
- The user-entered protein name,
- The date and time at which source program has been run.

The “Spectrum manipulation panel” is composed of three basic components: the toolbar, the spectrum, and the peak information panel. The toolbar consists of eight buttons to:

- Print the content of the spectrum area,
- Preview what will be printed,
- Move the spectrum on the horizontal axis,
- Zoom in to the spectrum on the horizontal axis,
- Zoom out of the spectrum on the horizontal axis,
- Select a region and zoom it on the horizontal axis,
- Compare two peaks on the spectrum.

The spectrum summarizes user-entered data, i.e., the m/Z ratio on the horizontal axis and the intensities on the vertical axis.

In the case where Biograph is called from Aldente, another graph is displayed, which is described in **Subheading 4.9.3**.

The peak information panel displays, if the mouse is moved over a peak, this peak’s properties, i.e., mass and intensity values on the one hand and data about the matched proteins on the other hand, including:

- Swiss-Prot or TrEMBL accession code and ID,
- matched peptide mass,
- difference between current mass and user-entered mass, number of missed cleavages,
- a symbol to indicate whether or not the matched sequence contains modifications,
- “from” and “to” positions of the match
- corresponding sequence.

4.9.3. Source Program-Specific Features

These features, which depend on the program from which BioGraph has been called, are summarized in the “Tool results panel.”

In the case where BioGraph was called from Aldente, the tool results panel displays a table of matched proteins whose rows can be selected to highlight the corresponding peaks on the spectrum.

Another Aldente-specific feature is a graph that aims at providing the user with a new visualization of data. This graph shows the m/Z ratio on the horizontal axis and the arithmetical difference between theoretical and experimental mass (in a user-defined range) on the vertical axis. The goal of such a representation is first to evaluate the spectrometer-intrinsic user-defined error rates in order to extract noise from signal, and then to validate true positive matches for the current run (as their corresponding points follow the same straight line). This is illustrated in [Fig. 8](#).

When BioGraph is called from FindMod or FindPept, the tool results panel consists of four checkboxes, whose selection leads to the highlighting of peaks corresponding to peptides with specified properties (e.g., matching or potentially modified peptides).

The option “Mass list creation” is available for both types of tool results panel: the user can export peaks of interest by using the “Create mass list” button.

5. Integration of the Tools With Each Other and With Swiss-Prot/TrEMBL

The ExPASy protein identification and characterization tools, in particular Aldente, FindMod, GlycoMod, and FindPept, are closely integrated and hyperlinked with Swiss-Prot and TrEMBL entries on ExPASy, and among each other. Navigation between database entries, data submission forms, and program results is made easy (in both directions), and the number of copy/paste and mouse click operations is minimized.

For example, in addition to its submission form, GlycoMod can be used, after the identification of a potential candidate protein from peptide-mass fingerprinting with Aldente, to further characterize this protein: GlycoMod may explain some of the unmatched, peaks which correspond to glycopeptides, i.e., the linkage of an oligosaccharide. FindMod, which allows the user to predict discrete posttranslational modifications or amino acid substitution, or FindPept, which detects unspecific cleavage or peaks due to contaminants or enzyme autolysis, may also be used before or after GlycoMod. Independently of the order in which you choose to apply these tools, direct submission forms are available with all the relevant data already filled in, and it is not necessary to go back to the original submission forms.

All tool submission pages (just like all other html pages on ExPASy) contain a search menu, which allows for easy keyword searches in the Swiss-Prot/TrEMBL as well as any other ExPASy database.

The NiceProt view of any Swiss-Prot/TrEMBL entry on ExPASy contains direct links to the results or, if additional parameters are required, to the submission forms of several protein-characterization tools. In the latter case, the submission forms already have the sequence information filled in, which again minimizes the number of copy/paste operations.

Sometimes it may be of interest to perform sequence analysis or prediction on a subsequence of the precursor molecule annotated in Swiss-Prot. This option is also supported on ExPASy: The positions (ranges) of certain regions of interest annotated in the Swiss-Prot feature tables (FT) are hyperlinked in the NiceProt view, giving access to a page that highlights the region in color, and that contains links allowing the user to submit just that region to the same analysis tools as those available for the complete sequence via NiceProt.

6. Notes

1. Protein pI is calculated using pK values of amino acids described in Bjellqvist et al. (25,26), which were defined by examining polypeptide migration between pH 4.5 and 7.3 in an immobilized pH gradient gel environment with 9.2 *M* and 9.8 *M* urea at 15°C or 25°C. Prediction of protein pI for highly basic proteins is yet to be studied, and it is possible that current Compute pI/Mw predictions may not be adequate for this purpose. The buffer capacity of a protein will affect the accuracy of its predicted pI, with poor buffer capacity leading to greater error in prediction (25,26). Because of this, pI predictions for small proteins can be problematic.

2. Protein Mw is calculated by the addition of average isotopic masses of amino acids in the protein and the average isotopic mass of one water molecule. This program does not account for the effects of posttranslational modifications; thus, modified proteins on a 2-D gel may migrate to a position quite different from that predicted. Protein glycosylation in particular can affect protein migration in both pI and Mw dimensions. Note, however, that the “GET REGION ON 2D PAGE” function in SWISS-2DPAGE (27) (accessed by selecting a “GET REGION ON 2D PAGE” hypertext link from a Swiss-Prot entry) uses the Compute pI/Mw algorithm to highlight the region on a 2-D gel to where an unmodified protein should run, and suggests a region where the modified protein might be found if it has modifications documented in the Swiss-Prot database.
3. Signal sequences or transit peptides of unknown length, however, become increasingly rare (currently 444 Swiss-Prot protein sequences out of 139,947): whenever signal sequences (and their length) are not experimentally determined, the manual annotation process includes the use of prediction programs (e.g., SignalP [28]), which results in annotation of potential signal sequences.
4. It is not possible to specify posttranslational modification for your protein, nor will ProtParam know whether your mature protein forms dimers or multimers. If you do know that your protein forms a dimer, you may just duplicate your sequence (i.e., append a second copy of the sequence to the first), as all computations performed by ProtParam are based on either compositional data, or on the N-terminal amino acid.
5. ProtParam sums the contributions of the different amino acids as if they were independent, not taking into account secondary or tertiary structure. Exact coefficients need to be measured experimentally.
6. The extinction coefficient is calculated using the equation:

$$E(\text{Prot}) = \text{Numb}(\text{Tyr}) \times \text{Ext}(\text{Tyr}) + \text{Numb}(\text{Trp}) \times \text{Ext}(\text{Trp}) + \text{Numb}(\text{Cystine}) \times \text{Ext}(\text{Cystine})$$

The absorbance (optical density) can be calculated using the following formula:

$$\text{Absorb}(\text{Prot}) = E(\text{Prot}) / \text{Molecular_weight}$$

The conditions at which these equations are valid are: pH 6.5, 6.0 M guanidium hydrochloride, 0.02 M phosphate buffer.

7. It has been shown (29–31) that the identity of the N-terminal residue of a protein plays an important role in determining its stability in vivo. It seems that the N-terminal residue plays a major role in the process of ubiquitin-mediated proteolytic degradation (for a review *see ref.* 32). The authors have, by site-directed mutagenesis, created beta-galactosidase proteins with different N-terminal amino acids. The β -gal proteins thus designed have strikingly different half-lives in vivo, from more than 100 h to less than 2 min, depending on the nature of the amino acid at the amino terminus and on the experimental model (yeast in vivo; mammalian reticulocytes in vitro, *E. coli* in vivo). The set of individual amino acids can thus be ordered with respect to the half-lives that they confer when present at the amino terminus of a protein (this is called the “N-end rule”).
8. Statistical analysis of 12 unstable and 32 stable proteins has revealed (33) that there are certain dipeptides, the occurrence of which is significantly different in the unstable proteins compared with those in the stable ones. The authors of this method have assigned a weight value of instability to each of the 400 different dipeptides (DIWV). Using these weight values, it is possible to compute an instability index (II), which is defined as:

$$i = L-1$$

$$II = (10/L) \times \sum_{i=1}^{L-1} DIWV(x[i]x[i+1])$$

$$i=1$$

where: L is the length of sequence

DIWV(x[i]x[i+1]) is the instability weight value for the dipeptide starting in position i.

9. The aliphatic index of a protein is calculated according to the following formula (34):

$$\text{Aliphatic index} = X(\text{Ala}) + a \times X(\text{Val}) + b \times [X(\text{Ile}) + X(\text{Leu})]$$

where $X(\text{Ala})$, $X(\text{Val})$, $X(\text{Ile})$, and $X(\text{Leu})$ are mole percent ($100 \times$ mole fraction) of alanine, valine, isoleucine, and leucine. The coefficients a and b are the relative volume of valine side chain ($a = 2.9$) and of Leu/Ile side chains ($b = 3.9$) to the side chain of alanine.

10. The “Monoisotopic Mass” option is useful in the mass prediction of small peptides (<3000 Da) that can often be isotopically resolved on mass spectrometers. The $(\text{M}+\text{H})^+$ option will calculate all peptide masses with an extra hydrogen atom, to give values for singly charged peptides as found in electrospray and MALDI-TOF mass spectrometers.
11. The program does take into account most types of *N*- or *O*-linked glycosylation or other complex modifications like glycan phosphatidyl-inositol anchors because of their unpredictable heterogeneity. However, the discrete *O*-GlcNAc and C-mannosylation modifications are considered.
12. The PeptideMass program does not *predict* new potential posttranslational modifications in user-entered sequences, and thus does not consider these in mass calculations. However, one can use the PROSITE database (<http://www.expasy.org/prosite/>), e.g., by using the ScanProsite tool (<http://www.expasy.org/tools/scanprosite/>, [35]) to predict the presence of posttranslational modifications in a sequence. A list of modifications documented in the PROSITE database can be found at: <http://www.expasy.org/prosite/browse/>. The ExPASy tools page further contains a section of prediction tools for sequence-based posttranslational modifications (e.g., for prediction of signal sequences, tyrosine sulfation, glycosylation, and so on). If PMF data is available for a protein, FindMod or GlycoMod can be used to predict potential posttranslational modifications.
13. PeptideCutter does not take into consideration any kind of modification, neither of the protein sequence (post-translational) nor of modifications evoked by the cleavage (e.g., conversion of Met into homoserine lactone upon cleavage with CNBr). If the user is interested in a more detailed analysis of the resulting peptide, we recommend using the PeptideMass program (see **Subheading 3.3.**).
14. An amino acid scale is defined by a numerical value assigned to each type of amino acid. The most frequently used scales are hydrophobicity scales, most of which were derived from experimental studies on partitioning of peptides in apolar and polar solvents, with the goal of predicting membrane-spanning segments that are highly hydrophobic, and secondary structure conformational parameter scales. In addition, many other scales exist, based on different chemical and physical properties of the amino acids.
15. Protein pI and Mw in TagIdent/AACompIdent are calculated as described for Compute pI/Mw (see **Subheading 3.1.**).
16. Care must be taken in the use of pI and Mw estimates from 2-D gels as part of protein-identification strategies. Windows around these estimates that are too narrow can exclude the correct identification from the list of candidate identifications. As a general rule, we use windows of $\text{pI} \pm 0.5$ units for proteins from bacteria and yeast, and $\text{pI} \pm 1.0$ units for mammalian proteins. We generally use a Mw window of $\pm 20\%$, but for proteins larger than 60,000 Da, a window of $\pm 10\%$ is sufficient, because of the more accurate estimations (in percentage terms) that can be made with higher mass proteins on gels. If proteins are thought to be highly posttranslationally modified, very large pI and/or Mw windows may be needed.
17. TagIdent is extremely useful for searching proteins in the database for the presence of sequence tags, as it can search in a species-specific manner and with pI and Mw param-

- eters. This is a valuable alternative to Basic Local Alignment Search Tool (BLAST) (36), which, although it can now be used with sequences as short as four amino acids (by increasing the E value), does not allow the restriction of a search with pI and Mw parameters.
18. Although TagIdent is certain to find all sequences present in Swiss-Prot and TrEMBL, it is still possible that the result misses some existing proteins if their coding sequences (CDS) have not been annotated in the nucleotide sequence database by the submitters. This would prevent a protein entry from being included in TrEMBL, and the sequence will only be integrated in Swiss-Prot if an annotator detects this previously unannotated CDS during the manual annotation process. The same is true if the original EMBL entry has an incorrectly predicted initiation site that has not yet been corrected by a Swiss-Prot annotator.
 19. If you specify parameters that generate an extremely large TagIdent output (>1 mega-byte), only the first 1000 lines will be sent by e-mail. This is to avoid problems that can arise when large messages arrive at some e-mail sites.
 20. If AA analyses yield unreliable data for one or more amino acids, such values can be ignored and matching undertaken only with “good” amino acids, using the AACompIdent free constellation (37). The free constellation also allows the user to modify the bias and weight for each AA, if desired.
 21. When calibration proteins are used, AACompIdent compares the experimental composition of the protein against the theoretical composition in the Swiss-Prot database to create a factor set. This factor set is then applied to the experimental composition of the unknown protein before it is matched against the Swiss-Prot database. Use of calibration proteins can increase identification efficiency dramatically, and is advised wherever possible. Note, however, that calibration proteins should be electrophoretically prepared in the same manner as unknown proteins, and subjected to AA analysis in parallel with unknown proteins. It is also essential that the complete sequence of calibration proteins be in the Swiss-Prot database, as calibration cannot be done if only a fragment of the calibration protein sequence is available.
 22. Protein AA composition and Mw are highly conserved across species boundaries and serve as useful parameters for cross-species protein identification (15,16). Protein pI is, however, poorly conserved between species. Cross-species protein identification in AACompIdent can be done by specifying “ALL” for the species of interest, or specifying the Swiss-Prot species code of a well-defined organism that is closely related to the species under study. It must be noted that high-confidence cross-species protein identification usually requires peptide mass data or sequence as well as AA composition (see **Subheading 4.4.**).
 23. AACompSim automatically uses the theoretical pI and Mw of the specified protein in the matching procedure. The pI and Mw are calculated as in Compute pI/Mw (see **Subheading 3.1.**).
 24. Default windows of $pI \pm 0.25$ and $Mw \pm 20\%$ are used by AACompSim in matching; however, matches undertaken without restriction to these windows are also included in the program output.
 25. While peptide masses for any protein type are not as well conserved across species boundaries as other parameters (38), they can be used for cross-species protein identification in conjunction with, for example, amino acid composition (16,17).
 26. The algorithm has no major difficulties when working with very crowded spectra (i.e., with a high number of input masses, >100) and with a large number of theoretical masses. Consequently, increasing the number of possible peptides by taking into account combinations of missed cleavages, posttranslational modifications, alternative splicing, and chemical modifications, is conceivable. However in addition to increasing the number of

- true-positive peptide matches, there is also a risk of increasing the number of false-positive hits.
- 27. Aldente does not do any *de novo* prediction of posttranslational modifications on proteins. All modified peptides shown in the results will be the verification of an event documented in Swiss-Prot. However, Aldente can match peptides whose modifications are documented in Swiss-Prot as “potential” or “by similarity,” and thus allows predicted post-translational modifications to be validated.
 - 28. If no number is specified for pI and/or Mw, the parameter will not be used in the Aldente score.
 - 29. The peptide masses should be specified with a high precision and can be supplied in the form (one per line) or uploaded in a file in plain-text format or in the .pkm (GRAMS) or .dta (SEQUEST) or .pkt (Data Explorer) formats used by peak identification software. These formats are described in detail in the online documentation on ExPASy.
 - 30. Multiple selection is possible by holding down the “Ctrl” key. We define “single species matching” where you, for example, have proteins from *E. coli* that you then match against only the *E. coli* proteins in the database. This is a good approach to use when the organism you are working with is molecularly well defined, or ideally, the subject of a genome project. If the source of your proteins is not molecularly well defined, it is best to do “cross-species matching.” Thus, for example, if you are working with proteins from *Candida albicans*, you may wish to either match your proteins against all proteins from fungi or against the fully sequenced yeast *Saccharomyces cerevisiae*. Note, however, that when cross-species matching, protein pI is frequently poorly conserved, but protein mass is generally very well conserved (38). You should take this into consideration when setting your pI and Mw values. On the contrary, peptide masses are not well conserved across species boundaries. The poor conservation of peptide mass data is expected, as a single-amino acid substitution in any peptide can drastically change its mass.
 - 31. Mass spectrometers typically have a mass-dependent error associated with mass measurements, which cannot be uniformly expressed in Daltons. The use of ppm can therefore be more accurate. If both Δ Da and Δ ppm have been specified, the program will combine them with a logical OR.
 - 32. Both MALDI and electrospray machines are now capable of achieving single decimal point mass resolution; however, this may depend on the care that has been taken in machine calibration and use of internal standards. We recommend the use of a tolerance of 0.2 Da or 200 ppm or better whenever it is possible. Electrospray ionization (ESI)-TOF mass spectrometers or MALDI-TOF apparatus equipped with delayed extraction and ion reflectors are ideal for this, since most can deliver monoisotopic masses below ± 40 ppm, when two-point internal calibration is used.
 - 33. Both MALDI and ES machines are now capable of achieving single decimal point mass resolution; however, this may depend on the care that has been taken in machine calibration and use of internal standards. We recommend the use of a tolerance of 0.2 Da or 200 ppm or better whenever it is possible. ESI-TOF mass spectrometers or MALDI-TOF apparatus equipped with delayed extraction and ion reflectors are ideal for this, because most can deliver monoisotopic masses below ± 40 ppm, when two-point internal calibration is used.
 - 34. Aldente supports any user-defined chemical modifications to amino acids, given the locus where the modification should appear and the chemical formula of the product to add/remove on this locus. Two types of modifications can be applied. Use fixed modification whenever amino acids should be modified, and specify in the tolerance field the number of exceptions allowed. For example, for carboxymethylation on cysteine (CAM) with fixed option and a tolerance of 1, the program will generate the theoretical peptide with all cysteines modified, and peptides with all but one cysteine modified. On the contrary, use

variable modification when residues may or may not be modified, and specify in the tolerance field the maximum number of modified residues expected. For example, for methionine oxidation (MSO) with variable option and a tolerance of 2, the program will generate the theoretical peptides with 0, 1, or 2 methionines modified. The program supports every combination of possible modifications (or PTMs, *see Note 11*) occurring on the same locus. In an advanced mode, the user can specify for each modification a factor to be applied on the score to penalize peptides with (un)modified locus.

34. For all types of PTMs annotated in Swiss-Prot, two modes of peptide modification are available: fixed or variable. In fixed mode, the program will generate the theoretical peptide with all modified loci. In variable mode, the program will generate theoretical peptides with all combinations of modified or unmodified loci. If several PTMs (or chemical modifications; *see Note 33*) are possible at the same sequence position, the program will generate the theoretical peptides corresponding to every possible combination.
35. Swiss-Prot annotation distinguishes between experimentally proven and computationally predicted posttranslational modifications, as well as those inferred by similarity (5). The Swiss-Prot document “How is biochemical information assigned to sequence entries” (<http://www.expasy.org/txt/annbioch.txt>) describes how these nonexperimental qualifiers are used.
36. The program performs two runs. First, it keeps the n best proteins within the user protein and peptide mass tolerance. In a second pass, the program finds the best line fitting the maximum of hits (matching peptide masses) only for those n best proteins. Because of a higher precision in the second step, it is possible that some hits from the first step are removed, which means that the number of hits in the output result can become smaller than the number of hits you requested to see.
37. In the case of a manually entered sequence, the user is required to specify the biological source of the query protein. This information is used to determine whether certain PTMs are likely to occur in the sequence.
38. Users should avoid using peptide masses known to be from autodigestion of an enzyme (e.g., trypsin!), or other artifactual peaks (e.g., matrix peaks). If you are not sure whether your set of masses contains such peaks, you may use FindPept (*see Subheading 4.8.*) to detect them.
39. Cysteine residues in proteins are usually subjected to reduction and then alkylation with different reagents before they are used to generate peptides. Such a reactant can be specified here as one of iodoacetamide, iodacetic acid, or 4-vinyl pyridene. If no reactant has been used and the protein has undergone polyacrylamide gel separation, acrylamide adducts are to be expected on part of the free cysteines, so the reactant “acrylamide” should be chosen.
40. You can request for all methionines in theoretical peptides to be oxidized. If this option is selected, the program will modify the theoretical masses of Met-containing peptides accordingly and consider both peptides with unmodified methionines and peptides with modified methionines. Note that proteins prepared by gel electrophoresis often show this modification.
41. Other possible causes for nonmatched masses include signal autosuppression on MALDI or ESI and modifications on peptides (natural or artifactual; FindMod and GlycoMod can identify some of the former).
42. A set of 20 different proteases of different sources are available, and their sequence and cleavage rules are used to detect specific cleavage, autolysis peaks, and also to digest frequent contaminants.
43. For example, the cleavage rule for trypsin that is widely used in identification programs is “cleavage after Lys or Arg, except if followed by Pro.” Experimental data (9) show, how-

ever, a much more complex specificity pattern—in particular, the negative influence of charged residues immediately adjacent to the targeted Lys or Arg. This rule has been used to refine the cleavage prediction and is available in ExPASy tools as “Trypsin, higher specificity.”

44. Protease specificity may be adversely affected by chemical alterations in the enzyme. Solution conditions and temperature can affect specificity. Longer incubation times increase the yield of unspecific cleavage. Specificity can be increased by the addition of inhibitors, short incubation times, and an appropriate ratio of protease to substrate, also to limit protease autolysis (39).
45. The program can take into account a notable number of posttranslational modifications of discrete mass (it shares the database used by the FindMod and Aldente tools), plus several chemical modifications resulting from the experimental treatment of the protein. Most of these modifications cannot be expected to be applied quantitatively, except some chemical treatments such as carbamidomethylation (by iodoacetamide) or methanolic esterification. Therefore, FindPept applies them combinatorially on each peptide, assuming that no two modifications can simultaneously occur on a single amino acid.
46. The combined effect of unspecific cleavage and posttranslational and artifactual modifications gives rise to a huge number of possible peptides that may match the input masses. Since unrelated peptides may have very similar masses, false attributions can be limited only if the measurements are done with a high precision. The experimental mass error should not exceed 20–25 ppm. Even then, several attributions are sometimes made for a single mass. In these cases, the most likely peptides are those that have a specific cleavage site at one of their ends (shown in the output with a red slash). Additional evidence should usually be sought by experimental methods to obtain a conclusive determination.
47. Side-chain and C-terminal carboxylic acid groups can be esterified to methyl esters. Additionally, the box “N-Acetylation and N-formylation” can be used to submit the N-terminal residue of the protein to possible acetylation or formylation. Both the modified and unmodified versions of N-terminal peptides are examined.
48. An intermediate page may appear instead of the output page if you have submitted the accession number of a Swiss-Prot/TrEMBL entry that contains several chains or mature peptides and/or a cleaved initiator methionine at the beginning of the sequence. You should select the sequence to be analyzed: either a single chain, the uncleaved precursor, or the sequence with the initiator methionine added. If you suspect that the characterized protein has not matured as expected *in vivo*, you should consider the precursor sequence for the analysis. The numbering of the residues will be the same as in the original Swiss-Prot/TrEMBL entry. If an initiator methionine is added before the beginning of the sequence, it is assigned the position 0.
49. The considered contaminant keratins are human cytokeratins 1, 2, 9, and 10 (Swiss-Prot accessions P04264, P35908, P35527, and P13645), which have been determined to be abundant in skin and dandruff (40) and are often encountered as contaminants in biological samples handled in the laboratory.

Acknowledgments

This work was supported by the National Institutes of Health (NIH) grant 1 U01 HG02712-01 and by the Swiss Federal Government through the Federal Office of Education and Science.

References

1. Boeckmann, B., Bairoch, A., Apweiler, R., et al. (2003) The Swiss-Prot protein knowledge-base and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 354–370.

2. Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R. D. and Bairoch, A. (2003). ExPASy—the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.* **31**, 3784–3788.
3. Apweiler, R., Bairoch, A., Wu, C. H., et al. (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.* **432**, D115–D119.
4. Jung, E., Gasteiger, E., Veuthey, A.-L., and Bairoch A. (2001) Annotation of glycoproteins in the SWISS-PROT database. *Proteomics* **1**, 262–268.
5. Farriol-Mathis, N., Garavelli, J. S., Boeckmann, B., et al. (2004), Annotation of post-translational modifications in the Swiss-Prot knowledgebase. *Proteomics*, in press.
6. Gill, S. C, von Hippel, P. H. (1989) Calculation of protein extinction coefficients from amino acid sequence data. *Anal. Biochem.* **182**, 319–326.
7. Kyte, J., and Doolittle, R. F. (1982) A simple method for displaying the hydrophobic character of a protein. *J. Mol. Biol.* **157**, 105–132.
8. Wilkins, M. R., Lindskog, I., Gasteiger, E., et al. (1997) Detailed peptide characterization using PEPTIDEMASS—a World-Wide-Web-accessible tool. *Electrophoresis* **18**, 403–408.
9. Keil, B. (1992) *Specificity of proteolysis*. Springer-Verlag Berlin-Heidelberg-New York, p. 335.
10. Wilkins, M. R., Gasteiger, E., Sanchez, J.-C., Appel, R. D., and Hochstrasser, D. F. (1996) Protein identification with sequence tags. *Curr. Biol.* **6**, 1543–1544.
11. Wilkins, M. R., Gasteiger, E., Tonella, L., et al. (1998) Protein identification with N- and C-terminal sequence tags in proteome projects. *J. Mol. Biol.* **278**, 599–608.
12. Ashburner, M., Ball, C. A., Blake, J. A., et al. (2000), Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29.
13. Wilkins, M. R., Pasquali, C., Appel, R. D., et al. (1996) From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis. *Bio/Technology* **14**, 61–65.
14. Wilkins, M. R., Ou, K., Appel, R. D., et al. (1996) Rapid protein identification using N-terminal “sequence tag” and amino acid analysis. *Biochem. Biophys. Res. Commun.* **221**, 609–613.
15. Hobohm, U. and Sander, C. (1995) A sequence property approach to searching protein databases. *J. Mol. Biol.* **251**, 390–399.
16. Cordwell, S. J., Wilkins, M. R., Cerpa-Poljak, A., et al. (1995) Cross-species identification of proteins separated by two-dimensional gel electrophoresis using matrix-assisted laser desorption time of flight mass spectrometry and amino acid composition. *Electrophoresis* **16**, 438–443.
17. Wheeler, C. H., Berry, S. L., Wilkins, M. R., et al. (1996) Characterisation of proteins from 2-D gels by matrix-assisted laser desorption mass spectrometry and amino acid compositional analysis. *Electrophoresis* **17**, 580–587.
18. Wilkins, M. R., Gasteiger, E., Wheeler, C., et al. (1998) Multiple parameter cross-species protein identification using MultIdent—a world wide web accessible tool. *Electrophoresis* **19**, 3199–3206.
19. Wilkins, M. R., Gasteiger E., Gooley, A. A., et al. (1999) High-throughput mass spectrometric discovery of protein post-translational modifications. *J. Mol. Biol.* **289**, 645–657.
20. Hulo, N., Sigrist, C. J., Le Saux, V., et al. (2004) Recent improvements to the PROSITE database. *Nucleic Acids Res.* **32**, D134–D137.
21. Henikoff, S., and Henikoff, J. G. (1993) Performance evaluation of amino acid substitution matrices. *Proteins* **17**, 49–61.
22. Cooper, C. A., Gasteiger, E., and Packer, N. (2001) GlycoMod—a software tool for determining glycosylation compositions from mass spectrometric data. *Proteomics* **1**, 340–349.

23. Cooper, C. A., Gasteiger, E., and Packer, N. (2003) Predicting glycan composition from experimental mass using GlycoMod. In: (Conn, P.M., ed.) *Handbook of Proteomic Methods*, Humana, Totowa, NJ: pp. 225–231.
24. Gattiker, A., Bienvenut, W. V., Bairoch, A., and Gasteiger, E. (2002) FindPept, a tool to identify unmatched masses in peptide mass fingerprinting protein identification. *Proteomics* **2**, 1435–1444.
25. Bjellqvist, B., Hughes, G., Pasquali, C., et al. (1993) The focusing positions of polypeptides in immobilized pH gradients can be predicted from their amino acid sequences. *Electrophoresis* **14**, 1023–1031.
26. Bjellqvist, B., Basse, B., Olsen, E., and Celis, J. E. (1994) Reference points for comparisons of two-dimensional maps of proteins from different human cell types defined in a pH scale where isoelectric points correlate with polypeptide compositions. *Electrophoresis* **15**, 529–539.
27. Hoogland, C., Sanchez, J.-C., Tonella, L., et al. (2000) The 1999 SWISS-2DPAGE database update. *Nucleic Acids Res.* **28**, 286–288.
28. Nielsen, H., Engelbrecht, J., Brunak, S., and von Heijne, G. (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* **10**, 1–6.
29. Bachmair, A., Finley, D., and Varshavsky, A. (1986) In vivo half-life of a protein is a function of its amino-terminal residue. *Science* **234**, 179–186.
30. Gonda, D. K., Bachmair, A., Wunning, I., Tobias, J. W., Lane, W. S., and Varshavsky, A. J. (1989) Universality and structure of the N-end rule. *J. Biol. Chem.* **264**, 16,700–16,712.
31. Tobias, J. W., Shrader, T. E., Rocap, G., and Varshavsky, A. (1991) The N-end rule in bacteria. *Science* **254**, 1374–1377.
32. Ciechanover, A. and Schwartz, A. L. (1989) How are substrates recognized by the ubiquitin-mediated proteolytic system? *Trends Biochem. Sci.* **14**, 483–488.
33. Guruprasad, K., Reddy, B. V. B., and Pandit, M. W. (1990) Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein Eng.* **4**, 155–161.
34. Ikai, A. J. (1980) Thermostability and aliphatic index of globular proteins. *J. Biochem.* **88**, 1895–1898.
35. Gattiker, A., Gasteiger, E., and Bairoch, A. (2002) ScanPosite: a reference implementation of a PROSITE scanning tool. *Applied Bioinform.* **1**, 107–108.
36. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.
37. Golaz, O., Wilkins, M. R., Sanchez, J.-C., Appel, R. D., Hochstrasser, D. F., and Williams, K. L. (1996) Identification of proteins by their amino acid composition: an evaluation of the method. *Electrophoresis* **17**, 573–579.
38. Wilkins, M. R. and Williams, K. L. (1997) Cross-species protein identification using amino acid composition, peptide mass fingerprinting, isoelectric point and molecular mass: a theoretical evaluation. *J. Theor. Biol.* **186**, 7–15.
39. Hara, S., Rosenfeld, R., and Lu, H. S. (1996) Preventing the generation of artifacts during peptide map analysis of recombinant human insulin-like growth factor-I. *Anal. Biochem.* **243**, 74–79.
40. Parker, K. C., Garrels, J. I., Hines, W., et al. (1998) Identification of yeast proteins from two-dimensional gels: working out spot cross-contamination. *Electrophoresis* **19**, 1920–1932.

Protein Sequence Databases

Michele Magrane, Maria Jesus Martin, Claire O'Donovan, and Rolf Apweiler

1. Introduction

With the availability of almost 150 completed genome sequences from both eukaryotic and prokaryotic organisms, efforts are now being focused on the identification and functional analysis of the proteins encoded by these genomes. The rapidly emerging field of proteomics, the large-scale analysis of these proteins, has started to generate huge amounts of data as a result of the new information provided by the genome projects and by a range of new technologies in protein science. For example, mass spectrometry approaches are being used in protein identification and in determining the nature of posttranslational modifications (1), and large-scale yeast two-hybrid screens provide valuable data about protein–protein interactions (2). These and other methods now make it possible to quickly identify large numbers of proteins in a complex, to map their interactions in a cellular context, to determine their location within the cell, and to analyze their biological activities. Protein sequence databases play a vital role as a central resource for storing the data generated by these efforts and making them freely available to the scientific community. Data from large-scale experiments are often no longer published in a conventional sense but are deposited in a database. This means that protein sequence databases are the most comprehensive resource of information on proteins available to scientists.

In order to exploit the various resources fully, it is essential to distinguish between them and to identify the types of data they contain. Universal protein databases cover proteins from all species, while specialized data collections contain information about a particular protein family or group of proteins, or related to a specific organism. Universal protein sequence databases can be further subdivided into two categories: simple archives of sequence data, or sequence repositories, where the data are stored with little or no manual intervention in the creation of the records; and expertly curated databases, in which the original data are enhanced by the addition of further information derived from sources such as published scientific literature. A number of the leading protein sequence databases will be presented here, and the characteristics of each database and the types of data they each provide to the scientific community will be discussed.

2. Sequence Repositories

A number of protein sequence databases act as repositories of protein sequences. These databases add little or no additional information to the sequence records they contain and generally make no effort to provide a nonredundant collection of sequences to users.

2.1. *GenPept*

The most basic example of this type of database is the GenBank Gene Products Data Bank, also known as GenPept, which is produced by the National Center of Biotechnology Information (3). The entries in the database are derived from translations of the sequences contained in the nucleotide database maintained collaboratively by DDBJ (4), the EMBL Nucleotide Sequence Database (5), and GenBank (6), and contain minimal annotation, which has been extracted primarily from the corresponding nucleotide entry. The entries lack any additional annotation, and the database does not contain proteins derived from amino acid sequencing. It presents a redundant view of the protein world, which means that each protein may be represented by multiple records and no attempt is made to group these records into a single database entry.

2.2. *NCBI's Entrez Protein*

The National Center for Biotechnology Information (NCBI)'s Entrez Protein (<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Protein>) is another example of a sequence repository. The database contains sequence data translated from the nucleotide sequences of the DDBJ/EMBL/GenBank database as well as sequences from Swiss-Prot (7), Protein Information Resource (PIR) (8), RefSeq (9), and the protein databank (PDB) (10). The database differs from GenPept in that many of the entries contain additional information, but much of the annotated data have been extracted from curated databases such as Swiss-Prot and PIR, so there is little novel information added to the entries that cannot be found in other data collections. And, as with GenPept, the sequence collection is redundant.

2.3. *RefSeq*

A more ambitious approach is taken by the Reference Sequence (RefSeq) collection, produced by the NCBI (<http://www.ncbi.nih.gov/RefSeq>). The aim of the project is to provide a nonredundant collection of reference protein sequences with links to gene and transcript information. RefSeqs exist for a limited set of species, including approx 1100 viruses and 150 bacteria, as well as a small number of higher organisms, such as human, mouse, rat, zebra fish, honeybee, sea urchin, cow, and several important plant species. The main features of the RefSeq collection include nonredundancy, explicitly linked nucleotide and protein sequences, updates to reflect current knowledge of sequence data and biology, data validation and format consistency, distinct accession series, and ongoing curation by NCBI staff and collaborators, with review status indicated on each record. However, the majority of the records are automatically generated with minimal manual intervention. In October 2003, the database contained 785,000 entries with approx 10,000 manually reviewed entries, so the database is closer to a sequence repository than to any of the curated databases discussed later.

3. Universal Curated Databases

Although repositories are an essential means of providing the user with sequences as quickly as possible, it is clear that, when additional information is added to a sequence, this greatly increases the value of the resource for users. The curated databases take basic sequence information and enrich it by adding additional information from a range of sources such as the scientific literature. This information is extracted and validated by expert biologists before being added to the databases, and this means that the data in these collections can be considered to be highly reliable. There is also a large effort invested in maintaining nonredundant datasets by compiling all reports for a given protein sequence into a single record.

3.1. PIR-PSD

The oldest universal curated protein sequence database is the Protein Information Resource Protein Sequence Database (PIR-PSD) (<http://pir.georgetown.edu/>). It was established in 1984 as a successor to the original NBRF Protein Sequence Database, developed over a 20-yr period by the late Margaret O. Dayhoff and published as the “Atlas of Protein Sequence and Structure” from 1965 to 1978 (11). It is now a joint effort by Georgetown University Medical Center and the National Biomedical Research Foundation in Washington, DC.

It compiles comprehensive, nonredundant protein sequence data, organized by superfamily and family, and annotated with functional, structural, bibliographic, and genetic data. In addition to the sequence data, the database contains the name and classification of the protein, the name of the organism in which it naturally occurs, references to the primary literature, function and general characteristics of the protein, and regions of biological interest within the sequence. The database is extensively cross-referenced with DDBJ/EMBL/GenBank nucleic acid and protein identifiers, PubMed and MEDLINE IDs, and unique identifiers from many other source databases. In October 2003, the database contained 273,339 annotated and classified entries, covering the entire taxonomic range and organized into 36,000 superfamilies and over 100,000 families.

3.1.1. Data Processing

The primary sources of PIR-PSD data are naturally occurring wild-type sequences from DDBJ/EMBL/GenBank translations, published literature, and direct submissions to PIR-International. Unique protein sequence reports are assigned an accession number and then undergo merging, annotation, and classification. To provide a truly nonredundant database, both identity and nonidentity merges are performed to generate reliable reference sequences for annotation. The original sequences can be regenerated from residue lines in accession blocks nested within reference blocks.

3.1.2. Annotation

PIR-PSD employs controlled vocabularies and adopts standard nomenclature whenever applicable. The use of status tags such as “validated” or “similarity” in entry titles and function and complex annotations, as well as tags such as “experimental,” “predicted,” “absent,” or “atypical” in feature annotations helps in distinguishing experi-

mental from predicted data. Rule-based and classification-driven procedures are used to propagate annotations among similar sequences and to perform integrity checks.

3.1.3. Family Classification

Protein family classification is central to the organization and annotation of PIR. Automated procedures have allowed the placement of more than 99% of sequences into families and more than 70% into superfamilies. Sequences in PIR-PSD are also classified with homology domains and sequence motifs. Homology domains, which are shared by more than one superfamily, may constitute evolutionary building blocks, while sequence motifs represent functional sites or conserved regions. The classification approach improves the sensitivity of protein identification, helps to detect and correct genome annotation errors systematically, and allows a more complete understanding of sequence-function-structure relationships.

3.2. Swiss-Prot

The leading universal curated protein sequence database is Swiss-Prot, which was established in 1986 and is maintained collaboratively by the European Bioinformatics Institute (EBI) (<http://www.ebi.ac.uk/Swiss-Prot/>) and the Swiss Institute of Bioinformatics (SIB) (<http://www.expasy.org/sprot/>). The database has grown steadily since its inception, with an average of almost 6500 new entries added each year. Swiss-Prot contains data from a wide variety of organisms: as of October 2003, release 42.0 contained 135,850 curated sequence entries from over 8000 different species.

3.2.1. Distinguishing Features

The database is nonredundant, which means that all reports for a given protein are merged into a single entry, thus summarizing many pages of scientific literature into a concise yet comprehensive record. It also provides a high level of integration with other databases (12). Cross-references link Swiss-Prot to more than 50 different sequence databases and specialized data collections, and this allows users to access a large amount of additional information related to a particular protein. The database strives to provide a high level of annotation through a process of literature-based manual curation, and this allows the addition of as much accurate and up-to-date information as possible about each protein. This includes descriptions of properties such as the function of a protein, posttranslational modifications, domains and sites, secondary and quaternary structure, similarities to other proteins, diseases associated with deficiencies of a protein, developmental stages in which the protein is expressed, in which tissues the protein is found, pathways in which the protein is involved, and sequence conflicts and variants.

3.2.2. Entry Format

The database consists of sequence entries composed of different line types, each with its own specified format. The large amounts of different data types found in the databases are stored in a highly structured and uniform manner, which simplifies data access for users and data retrieval by computer programs. The entries contain core data, which are generally provided by the submitter of the sequence. These consist of sequence data, citation information, and taxonomic data. In addition, further information is added by a team of biologists during a process of literature-based curation and

rigorous sequence analysis. The annotation added during this process is stored mainly in the description (DE) and gene (GN) lines, the comment (CC) lines, the feature table (FT) lines, and the keyword (KW) lines.

The DE line lists all the names by which a protein is known, and includes standardized names assigned by official nomenclature bodies as well as Enzyme Commission (EC) numbers where applicable. An example of a Swiss-Prot description showing the main protein name, EC number, and additional synonyms is as follows:

DE Acid ceramidase precursor (EC 3.5.1.23) (Acylsphingosine deacylase)
DE (*N*-acylsphingosine amidohydrolase) (AC) (Putative 32 KDA heart
DE protein) (PHP32).

To promote interoperability, Swiss-Prot includes the unique gene identifiers assigned by genome sequencing projects, and these are used to link to genome databases where possible. Use is also made of authoritative gene name sources such as Genew, the database of the HUGO gene nomenclature committee (13), FlyBase (14), and the Mouse Genome Database (MGD) (15).

The CC lines are free text comments, which are used to convey any useful information about a protein. The information in the CC lines is contained in a number of defined topics that allow the easy retrieval of specific categories of data from the database. Although free text is permissible within the comments, as this is often necessary to convey detailed and complex information about a protein, a number of comments have a standardized syntax. A list of the currently used comment topics and their definitions is shown in **Table 1**. There are over 460,000 comments in SWISS-PROT release 42.0, with an average of three comments per entry. An example of the comment lines found in a single entry is shown in **Fig. 1**.

The FT lines provide position-specific data relating to the sequence. The lines have a fixed format and a defined set of feature keys. These feature keys describe domains and sites of interest within a sequence, such as posttranslationally modified residues, binding sites, enzyme active sites, secondary structure, and any other regions of interest. The full list of currently defined feature keys is available in the SWISS-PROT user manual at <http://www.expasy.ch/txt/userman.txt>. There are 720,924 sequence features in SWISS-PROT release 42.0, with an average of five features per entry.

The keywords are found in the keyword or KW lines of an entry. They serve as a subject reference for each sequence and assist in the retrieval of specific categories of data from the database. SWISS-PROT maintains a controlled keyword list, which currently contains almost 900 keywords, each with a definition to clarify its biological meaning and intended usage. This list is available at <http://www.expasy.org/cgi-bin/keylist.pl> and is updated on a regular basis.

3.2.3. Manual Annotation

Each entry in SWISS-PROT is thoroughly analyzed and annotated by biologists to ensure a high standard of annotation and to maintain the quality of the database. A process of literature-based curation is used to extract experimental and validated data that will improve the content of the knowledgebase. This experimental knowledge is supplemented by manually confirmed results from various sequence-analysis programs. At the time of annotation of a record, use is made of all relevant literature to ensure that

Table 1
Comment Topics Used in the Swiss-Prot Database

Comment topic	Description
ALLERGEN	Information relevant to allergenic proteins
ALTERNATIVE PRODUCTS	Description of the existence of protein sequences produced by alternative splicing of the same gene or by the use of alternative initiation codons.
BIOTECHNOLOGY	Description of the biotechnological use(s) of a protein
CATALYTIC ACTIVITY	Description of the reaction(s) catalyzed by an enzyme
CAUTION	Warns about possible errors and/or grounds for confusion
COFACTOR	Description of an enzyme cofactor
DATABASE	Description of a cross-reference to a database for a specific protein
DEVELOPMENTAL STAGE	Description of the developmental-specific expression of a protein
DISEASE	Description of disease(s) associated with a deficiency of a protein
DOMAIN	Description of the domain structure of a protein
ENZYME REGULATION	Description of an enzyme regulatory mechanism
FUNCTION	Description of the function(s) of a protein
INDUCTION	Description of compound(s) which stimulate the synthesis of a protein
MASS SPECTROMETRY	Reports the exact molecular weight of a protein or part of a protein as determined by mass spectrometric methods
MISCELLANEOUS	Any comment which does not belong to any of the other defined topics
PATHWAY	Description of the metabolic pathway(s) with which a protein is associated
PHARMACEUTICAL	Description of the use of a protein as a pharmaceutical drug
POLYMORPHISM	Description of polymorphism(s)
PTM	Description of a posttranslational modification
SIMILARITY	Description of the similarity (sequence or structural) of a protein with other proteins
SUBCELLULAR LOCATION	Description of the subcellular location of the mature protein
SUBUNIT	Description of the quaternary structure of a protein
TISSUE SPECIFICITY	Description of the tissue specificity of a protein

the functional information included is complete and up to date. As new information arises, entries are updated so that they always reflect the current state of knowledge in the literature. The addition of a number of qualifiers in the comment and feature table lines during the annotation process allows users to distinguish between experimentally verified data, data which have been propagated from a characterized protein based on sequence similarity, and data for which no experimental evidence currently exists (16).

3.3. TrEMBL

To produce a fully curated Swiss-Prot entry is a highly labor-intensive process and is the rate-limiting step in the growth of the database, as new sequences are submitted

- **FUNCTION** 3BETA-HSD IS A BIFUNCTIONAL ENZYME, THAT CATALYZES THE OXIDATIVE CONVERSION OF DELTA(5)-ENE-3-BETA-HYDROXY STEROID, AND THE OXIDATIVE CONVERSION OF KETOSTEROIDS. THE 3BETA-HSD ENZYMATIC SYSTEM PLAYS A CRUCIAL ROLE IN THE BIOSYNTHESIS OF ALL CLASSES OF HORMONAL STEROIDS.
- **CATALYTIC ACTIVITY** 3-beta-hydroxy-delta ⁵-steroid + NAD ⁺ = 3-oxo-delta ⁵-steroid + NADH.
- **CATALYTIC ACTIVITY** A 3-oxo-delta ⁵-steroid = a 3-oxo-delta ⁴-steroid.
- **PATHWAY**: Steroid biosynthesis.
- **SUBCELLULAR LOCATION** Endoplasmic reticulum and mitochondrial membrane-bound protein.
- **TISSUE SPECIFICITY** PLACENTA AND SKIN. PREDOMINANTLY EXPRESSED IN MAMMARY GLAND TISSUE.
- **DISEASE**: Congenital deficiency of 3beta-HSD activity causes a severe depletion of steroid formation frequently lethal in early life. The classical form of this disease includes the association of severe salt-losing adrenal insufficiency and ambiguity of external genitalia in both sexes.
- **SIMILARITY**: Belongs to the 3beta-HSD family.

Fig. 1. The comment lines from a single Swiss-Prot entry, P14060 (3BH1_HUMAN).

more quickly than they can be manually annotated and integrated into the database. To address this, the TrEMBL (Translation from EMBL) database (<http://www.ebi.ac.uk/trembl/>) was introduced as a supplement to Swiss-Prot in 1996 to make new sequences available as quickly as possible while preventing the dilution of the high-quality annotation in Swiss-Prot (7). It consists of computer-annotated entries derived from the translation of all coding sequences in the DDBJ/EMBL/GenBank nucleotide sequence database that are not yet included in Swiss-Prot. To ensure completeness, it also contains a number of protein sequences extracted from the literature or submitted directly by the user community. There are more than 58,000 different species represented in the database. TrEMBL follows the Swiss-Prot format and conventions described above as closely as possible.

The production of TrEMBL starts with the translation of coding sequences in the DDBJ/EMBL/GenBank nucleotide sequence database. At this stage, all annotation in a TrEMBL entry derives from the corresponding nucleotide entry. The next step involves redundancy removal through merging of multiple records that correspond to the same protein into a single database entry (17). The second post-processing step is the automated enhancement of the information content in TrEMBL (18). The process is based on a system of standardized transfer of annotation from well-characterized proteins in Swiss-Prot to unannotated TrEMBL entries belonging to defined groups (19). To assign entries to these groups, the highly diagnostic protein family signature database InterPro (20) is used. This is an integrated resource of protein families, domains, and functional sites that amalgamates the efforts of the member databases, which are currently PROSITE (21), PRINTS (22), Pfam (23), ProDom (24), SMART (25), TIGRFAMS (26), PIR SuperFamilies (27), and SUPERFAMILY (28). This process brings the standard of annotation in TrEMBL closer to that found in Swiss-Prot through the addition of accurate, high-quality information to TrEMBL entries, thus improving the quality of data available to the user.

4. Specialized Databases

As well as the larger universal sequence collections, there are a number of more specialized databases, which focus on a particular protein group or family, or on a specific organism. These vary in size and in the scope of the data they contain.

4.1. Protein Family Databases

Protein family databases contain information related to a specific family or group of proteins. These databases are generally maintained by experts in the field, and due to the restricted nature of the data they contain, they are able to offer a finer level of granularity than may be possible in the universal databases. An example of such a database is MEROPS (<http://merops.sanger.ac.uk/>), an information resource for peptidases and their inhibitors (29). A summary page is provided for each peptidase, describing the classification and nomenclature of the protein and offering links to supplementary pages that show sequence identifiers, any known structures, and literature references. The proteins are classified using a hierarchical, structure-based approach in which each peptidase is assigned to a family on the basis of statistically significant similarities in amino acid sequence, and families that are thought to be homologous are grouped together in a clan. Information is also provided about naturally occurring peptidase inhibitors.

4.2. Organism-Specific Databases

Model organisms are an invaluable tool for understanding the basis of human diseases and biological processes. Increasing amounts of genetic, phenotypic, and protein-related information are being generated in a variety of model organism systems, and a number of databases dealing specifically with the biology of these organisms have been established to capture this information and provide it to the scientific community. Examples include FlyBase (<http://flybase.bio.indiana.edu/>), the database of the *Drosophila* genome, and the Mouse Genome Database (<http://www.informatics.jax.org/>), which contains information related to the laboratory mouse. These databases play an important role in providing integrated access to the data available about these organisms.

5. Protein Sequence Databases—The Next Generation

The increasing volume and complexity of protein data has meant that the protein databases have had to find ways to adapt to this data influx so that they can continue to play a central role as we enter the proteomics era. One of the most significant developments in this regard is the recent decision by the National Institutes of Health to award a \$15 million grant over the next 3 yr (30) to combine the Swiss-Prot, TrEMBL, and PIR-PSD databases into a single resource, the United Protein Database (UniProt). UniProt comprises three components: (1) the UniProt Knowledgebase, which will continue the work of Swiss-Prot and PIR by providing an expertly curated database; (2) the UniProt Archive (UniParc), into which new and updated sequences are loaded on a daily basis from Swiss-Prot, TrEMBL, PIR-PSD, EMBL, Ensembl (31), International Protein Index (IPI) (<http://www.ebi.ac.uk/IPI>), PDB, RefSeq, FlyBase, WormBase (32), and the patent offices in Europe, the United States, and Japan, making it the most comprehensive protein sequence database available; and (3) the UniProt nonredundant

reference databases (UniProt NREF), which provide nonredundant views of UniParc. The UniProt resource (<http://www.uniprot.org>) will provide a comprehensive, fully classified, and richly annotated protein sequence knowledgebase, and will build upon the foundations laid down by the consortium members.

References

1. Sickmann, A., Mreyen, M., and Meyer, H. E. (2003) Mass spectrometry—a key technology in proteome research. *Adv. Biochem. Eng. Biotechnol.* **83**, 141–76.
2. Coates, P. J. and Hall, P. A. (2003) The yeast two-hybrid system for identifying protein-protein interactions. *J. Pathol.* **199**, 4–7.
3. Wheeler, D. L., Church, D. M., Federhen, S., et al. (2003) Database resources of the National Center for Biotechnology. *Nucl. Acids Res.* **31**, 28–33.
4. Miyazaki, S., Sugawara, H., Gojobori, T., and Tateno, Y. (2003) DNA Data Bank of Japan in XML. *Nucleic Acids Res.* **31**, 13–16.
5. Stoessner, G., Baker, W., van den Broek, A., et al. (2003) The EMBL Nucleotide Sequence Database: major new developments. *Nucleic Acids Res.* **31**, 17–22.
6. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Wheeler, D. L. (2003) GenBank. *Nucleic Acids Res.* **31**, 23–27.
7. Boeckmann, B., Bairoch, A., Apweiler, R., et al. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 365–370.
8. Wu, C. H., Yeh, L. S., Huang, H., et al. (2003). The Protein Information Resource. *Nucleic Acids Res.* **31**, 345–347.
9. Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2003) NCBI Reference Sequence Project: update and current status. *Nucleic Acids Res.* **31**, 4–37.
10. Westbrook, J., Feng, Z., Chen, L., Yang, H., and Berman, H. M. (2003) The Protein Data Bank and structural genomics. *Nucleic Acids. Res.* **31**, 489–491.
11. Dayhoff, M. O. (1978) *Atlas of Protein Sequence and Structure* Vol. 5, Supplement 3. National Biomedical Research Foundation, Washington, DC.
12. Gasteiger, E., Jung, E., and Bairoch, A. (2001) SWISS-PROT: connecting biomolecular knowledge via a protein database. *Curr. Issues Mol. Biol.* **3**, 47–55.
13. Wain, H. M., Lush, M., Ducluzeau, F., and Povey, S. (2002) Genew: the human gene nomenclature database. *Nucleic Acids Res.* **30**, 169–171.
14. FlyBase consortium. (2003) The FlyBase database of the *Drosophila* genome projects and community literature. *Nucleic Acids Res.* **31**, 172–175.
15. Blake, J. A., Richardson, J. E., Bult, C. J., Kadin, J. A., and Eppig, J. T. (2003) MGD: the Mouse Genome Database. *Nucleic Acids Res.* **31**, 193–195.
16. Junker, V., Apweiler, R., and Bairoch, A. (1999) Representation of functional information in the Swiss-Prot data bank. *Bioinformatics* **15**, 1066–1067.
17. O'Donovan, C., Martin, M. J., Glemet, E., Codani, J., and Apweiler, R. (1999) Removing redundancy in Swiss-Prot and TrEMBL. *Bioinformatics* **15**, 258–259.
18. Apweiler, R. (2001) Functional information in SWISS-PROT: the basis for large-scale characterisation of protein sequences. *Briefings in Bioinformatics* **2**, 9–18.
19. Fleischmann, W., Moeller, S., Gateau, A., and Apweiler, R. (1998) A novel method for automatic and reliable functional annotation. *Bioinformatics* **15**, 228–233.
20. Mulder, N. J., Apweiler, R., Attwood, T. K., et al. (2003) The InterPro database, 2003 brings increased coverage and new features. *Nucleic Acids Res.* **31**, 315–318.
21. Falquet, L., Pagni, M., Bucher, P., et al. (2002) The PROSITE database, its status in 2002. *Nucleic Acids Res.* **30**, 235–238.
22. Attwood, T. K., Bradley, P., Flower, D. R., et al. (2003) PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res.* **31**, 400–402.

23. Bateman, A., Birney, E., Cerruti, L., et al. (2002) The Pfam protein families database. *Nucleic Acids Res.* **30**, 276–280.
24. Corpet, F., Servant, F., Gouzy, J., and Kahn, D. (2000) ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Res.* **28**, 267–269.
25. Letunic, I., Goodstadt, L., Dickens, N. J., et al. (2002) Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic Acids Res.* **30**, 242–244.
26. Haft, D. H., Selengut, J. D., and White, O. (2003) The TIGRFAMs database of protein families. *Nucleic Acids Res.* **31**, 371–373.
27. Huang, H., Barker, W. C., Chen, Y., and Wu, C. H. (2003) iProClass: an integrated database of protein family, function and structure information. *Nucleic Acids Res.* **31**, 390–392.
28. Gough, J., Karplus, K., Hughey, R., and Chothia, C. (2001) Assignment of homology to genome sequences using a library of Hidden Markov Models that represent all proteins of known structure. *J. Mol. Biol.* **313**, 903–919.
29. Rawlings, N. D., O'Brien, E., and Barrett, A. J. (2002) MEROPS: the protease database. *Nucleic Acids Res.* **30**, 343–346.
30. Butler, D. (2002) NIH pledges cash for global protein database. *Nature* **419**, 101.
31. Clamp, M., Andrews, D., Barker, D., et al. (2003) Ensembl 2002: accommodating comparative genomics. *Nucleic Acids Res.* **31**, 38–42.
32. Harris, T. W., Lee, R., Schwarz, E., et al. (2003) WormBase: a cross-species database for comparative genomics. *Nucleic Acids Res.* **31**, 133–137.

In Silico Characterization of Proteins

InterPro and Proteome Analysis

Nicola Jane Mulder, Manuela Pruess, and Rolf Apweiler

1. Introduction

The main problem we aim to solve in this chapter is the quick and reliable elucidation of protein function and large-scale analysis of whole proteomes (protein component of genomes). This problem arose with the advancement of DNA sequencing technologies and the dawning of the genome sequencing era. Previously, unclassified DNA sequences trickled into the public databases from bench scientists working on experimental investigation of the function of the gene products. However, currently the raw sequences are flooding in with a distinct lack of accompanying annotation, resulting in a requirement for automatic *in silico* protein sequence analysis tools. Traditionally, scientists use sequence similarity searches to compare a query sequence to those of known function, but this method has its limitations and relies on the quality of existing data. Here we describe improved methods for protein sequence classification using protein signatures.

A number of different databases developing protein signatures diagnostic for known protein families or domains have arisen. Each has its own focus, criteria, and method for creating the signatures, and as a result also its own strengths and limitations. In addition, the shear number of these databases is daunting for a bench scientist who is not necessarily qualified to understand the similarities, differences, and idiosyncrasies of each. The aim of this chapter is to explain these, as well as to provide a full description of an integrated resource that aims to solve the abovementioned problem. InterPro (1) unifies the major protein signature databases into a single, comprehensive resource with manual intervention by trained biologists to turn a database and software application into an understandable, usable tool for scientists world-wide. InterPro has been used by bench biologists analyzing a single gene product and by genome sequencing centers for the annotation of entire genomes. The analysis and comparison of whole genomes provides a powerful tool for identifying unique or target genes in the reference organism. For example, a comparison between pathogenic and nonpathogenic organisms may shed light on which genes or proteins are responsible for pathogenesis if they are found only in the former organisms. The Proteome Analysis database (2), another application of InterPro, aims to present a statistical analysis of completely sequenced proteomes for this purpose. This database will also be described in more detail below.

From: *The Proteomics Protocols Handbook*
Edited by: J. M. Walker © Humana Press Inc., Totowa, NJ

2. Methods

2.1. Protein Signature Methods

To create a protein signature, a number of related protein sequences are required. It is not possible to identify common features of a protein family or domain from one sequence; however, an alignment of related sequences can be used to create a consensus for the protein family, or identify conserved domains or residues. The highly conserved areas are likely to be involved in the common function of the related sequences—for example, highly conserved residue triads may indicate an active site or binding site. These conserved areas diagnostic of a protein family, domain, or functional site can be used to develop protein signatures using several different methods. These are then compared to a query protein sequence to identify which family the protein belongs to.

The simplest protein signature method uses regular expressions to show patterns of conserved amino acid residues. The regular expression specifies which amino acid(s) may or may not occur at each position. This core pattern is tested against a set of annotated sequences and optimized until it hits only the correct sequences in the test set (3). Regular expressions are useful for identifying highly conserved active sites and binding sites, but are limited in their ability to find more distantly related sequences due to their lack of flexibility in recognizing single-residue variations.

Another widely used protein signature method is a profile. Profiles are built from multiple sequence alignments, and are tables of position-specific amino acid weights and gap costs, or matrices describing the probability of finding an amino acid at a given position in the sequence (4). The numbers in the table (scores) are used to calculate similarity scores between a profile and a sequence for a given alignment. For each set of sequences, a threshold score is calculated to determine whether the query sequence is related to the original set of sequences in the alignment. An additional method derived from profiles is a hidden Markov model (HMM), which essentially is a statistical profile based on probabilities rather than scores (5,6). HMMs are most commonly derived from the HMMER package, written by Sean Eddy (7), which allows a user to create an HMM from a sequence alignment and to search a database of sequences against the HMM without the requirement of understanding how the HMMs work. Profiles and HMMs are powerful protein signature methods and compensate for the limitations of regular expressions in that they generally cover larger areas of the sequence, and are capable of identifying more divergent family members.

2.2. Protein Signature Databases

There are a number of protein signature databases in the public domain that use the methods described above to produce diagnostic signatures for protein families and domains, including the sequence-based PROSITE (8), Prints (9), Pfam (10), SMART (11), TIGRFAMs (12), and Protein Information Resource (PIR) SuperFamily (13), as well as the structure-based SUPERFAMILY (14). There are also databases that use sequence clustering and alignment methods—for example, ProDom (15) and CluSTr (16), which cluster all proteins in the database into families.

PROSITE is a database of both regular expressions and profiles, and has a primary focus on signatures for the annotation of Swiss-Prot (17) proteins. Prints uses a variation on profiles to produce fingerprints, which are a collection of motifs along the conserved regions of a protein sequence. Prints are particularly successful at diagnosing receptors and ion channels, and for their high granularity, showing several different levels in a protein family hierarchy.

Pfam, SMART, TIGRFAMs, PIR SuperFamily, and SUPERFAMILY all use HMMs, but in different ways. For example, each database may have different criteria for assembling their multiple sequence alignments, different means of calculating their thresholds, or different methods for postprocessing their results. Pfam is driven by coverage and aims to have HMMs to represent all sequence space. Pfam includes both families and domains in their sets, and where possible they base their HMMs on protein structural information. There are two parts to Pfam—the curated PfamA and the automatically generated PfamB, which covers those families not represented in PfamA. SMART and SUPERFAMILY tend to create HMMs for protein domains rather than families. SMART focuses predominantly on signaling, extracellular, and chromatin-associated proteins, while SUPERFAMILY bases its domains on SCOP (18) superfamilies, thus serving as the only fully “structure-based” signature database mentioned here.

TIGRFAMs and PIR SuperFamily are protein family oriented, although TIGRFAMs do contain some domain HMMs. TIGRFAMs are also used for annotation, particularly of microbial genomes, and as a result, their HMMs hit mostly bacterial proteins. They focus on creating HMMs based on actual function rather than just sequence similarity, so all proteins belonging to a TIGRFAMs family must be “equivalogs,” i.e., have the same function. They are therefore quite specific. PIR SuperFamily, on the other hand, only produces HMMs representing protein families, mostly at the superfamily level. The HMMs cover the full length of the protein sequences and are derived for those proteins containing the same domain composition with very similar sequence lengths. Multifunctional proteins are represented by different HMMs from those representing each part of the multifunctional protein, and no protein should match more than one unique PIR SuperFamily HMM.

2.3. Integration of These Databases into InterPro

While all the databases described above have significant overlaps in the protein families and domains they predict, they arrive at these overlaps by different means. Using just one of the databases to analyze a query sequence could result in no hits if the sequence is outside their range of coverage, and also makes one vulnerable to any limitations the chosen database may have. Conversely, however, trying to use all of them at the same time but from the separate sites may lead to confusion in trying to rationalize the different results obtained at each. This problem was resolved by the InterPro resource, which integrates all the protein-signature databases into one. Signatures from PROSITE, Prints, Pfam, ProDom, SMART, TIGRFAMs, PIR SuperFamily, and SUPERFAMILY that describe the same domain, family, repeat, active site, binding site, or post-translational modification, are grouped into single InterPro entries with unique accession numbers. InterPro entries that match a subset of proteins matched by other entries, are related to each other as parent/child (family and subfamily) or contains/found in (domain composition) relationships.

2.3.1. *InterPro Annotation*

Each InterPro entry contains high-quality manual annotation, providing useful information on the protein family, domain, and so on. This is in the form of a name (short description), abstract, and cross-references. Literature references cited in the abstract are stored in a reference field in each entry. Mapping of InterPro entries to Gene Ontology (GO) terms (19), where possible, provides additional functional annotation. The GO project is an effort to provide a universal ontology for describing gene products across all species, and has developed a set of terms in a directed acyclic graph under the three ontologies: molecular function, biological process, and cellular component. Where all proteins matching an InterPro entry have the same function, then the entry can be mapped to the appropriate GO term describing that function. InterPro entries are manually mapped to GO terms taking into account information in the abstract of the entries and annotation of proteins in the match lists. The associated GO terms should also apply to all proteins with true hits to all signatures in the InterPro entry. The mappings provide an automatic means of large-scale GO characterization of proteins.

Two new fields exist in InterPro entries—the “Structural links” field and the “Taxonomy” field. The former provides information on curated structure links based on the correspondence between the proteins matching the InterPro entry and those proteins of known structure and belonging to SCOP or CATH (20) superfamilies. The links are included only when the structural domains overlap considerably with one or more of the InterPro signatures on the protein sequence. The “Taxonomy” field aims to provide an “at a glance” view of the taxonomic range of the sequences associated with each InterPro entry. The lineages were carefully selected to provide a view of the major groups of organisms. The circular display has the taxonomy-tree root as its center, with the model organisms populating the outermost circle and nodes of the taxonomy tree placed on the inner circles. The nodes themselves are either true taxonomy nodes and have a National Center for Biotechnology Information (NCBI) taxonomy ID, or are artificial nodes created for this display, of which there are three: “Unclassified,” “Other Eukaryota (Non-Metazoa),” and the “Plastid Group.” The number of sequences associated with each lineage is displayed, and in the future this will be extended to facilitate downloading of the sequences.

2.3.2. *Protein Matches and Software*

In addition to annotation, each InterPro entry contains a list of precomputed matches to Swiss-Prot and TrEMBL proteins (17). Protein matches are calculated using the InterProScan (21) software package, which combines different protein signature recognition methods from each of the InterPro member databases into one package. The software is used to compute all matches for each entry, and is also available for user query sequence searches. The list of protein matches in InterPro entries may be viewed in tabular and different graphic formats. The table lists the protein accession numbers and the positions in the amino acid sequence where each signature from that InterPro entry hits. The match list may be displayed in a detailed graphic view in which the sequence is split into several lines, one for each hit by a unique signature (the bars are colored coded according to the member database). The graphic overview splits the protein sequence into different lines for each InterPro entry matched, and displays the consensus domain boundaries of all signatures within each entry. The graphic views include hits to all signatures from the same and other InterPro entries; thus, for

each sequence, the domain and/or motif organization can be seen at a glance. Where structures are available for proteins, there is a link from the graphic view to the corresponding protein databank (PDB) structures, and a separate line in the display showing the SCOP and/or CATH curated matches on the sequence as white-striped bars. Clicking on the protein accession number takes the user to the detailed view for that protein (see **Fig. 1**).

2.3.3. Using InterPro

InterPro is accessible via the Web interface at: <http://www.ebi.ac.uk/interpro> and through Sequence Retrieval System (SRS). The database can be searched via a text search or by inputting a query sequence and running it through InterProScan. All data and software are also available on the FTP site for downloading. A user manual and frequently asked questions are provided to help users get started.

2.4. Analysis of Whole Proteomes

The computer-derived analysis and characterization of proteins, as shown above, can also be applied to the analysis of whole proteomes. More and more complete proteomes of organisms are becoming available, and they represent an important source for meaningful comparisons between species. Specific databases and tools help to carry out the analysis of proteomes of completely sequenced organisms according to their protein composition, thus aiming at overcoming the lack of *in vivo* gathered knowledge about the functions of predicted proteins.

The Proteome Analysis database (2) has been set up to provide comprehensive statistical and comparative analyses of the predicted proteomes of fully sequenced organisms. It aims to integrate information from a variety of sources that will together facilitate the classification of the proteins in complete proteome sets. The analysis is compiled using InterPro, the CluSTr database (16), and GO Slim, and is performed on the nonredundant complete proteome sets of Swiss-Prot and TrEMBL entries. Proteome sets are built from the Swiss-Prot and TrEMBL protein sequence databases, which provide reliable, well-annotated data as the basis for the analysis. All completely sequenced organisms are included, with the exception of viruses. Currently (October 2003) proteome sets are available for a total of 150 organisms from archaea, bacteria, and eukaryota. Individual Web pages are available for each organism, on which different statistical analyses, protein classifications, taxonomic information, comparisons with other organisms, and additional links are provided. In addition, the whole, nonredundant proteome set, downloadable in Swiss-Prot or FASTA format, chromosome tables, and the possibility of doing a FASTA search against this proteome are available. For individual proteins from each of the proteomes, links to structural information databases like the homology-derived structures of proteins (HSSP) database (22) and the PDB (23) exist. Statistics like the protein length distribution and the amino acid composition of a whole proteome, are represented graphically, and structures of proteins can be viewed with molecular visualization software.

2.4.1. InterPro Analysis

The InterPro-based statistical analysis uses the InterPro hits for the proteins in a proteome and displays the data in different tables sorted by the frequency of occurrence or by type of InterPro entry (family, domains, repeats) matched. The percentage of proteins in the proteome matched by each entry is shown. There are also precom-

Interpro Entry	Method accession	Graphical match <input checked="" type="checkbox"/>	Method name
IPR000051:	P350193		SAM_BIND
IPR000780:	PF01739		CheR
IPR000780:	PF03705		CheR_M
IPR000780:	PR00998		CHERMTFRASE
IPR000780:	P350123		CHER
IPR000780:	SM00138		MeTrc
IPR001801:	P350124		MET_TRANS
NONE:	1bc5a1		1bc5a1
NONE:	1bc5a2		1bc5a2
NONE:	d1bc5a1		d1bc5a1
NONE:	d1bc5a2		d1bc5a2

Fig. 1. Example of the InterPro detailed graphic view for the *Salmonella typhimurium* chemotaxis protein methyltransferase. The purple, white, and black and white-striped bars represent the CATH and SCOP domains, respectively. The member database signatures are color coded according to database (see the InterPro website for the colors).

puted InterPro-based proteome comparisons available, in which different organisms, chosen by taxonomic relationship or general interest, can be compared at a glance (see Fig. 2 for an example). Furthermore, users can perform their own interactive proteome comparisons between any combination of organisms in the database. An easy-to-navigate search interface allows the user to choose one reference organism and one or many other organisms to perform the comparison with their chosen analysis output.

2.4.2. CluSTr Analysis

The CluSTr-based statistical analysis makes use of data that are provided by the CluSTr (Clusters of Swiss-Prot+TrEMBL proteins) database, which offers an automatic classification of Swiss-Prot and TrEMBL proteins into groups of related proteins. The clustering is based on the analysis of all pairwise comparisons between protein sequences. In the Proteome Analysis database, protein clusters within a proteome are shown (at different z scores), sorted by size, and also a list of singletons that do not belong to any clusters is available. Clusters of proteins that do not have links to HSSP or InterPro are disclosed too, since they display groups of proteins with unknown structures or functions that may be of interest for closer examination. The proteins in the clusters are sorted according to their GO Slim classification to allow a quick assessment of the functional composition of the cluster.

2.4.3. GO Slim Analysis

A functional classification of proteomes using GO (19) is available for each organism. This classification shows the general statistics for all proteins in the proteome that are assigned to a special selection of high-level terms from each of the three GO sections—molecular function, biological process, and cellular component. These selected terms are collectively called “GO Slim,” and were chosen to cover most aspects of each of these three ontologies without overlapping in the GO hierarchy. The data are derived from GO assignments based on manual assignment of GO terms to Swiss-Prot

Proteome Analysis @EBI

InterPro statistics [Help]

Oscode	Number of proteins in proteome	Proteins with InterPro matches (% of all proteins)	Number of signatures	Number of InterPro entries
HUMAN	29838	23036 (74.4%)	7296	4163
MOUSE	21041	16652 (79.1%)	7123	4053
RAT	6479	5952 (91.9%)	5311	2866

	<i>H. sapiens</i>	<i>M. musculus</i>	<i>R. norvegicus</i> (Rat)	
InterPro	Matches per genome [Proteins matched]	Matches per genome [Proteins matched]	Matches per genome [Proteins matched]	Name
PR0000001	501(26)	184(16)	125(8)	Kingpin
PR0000002	13(7)	4(2)	2(1)	Cdc20/Fizzy
PR0000003	65(10)	60(9)	39(6)	Retinol X receptor
PR0000005	6(6)	4(4)	0(0)	Helix-turn-helix, AraC type
PR0000006	79(21)	17(5)	12(3)	Vertebrate metallothionein
PR0000007	78(8)	55(6)	11(1)	Tubby
PR0000008	877(182)	620(116)	452(76)	C2 domain
PR0000009	68(6)	49(4)	52(4)	Protein phosphatase 2A regulatory subunit PR55
PR0000010	56(20)	70(25)	65(19)	Cysteine protease inhibitor
PR0000011	6(3)	9(5)	0(0)	Ubiquitin-activating enzyme
PR0000014	226(53)	140(33)	113(25)	PAS domain
PR0000018	9(1)	9(1)	9(1)	P2Y4 purinceptor
PR0000020	39(6)	43(7)	20(4)	Anaphylatoxin/Cribulin
PR0000022	6(6)	5(5)	3(3)	Carboxyl transferase
PR0000023	73(5)	60(4)	45(3)	Phosphofructokinase
PR0000024	112(32)	80(26)	55(14)	Fzoned CRD region
PR0000025	20(3)	20(3)	18(3)	Melatonin receptor

Fig. 2. Example of a static InterPro-based comparison between proteomes: InterPro matches for *Homo sapiens*, *Mus musculus*, and *Rattus norvegicus*. Clicking on the number of proteins matched brings up a table listing the protein sets matching the reference InterPro entry. This table shows the protein accession number and the description taken from the Swiss-Prot/TrEMBL entry, and has links to the InterPro graphic view and the CluSTr data.

and TrEMBL entries, as well as through the application of mappings from GO to InterPro entries, Swiss-Prot keywords, and Enzyme Commission numbers. The process of mapping GO to InterPro, keywords, and EC numbers is manual and ongoing due to the dynamic nature of GO.

2.4.4. Chromosome Tables

For most of the proteomes in the database, chromosome tables are available, providing an ordered list of genes, together with their chromosomal location, information about the proteins they encode, and useful links to other databases. Tables are provided for each chromosome, and for organelle genomes and plasmids, where these are considered to comprise part of the normal genome of a fully sequenced organism. For archaeal and bacterial proteomes, the chromosomal location for the start of each gene is given in nucleotides, using the DNA sequence in the corresponding EMBL Nucleotide Sequence database (24) genomes record as a reference. The length of the gene (the “offset” from the start position) and its location on the forward or reverse strands are also given. Two views for each chromosome are available—a full view, listing all genes in the specified chromosome together with additional links to specialized databases, and a gene-disease view, which displays only genes that are annotated in Swiss-Prot and TrEMBL as being linked to one or more diseases. In addition to the chromosome tables for the human proteome, a gene search facility allows the user to search using the gene name, chromosome location, keyword, and/or text.

3. Notes

The tools and resources mentioned in this chapter are intended, and already have begun, to relieve some of the burden of the influx of raw sequencing data. The integrated resources for protein family and domain signature databases, like InterPro, have several uses not only for the scientific community, where they exploit the strengths of the different methods and databases, but also for the member databases that contribute. Overlaps in integrating the different signatures means removal of redundancy in writing annotation, and an internal quality check of each signature when compared to the others. The integration has also led to better communication and collaboration between the member databases. The increasing availability of complete genome sequences is serving to identify uncharacterized protein families that may be unique to single or groups of related organisms. In this way the databases, and therefore InterPro, will move towards higher coverage of all protein sequences.

InterPro provides a one-stop shop for protein sequence classification, so that the user does not have to go to each database separately and rationalize the different results in varying formats. Rationalizing InterProScan results may seem daunting if multiple hits are detected, but it need not be. If multiple hits are recorded within overlapping positions on the sequence, it is important to look further for relationships between the entries that are hit. It may be that they are related through parent/child or contains/ found in relationships. The former case facilitates characterization of your protein on different levels of family and superfamily, whereas the latter provides domain composition information. It is then possible to find other proteins that are similar on the superfamily level or more specifically at the family or subfamily levels, or alternatively those that share a domain or domain composition in common with your query protein. If you have hits to multiple InterPro entries that are not related to each other and are not overlapping on the sequence, this may be indicative of multifunctional proteins.

There are a number of applications of InterPro, including automatic annotation and genome annotation. It has been used for annotation of numerous completely sequenced genomes, notably that of human. A useful application, as we have seen here, is in the analysis and comparison of complete proteomes as provided in the Proteome Analysis database. The Proteome Analysis database provides a useful source of whole genome analysis and comparisons. It facilitates searching, downloading, or simply viewing of precomputed and user-chosen data. It is hoped that a combination of tools such as these will help biologists shed some light on the biological function of newly discovered proteins. Only after this is done is it possible to use the data to their full potential in medical or commercial applications.

References

1. Mulder, N. J., Apweiler, R., Attwood, T. K., et al. (2002) InterPro—an integrated documentation resource for protein families, domains and functional sites. *Brief. Bioinform.* **3**(3), 225–235.
2. Pruess, M., Fleischmann, W., Kanapin, A., et al. (2003) The Proteome Analysis database: a tool for the *in silico* analysis of whole proteomes. *Nucl. Acids Res.* **31**, 414–417.
3. Sigrist, C. J. A., Cerutti, L., Hulo, N., et al. (2002) PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief. Bioinform.* **3**, 265–274.
4. Gribskov, M., Luthy, R., and Eisenberg, D. (1990) Profile analysis. *Methods Enzymol.* **183**, 146–159.

5. Krogh, A., Brown, M., Mian, I. S., Sjolander, K., and Haussler, D. (1994) Hidden Markov models in computational biology. Applications to protein modeling. *J. Mol. Biol.* **235**(5), 1501–1531.
6. Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998) *Biological Sequence Analysis: Probabilistic models of proteins and nucleic acids*. Cambridge University Press, Cambridge, UK.
7. Eddy, S. HMMER2 Profile hidden Markov models for biological sequence analysis. [<http://hmmer.wustl.edu/>]
8. Falquet, L., Pagni, M., Bucher, P., et al. (2002) The PROSITE database, its status in 2002. *Nucl. Acids Res.* **30**, 235–238.
9. Attwood, T. K., Bradley, P., Flower, D. R., et al. (2003) PRINTS and its automatic supplement pre-PRINTS. *Nucl. Acids Res.* **31**(1), 400–402.
10. Bateman, A., Birney, E., Cerruti, L., et al. (2002) The Pfam protein families database. *Nucl. Acids Res.* **30**(1), 276–280.
11. Letunic, I., Goodstadt, L., Dickens, N. J., et al. (2002) Recent improvements to the SMART domain-based sequence annotation resource. *Nucl. Acids Res.* **30**(1), 242–244.
12. Haft, D. H., Selengut, J. D., and White, O. (2003) The TIGRFAMs database of protein families. *Nucl. Acids Res.* **31**, 371–373.
13. Barker, W. C., Pfeiffer, F., and George, D. G. (1996) Superfamily classification in PIR—International Protein Sequence Database. *Methods Enzymol.* **266**, 59–71.
14. Gough, J. and Chothia, C. (2002) SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucl. Acids Res.* **30**(1), 268–272.
15. Corpet, F., Servant, F., Gouzy, J., and Kahn, D. (2000) ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucl. Acids Res.* **28**, 267–269.
16. Kriventseva, E. V., Servant, F., and Apweiler, R. (2003) Improvements to CluSTR: the database of SWISS-PROT+TrEMBL protein clusters. *Nucl. Acids Res.* **31**, 388–389.
17. Boeckmann, B., Bairoch, A., Apweiler, R., et al. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucl. Acids Res.* **31**, 365–370.
18. Lo Conte, L., Brenner, S. E., Hubbard, T. J., Chothia, C., and Murzin, A. G. (2002) SCOP database in 2002: refinements accommodate structural genomics. *Nucl. Acids Res.* **30**, 264–267.
19. The Gene Ontology Consortium (2001) Creating the gene ontology resource: design and implementation. *Genome Res.* **11**, 1425–1433.
20. Pearl, F. M., Lee, D., Bray, J. E., Buchan, D. W., Shepherd, A. J., and Orengo, C. A. (2002) The CATH extended protein-family database: providing structural annotations for genome sequences. *Protein Sci.* **11**, 233–244.
21. Zdobnov, E. M. and Apweiler, R. (2001) InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*. **17**(9), 847–848.
22. Sander, C. and Schneider, R. (1991) Database of homology derived protein structures and the structural meaning of sequence alignment. *Proteins*. **9**, 56–68.
23. Westbrook, J., Feng, Z., Jain, S., et al. (2002) The Protein Data Bank: Unifying the Archive. *Nucl. Acids Res.* **30**, 245–248.
24. Stoessner, G., Baker, W., van den Broek, A., et al. (2003) The EMBL Nucleotide Sequence Database: major new developments. *Nucl. Acids Res.* **31**, 17–22.

Computational Prediction of Protein–Protein Interactions

Anton J. Enright, Lucy Skrabanek, and Gary D. Bader

1. Introduction

One of the current goals of proteomics is to map the protein interaction networks of a large number of model organisms (1). Protein–protein interaction information allows the function of a protein to be defined by its position in a complex web of interacting proteins. Access to such information will greatly aid biological research and potentially make the discovery of novel drug targets much easier. Previously, the detection of protein–protein interactions was limited to labor-intensive experimental techniques such as co-immunoprecipitation or affinity chromatography. High-throughput experimental techniques such as yeast two-hybrid and mass spectrometry have now also become available for large-scale detection of protein interactions. These methods, however, may not be generally applicable to all proteins in all organisms, and may also be prone to systematic error. Recently, a number of complementary computational approaches have been developed for the large-scale prediction of protein–protein interactions based on protein sequence, structure, and evolutionary relationships in complete genomes.

Initially, computational prediction of protein–protein interactions was strictly limited to proteins whose three-dimensional (3-D) structures had been determined. These methods predicted protein–protein interaction based on the structural context of proteins. Recent advances in complete genome sequencing have, however, provided a wealth of genomic information. It is now possible to establish the genomic context of a given gene in a complete genome (2,3). A gene is no longer thought of as a single protein-coding entity, but as part of a coordinated network of interacting proteins. The potential for two proteins to interact is not only specified by the physical and structural properties of their structures, but is also encoded at a genomic level. For example, interacting genes are generally co-expressed (4–6) (both temporally and spatially). In other words, the fact that two proteins have the physical potential to interact is meaningless unless they are present in the same part of the cell at the same time. Other examples of genomic context include the co-localization of genes on chromosomes, the complete fusion of pairs of genes, correlated mutations between interacting protein families, and phylogenetic gene profiles. Even in the absence of structural or sequence information, one can detect the evolutionary fingerprints of pairs of interacting proteins from their genomic context. A number of these computational approaches also take advantage of high-throughput experimental information such as gene-expression data, cellular locality, and molecular complex information (7,8). These hybrid

From: *The Proteomics Protocols Handbook*
Edited by: J. M. Walker © Humana Press Inc., Totowa, NJ

computational approaches exploit both the genomic and biological context of genes and proteins in complete genomes in order to predict interactions.

In this chapter, we will describe computational methods and resources available for protein–protein interaction prediction that exploit the structural, genomic, and biological contexts of proteins in complete genomes. In addition to algorithms and methods for interaction prediction, a number of useful databases pertaining to protein–protein interaction will be described. These databases combine a large amount of data from both computational and experimental techniques. Finally, a number of tools for protein interaction network visualization and analysis will be described. Methods are presented in historical order, together with online access information. Where available, detailed computational protocols will also be provided for each method.

1.1. Structural Context Approaches

Computational prediction of protein–protein interactions consists of two main areas: (1) the mapping of protein–protein interactions, i.e., determining whether two proteins are likely to interact, and (2) the understanding of the mechanism of protein–protein interactions and the identification of residues in proteins that are involved in those interactions. The first successful computational analyses of protein–protein interactions used the structural context of proteins for the analysis of known protein interaction interfaces in order to determine physical rules determining protein–protein interaction specificity. Unlike other computational methods that use an evolutionary or genomic context to predict interaction, structural approaches tend to be more limited in terms of scale, as only a small proportion of protein sequences have accurate 3-D structures deposited in the Protein Databank (PDB) (9). However, structural approaches allow for a much more detailed analysis of protein interactions than the genome context-based approaches. Structural approaches can determine not only whether two proteins interact, but also the physical characteristics of the interaction, and residues (sites) at the protein interface that mediate the interaction.

The identification of protein interaction sites is important for functional genomics, analysis of metabolic and signal transduction networks, and also rational drug design. The first attempt to describe the characteristics of protein interaction sites was undertaken in 1975 by Chothia and Janin. With data from only three complexes, they suggested that the residues that form the interface are closely packed, tend to be hydrophobic, and that complementarity may be an important factor in predicting which proteins can interact (10).

Later studies with larger datasets extended and developed their work to try to identify other characteristics of the interaction site that are sufficiently different from the rest of the protein to be identifiable, and thus predictive. Further analysis of the hydrophobicity distribution of amino acids can be used to predict interaction sites, since interacting regions tend to be the most hydrophobic clusters on the surface of the protein (11–14). This type of analysis yields a 60% success rate at predicting interacting sites. In general, hydrophobic residues such as Leu, Ile, Val, Phe, Tyr, and Met are over-represented at interaction sites, whereas polar residues such as Lys, Asp, and Glu (but not Arg) are under-represented (15,16). Other parameters that have been analyzed for their importance in identifying those residues in a protein that form the interaction site include the accessible surface area and residue composition (17,18). It has also

become apparent that a distinction must be made between different types of complexes. Interaction sites on stable and transient complexes have different properties (18). A recent study (19) indicates that the residue composition can be used to identify six different types of protein–protein interfaces, from domain-domain interfaces in the same protein to inter-protein contact surfaces.

Further studies (18,20) using a six-parameter analysis (solvation potential, residue interface potential, hydrophobicity, planarity, protrusion, and accessible surface area) have indicated that none of these parameters individually could be definitively used as a prediction method. Using a combined score from all six parameters yielded accurate predictions for 66% of 59 structures (21). All interfaces tend to be planar and surface accessible; the other parameters differed between complex types. Computational resources for the prediction and analysis of protein–protein interactions in this way are described in **Subheading 2.1.** and **Table 1.**

Shape complementarity is primarily used in docking studies, which focus on finding the best fit of the two interacting proteins using rigid- and soft-body searches (22–24). Electrostatic complementarity between interfaces (**Fig. 1**) plays an important role in determining the best fit of two interacting proteins (23). Interfaces between antibody–antigen complexes and transient heterodimers tend to have the least shape complementarity, while homodimers, enzyme-inhibitor complexes, and permanent heterodimers are the most complementary (18).

However, other research (25) has indicated that the chemico-physical properties of interacting surfaces are difficult to distinguish from those of the whole protein surface. Recently, it has been suggested that instead of using patch analysis, it may be better to use interface contacts (19), i.e., residues whose closest atoms are annotated in PDB as being less than 6 Å apart. They argue that the analysis of surface patches may miss slightly buried residues with long side-chains, while other residues identified as being part of a patch may in fact not be important, or may not form contacts at all.

Other methods of predicting protein interaction sites include multiple sequence alignment and analysis of amino acid characteristics of neighboring residues using neural networks. Multiple sequence alignments can help identify specific family structures that are conserved within a subfamily but differ between subfamilies. These regions are interpreted as being interaction sites that may endow specificity of ligand interaction (26–28). Two groups (29,30) have trained neural networks with sequence profiles of spatial neighbors of a target residue with solvent exposure to predict whether a residue will be part of an interaction site. Both of their methods gave approximately a 70% accurate prediction rate. The validity of using sequence profiles has been verified by results that demonstrate that the majority of interacting residues are clustered in sequence segments of several contacting residues (31).

Recently, methods have been developed to validate predicted protein–protein interactions against experimentally determined 3-D structures (32,33). Given a known 3-D structure, they map homologs of the interacting proteins onto the structure and, using empirical potentials, test whether the homologous proteins preserve the interactions from the known structure. However, the number of experimentally determined structures for complexes is small, and of the 2590 interactions predicted by large-scale methods, only 59 could be mapped onto their set of interacting complexes. Of these, 59% had domains that appeared to be in direct contact, thus increasing the probability that

Table 1
Methods and Databases for Computational Prediction of Protein-Protein Interactions

Resource	Type of resource	WWW Address (URL)	Ref.
Structural Context Interaction Prediction			
Protein–Protein Interaction Server	Structure based interaction prediction	http://www.biochem.ucl.ac.uk/bsm/PP/server/	(18)
InterPreTS	Structure based interaction prediction	http://www.russell.embl.de/interprets/	(33)
Genomic Context Interaction Prediction			
AlIFUSE	Gene fusions	http://www.ebi.ac.uk/research/cgg/allfuse/	(51)
STRING	Gene Co-Localization, gene-fusion, phylogenetic profiles	http://www.bork.embl-heidelberg.de/STRING/	(55)
WIT	Orthology/phylogenetic profiles/gene co-localization	http://wit.mcs.anl.gov/WIT2/	(56)
Predictome	Gene Co-Localization, gene-fusion, phylogenetic profiles	http://predictome.bu.edu/	(58)
COGs	Orthology/phylogenetic profiles	http://www.ncbi.nlm.nih.gov/COG/	(59)
Biological Context Interaction Prediction			
GeneCensus	Combined predictions (bayesian network)	http://genecensus.org/initint/	(8)
Pathway Databases	EcoCyc Metabolic pathway analysis	http://ecocyc.cpaneasystems.com/ecocyc/	(34)
KEGG	Metabolic / regulatory pathway analysis and reconstruction	http://www.genome.ad.jp/kegg/	(72)
SigPath	Signalling pathways	http://www.sigpath.org/	(73)
MIPS	Pathways, complexes, cellular locations	http://www.mips.biochem.mpg.de/proj/yeast/pathways/	(79)
Protein Interaction Databases			
BIND	Interactions, complexes, pathways	http://www.bind.ca/	(76)
DIP	Database of protein interactions	http://dip.doe-mbl.ucla.edu/	(77)
INTACT	Database of protein interactions	http://www.ebi.ac.uk/intact/index.html	(80)
MINT	Database of protein interactions	http://160.80.34.4/mint/	(81)
Gene-Expression Databases			
SMD	Gene expression data	http://genome-www5.stanford.edu/	(67)
Array Express	Gene expression data	http://www.ebi.ac.uk/arrayexpress/	(68)
GEO	Gene expression data	http://www.ncbi.nlm.nih.gov/geo/	(69)
Visualization Tools for Protein Interactions			
BioLayout	Interaction Network Visualization	http://www.biologayout.org/	(74)
Cytoscape	Interaction Network Visualization	http://www.cytoscape.org/	(75)

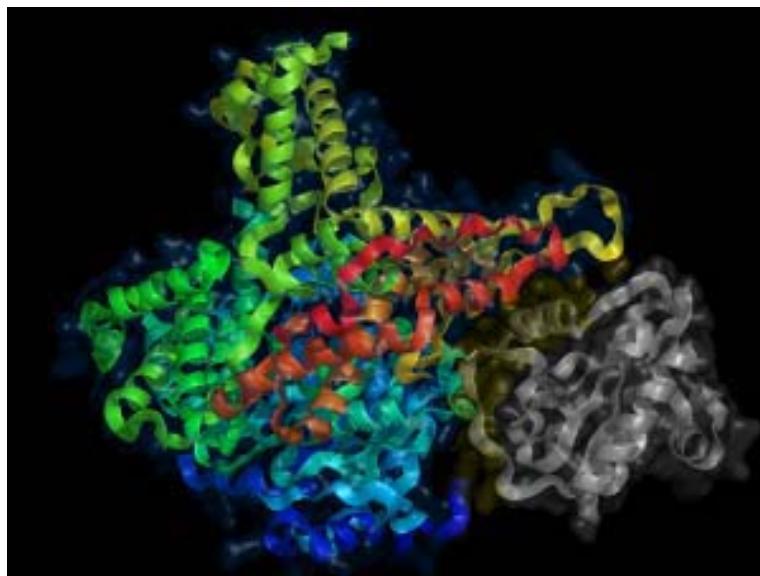


Fig. 1. Three-dimensional structure of the T7 bacteriophage RNA polymerase complexed with T7 lysozyme. The structure on the left is RNA polymerase. The lysozyme is shown on the right. This figure was produced using protein databank structure 1ARO and PyMol (see Note 8).

these predicted protein–protein interactions are biologically correct. Computational methods (34) for the prediction of protein–protein interactions based on this (and other structural approaches) are described in **Subheading 2.1**.

1.2. Genomic Context Approaches

1.2.1. Co-Localization

One of the first methods for predicting protein–protein interactions from the genomic context of genes utilizes the idea of co-localization, or gene neighborhood (Fig. 2A). Such methods exploit the notion that genes that physically interact (or are functionally associated) will be kept in close physical proximity to each other on the genome (35–37). The most apparent case of this phenomenon involves bacterial and archaeal operons, where genes that work together are generally transcribed on the same polycistronic mRNA. In these cases, proteins involved in the same process or pathway are frequently encoded on the same polycistronic messenger.

Operons are rare in eukaryotic species (38,39). However, genes involved in the same biological process or pathway are frequently situated in close genomic proximity (36). It is hence possible to predict functional or physical interaction between genes that are repeatedly observed in close proximity (e.g., within 500 bp) across many genomes. This method has been successfully used to identify new members of metabolic pathways (36). Like many of the genome-context approaches, this method becomes more

powerful with larger numbers of genomes. This approach and a number of online resources that implement it are described in **Subheading 2.2**.

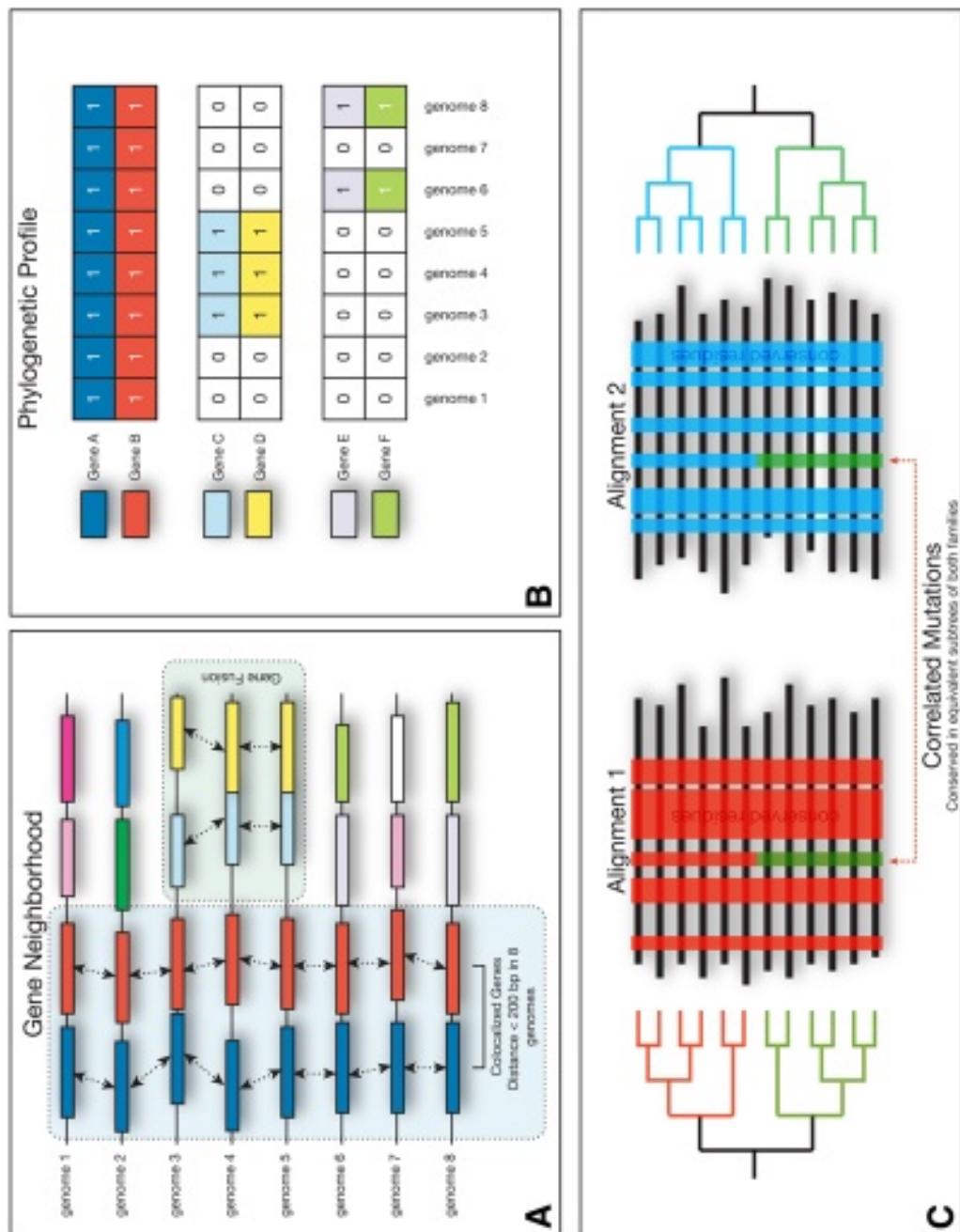
1.2.2. Phylogenetic Profiles

A relatively simple, yet powerful, form of genomic context is the co-occurrence of pairs of genes across multiple genomes. Two of the main driving forces in genome evolution are gene genesis and gene loss (40,41). The fact that a pair of genes remains together across many disparate species represents a concerted evolutionary effort that suggests that these genes are functionally associated (i.e., same biological process or pathway) or physically interacting. This criterion is less stringent than that of gene co-localization, where gene pairs must not only be present, but also situated close to each other on the genome. Homologous genes can be termed either *orthologs* or *paralogs*. In general, the term *ortholog* is used to describe genes that are related by a speciation event, i.e., perform analogous functions in different organisms and are related to a single common ancestor gene in an ancestor species. The term *paralog* is used to describe homologous genes that have arisen following a gene duplication event, i.e., perform similar functions in the same organism. Classifying homologous genes as either paralogs or orthologs is difficult in the absence of accurate phylogenetic or speciation information (42). Classification of genes in this way allows the inference of a phylogenetic context for a given gene.

The analysis of phylogenetic context in this fashion has been termed *phylogenetic profiling* (43). These profiles can be as simple as a binary representation of the presence or absence of a gene across multiple genomes (43–45) (Fig. 2B). A library of these profiles may then be scanned to find genes that exhibit identical (or highly similar) phylogenetic patterns to each other. Pairs of genes detected in this fashion are hence candidates for physical interaction or functional association. This method has been used not only to infer physical interaction (43), but also to predict the cellular localization of gene products (46).

This system is not, however, without flaw. Firstly, the strength of any inference made using such profiles is heavily dependent on the number and distribution of genomes used to build the profile. A pair of genes with similar profiles across many bacterial, archaeal, and eukaryotic genomes are much more likely to interact than genes found to co-occur in a small number of closely related species. Secondly, evolutionary processes such as lineage-specific gene loss, horizontal gene transfer, nonorthologous gene-displacement (47), and the extensive expansion of many eukaryotic gene families can make orthology assignment across genomes very difficult. However, given the increasing number of completely sequenced genomes, the accuracy of these predictions is expected to improve over time. The details of this approach and online-

Fig. 2. (opposite page) Overview of genome context approaches. (A) Gene neighborhood plots for eight complete genomes, showing a pair of genes that are in close physical proximity in all eight genomes. A gene fusion event between two genes in two genomes is also shown. (B) Example of phylogenetic profiles of selected genes from the previous panel. These three pairs of genes have the same patterns of co-occurrence in all eight genomes, and may physically interact based on this evidence. (C) Two protein family alignments are shown with conserved regions highlighted. Correlated mutations are present in two identical subtrees for each family, which indicates that these sites may be involved in mediating interactions between proteins from each family.



resources for phylogenetic profile-based prediction of protein interaction are described in **Subheading 2.3**.

1.2.3. Gene Fusion

Genome context approaches to the prediction of protein–protein interaction also include the analysis of gene fusion across complete genomes. This method is complementary to both co-localization of genes and phylogenetic profiles, and uses both gene location and phylogenetic analysis to infer function or interaction. A gene fusion event represents the physical fusion of two separate parent genes into a single multifunctional gene. This is the ultimate form of gene co-localization, i.e., interacting genes are not just kept in close proximity on the genome, but are physically joined into a single entity (**Fig. 2A**). It has been suggested that the driving force behind these events is to lower the regulational load of multiple interacting gene products (48). Gene fusion events hence provide an elegant way to computationally detect functional and physical interactions between proteins (48,49).

Gene fusion events are detected by cross-species sequence comparison. Fused (composite) proteins in a given reference genome are detected by searching for unfused component protein sequences that are homologous to the reference protein, but not to each other. These un-fused query sequences align to different regions of the reference protein, indicating that it is a composite protein resulting from a gene fusion event (48). Once again, predictions of this type are complicated by a number of issues. The largest hindrance is the presence of so-called promiscuous domains. These domains (such as helix-turn-helix [HTH] and DnaJ) are highly abundant in eukaryotic organisms. The domain complexity of eukaryotic proteins coupled with the presence of promiscuous domains and large degrees of paralogy can hamper the accurate detection of gene fusion events (50).

Although the method is not generally applicable to all genes—i.e., it requires that an observable fusion event can be detected between gene pairs—it has been successfully applied to a large number of genomes (including eukaryotes) (51). The basic gene fusion detection method and online resources such as the AllFUSE database (51), will be described in **Subheading 2.4**.

1.2.4. In Silico Two-Hybrid

The *in silico* two-hybrid (i2h) approach has much in common with the other genome-context approaches, but also indirectly assesses structural properties of proteins that potentially interact. It has previously been shown that a mutation in the sequence of one protein in a pair of interacting proteins is frequently mirrored by a compensatory mutation in its interacting partner (**Fig. 2C**). The detection of such correlated mutations can not only be used to predict protein–protein interactions, but also has the potential to identify specific residues involved at the interaction sites (52).

Previous analyses (53) involved searching for correlation of residue mutations between sequences in the same protein family alignment (intra-family). The *in silico* two-hybrid method extends this approach by searching for such mutations across different protein families (interfamily). Prediction of protein–protein interactions using this approach is achieved by taking pairs of protein family alignments and concatenating these alignments into a single cross-family alignment. A position-specific matrix is then built from this alignment, and a correlation function is then applied to detect residues that are correlated both within and across families. Correlated sites that poten-

tially indicate protein interaction are returned with a score. The method suffers as a result of the computational complexity of constructing the large numbers of alignments needed, and poor quality alignments can dramatically increase noise in the procedure (52). However, the method is similar to the gene fusion approach, as a single accurate prediction between two proteins can infer interaction between all members of both families used. This approach is not discussed in the Methods section, as currently the method is not freely available (*see Note 1*).

1.3. Biological Context Approaches

High-throughput experimental techniques now provide access to a more detailed view of biological processes at a genomic level. Gene expression analysis allows one to not only determine which genes are active in a given state, but also sets of genes that are co-regulated in many different states. It has been shown that many interacting proteins are co-expressed according to microarray analyses (4–6). Current gene-expression methods now allow for every coding gene of a genome to be placed on a single microarray, allowing the activity of every gene to be monitored across different states or time-points. Although these methods cannot directly be used to determine whether or not two proteins interact, a number of computational approaches have been developed that use this information towards the prediction of protein–protein interaction and gene regulatory networks (4–6). Other high-throughput experimental techniques such as yeast two-hybrid specifically test a bait protein for interactions against a set of prey proteins. The bait and prey consist of fusion constructs that activate a reporter gene if they interact with each other. While this method is not as accurate as other techniques such as co-immunoprecipitation, affinity chromatography, or gel-overlay assays, it can be applied rapidly to genome-scale studies of protein–protein interactions.

Many of these high-throughput methods for investigating the biological context of genes and proteins are inherently noisy. For example, some proteins in yeast two-hybrid assays appear to detect a large number of spurious interactions (false-positives). Gene expression techniques suffer from a number of problems also, such as cross-hybridization and poor signal-to-noise ratios. Recently, however, research has shown that multiple datasets pertaining to the biological context of genes and proteins can be combined using machine learning techniques (8). Using Bayesian network analysis, it is hence possible to computationally combine multiple noisy datasets in such a way that protein–protein interactions can be more reliably predicted. In this method, each source of interaction evidence is compared against samples of known positive (proteins in the same complex) and negative (proteins in different cellular locations) interactions, allowing a statistical reliability index to be built for each data source. When this information is applied genome-wide, a prediction can be made for every protein pair in a genome by combining different sets of independent evidence according to their calculated reliability. Protein interactions predicted in this way have been shown to be as reliable as pure experimental techniques, while simultaneously covering a larger proportion of genes than most experimental methods (8).

A number of available resources for protein–protein interaction data, gene expression data, and Bayesian network analysis of multiple interaction datasets are described in **Subheading 2.5**.

1.4. Data Sources and Visualization Techniques

Computational biology is a data-rich research field. The advent of complete genome sequencing and high-throughput experimental techniques has created an enormous amount of data. In order for these data to be both informative and useful, they must be stored in a sensible and accessible way, and tools must be made available to visualize and exchange this information. A number of initiatives are tackling these problems by creating freely accessible databases storing a wide variety of biological information, including protein–protein interactions. Recently, a number of research groups have created visualization tools for biological networks. These tools provide a new way to analyze protein–protein interaction networks, provide a multitude of different ways to represent interactions, and can overlay other biological information onto these networks. A number of databases that store protein–protein interactions, molecular complexes, and pathways will be described later (see **Subheading 2.6.** and **Table 1**). Finally, we will detail methods for the visualization and analysis of protein–protein interaction networks. (see **Subheading 2.7.**).

2. Methods

In this section, we will describe computational resources and methods for the prediction of protein–protein interactions. These methods will be detailed in chronological order. Within each section, a number of online computational resources are described that allow one to perform this type of analysis interactively. Where possible (mostly for genomic context-based approaches), detailed computational protocols will also be provided. Resources mentioned in this section are further summarized in **Table 1**.

2.1. Structure-Based Prediction of Interactions

The Protein–Protein Interaction Server at University College London (UCL) provides a simple, Web-based interface for exploring protein–protein interaction interfaces, given 3-D structures (18). This server takes into account the following information for interaction analysis: accessible surface area, planarity, length and breadth, secondary structure, hydrogen bonds, salt bridges, gap volume, gap volume index, bridging water molecules, and interface residues. This resource (**Table 1**) is very useful for exploring the protein–protein interaction potential of two protein structures identified through docking or shape complementarity.

The structural bioinformatics group at EMBL Heidelberg provides the InterPreTS server for protein–protein interaction prediction (33). Using this resource (**Table 1**), one can submit pairs of sequences that are then compared to the 3-D structures of known protein–protein interactions. This resource utilizes a prebuilt Database of Interacting Domains (DBID) and an empirical scoring system to test whether a sequence pair fits a known 3-D structure of an interacting pair of proteins.

2.2. Gene-Neighborhood-Based Interaction Prediction

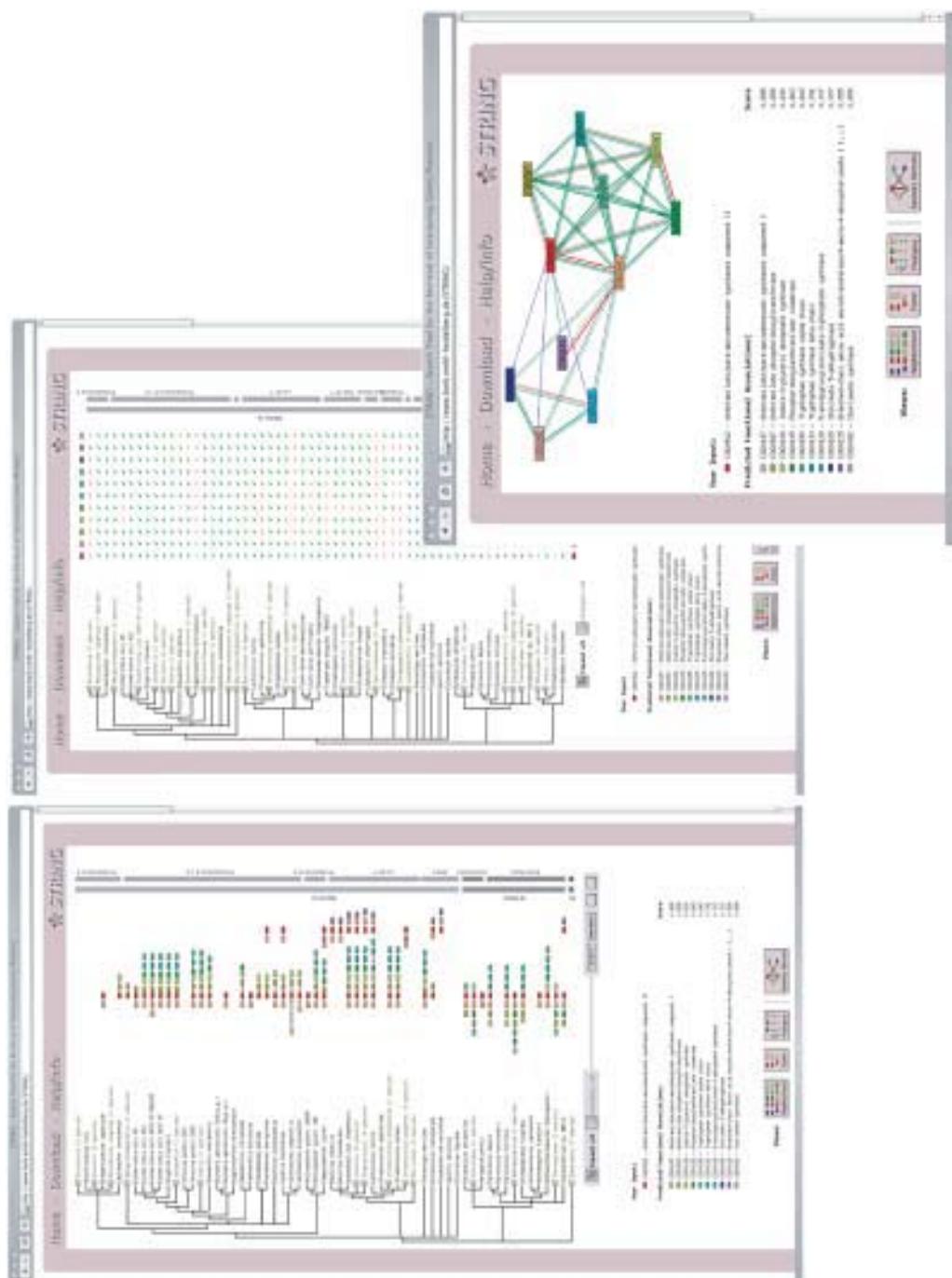
Co-localization of genes across multiple genomes provides a fingerprint suggesting that they may physically interact (36). Analysis of conserved gene locations across multiple genomes (**Fig. 2A**) can hence be used to predict protein interaction networks and metabolic pathways (54). A number of excellent resources exist that allow one to

determine whether two proteins may interact using this approach. The most notable of these are Search Tool for Recurring Instances of the Neighborhood of Genes (STRING) (55) and What is There? (WIT) (56). The STRING database (Fig. 3) provides a Web interface giving comprehensive access to gene neighborhood information (57) for 356,775 genes in 110 complete genomes. Similarly, the Predictome database at Boston University (58) provides a comprehensive Web interface to predictions of this type. The WIT database provides access to protein family information, metabolic pathway reconstruction, and gene co-localization information. Using these resources allows detailed precomputed gene neighborhood information to be analyzed for evidence of protein–protein interaction (Table 1). The actual protocols used for these analyses can vary considerably; a general protocol adapted from WIT (56) is described below:

1. In order to assess whether pairs of orthologous genes share a common gene neighborhood across multiple genomes, one needs (a) protein sequences/genomic locations and (b) orthology mappings between proteins from multiple genomes.
2. Orthology mappings are generated by searching for pairs of close bi-directional best hits (PCBBH). These are a specific form of bi-directional best hit (see Subheading 2.3.), a commonly used method for orthology assignment. For a given pair of proteins α and β in genome X, a bi-directional best hit to genes α' and β' in genome Y is defined as follows:
 - a. The best Basic Local Alignment Search Tool (BLAST) hit for protein α in genome X is protein α' in genome Y.
 - b. The best BLAST hit for protein β in genome X is protein β' in genome Y.
 - c. The genes of proteins α and β are situated within 300 bp in genome X.
 - d. The genes of proteins α' and β' are situated within 300 bp in genome Y.
3. Genes that satisfy the above criteria can be considered as having a conserved gene neighborhood across two genomes. When this procedure is repeated across multiple genomes, it becomes possible to identify genes that are significantly co-localized across many genomes, and are hence likely to either physically interact or be functionally associated.
4. The PCBBH criteria are quite strict, and it is also possible to perform the procedure using pairs of close homologs (PCHs) (see Note 2).
5. Sets of PCBBHs or PCHs in multiple genomes are typically scored for significance based on the number and phylogenetic distribution of genomes in which they are co-localized. Phylogenetic distance can be estimated by examining a 16S rRNA phylogenetic tree.
6. A common score (coupling score) for the likelihood that two genes interact based on summing individual scores from multiple genomes is then calculated.
7. Finally, candidate genes that have significant coupling scores are candidates for either physical interaction or functional association.

2.3. Phylogenetic Profile-Based Prediction of Interaction

Phylogenetic profile-based prediction of protein interactions (Fig. 2B) has been shown to be an accurate and widely applicable method. Perhaps the easiest way to utilize this information for prediction of protein interaction is to use precomputed phylogenetic profiles for proteins of interest. The Clusters of Orthologous Groups (COGs) resource at the National Center for Biotechnology Information (NCBI) contains large numbers of profiles for a variety of bacterial and archaeal organisms, and also *S. cerevisiae* (59,60). Other excellent resources for combined computational predictions of protein interactions using phylogenetic profiles are available from the STRING (55)



resource at EMBL Heidelberg and from Predictome (58) (Table 1). Using the Web interfaces to these resources, it is relatively straightforward to find groups of proteins with similar or identical phylogenetic profiles, indicating proteins that physically interact or are functionally associated (Fig. 3). For a more detailed analysis of specific proteins of interest, a general protocol is described below:

1. For each genome to be analyzed, a FASTA sequence file containing all protein sequences is assembled.
2. All protein sequences in each genome are compared against all other sequences using a sequence similarity search algorithm such as BLASTp (61). A variety of other sequence similarity search tools could also be used at this step (see Note 3).
3. Orthology between proteins in different genomes is assigned as follows:
 - Two proteins (from different genomes) are orthologous if they were each other's highest-scoring BLAST hit when searched against the other genome. This is frequently referred to as a bidirectional best hit (BBH).
 - This process is repeated to assign (if possible) an ortholog for each protein in a given genome, to a protein in all other genomes.
4. All orthology assignments made in this way are stored for post-processing.
5. A phylogenetic profile for a protein can then be constructed by representing the presence or absence of an ortholog for that protein across all genomes analyzed. Frequently, this is represented by a simple binary vector with “1” indicating presence and “0” representing absence of a gene in each genome (Fig. 2B) (see Note 4).
6. All profiles are compared to all other profiles using a clustering procedure. A distance measure (such as Pearson correlation or Euclidean distance) between each profile and all other profiles is used to group profiles according to how similar they are (see Note 5).
7. Finally, protein profiles that are highly similar or identical to each other represent candidate proteins that physically or functionally interact.

2.4. Gene Fusion Prediction of Protein Interactions

Gene fusion is a relatively common evolutionary phenomenon (51). A detected gene fusion between two genes indicates that their protein products may physically interact or be involved in the same biological process or pathway (48,49). One extreme example of this is the aromatic amino acid biosynthesis pathway in *Saccharomyces cerevisiae*. In yeast, a single fused gene encodes the entire pathway of these five normally separate genes (48). Prediction of protein interactions using gene fusion has been successful in a number of areas, including the prediction of novel protein interactions involved in important biological processes in *Drosophila melanogaster* (62).

A comprehensive set of fused genes and inferred protein–protein interactions is available from the AllFUSE database (51) at the European Bioinformatics Institute (EBI), the STRING database at EMBL Heidelberg (55) (Fig. 3), and the Predictome database at Boston University (58) (Table 1). Using the AllFUSE resource, one can search for potential interactions for a given protein sequence from a database of 24

Fig. 3. (opposite page) Screenshots from the STRING web resource. The left panel illustrates the STRING representation of gene neighborhood and gene fusion. The right panel shows a typical phylogenetic profile for multiple genes and genomes. Finally, the inset shows a predicted protein interaction map generated from gene neighborhood, gene fusion, and phylogenetic profile methods combined.

complete genomes. A general protocol for gene fusion-based prediction of protein–protein interactions can be described as follows (48):

1. This analysis requires two genomes, a query, and a reference. One searches for gene fusion (composite) proteins in the reference genome using protein sequences from the query genome. Sequences from both genomes need to be assembled into FASTA format for this analysis.
2. Each protein in the query genome is then interactively searched against each protein from the reference genome using a sequence similarity search tool such as BLASTp (61), using an expectation value (E value) threshold to eliminate similarities that may have arisen by chance.
3. All significant similarities detected in this way are then stored in a binary matrix, which for each protein pair stores “1” for significant similarity or “0” for no detectable similarity. The matrix may be symmetrified by postprocessing with a more sensitive sequence search tool, such as Smith–Waterman (63), to clear up ambiguities.
4. Finding evidence of a gene fusion event in the reference species extends the previous symmetrification problem to one of transitivity (48). In this case, one searches for instances where query proteins A and B match a reference protein C, but do not match each other—i.e., $A \leftrightarrow C$; $B \leftrightarrow C$ but $A \neq B$. These triangular inequalities are resolved once again by using the more accurate Smith–Waterman algorithm to double check that no detectable significant similarity exists between A and B. Further analysis using alignment geometry can then verify that proteins A and B are orthologous to different regions of a composite fusion protein but not to each other (64).
5. Candidate fusion proteins detected in this way provide evidence that proteins A and B may physically interact.

Although this method is not generally applicable to all genes, and suffers from the high levels of paralogy usually present in eukaryotic genomes (65), this approach has been shown to have an accuracy as high as 90% and readily detects well-known interacting proteins (e.g., tryptophan synthase α and β subunits) and many proteins previously shown to form complexes. As such, this method represents a useful way to build interaction networks for proteins of interest within and across genomes.

2.5. Prediction of Protein Interactions From High-Throughput Biological Datasets

Gene expression analysis allows for all genes from a given genome to be placed on a single microarray, allowing many gene-expression experiments to be carried out rapidly and in parallel. Recently, efforts have been made to standardize data formats for reporting the results of gene expression experiments. The Minimum Information About a Microarray Experiment (MIAME) (66) standard allows different laboratories to effectively and accurately exchange microarray expression information. Using such standards, it has become easier for a number of publicly accessible resources to distribute microarray data (**Table 1**).

The Stanford Microarray Database (SMD) (67) provides access to raw data from public microarray experiments, as well as a number of software tools for utilizing these data. Currently, 140 experiments are indexed in the SMD Web resource. The MicroArray group at the European Bioinformatics Institute provides ArrayExpress (68), publicly available gene expression data in MIAME format for over 66 publicly available experiments and also integrated tools for expression profile analysis. Finally, the Gene Expression Omnibus (GEO) (69) database at the NCBI contains data from

over 300 large-scale publicly available microarray and SAGE experiments, for which all data are linked into the NCBI protein, nucleotide, and genomic databases.

Using these resources, it is hence possible to select a number of datasets for an organism of interest and extract gene expression profiles for some or all genes. Proteins whose genes exhibit very similar patterns of expression across multiple states or experiments (70) may then be considered candidates for functional association and possibly direct physical interaction (4–6). Gene expression analysis becomes much more reliable with more expression data. For example, genes that have high correlation across 10 experiments are much more likely to be related functionally than genes correlating across 2 experiments. Gene expression data are relatively susceptible to noise, and great care must be taken to minimize and filter this from any analysis. These data can, however, be very powerful when combined with analyses involving regulatory network reconstruction, and with other methods of detection of functional association and interaction of proteins (8).

The Bayesian networks approach (see **Subheading 1.3.**), which combines data from multiple biological datasets, is a useful way to minimize this noise and perform reliable protein–protein interaction prediction in *S. cerevisiae* (8). Validation of the method indicates that it can successfully recover large numbers of previously known protein–protein interactions (Fig. 4) and many novel interaction predictions. The results of this analysis are available from the GeneCensus Web site at Yale University (Table 1). These predictions are remarkable as they illustrate that combining multiple independent and noisy datasets in an intelligent way does not necessarily increase noise in the combined protein interaction predictions (assuming orthogonal error between datasets). This is also an excellent example of a combined computational and experimental approach, as interactions predicted using this approach appear to be more reliable than many pure experimental approaches (8).

2.6. Tools for Protein–Protein Interaction Visualization

Network and pathway visualization tools are computer programs that can automatically generate a diagram of a network or pathway. Perhaps the simplest such representation of a protein–protein interaction network is a graph composed of nodes (proteins) connected by edges (interactions). Some of the first visualization tools were developed for browsing metabolic pathways. For example, a pathway-drawing tool is present in the ACeDB database (71) and in EcoCyc (34). In many cases these representations are clickable, so that one can select a member of a pathway or a small molecule and get further information about that entity. Many of these initial visualization tools are static, and generated semiautomatically. The Kyoto Encyclopedia of Genes and Genomes (KEGG) (72), BioCarta, and SigPath (73) Web sites (Table 1) are examples of this type of visualization. Other more advanced methods can dynamically generate pathway diagrams from raw information in a biological database, such as the EcoCyc and WIT databases (see Table 1).

Recently, a number of purely automatic and general algorithms have been developed for visualizing biological networks. These tools rely on a layout algorithm to organize a graph of nodes and edges into an aesthetically pleasing layout. In graphing terms, this usually means minimizing the number of edges that cross each other, and grouping groups of nodes that are highly connected to each other. Typically, a well-

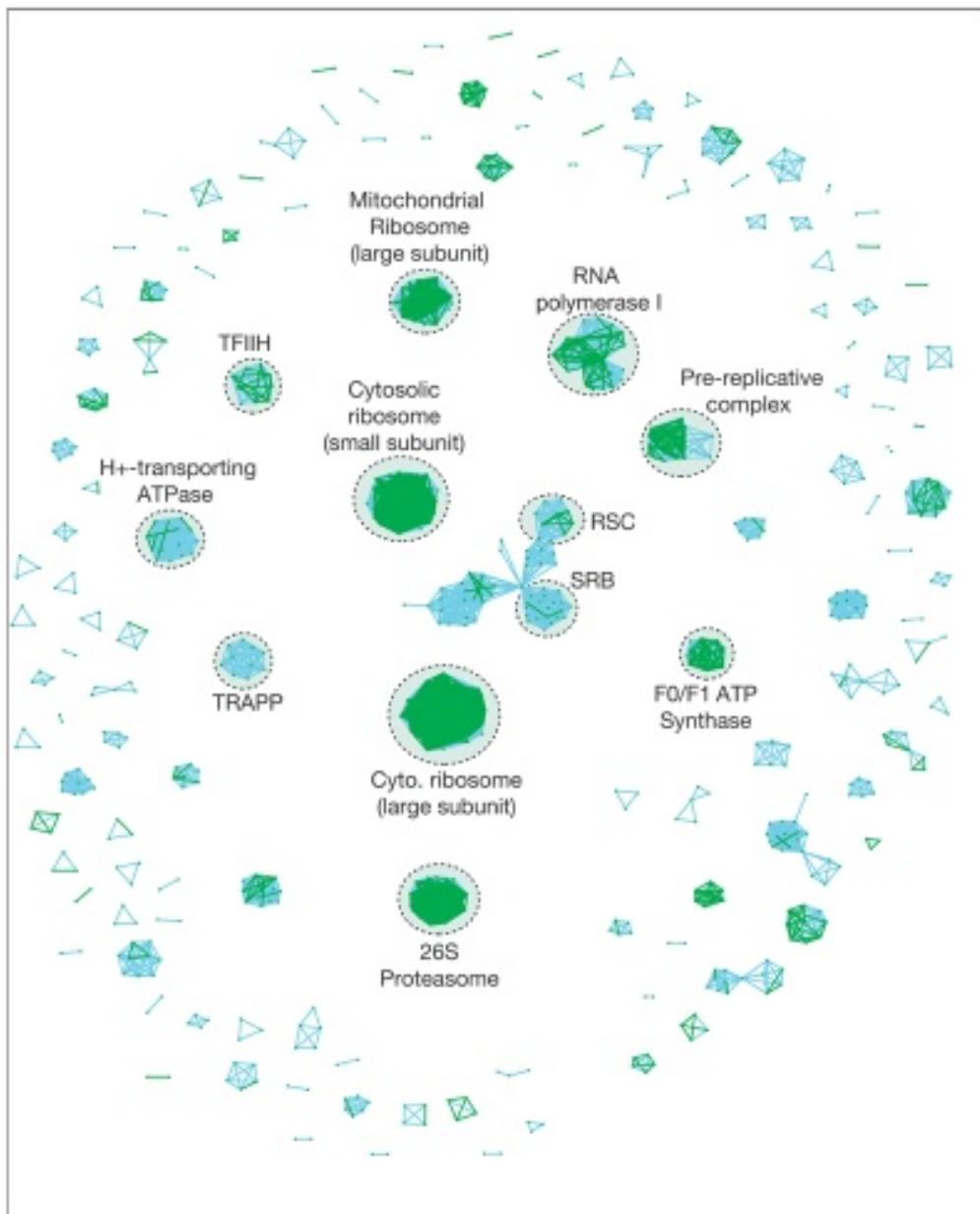


Fig. 4. Bayesian network predictions of protein–protein interactions. Experimentally validated gold-standard protein–protein interactions between *Saccharomyces cerevisiae* proteins are shown as an interaction network. Bayesian network analysis prediction of protein–protein interaction successfully recovers a significant subset of these interactions. The gold-standard interactions are derived from MIPS, and well-known complexes are annotated.

organized graph layout will allow the user to identify global features of their data that may not have been previously apparent. An example of a layout algorithm is the Spring Embedder algorithm. This method models the graph as a physical system where nodes are spheres connected by springs (edges). Nodes are initially organized in a random

state, and forces between connected spheres (as a result of springs) push the system into a lower-energy, more stable state. Other methods, such as the Weighted Fruchterman–Rheingold algorithm (64), represent the graph as a system of nodes that exert an attractive force (similar to a spring) between nodes connected by an edge, and a distance-dependent repulsive force between all nodes. Additionally, the weighted algorithm allows the attractive forces between nodes to be modulated using weights, and the energy of the entire system is controlled using a temperature function. Other layout algorithms can involve arranging nodes hierarchically, in a circular fashion, or in less structured formats. It is important to choose the best layout algorithm for the type of graph being visualized. For example, a highly connected interaction network will not assume a meaningful graph layout when a hierarchical layout algorithm is used.

Two of the most commonly used visualization tools for biological networks are BioLayout (74) and Cytoscape (75) (Table 1). Both of these tools are written using the JAVA (see Note 6) programming language and are hence portable across a wide variety of computer environments. Both tools also allow the interactive editing of graphs, through the movement of nodes, node labeling, and the ability to change the appearance of nodes and edges. Additionally, both tools can export publication-quality high-resolution graph images (see Note 7). BioLayout utilizes the weighted Fruchterman–Rheingold layout algorithm, and has a number of options for graph customization, data overlay, export, and graph analysis (Fig. 5). Cytoscape provides a number of different layout algorithms for producing useful visualizations and a number of plug-ins and import options for representing data such as gene expression (Fig. 6). Specifically, circular, hierarchical, organic, embedded, and random layouts are available. Circular and hierarchical algorithms try to lay out a network as their names suggest. Organic and embedded are two versions of a force-directed layout algorithm. Types of plug-ins that are currently available for Cytoscape include one that allows reading Proteomics Standards Initiative (PSI) files (see Subheading 2.7.) and one called ActiveModules that finds regions of a molecular interaction network that are correlated across multiple gene expression experiments. Both of these methods are suitable for small to medium-sized networks (less than 1000 nodes), although it may not be long before both layout and visualization techniques become available for the analysis of much larger graphs.

2.7. Data Resources for Protein–Protein Interactions

Current computational and experimental methods for protein–protein interaction prediction have been generating large amounts of data. It is imperative that these data be stored in a consistent and reliable way so that it may be useful for biological research. A number of databases are now publicly available for making this information accessible. Two of the largest and most comprehensive interaction databases now available are the Biomolecular Interaction Network Database (BIND) (76) and the Database of Interacting Proteins (DIP) (77). DIP is based at UCLA and currently contains over 18,000 experimentally determined protein–protein interactions (mostly from high-throughput *S. cerevisiae* experiments) for over 7,000 proteins in 104 organisms. Interactions in DIP are curated both manually (by expert curators) and automatically (text-mining approaches). BIND, at the University of Toronto, not only stores and curates pairwise protein–protein interactions, but also molecular complex information and biological pathways. Currently, BIND contains over 21,349 protein–protein and

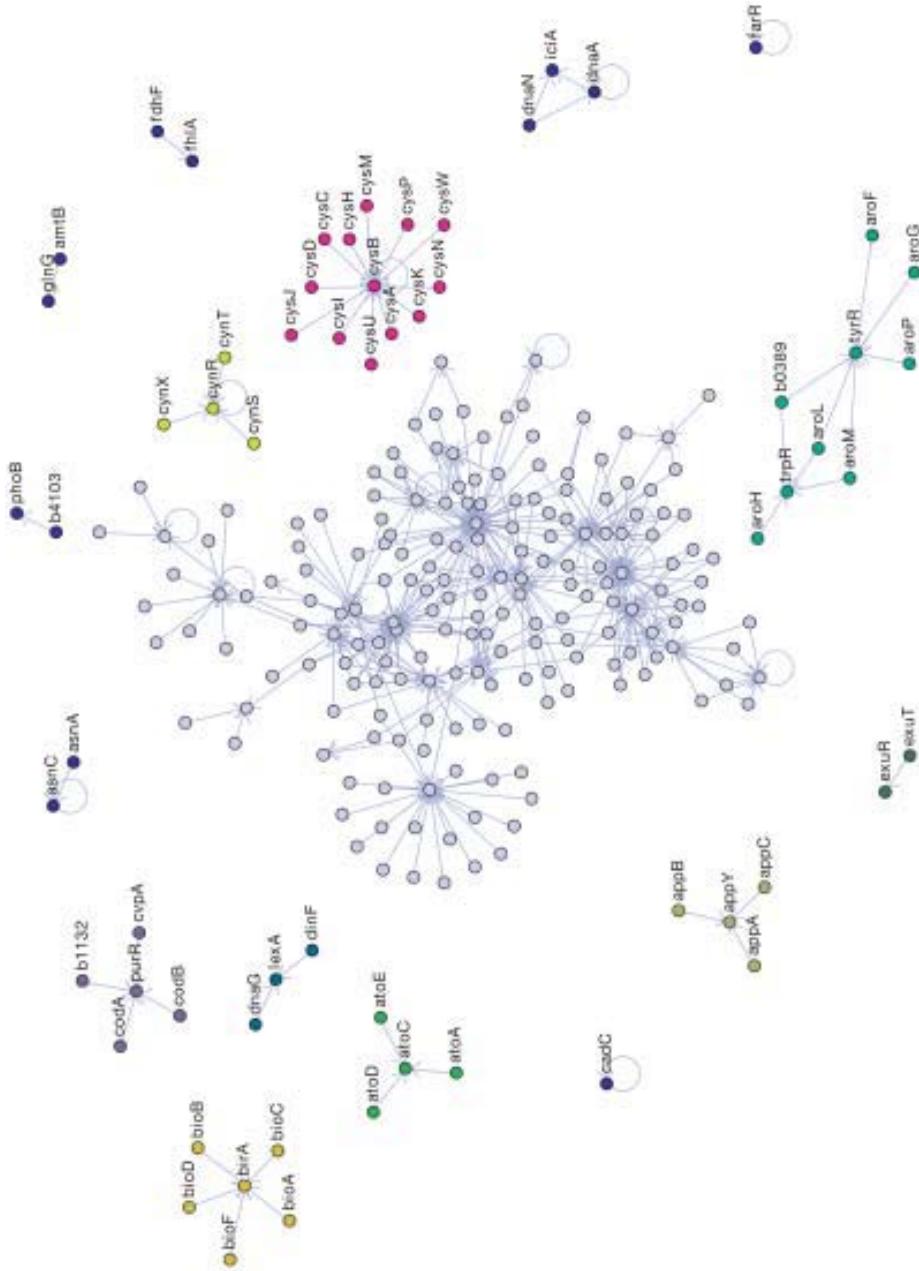


Fig. 5. Sample graph from BioLayout. This graph illustrates a genetic regulatory network of *Escherichia coli* genes. Genes are represented by circles (nodes) connected by regulatory interactions represented by lines (edges). Nodes are colored according to biochemical pathway assignments. Nodes in the center of the graph are not labeled for clarity.

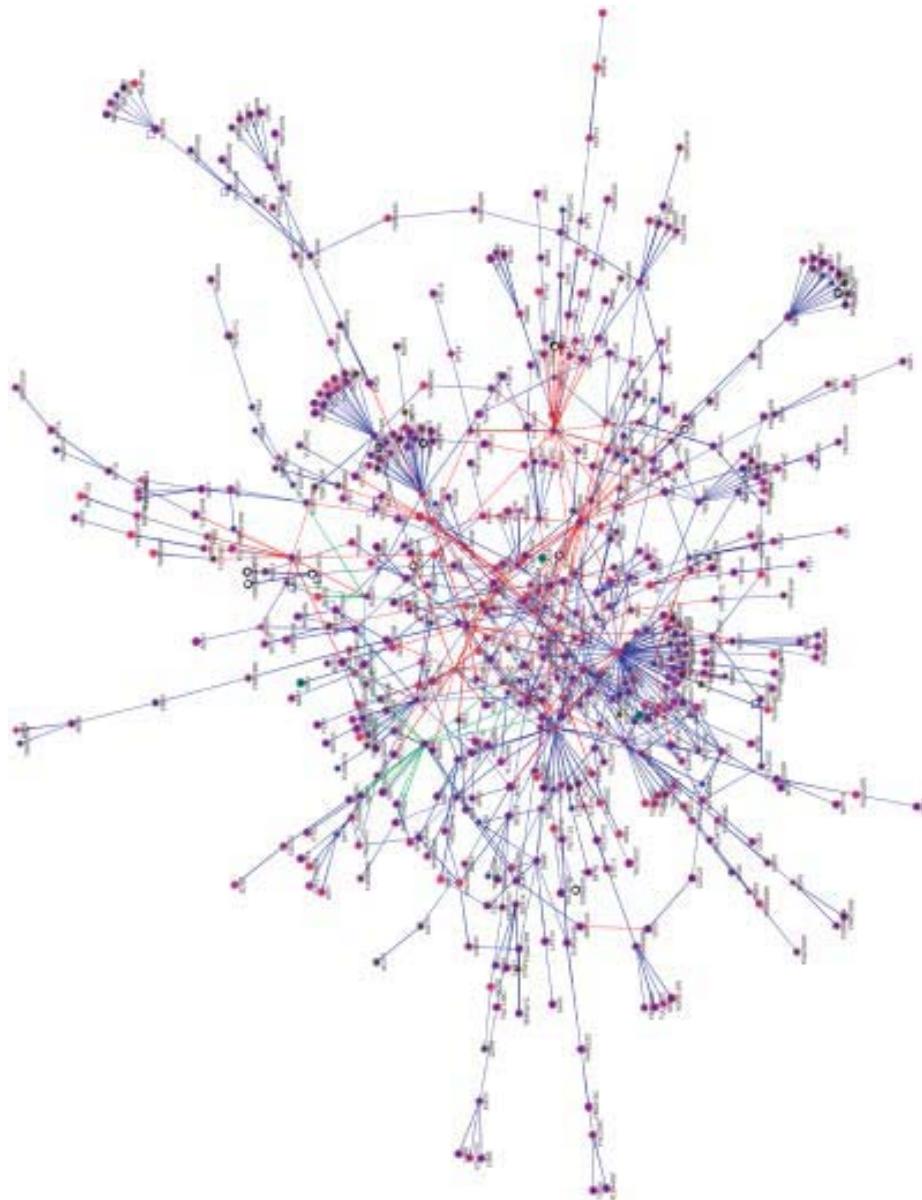


Fig. 6. Sample graph from Cytoscape. A number of the important features of Cytoscape are represented in this graph layout. Nodes in this case represent genes, and edges represent either genetic interactions, protein–protein interactions, or protein–DNA interactions. Node shapes are determined by the annotation of each gene, diamonds for signal transduction genes, triangles for meiosis, Pol III transcription, and DNA repair. Circles represent genes that were not assigned to any of these categories.

protein–DNA interactions, 1334 molecular complexes, and 8 pathways encompassing 28 genomes and over 6000 proteins.

A number of initiatives are currently underway to ensure that these data from different interaction databases are stored in a consistent and exchangeable format. PSI (78) has created a standard format for the exchange of protein–protein interaction data, while the BioPAX format aims to capture protein–protein interactions, molecular complexes, and pathway information in a single, consistent ontology and exchange format. Access information for DIP, BIND, and a number of other interaction databases is detailed in **Table 1**.

3. Notes

1. Unfortunately, the online Web server for *in silico* two-hybrid predictions at <http://www.pdg.cnb.uam.es/i2h/> is not presently available, although the Plotcorr program for analysis of correlated mutations is available at: <http://www.pdg.cnb.uam.es/pazos/plotcorr.html>.
2. Pairs of close homologs (PCHs) can be defined as follows: (a) A significant BLAST hit exists for protein α in genome X and protein α' in genome Y; (b) a significant BLAST hit exists for protein β in genome X and protein β' in genome Y; (c) the genes of proteins α and β are situated within 300 bp in genome X; (d) the genes of proteins α' and β' are situated within 300 bp in genome Y.
3. Although BLAST is useful for many analyses of this type, it is also possible to use more sensitive algorithms such as PSI-BLAST, HMMER, or Smith–Waterman. Although these methods tend to be far more computationally intensive, they may produce more accurate predictions.
4. Phylogenetic profiles do not necessarily have to be binary representations. It would also be possible to generate profiles that express a score or expectation value that a homolog is present in a given genome instead of simply “1” and “0.”
5. This type of analysis is very easy to perform using common mathematical analysis tools or the PEARSON function from Microsoft Excel™.
6. The Java™ environment is commonly preinstalled on many computer systems. If not already installed, it can be obtained at: <http://java.sun.com/>.
7. Graph images of protein–protein interaction networks can be exported in a number of ways. Capturing screenshots of either application will most likely result in a poor-quality, low-resolution image. For publication-quality images, it is generally best to export images as a vector graphics format such as PDF.
8. The PyMOL molecular graphics package is freely available for a variety of platforms at: <http://pymol.sourceforge.net/>.

Acknowledgments

The authors would like to thank Ronald Jansen for providing information about Bayesian network-based prediction of protein–protein interactions and the graph used for **Fig. 4**.

References

1. Mendelsohn, A. R. and Brent, R. (1999) Protein interaction methods—toward an endgame. *Science* **284**, 1948–1950.
2. Eisenberg, D., Marcotte, E. M., Xenarios, I., and Yeates, T. O. (2000) Protein function in the post-genomic era. *Nature* **405**, 823–826.

3. Huynen, M., Snel, B., Lathe, W., and Bork, P. (2000) Exploitation of gene context. *Curr. Opin. Struct. Biol.* **10**, 366–370.
4. Grigoriev, A. (2001) A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast *Saccharomyces cerevisiae*. *Nucleic Acids Res.* **29**, 3513–3519.
5. Ge, H., Liu, Z., Church, G. M., and Vidal, M. (2001) Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. *Nat. Genet.* **29**, 482–486.
6. Jansen, R., Greenbaum, D., and Gerstein, M. (2002) Relating whole-genome expression data with protein-protein interactions. *Genome Res.* **12**, 37–46.
7. Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O., and Eisenberg, D. (1999) A combined algorithm for genome-wide prediction of protein function. *Nature* **402**, 83–86.
8. Jansen, R., Yu, H., Greenbaum, D., et al. (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* **302**, 449–453.
9. Sussman, J. L., Lin, D., Jiang, J., et al. (1998) Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. *Acta Crystallogr. D. Biol. Crystallogr.* **54**, 1078–1084.
10. Chothia, C. and Janin, J. (1975) Principles of protein-protein recognition. *Nature* **256**, 705–708.
11. Gallet, X., Charlotteaux, B., Thomas, A., and Brasseur, R. (2000) A fast method to predict protein interaction sites from sequences. *J. Mol. Biol.* **302**, 917–926.
12. Korn, A. P. and Burnett, R. M. (1991) Distribution and complementarity of hydropathy in multisubunit proteins. *Proteins: Struct. Funct. Genet.* **9**, 37–55.
13. Young, L., Jernigan, R. L., and Covell, D. G. (1994) A role for surface hydrophobicity in protein-protein recognition. *Protein Sci.* **3**, 717–729.
14. Mueller, T. D. and Feigon, J. (2002) Solution structures of UBA domains reveal a conserved hydrophobic surface for protein-protein interactions. *J. Mol. Biol.* **319**, 1243–1255.
15. Lijnzaad, P. and Argos, P. (1997) Hydrophobic patches on protein subunit interfaces: Characteristics and prediction. *Proteins: Struct. Funct. Genet.* **28**, 333–343.
16. Janin, J., Miller, S., and Chothia, C. (1988) Surface, subunit interfaces and interior of oligomeric proteins. *J. Mol. Biol.* **204**, 155–164.
17. Argos, P. (1988) An investigation of protein subunit and domain interfaces. *Prot. Eng.* **2**, 101–113.
18. Jones, S. and Thornton, J. M. (1996) Principles of protein-protein interactions. *Proc. Natl. Acad. Sci. USA* **93**, 13–20.
19. Ofran, Y. and Rost, B. (2003) Analysing six types of protein-protein interfaces. *J. Mol. Biol.* **325**, 377–387.
20. Jones, S. and Thornton, J. M. (1997) Analysis of protein-protein interaction sites using surface patches. *J. Mol. Biol.* **272**, 121–132.
21. Jones, S. and Thornton, J. M. (1997) Prediction of protein-protein interaction sites using patch analysis. *J. Mol. Biol.* **272**, 133–143.
22. Lawrence, M. C. and Colman, P. M. (1993) Shape complementarity at protein-protein interfaces. *J. Mol. Biol.* **234**, 946–950.
23. Gabb, H. A., Jackson, R. M., and Sternberg, M. J. E. (1997) Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J. Mol. Biol.* **272**, 106–120.
24. Shoichet, B. K. and Kuntz, I. D. (1991) Protein docking and complementarity. *J. Mol. Biol.* **221**, 327–346.
25. Aloy, P. and Russell, R. B. (2002) Potential artefacts in protein-interaction networks. *FEBS Lett.* **530**, 253–254.

26. Casari, G., Sander, C., and Valencia, A. (1995) A method to predict functional residues in proteins. *Nat. Struct. Biol.* **2**, 171–178.
27. Lichtarge, O., Bourne, H. R., and Cohen, F. E. (1996) An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257**, 342–358.
28. Pazos, F., Helmer-Citterich, M., Ausiello, G., and Valencia, A. (1997) Correlated mutations contain information about protein-protein interaction. *J. Mol. Biol.* **271**, 511–523.
29. Zhou, H. X. and Shan, Y. B. (2001) Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins: Struct. Funct. Genet.* **44**, 336–343.
30. Fariselli, P., Pazos, F., Valencia, A., and Casadio, R. (2002) Prediction of protein-protein interaction sites in heterocomplexes with neural networks. *Eur. J. Biochem.* **269**, 1356–1361.
31. Ofran, Y. and Rost, B. (2003) Predicted protein-protein interaction sites from local sequence information. *FEBS Lett.* **544**, 236–239.
32. Aloy, P. and Russell, R. B. (2002) Interrogating protein interaction networks through structural biology. *Proc. Natl. Acad. Sci. USA* **99**, 5896–5901.
33. Aloy, P. and Russell, R. B. (2003) InterPreTS: protein Interaction Prediction through Tertiary Structure. *Bioinformatics* **19**, 161–162.
34. Karp, P. D., Riley, M., Saier, M., Paulsen, I. T., Paley, S. M., and Pellegrini-Toole, A. (2000) The EcoCyc and MetaCyc databases. *Nucleic Acids Res.* **28**, 56–59.
35. Tamames, J., Casari, G., Ouzounis, C., and Valencia, A. (1997) Conserved clusters of functionally related genes in two bacterial genomes. *J. Mol. Evol.* **44**, 66–73.
36. Dandekar, T., Snel, B., Huynen, M., and Bork, P. (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.* **23**, 324–328.
37. Overbeek, R., Fonstein, M., D’Souza, M., Pusch, G. D., and Maltsev, N. (1999) The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. USA* **96**, 2896–2901.
38. Zorio, D. A., Cheng, N. N., Blumenthal, T., and Spieth, J. (1994) Operons as a common form of chromosomal organization in *C. elegans*. *Nature* **372**, 270–272.
39. Blumenthal, T. (1998) Gene clusters and polycistronic transcription in eukaryotes. *Bioessays* **20**, 480–487.
40. Snel, B., Bork, P., and Huynen, M. A. (2002) Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Res.* **12**, 17–25.
41. Kunin, V., Cases, I., Enright, A. J., de Lorenzo, V., and Ouzounis, C. A. (2003) Myriads of protein families, and still counting. *Genome Biol.* **4**, 401.
42. Ouzounis, C. (1999) Orthology: another terminology muddle. *Trends Genet.* **15**, 445.
43. Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D., and Yeates, T. O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. USA* **96**, 4285–4288.
44. Ouzounis, C. and Kyriopoulos, N. (1996) The emergence of major cellular processes in evolution. *FEBS Lett.* **390**, 119–123.
45. Rivera, M. C., Jain, R., Moore, J. E., and Lake, J. A. (1998) Genomic evidence for two functionally distinct gene classes. *Proc. Natl. Acad. Sci. USA* **95**, 6239–6244.
46. Marcotte, E. M., Xenarios, I., van Der Blieck, A. M., and Eisenberg, D. (2000) Localizing proteins in the cell from their phylogenetic profiles. *Proc. Natl. Acad. Sci. USA* **97**, 12,115–12,120.
47. Galperin, M. Y. and Koonin, E. V. (2000) Who’s your neighbor? New computational approaches for functional genomics. *Nat. Biotechnol.* **18**, 609–613.
48. Enright, A. J., Iliopoulos, I., Kyriopoulos, N. C., and Ouzounis, C. A. (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402**, 86–90.
49. Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W., Yeates, T. O., and Eisenberg, D. (1999) Detecting protein function and protein-protein interactions from genome sequences. *Science* **285**, 751–753.

50. Enright, A. J., Van Dongen, S., and Ouzounis, C. A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575–1584.
51. Enright, A. J. and Ouzounis, C. A. (2001) Functional associations of proteins in entire genomes by means of exhaustive detection of gene fusions. *Genome Biol.* **2**, RESEARCH0034.
52. Pazos, F. and Valencia, A. (2002) *In silico* two-hybrid system for the selection of physically interacting protein pairs. *Proteins* **47**, 219–227.
53. Gobel, U., Sander, C., Schneider, R., and Valencia, A. (1994) Correlated mutations and residue contacts in proteins. *Proteins* **18**, 309–317.
54. Snel, B., Bork, P., and Huynen, M. A. (2002) The identification of functional modules from the genomic association of genes. *Proc. Natl. Acad. Sci. USA* **99**, 5890–5895.
55. Snel, B., Lehmann, G., Bork, P., and Huynen, M. A. (2000) STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res.* **28**, 3442–3444.
56. Overbeek, R., Larsen, N., Pusch, G. D., et al. (2000) WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res.* **28**, 123–125.
57. von Mering, C., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., and Snel, B. (2003) STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res.* **31**, 258–261.
58. Mellor, J. C., Yanai, I., Clodfelter, K. H., Mintseris, J., and DeLisi, C. (2002) Predictome: a database of putative functional links between proteins. *Nucleic Acids Res.* **30**, 306–309.
59. Tatusov, R. L., Koonin, E. V., and Lipman, D. J. (1997) A genomic perspective on protein families. *Science* **278**, 631–637.
60. Tatusov, R. L., Fedorova, N. D., Jackson, J. D., et al. (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41.
61. Altschul, S. F., Madden, T. L., Schaffer, A. A., et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402.
62. Iliopoulos, I., Enright, A. J., Poulet, P., and Ouzounis, C. (2003) Mapping functional associations in the entire genome of *Drosophila melanogaster*. *Comp. Funct. Genomics* **4**, 337–341.
63. Smith, T. F. and Waterman, M. S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197.
64. Enright, A. J. (2002) Computational Analysis of Protein Function in Complete Genomes. Ph.D. Thesis, University of Cambridge, p. 241.
65. Enright, A. J., Kunin, V., and Ouzounis, C. A. (2003) Protein families and TRIBES in genome sequence space. *Nucleic Acids Res.* **31**, 4632–4638.
66. Brazma, A., Hingamp, P., Quackenbush, J., et al. (2001) Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nat. Genet.* **29**, 365–371.
67. Gollub, J., Ball, C. A., Binkley, G., et al. (2003) The Stanford Microarray Database: data access and quality assessment tools. *Nucleic Acids Res.* **31**, 94–96.
68. Brazma, A., Parkinson, H., Sarkans, U., et al. (2003) ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.* **31**, 68–71.
69. Edgar, R., Domrachev, M., and Lash, A. E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30**, 207–210.
70. Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **95**, 14,863–14,868.
71. Walsh, S., Anderson, M., and Cartinhour, S. W. (1998) ACEDB: a database for genome information. *Methods Biochem. Anal.* **39**, 299–318.
72. Kanehisa, M. and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30.

73. Campagne, F., Neves, S., Chang, C., et al. (2003), In Press.
74. Enright, A. J. and Ouzounis, C. A. (2001) BioLayout—an automatic graph layout algorithm for similarity visualization. *Bioinformatics* **17**, 853–854.
75. Shannon, P., Markiel, A., Ozier, O., et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**(11), 2498–2504.
76. Bader, G. D., Betel, D., and Hogue, C. W. (2003) BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.* **31**, 248–250.
77. Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K., Marcotte, E. M., and Eisenberg, D. (2000) DIP: the database of interacting proteins. *Nucleic Acids Res.* **28**, 289–291.
78. Orchard, S., Hermjakob, H., and Apweiler, R. (2003) The proteomics standards initiative. *Proteomics* **3**, 1374–1376.
79. Mewes, H. W., Frishman, D., Guldener, U., et al. (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* **30**, 31–34.
80. Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., et al. (2004) IntAct: an open source molecular interaction database. *Nucleic Acids Res.* **32**, Database issue:D452–455.
81. Zanzoni, A., Montecchi-Palazzi, L., Quondam, M., Ausiello, G., Helmer-Citterich, M., and Cesareni, G. (2002) MINT: a Molecular INTeraction database. *FEBS Lett.* **513**, 135–140.

The Yeast Two-Hybrid System for Detecting Interacting Proteins

Ilya G. Serebriiskii, Erica A. Golemis, and Peter Uetz

1. Introduction

Protein–protein interactions play an essential role in all living systems, and hence their analysis is of foremost importance in molecular biology. Although there are number of methods to detect protein–protein interactions, the yeast two-hybrid system is probably the most successful method. Recently established protein interaction databases draw their data to a large extent from the summed input of small and large-scale two-hybrid screens.

1.1. The Principle

The two-hybrid system is based on the observation that protein domains can be separated, recombined, and still retain their properties. In particular, transcription factors can frequently be split into DNA-binding and activation domains. In the two-hybrid system, a DNA-binding domain (in this case, from the LexA protein) is fused to a protein B (for bait) for which one wants to find interacting partners (Fig. 1). A transcriptional activation domain is then fused to one or more proteins P (for prey), and the bait and prey fusion proteins are co-expressed in the same cell. Usually both protein fusions are expressed from plasmids that can be manipulated easily and then transformed into yeast cells. If the two proteins B and P interact, a transcription factor is reconstituted, which in turn activates one or more appropriate reporter genes. The expression of the most critical reporter allows the cell to grow only under certain conditions: e.g., the HIS3 reporter encodes imidazoleglycerolphosphate (IGP) dehydratase, a critical enzyme in histidine biosynthesis. In a screening strain lacking an endogenous copy of HIS3, expression of a HIS3 reporter gene is driven by a promoter that contains a LexA-binding site, so the bait protein fusion can bind to it. However, since the bait fusion does not contain a transcriptional activation domain, it remains inactive. If a protein P with an attached activation domain binds to the bait, this activation domain can recruit the basal transcription machinery, and expression of the reporter gene ensues. These cells can now grow in the absence of histidine in the medium because they can synthesize their own amino acid.

1.2. Variations on a Theme

Although the protocols in this chapter are based on the DNA-binding domain of LexA, other DNA-binding domains can be used. A popular DNA-binding domain has

From: *The Proteomics Protocols Handbook*
Edited by: J. M. Walker © Humana Press Inc., Totowa, NJ

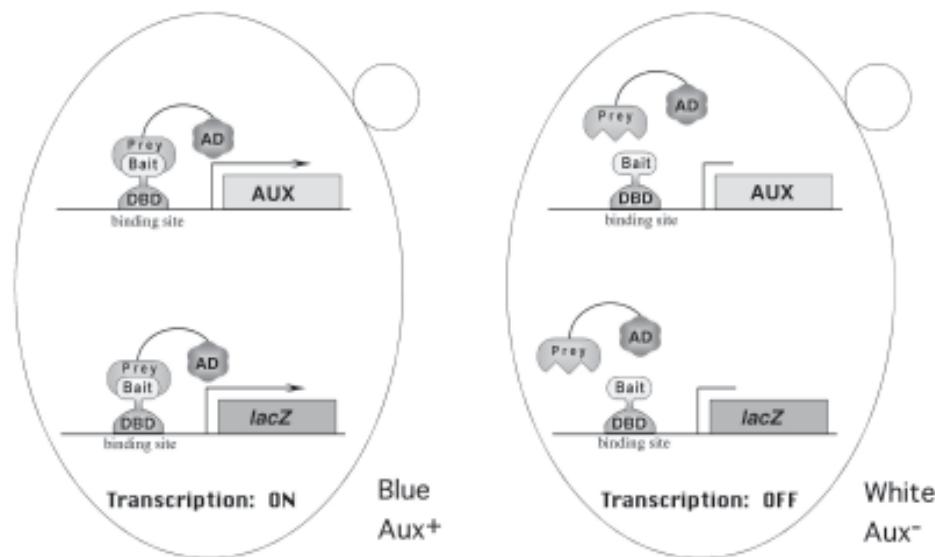


Fig. 1. Schematic of two-hybrid assay (modified from ref. 19). Paradigm for two-hybrid system, in which interaction between activation-domain-fused protein Y and DNA-binding-domain-fused protein X causes activation of a reporter gene under transcriptional control of binding sites for the DNA-binding domain. In the case of the interaction trap, the activation domain is provided by the “acid blob” B42, while the DNA-binding domain is provided by the bacterial protein LexA; activated reporter genes are LacZ and LEU2.

been taken from the yeast Gal4p protein. The same is true for the activation domain: our protocols use the B42 activation domain, while many commercial yeast two-hybrid systems are based on the Gal4p activation domain. In general, every component of the “classic” two-hybrid system can be replaced by different components: For example, the reporter gene does not need to be HIS3: in the protocols described below, LEU2, involved in leucine biosynthesis, is used. The reporter does not have to be a biosynthetic enzyme at all. Green fluorescent protein (GFP) has been successfully used as a reporter gene, β -galactosidase (*lacZ*) is common, and many others are under investigation. Finally, the two-hybrid system does not need to be based on transcription. Nils Johnsson and Alexander Varshavsky developed a related system that is based on reconstituting artificially split ubiquitin, a protein that tags other proteins for degradation (1). As long as the function of a protein can be used as a selective marker, it is theoretically possible to divide it into fragments, and drive the reassociation of the two fragments by exogenous “B and P” proteins that are attached to each half. Many such variations have been developed and are described elsewhere (2–4).

1.3. Applications

Originally, the two-hybrid system was invented to demonstrate the association of two proteins (5). Later, it was demonstrated that completely new protein interactions can be identified with this system, even when there are no candidates for interaction with a given bait. Over time, it has become clear that the ability to conveniently per-

form unbiased library screens is the most powerful application of the system, leading to new insights into cell signaling. Since its inception, the two-hybrid system has been adapted to a variety of related questions, such as the identification of mutants that prevent or allow interactions (6,7), the screening for drugs that affect protein interactions (8), the identification of RNA-binding proteins (9), or the semiquantitative determination of binding affinities (10). The system can also be exploited to map binding domains (11), to study protein folding (12), or to map interactions within a protein complex (13). Finally, recent large-scale projects have begun to map all interactions within whole proteomes (11,14–16). These studies have showed for the first time that most proteins in a cell are actually connected to each other. In combination with structural genomics, gene expression data, and metabolic profiling, the enormous amount of data in these interaction networks should allow us eventually to model complex biological phenomena in molecular detail. An ultimate goal of this work is to understand the interplay of DNA, RNA, and proteins, together with small molecules, in a dynamic and realistic way.

1.4. Limitations

Some classes of proteins are not suitable to analysis by the classic, transcription-based, yeast two-hybrid system. For example, transcriptional activators may activate transcription without any interaction, and hence cannot be used as baits. Another class of troublesome proteins is comprised of proteins containing hydrophobic transmembrane domains that may prevent the proteins from reaching the nucleus. To overcome this limitation, one of the alternative membrane-associated two-hybrid systems may be used (17). Other proteins may require modification by cytoplasmic or membrane-associated enzymes such as kinases in order to interact with binding partners. Alternative methods could also help in such cases.

False positives. As with many assay systems, the two-hybrid system has the potential to produce false positives (that is, reporter gene activity where no specific protein–protein interaction is involved). Frequently, such false positives are associated with bait proteins that act as transcriptional activators. False positives may also be caused by proteins that have the propensity to take part in nonspecific interactions (for largely unknown reasons). Some bait or prey proteins may affect general colony viability, and hence enhance the ability of a cell to grow under selective conditions and activate reporter gene activation. Mutations or other random events of unknown nature may be invoked as potential explanations as well. Overall, extremely few cases of false positives can be explained mechanistically. A number of procedures have been developed in order to identify or avoid false positives, including the utilization of multiple reporters, independent methods of specificity testing, or simply repeating assays to make sure a result is reproducible.

False negatives. In other cases, physiological protein–protein interactions are not detected by two-hybrid assays. False negatives may arise from steric hindrance of the two fusion proteins so that physical interaction or subsequent transcriptional activation is prevented. Other explanations for false negatives include instability of proteins or failure of nuclear localization; absence of a prey protein from a library; and inappropriate posttranslational modification of a bait or prey, prohibiting an interaction.

1.5. Scope of This Chapter

The provided protocols first describe the execution of a directed two-hybrid library screen utilizing a single bait protein. These protocols conventionally divide the execution of an interaction trap/two-hybrid screen into three stages, as illustrated in **Fig. 2**. First, initial characterization of a bait protein is described, with emphasis on the controls to be employed to increase the likelihood that the bait will function properly in a two-hybrid system. Then, transformation of a cDNA library, interaction mating with a pretransformed library, and selection of positive interactors is detailed. Subsequently, a number of control experiments aiming to establish significance of any detected interactions are outlined. Finally, protocols aiming to adapt these basic screening tools to genomic-level applications are provided.

2. Materials

Interaction trap reagents represent the work of many contributors: R. Brent, J. Kamens, S. Hanes, J. Gyuris, R. Finley, E. Golemis, I. York, M. Sainz, S. Nottwehr, D. Shaywitz, and others. Further contributions have been made by workers at Glaxo-SmithKline in Research Triangle Park, and at Invitrogen. A more complete table with all reagents compatible with interaction trap and a derivative system (the dual bait system, (18)) is available at <http://www.fccc.edu/research/labs/golemis/interaction-trapinwork.html>. Many of these reagents are available commercially, or can be acquired by request from Ilya Serebriiskii at Fox Chase Cancer Center, (215) 728-3885 phone, (215) 728-3616 fax, IG_Serebriiskii@fccc.edu. All the protocols utilize a common set of reagents. Thus, all materials necessary for the three basic protocols are presented here.

2.1. Plasmids

1. pMW103, a plasmid for making LexA fusion proteins. Expression is from the strong ADH1 promoter, while the plasmid is selected for in yeast with a *HIS3* marker. The bacterial selective marker is for kanamycin resistance (Km^R). The polylinker is shown below (only sites suitable for cloning are indicated in bold)

<i>Eco</i> RI	<i>Sal</i> I	<i>Not</i> I	<i>Sal</i> I
<i>Bam</i> HI		<i>Xba</i> I	
GAA TTC CCG GGG ATC CGT CGA CCA TGG CGG CCG CTC GAG TCG AC			

2. pJG4-5, a plasmid for making a nuclear localization sequence-activation domain-hemagglutinin epitope tag fusion to a unique protein or a cDNA library. Expression in yeast is from the GAL1 galactose-inducible promoter, while the plasmid is selected for with a *TRP1* marker. The bacterial selective marker is for ampicillin resistance (Ap^R).
3. pMW109, a plasmid containing two LexA operators upstream of the LacZ reporter gene. Yeast selection is for URA3, whereas the bacterial selective marker is Km^R .
4. pSH17-4, a plasmid encoding LexA-GAL4, a strong positive control for activation. Expression is from the strong ADH1 promoter, whereas the plasmid is selected for in yeast with a *HIS3* marker. The bacterial selective marker is Ap^R .
5. pEG202-hsRPB7, *HIS3* plasmid encoding LexA-hsRPB7, a weak positive control for activation. Expression is from the strong ADH1 promoter, whereas the plasmid is selected for in yeast with a *HIS3* marker. The bacterial selective marker is Ap^R .
6. pEG202-Ras, *HIS3* plasmid encoding LexA-Ras, a negative control for activation and positive control for interaction. Expression is from the strong ADH1 promoter, while the

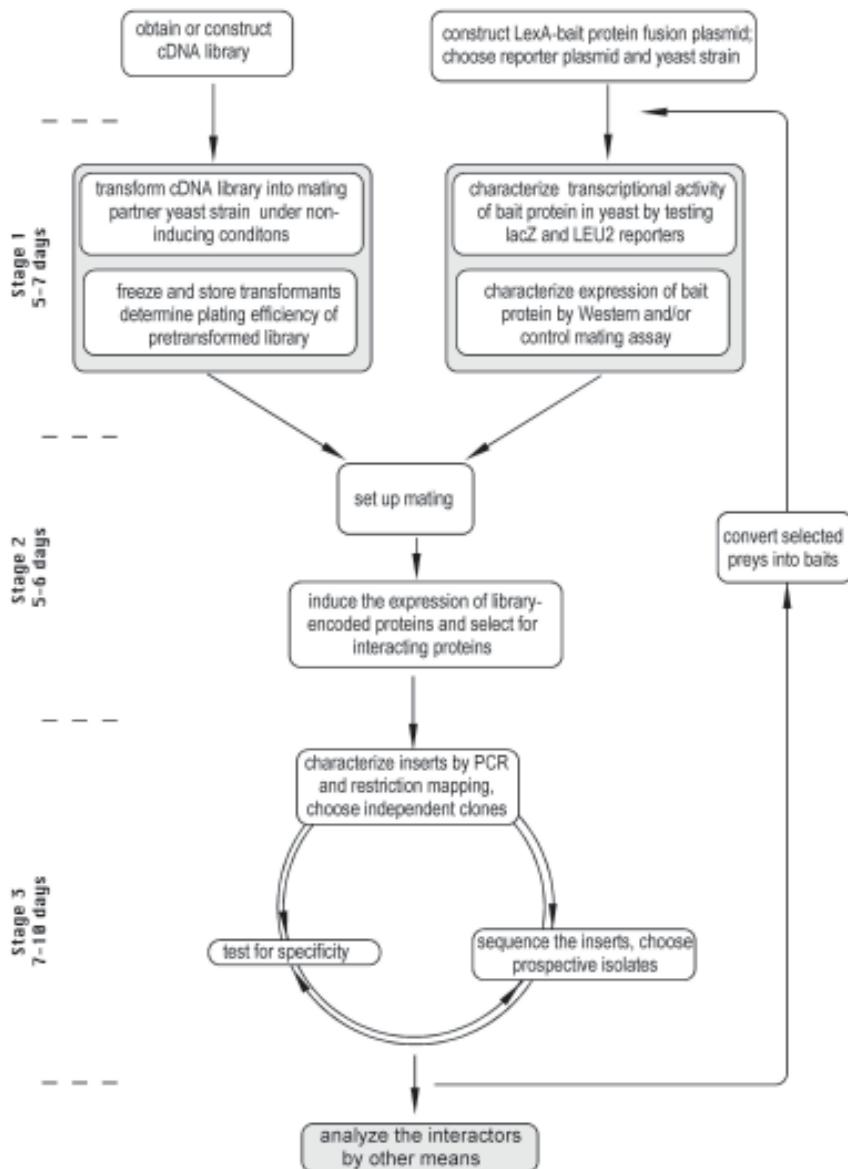


Fig. 2. Flowchart of the two-hybrid screen done by interaction mating (modified from ref. 19). Stage 3 allows flexibility in step order, shown in greater detail in [Fig. 4](#). See text for details.

- plasmid is selected for in yeast with a *HIS3* marker. The bacterial selective marker is Ap^R .
7. pJG4-5:Raf, a prey plasmid encoding a positive control for interaction with Ras. Markers and expression as for pJG4-5.
 8. (Optional) pJK202, pGilda, pNLexA; plasmids related to pEG202 that incorporate a nuclear localization sequence into the LexA-fusion construct, are expressed from the inducible *GAL1* promoter, or fuse LexA to the carboxy-terminal end of the test protein, respectively.

Maps and sequences for most of these plasmids are available at <http://www.fccc.edu/research/labs/golemis/plasmids.html>.

2.2. Strains

Yeast strain SKY 48 (*MAT α ura3 trp1 his3 3LexAop- leu2 3cIop-lys2*): for baits.

Yeast strain SKY473 (*MAT α ura3 trp1 his3 2LexAop-leu2 3cIop-lys2*): for a prey library.

The *ura3*, *trp1*, and *his3* markers in these strains are selective markers to maintain colorimetric reporter, library, and bait plasmids, respectively. *LexAop-leu2* is an auxotrophic reporter that is used to select clones expressing library proteins that interact with the LexA-fused bait. *3cIop-lys2* is an auxotrophic reporter that may be used with an optional cI-fused bait in the dual bait system (discussed in [ref. 18](#)).

2.3. Reagent for β -Galactosidase Overlay Assays

X-Gal-agarose: 1% low-melting agarose in 100 mM KHPO₄, pH 7.0; add XGal to 0.25 mg/mL when cooled to approx 60°C.

2.4. Plates for Growing Bacteria (100 mm)

LB supplemented with 50 μ g/mL ampicillin.

2.5. Defined Minimal Yeast Medium

All minimal yeast media, liquid and plates, are based on the following three ingredients, which are sterilized by autoclaving 15–20 min: per liter, 6.7 g Yeast Nitrogen Base-amino acids (Difco 0919-15); 20 g glucose or 20 g galactose + 10 g raffinose; 2 g appropriate nutrient “dropout” mix (see below). For plates, 20 g Difco bacto-agar (Difco 0140-01) are also added.

A complete minimal (CM) nutrient mix includes the following: adenine (2.5 g), L-arginine (1.2 g), L-aspartic acid (6.0 g), L-glutamic acid (6.0 g), L-histidine (1.2 g), L-isoleucine (1.2 g), L-leucine (3.6 g), L-lysine (1.8 g), L-methionine (1.2 g), L-phenylalanine (3.0 g), L-serine (22.0 g), L-threonine (12.0 g), L-tryptophan (2.4 g), L-tyrosine (1.8 g), L-valine (9.0 g), uracil (1.2 g).

Leaving out one or more nutrients selects for yeast able to grow in its absence, i.e., containing a plasmid that covered the deficiency. Thus, “dropout medium” lacking uracil (denoted –Ura in the following recipes) would select for the presence of plasmids with the URA3 marker, etc. Note, preparing a complete minimal mix with the above quantities of nutrients produce a quantity of dropout powder sufficient to make 40 L of medium: it is advisable to scale down for most of the below dropout combinations. Note, premade dropout mixes are available from some commercial suppliers.

2.6. Plates for Growing Yeast (100 mm)

- Defined minimal dropout plates, with glucose as a carbon source (Glu/CM): –Ura, –His, –Ura, –His, –Trp; –Ura, –His, –Trp, –Leu.
- Defined minimal dropout plates, with galactose + raffinose as a carbon source (Gal/Raf/CM): –Ura, –His, –Trp, –Leu.
- YPD (rich medium): per liter, 10 g yeast extract, 20 g peptone, 20 g glucose, 20 g Difco Bactoagar; autoclave approx 18 min, pour approx 40 plates.

2.7. Liquid Medium for Growing Yeast

1. Defined minimal dropout medium, with glucose as a carbon source:
–Ura, –His; –Trp.
2. Defined minimal dropout medium, with galactose + raffinose as a carbon source:
–Ura, –His, –Trp.
3. YPD: per liter, 10 g yeast extract, 20 g peptone, 20 g glucose; autoclave approx 15 min.

2.8. Plates for Growing Yeast Library Transformations (240 x 240 mm)

Defined minimal dropout plates, with glucose as a carbon source: –Trp. Each plate requires approx 250 mL medium.

2.9. Primers

1. For LexA-fusion plasmids: forward primer, to confirm correct reading frame after cloning a bait into the polylinker, CGT CAG CAG AGC TTC ACC ATT G.
2. For JG4-5 plasmid, to confirm identity of prey proteins: forward primer, FP1, 5'-CTG AGT GGA GAT GCC TCC; reverse primer, FP2, 5' CTG GCA AGG TAG ACA AGC CG.
3. To transfer a cDNA from pJG4-5 to pEG202 (by polymerase chain reaction [PCR] recombination): forward primer,
5' GGG CTG GCG GTT GGG GTT ATT CGC AAC GGC GAC TGG CTG GTG CCA GAT TAT GCC TCT CCC G 3'; reverse primer, 5' GAG TCA CTT TAA AAT TTG TAT ACA C 3'.

2.10. Miscellaneous

1. Sterile glass balls, 3–4 mm, #3000, Thomas Scientific 5663L19 or Fisher #11-312A.
2. Sterile glycerol solution for freezing transformants (65% sterile glycerol, 0.1 M MgSO₄, 25 mM Tris-HCl, pH 8.0).
3. “Insert grid” from a rack of pipet tips (Rainin RT series, 200 μ L capacity).
4. A metal frogger (e.g., Dankar Scientific #MC48).
5. A plastic replicator (Bel-Blotter, Bel-Art Products #378776-0002 or Fisher # 1371213).
6. Nunc omnitrays (Nunc #242811).

3. Methods

In the interest of space only, we refer the reader to the commercial kits in many instances (e.g., for yeast transformation, or isolation of DNA from yeast, and so on). Homemade reagents work equally well; standard protocols are described (e.g., in [ref. 19](#)). All methods described below should use aseptic techniques; yeast are grown at 30°C, and *Escherichia coli* at 37°C.

3.1. Constructing and Transforming a Bait Protein

A prerequisite for an interactor hunt is the construction of plasmids that express the protein of interest as fusions to the bacterial protein LexA (or some other DNA-binding domain). These plasmids are transformed into a yeast reporter strain to assess the suitability of the bait proteins for library screening. Comparison to some established controls allows determination of whether the baits are appropriately synthesized, not transcriptionally active, and not toxic. If any of these conditions is not met, strategies for modifying bait or screening conditions are suggested ([Table 1](#)). To minimize the chance of artifactual results or other difficulties, it is a good idea to move rapidly through the suggested characterization steps before undertaking a library screen. Although plasmids will be retained for extended periods in yeast maintained at 4°C on

Table 1
Troubleshooting and Alternative Plasmids: Possible Modifications to Enhance Bait Performance in Specific Applications

Bait problem response	Strongly activating	Weakly activating	expression level	Continuous expression of LexA-fusion is toxic to yeast	Bait protein requires unblocked amino-terminal end for function	interacts promiscuously	Potential new problem*
Truncate/modify bait	+	—	—	—	—	—	It may be necessary to subdivide bait into two or three overlapping constructs, each of which must be tested independently
Use more stringent strain/reporter combination	+	+	—	—	?	?	Use of very stringent interaction strains may eliminate detection of biologically relevant interactions
Fuse to nuclear localization sequence	—	—	+	—	—	—	
pJK202	?	—	—	—	+	?	Can no longer use GAL4-dependence of reporter phenotype to indicate cDNA-dependent interaction
Put LexA-fused protein under GAL4-inducible promoter pGilda	?	—	—	—	—	—	Generally, LexA poorly tolerates attachment of the N-terminal fusion
Fuse LexA to the carboxy terminus of the bait pNLexA	—	—	—	—	—	+	

Integrate bait, reduce concentration pEG2021	+?	—	—	+?	—	+?	+
		—	—	—	—	—	—

+, Would usually help; +?, may help; -, will not help.
 *Most of the alternative bait expression vectors remain on an Amp^R selection for bacteria. If using them as is, the investigator may need to use a KC8 bacteria to isolate the library plasmid after a library screen.

stock plates, it is recommended to perform the next step in the protocol (e.g., mating or characterization) within a week from the initial introduction of the plasmid(s) in the yeast, to avoid variable protein expression and transcriptional activation problems.

1. Clone the DNA encoding the protein of interest into the polylinker of pMW103 (**Subheading 2.1.**) to enable synthesis of an in-frame protein fusion to LexA (*see Notes 1, 2*).
2. Select a colony of SKY48 and prepare competent yeast (*see Notes 3, 4*).
3. Transform competent yeast cells with the following combinations of LexA fusion and pMW109 plasmids (100–500 ng each):
 - a. pBait + pMW109 (test for activation)
 - b. pEG202-hsRPB7 + pMW109 (weak positive control for activation)
 - c. pSH17-4 + pMW109 (strong positive control for activation)
 - d. pEG202-Ras + pMW109 (negative control for activation)
4. Plate each transformation mixture on Glu/CM –Ura–His dropout plates, and maintain for 2 to 3 d to select for yeast colonies containing transformed plasmids (*see Note 5*).

3.2. Assessing Bait Activation and Expression

3.2.1. Replica Technique/Gridding Yeast: Assessing Activation of Auxotrophic Reporter

For each combination of plasmids, assay at least six (*see Note 6*) independent colonies for activation phenotype of auxotrophic and colorimetric reporters. Assessment of transcriptional activation requires the transfer of yeast from master plates to a variety of selective media. This transfer can be accomplished simply by using a sterile toothpick to move cells from individual patches on the master plate to each of the selective media. However, in cases in which large numbers of colonies and combinations of bait and prey are to be examined, and particularly in genomic-scale applications, it is useful to use a transfer technique that facilitates high-throughput analysis. The following technique, based on microtiter plates, is an example of such an approach.

1. Add approx 50 μ L of sterile water to each well of one-half (wells A1–H6) of a 96-well microtiter plate (a syringe-based repeater or a multichannel pipettor can be used). Place an insert grid from a rack of pipet tips over the top of the microtiter plate and attach it with tape: the holes in the insert grid should be placed exactly over the wells of the microtiter plate (placing the grid is essential to stabilize the tips in the plate, and allow their simultaneous removal) (*see Fig. 3* for details).
2. Using sterile plastic pipet tips, pick six yeast colonies (1 to 2 mm diameter) from each of the transformation plates a–d (**Subheading 3.1., step 3**). Leave the tips supported by the insert grid until all the colonies have been picked.
3. Swirl the plate gently to mix the yeast into suspension, remove the tape attaching the grid to the plate, and lift the insert grid, thereby removing all the tips at once.
4. Use a replicator (*see Note 7*) to plate yeast suspensions (each spoke will leave a drop approximately equal to a 3- μ L vol) on the following plates: Glu/CM –Ura–His (a new master plate); Gal-Raff/CM –Ura–His–Leu, and Gal- Raff/CM –Ura–His, (for X-Gal overlay assay, *see Subheading 3.2.2.* below). Incubate the plates for up to 4 d, and save the Glu/CM –Ura–His master plate at 4°C.
5. Yeast containing the strong positive control (c) should be detectably growing on –Ura–His–Leu plate within 1–2 d; yeast with the weak positive control (b), within 4 d; and the negative control (d) should not grow (*see Fig. 3* for illustration). If the bait under test (a) shows no growth in this period, it is probably suitable for library screening; if it gives a profile similar to (b), it may be suitable, but is likely to have a high background in library

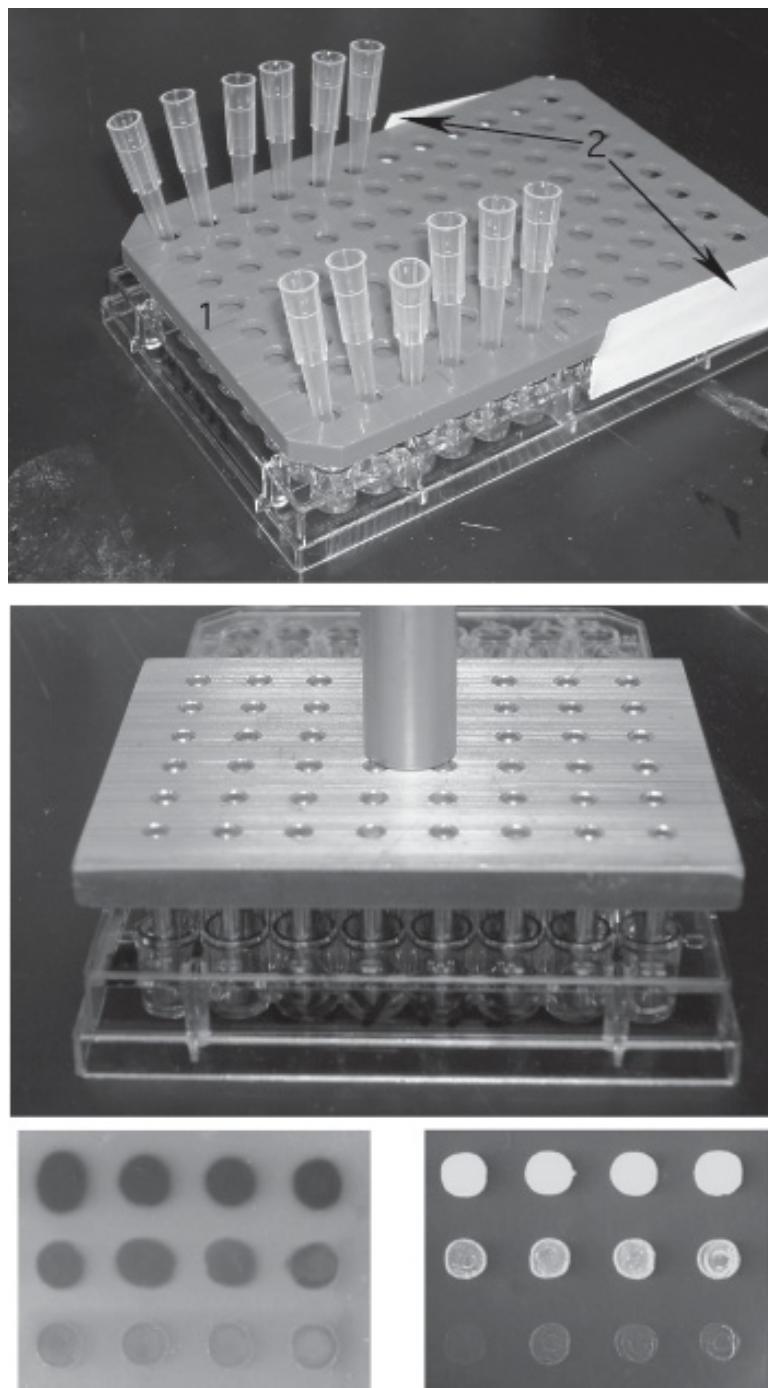


Fig. 3. Replica plating technique. Top, the 96-well plate with a pipet tip insert (1) attached with a sticky tape (2). Pipet tips are used to transfer yeast colonies in the liquid in the wells (only two rows are shown for clarity). Middle, a replicator is inserted in the wells after the insert and the tips are removed. Bottom, representative data for colorimetric (left) and auxotrophic (right) reporter activation. From top to bottom: strong, medium, weak (background). See text for details

screening, suggesting that use of a different screening strain may be appropriate; if it is similar to (c), it must be reconfigured (see **Table 1**).

3.2.2. Assessing Activation of Colorimetric Reporter

Approx 24 to 30 h after the plating, overlay the Gal-Raff/CM –Ura–His plate (**Subheading 3.2.1., step 4**) with X-Gal agarose (**20**) as follows (see **Note 8**):

1. Gently overlay each plate with chloroform, pipetting slowly in from the side so as not to smear colonies. Leave colonies completely covered for 5 min. Caution: CHCl_3 is quite toxic, and should neither be inhaled nor come into contact with skin. Wear gloves and work in a chemical hood. Try to avoid extensive contact with the walls of the plate, as the plastic dissolved in CHCl_3 may leave a film on the agar/colonies surface.
2. Briefly overlay the plates with another approx 5 mL chloroform (optional), drain, and let dry, uncovered, for another 5 min at 37°C or for 10 min in the chemical hood.
3. Overlay the plate with approx 10 mL of X-Gal-agarose, making sure that all yeast spots are completely covered.
4. Incubate plates at 30°C and monitor for color changes. It is generally useful to check the plates after 20 min, and again after 1 to 3 h. Strongly activating baits will be detectable as dark blue colonies in 20 to 60 min, whereas negative controls should remain as faint blue or white colonies; an optimal bait would either mimic the negative control or only develop faint blue color (see **Fig. 3** for illustration and **Note 8**).

In an optimal result, all six colonies assayed representing the same transformation would have essentially the same phenotype. For a small number of baits, this is not the case. The most typical deviation is that of six colonies assayed for the bait, some fraction appears to be inactive (white in colorimetric assay, and not growing on auxotrophic selection medium), while the remaining fraction displays some degree of blueness and growth. Do not select the white, nongrowing colonies as the starting point in a library screen; often, these colonies are synthesizing little or no bait protein (as can be assayed by Western blot, **Subheading 3.2.3.**).

3.2.3. Detection of Bait Protein Expression

One excellent confirmation that a bait protein is correctly expressed would be its specific interaction with a known partner, expressed as an activation-domain fusion protein. In the absence of such confirmation, Western analysis of lysates of yeast containing DBD-fused baits is helpful in characterization of the bait's expression level and size. Some proteins (especially where the fusion domain is approx 60–80 kDa or larger) may either be synthesized at very low levels, or be posttranslationally clipped by yeast proteases. Proteins expressed at low levels, and apparently inactive in transcriptional activation assays, can be epigenetically upregulated to much higher levels under the auxotrophic selection and suddenly demonstrate a high background of transcriptional activation. Where proteins are proteolytically clipped, screens might inadvertently be performed with DBD fused only to the amino-terminal end of the larger intended bait. Either of the above two problems can lead to complications in library screens. Western analysis should be performed as follows:

1. From the master plate (**Subheading 3.2.1., step 5**), inoculate at least two representative bait/reporter transformants for each bait to be tested into liquid medium (see **Note 9**). Grow overnight cultures and then dilute into fresh tubes containing 2 mL of the same medium to a density of approx 0.15 at OD 600 nm (see **Note 10**).

2. Harvest cells from 1.5 mL of each exponentially growing culture (OD 600 nm approx 0.45–0.7) by centrifuging at 13,000g for 3–5 min and carefully removing supernatant.
3. Resuspend each pellet in 50 μ L of 2X Laemmli sample buffer (*see Note 11*). Heat the samples at 100°C for 5 min, chill on ice, and centrifuge for 30 s at 13,000g to pellet large-cell debris. Load 10–25 μ L of each sample onto a 0.1% (w/v) sodium dodecyl sulfate (SDS)-polyacrylamide gel electrophoresis (PAGE) gel.
4. Prepare a Western blot and screen LexA fusions using an antibody to LexA; compare expression levels of the bait protein under test with other standard bait proteins (*see Note 12*). Note which colonies on the master plate express bait appropriately, and use one of these colonies as a founder to propagate for library mating.

3.3. Transforming a Library, and Characterizing Interactors From a Screen

A partial list of available libraries compatible with the interaction trap is found at <http://www.fccc.edu/research/labs/golemis/InteractionTrapInWork.html>; currently, the most convenient source of libraries suitable for the interaction trap is commercial. The following protocols are designed with the goal of saturation screening of a cDNA library derived from a genome of mammalian complexity. Fewer plates will be required for screens with libraries derived from organisms with less complex genomes, and researchers should scale back accordingly. A protocol for mating the library against the bait of interest is provided in (21); it is generally a good idea to additionally mate new bait strains with a negative control strain. The control strain is the same strain used for the library but containing the library vector with no cDNA insert. This control will provide a clear estimate of the frequency of cDNA-independent false positives, which is important to know when deciding how many positives to pick and characterize.

3.3.1. Transforming the Library

1. Select a colony of SKY473 and prepare competent yeast (*see Notes 3–5*). There should be enough yeast to obtain at least approx 10^6 transformants. Add up to 30 μ g of library DNA, subdivide the mixture into 30 separate tubes, and complete transformation (*see Note 13*).
2. Pipet the contents of each tube onto a separate 240 \times 240 mm Glu-Trp dropout plate, and spread the cells evenly using 12–24 sterile glass beads (*see Subheading 2.10., item 1*). Invert the plates (*see Note 14*) and incubate until all colonies appear (3–4 d).
3. Select two representative transformation plates, draw a 23 \times 23 mm square (1% of the plate bottom surface) over an average density spot, count the colonies in each grid section, and recalculate for the whole transformation. A good transformation should yield approx 20,000–40,000 colonies per plate.
4. Use a small aliquot of competent yeast from step 1 to transform the empty library plasmid (pJG4-5 or pYesTrp2). Plate on 100-mm plate, and collect the transformed cells as for the library (protocol outlined below), scaling down accordingly. The control strain with an empty library plasmid can be safely reamplified in liquid medium.

3.3.2. Harvesting and Pooling Primary Transformants

In the next step, a homogenized slurry is prepared from the pool of primary transformants (approx 3×10^5 – 10^6 colonies), aliquoted, and frozen. Each of these aliquots is representative of the complete set of primary transformants and can be used in subsequent mating (*see Note 15*). To prepare this slurry:

1. Pour 10–15 mL of sterile water onto each of five 240 \times 240 mm plates containing transformants. Stack the five plates on top of each other. Holding on tightly, shake the

stack horizontally until all the colonies are resuspended (1 to 2 min). Using a sterile pipet, collect yeast slurry from each plate (by tilting the plates) and pool in a sterile 50-mL conical tube (see Note 16).

2. Repeat for further sets of five plates of transformants, resulting in a total of up to 150–200 mL of suspension split between several 50-mL tubes (see Notes 17, 18).
3. Fill each tube containing yeast to the top with sterile TE or water, and vortex/invert to suspend the cells. Centrifuge the tubes at 1000–1500g for 5 min at room temperature, and discard the supernatants. Repeat this step. After the second wash, the cumulative pellet volume should be approx 25 mL of cells derived from up to 10^6 transformants.
4. Resuspend each packed cell pellet in 1 vol of glycerol solution. Combine the contents of the three tubes and mix thoroughly. Dispense as 0.2- to 1.0-mL aliquots in a series of sterile Eppendorf tubes and freeze at -70°C (stable for at least 1 yr).

3.3.3. Mating the Bait Strain and the Pretransformed Library

Once the bait strain has been made and characterized, and the library strain has been transformed and frozen in aliquots, the next step is to mate the two strains. To mate the two strains, the bait strain is grown in liquid culture and mixed with a thawed aliquot of the pretransformed library strain. The mixture is plated on rich medium and grown overnight. During this time, individual cells of the bait strain will fuse with individual cells of the library strain to form diploid cells. The diploids, along with unmated haploids, are collected and plated on media to select for interactors (as described in Subheading 3.3.4.). In practice, the diploid/haploid mixture is generally frozen in a few aliquots to allow titering and repeated platings at various dilutions. Mating with the negative control strain (see Subheading 3.3.1., step 4) should be performed at the same time as the library mating, and both matings can be treated identically in the next step, selecting interactors.

1. Start a 30-mL Glu/CM –Ura–His liquid culture of the bait strain (SKY48/pBait/placZ) from the Glu/CM –Ura–His master plate. Grow with shaking to mid- to late-log phase ($\text{OD}_{600} = 1.0\text{--}2.0$) (see Note 19).
2. Collect the cells by centrifuging at 1000g for 5 min at room temperature. Resuspend the cell pellet in 1 mL of sterile water and transfer to a sterile 1.5-mL microfuge tube. This will yield yeast suspension of about 1×10^9 cells/mL.
3. Thaw an aliquot of the pretransformed library strain and an aliquot of the negative control strain at room temperature. Mix approx 2×10^8 cells of the bait strain (approx 200 μL) with approx 10^8 colony-forming units (CFU) of the pretransformed library strain (see Subheading 3.3.2., step 4) on a single 100-mm diameter YPD plate and incubate overnight. In parallel, set up a mating with the negative control strain.
4. Add 1.5 to 2 mL of sterile water to the surface of each YPD plate and suspend the cells using sterile glass beads. Transfer the suspension to a sterile tube and vortex gently for 2 min. Collect the cells by centrifugation at 1000g for 5 min and resuspend in 1 vol of sterile glycerol solution. Distribute into 200- μL aliquots and freeze at -80°C (see Note 20). An option is to leave one aliquot unfrozen and proceed directly to plating on selective medium (see Subheading 3.3.4.); approx 5 h required to complete process.
5. Titer the mated cells by plating serial dilutions on Glu/CM –Trp–His–Ura plates (unmated haploids will not grow on this medium). Count the colonies that grow after 2 to 3 d, and determine the titer of the frozen mated cells (see Note 21).

3.3.4. Screening for Interacting Proteins

This section describes how interactors are selected by plating the mated cells onto auxotrophic selection plates. It is important to know how many viable diploids were plated onto these selection plates to gain a sense of how much of the library has been screened and to determine the false-positive frequency. This information is provided by the titer (colony-forming units per milliliter) of the frozen mated cells (*see Subheading 3.3.3., step 5*).

1. Thaw an aliquot of the mated yeast. Dilute 100 μ L into 10 mL of Gal-Raff/CM –Ura–His–Trp liquid dropout medium and incubate with shaking for 5 h.
2. On the assumption that a culture at OD 600 nm = 1.0 contains 1×10^7 cells/mL, plate 10^6 cells on five 100-mm plates with the appropriate auxotrophic selection medium. Plate 10^7 cells on each of five additional plates with the same medium. Plating 10^7 cells/plate allows screening more diploids on a smaller number of plates, but may or may not result in higher levels of background growth.
3. Incubate for up to 6 d at 30°C (*see Note 22*). Depending upon the individual bait used, good candidates for positive interactors will generally produce LEU+ colonies over this time period, with the most common appearance of colonies at 2–4 d (*see Note 23*).
4. Inspect the plates on a daily basis. Mark the location of colonies visible on d 1 with dots of a given color on the plate. Each day, mark further colonies arising with different colors. At d 4 or 5, streak colonies in a microtiter plate (i.e., 6×8) format onto a solid master plate (Glu/CM–Ura–His–Trp), in which colonies are grouped by day of appearance (*see Note 24*). If many apparent positives appear, pick separate master plates for colonies arising on d 2, 3, and 4, respectively (*see Note 25*).
5. Include a few colonies from the titer plate of control mating on each of the master plates. Because they contain empty library plasmids, the phenotype of these colonies should be negative. Incubate the master plates until patches/colonies form.

3.3.5. First Confirmation of Positive Interactions

The following steps test for galactose-inducible transcriptional activation of both the auxotrophic (*LEU2*) and colorimetric (*lacZ*) reporters. Simultaneous activation of both reporters in a galactose-specific manner generally indicates that the transcriptional phenotype is attributable to expression of library-encoded proteins, rather than derived from mutation of the yeast.

1. Invert a frogger on a flat surface and place a master plate upside down on the spokes, making sure that the spokes and colonies are properly aligned. Remove the plate and insert the frogger into a microtiter plate containing 50 μ L of sterile water in each well. Let the plate sit for 5–10 min, shaking from time to time to resuspend the cells left on the spokes. When all yeast are resuspended, print (*see Note 7*) on the following plates:
 - a. Master plate: Glu/CM –Ura–His–Trp
 - b. Test for activation of *LEU2*: Gal-Raff/CM (–Ura–His–Trp)–Leu
 - c. Glu/CM (–Ura–His–Trp)–Leu
 - d. Plates, to be assayed for *LacZ* activation: Glu/CM –Ura–His–Trp
(*see Note 26*) Gal-Raff /CM –Ura–His–Trp

2. Incubate the plates for 3 to 4 d. After 20–30 h of incubation, take out all –Ura–His–Trp plates. Retain one Glu/CM –Ura–His–Trp plate as a fresh master plate, and overlay the remaining set with X-Gal agarose, as described in **Subheading 3.2.2**. Score growth on the –Leu plates 48–72 h after plating (see **Fig. 3** for illustration).

For interpretation of the results, refer to **Table 2**.

3.4. DNA Isolation and Second Confirmation of Positive Interactions

Execution of the above protocols for a given bait will result in the isolation of between zero and hundreds of potential “positive” interactors (see **Note 27**). These positives must next be evaluated for reproducible phenotype, and for specificity of interaction with the bait used to select them. If a large number of positives are obtained, these subsequent characterizations require prioritization. In this case, select up to approx 24–48 independent colonies with robust phenotype for the first round of characterization, while maintaining a master plate of additional positives at 4°C. This first analysis set will be tested for specificity, and screened by PCR/restriction analysis and/or sequencing to determine whether clusters of frequently isolated cDNAs are obtained: such clusters are generally a good indication for a specific interaction. Two strategies for analyzing positives are provided below, and summarized in the flow chart in **Fig. 4**. While both utilize similar methods, the order with which techniques are applied differs; the choice between strategies depends on whether the individual investigator would rather spend time and money doing bulk yeast plasmid recovery, or bulk PCR. The latter protocol is generally 1–3 d faster, but is not as reproducibly accomplished in some investigators’ hands.

3.4.1. PCR Approach: Rapid Screen for Interaction Trap Positives

A major strength of this protocol is that it will identify redundant clones prior to plasmid isolation and bacterial transformation, which in some cases greatly reduces the amount of work required.

For the sake of simplicity, the protocol outlined below relies on a commercially available kit to isolate plasmid DNA from yeast; however, PCR product can be obtained directly from the yeast colonies (e.g., by introducing a 10-min, 94°C step at the beginning of the PCR program).

3.4.1.1. PCR Characterization of Initial Positives

1. Isolate plasmid DNA from yeast (e.g., using commercially available kit (Zymoresearch, Cat. No. D2001). Alternatively, a simple enzymatic treatment can generate crude yeast lysates (**19**), later used as template for the PCR reaction or for bacterial transformation.
 2. (a) Use primers specific for the library plasmid (see **Subheading 2.9.**). Perform a PCR amplification (in approx 30 µL volume) as follows:
 - 2' – 94°C
 - Thirty-one cycles of steps b–d:
 - 45" – 94°C
 - 45" – 56°C 31 cycles of steps b through d
 - 45" – 72°C
- (see **Note 28**).
- (b) In parallel, set up PCR reactions from the following control templates: (i) pJG4-5:Raf plasmid (diluted!); (ii) DNA prep isolated from the yeast from the control mating’s titer (**Subheading 3.3.3.**, **step 5**, and **Subheading 3.3.5.**, treated along with experimental

Table 2
Interpretation of Primary Isolates' Behavior

Phenotype		Explanation			Suggestion
-Leu growth	X-Gal color	Glu	Gal	Traditional	Optimistic
-	+	-	+	Very good sign	Work with those clones first
(+)	+	(+)	+	Bait is upregulated / mutated to a high background of transcriptional activation.	<ul style="list-style-type: none"> • <i>GAL1</i> promoter is slightly leaky • Both proteins are very stable • Interaction occurs with high affinity
-	+	-	-	Yeast mutation occurred that favors growth or transcriptional activation on galactose medium	<ul style="list-style-type: none"> Some bait-interactor combinations are known to preferentially activate <i>lacZ</i> versus <i>LEU2</i>, or vice versa Something really new
All other phenotype		Contamination/ plasmid rearrangements/ mutations			Trash

From ref. 19

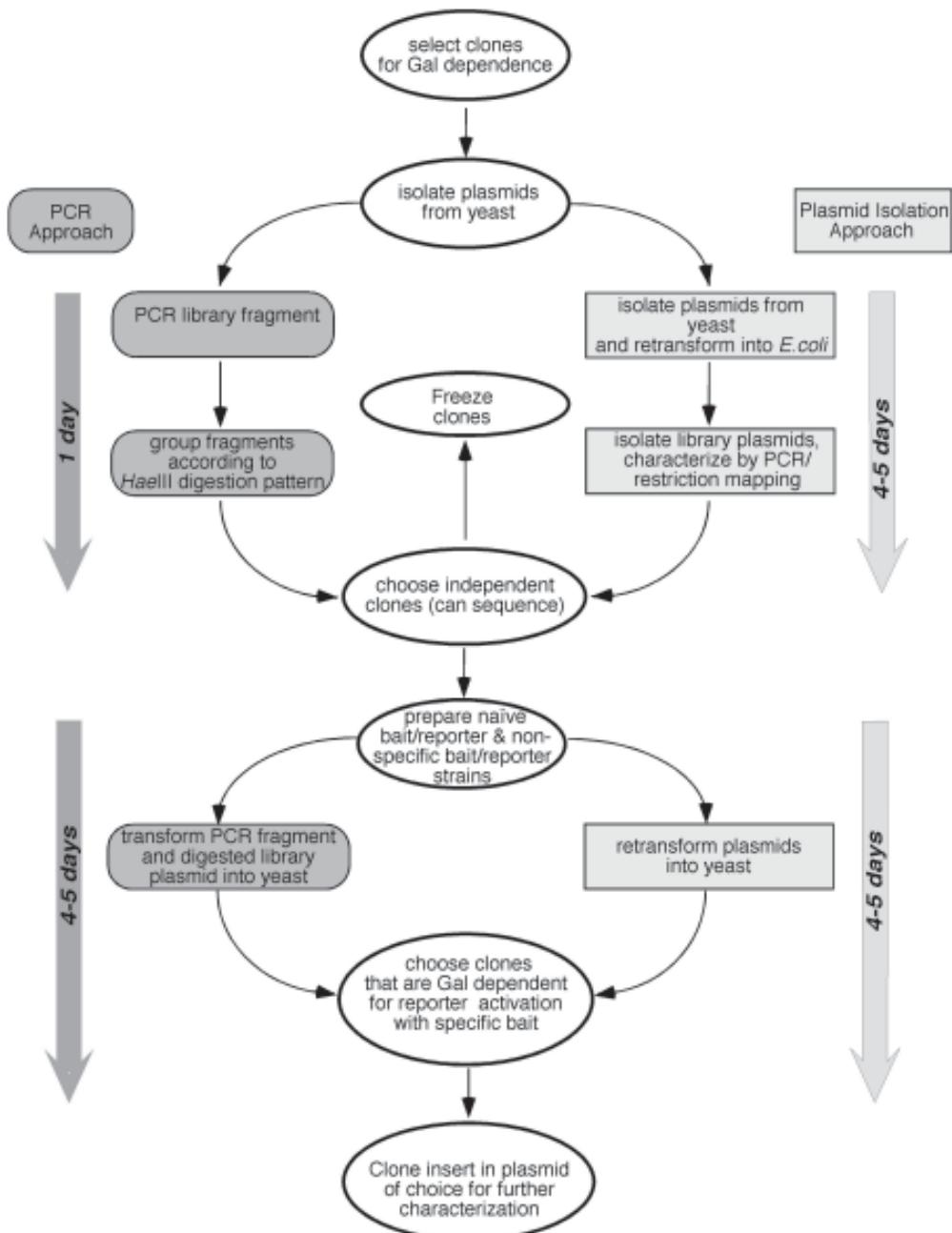


Fig. 4. Detailed library screening flow chart (modified from ref. 19). See text for details.

clones as above); (iii) and the same amounts of (i) and (ii), mixed together. For analysis of results, see **Table 3**.

3. Take 10 μ L of the PCR product for the *Hae* III digestion (below), and run out the remainder of the PCR reaction (about 20 μ L) on a 0.7% agarose gel. Identify fragments that appear to be of the same size; *Hae* III digests of these fragments should be run side-by-side. Put gel in a refrigerator until you are ready to isolate fragments (see Note 29).

Table 3
Suggested Polymerase Chain Reactions (PCRs)
and Interpretation of Results

Template			Possible outcomes				
Plasmid prep*	Control yeast **	Test clones	–	+	+	+	
✓			–	+	+	+	
	✓		–	–	–	–	+/–
✓	✓		–	+	–	–	+/–
	✓		–	–	–	–	+/–
Interpretation	Bad mastermix/ wrong settings/ faulty amplifier		Not enough template	Lysed yeast inhibited PCR	Uneven template load/digestion		
Recommendation	Double-check/ Repeat		Add more template/ improve lysis	Add less template	Adjust template load/re-PCR from obtained bands		

* JG4-5 - Raf1 is a suitable template; be sure to use as control library plasmid (with the insert) of the same type as you have your library.

** PCR from the empty vector yields a product of approx 130 bp for JG4-5 (FP1 and FP2 primers); Note, double bands may be NOT observed on the PCR done on the mixture of the plasmid prep with yeast, since smaller product from the empty vector may outcompete the bigger one from the plasmid. (From **ref. 19**.)

4. Perform a restriction digest of 10 µL of the PCR product with *Hae* III in a total volume of 20 µL. Rearrange the loading order according to the results obtained with nondigested PCR, and load the digestion products on a 1.5% agarose gel. Run out the DNA a sufficient distance to get good resolution of DNA products in the 200- to 1000-bp size range. This will generally yield distinctive and unambiguous groups of inserts, confirming whether multiple isolates of a small number of cDNAs have been obtained.
5. Purify fragments from the agarose gel using standard molecular biology techniques. In cases where a very large number of isolates representing a small number of cDNA classes have been obtained, the investigator may choose to directly sequence the PCR product (see **Note 30**).

3.4.1.2. CONFIRMATION OF INTERACTION PHENOTYPE USING PCR PRODUCTS

The next step is to determine whether isolated cDNAs reproduce interaction phenotypes specifically with the bait(s) of interest, and to exclude library-encoded cDNAs that interact with the baits in a nonspecific manner, and clones isolated because of mutations in the initial bait strain that result in growth and transcriptional activation nonspecifically. This can be done using a PCR-recombination approach (derived from **ref. 22**) in a single step, after which confirmed specific positive clones can be worked up through conventional plasmid purification. In **steps 1–3** below, the quality of the prepared reagents is assessed; in **steps 4–7**, specificity of the isolated clones is examined by comparison of their interaction with the original and an unrelated (pEG202-Ras) baits:

1. Digest an empty library plasmid with two enzymes producing incompatible ends in the polylinker region (e.g., *Eco*RI and *Xba*I) (see **Note 31**).

2. Prepare three separate transformations of SKY48 containing pEG202-Ras + pMW109 (*see Note 19*), with: (a) digested library plasmid; (b) digested library plasmid (50–100 ng) and control PCR product (0.5–1 µg) (from pJG4-5:Raf plasmid, *see Subheading 3.4.1.1., step 2(b)*); (c) uncut library plasmid. Save the digested library plasmid and the control PCR product for **step 4** (*see Note 32*).
3. Plate the transformations on Glu/CM –Ura, –His, –Trp dropout plates and incubate until colonies grow (2–3 d). Store at 4°C (*see Note 33*).

When transformed together, the PCR-amplified cDNA fragment from prey-control PCR product and the digested library plasmid will undergo homologous recombination *in vivo* in up to 97% of the transformants that acquired both vector and insert, as in (22). This is due to the identity between the cDNA PCR fragment and the plasmid at the priming sites. If transformation efficiency in (b) is better than in (a) by 5- to 20-fold, it is acceptable to proceed to the next steps. (c) Is a positive control for the transformation.

4. Using same ratios as in **step 2b** above, transform digested library vector in combination with selected PCR products (again, include positive control[s] from **step 2** above) to the following:
 - a. pMW103-Bait + pMW109
 - b. pEG202-Ras + pMW109
 Plate each transformation mix on Glu/CM –Ura, –His, –Trp dropout plates and incubate until colonies grow (2–3 d).
5. Prepare master plate(s) on Glu/CM –Ura, –His, –Trp. For each library plasmid being tested, include at least 10 colonies from **step 4** above (each of a and b).
6. Test for coloration and for auxotrophic requirements exactly as described for first confirmation of positive interactions, **Subheading 3.3.5**. True positives should show an interaction phenotype with (a), but not with (b). Clones transformed with control PCR product (Raf) will provide both positive and negative controls: (a) should be negative, whereas (b) should be positive when assayed for both color and growth on the corresponding plates.
7. Proceed with sequencing and biological characterization. Most often, PCR provides an ample source of DNA for all subsequent cloning. If needed, transform selected positives into *E. coli* by electroporation, using 1–2 µL of the DNA prep from yeast (*see Subheading 3.4.1.1., step 1*), and isolate plasmid DNA from Amp^R colonies (*see Note 34*).

3.4.2. Plasmid Isolation Approach: Isolation of Plasmids, Transfer to Bacteria

The following option is suggested as an alternative to the basic protocol in case PCR technology is not readily available for use, or in case of failure to obtain a specific PCR product using the library vector primers (as some investigators have difficulty with this procedure). It is based on the extraction of the DNA from yeast and transformation in *E. coli*, followed by plasmid isolation and plasmid retransformation into yeast. Some companies (e.g., Qbiogene, <http://www.qbiogene.com/services/two-hybrid.html>) provide a service in which they will isolate plasmid from the yeast cells, transform, and amplify the plasmid in *E. coli* to produce a sequencing template.

1. Isolate plasmid DNA from yeast (as in **Subheading 3.4.1.1., step 1**) and transform by electroporating *E. coli* strain DH5a. Select on medium containing ampicillin, because only bacteria that have taken up a library plasmid will grow (*see Note 34*).
2. Select at least two bacterial clones for each yeast clone, and prepare a small quantity of plasmid DNA from each bacterial clone (*see Note 35*).

3. Follow **Subheading 3.4.1.2.** from **step 4** to the end essentially as described, except transform with purified library plasmids, instead of mixture of PCR product + digested library vector.

3.5. Reiterative Scale-Up

One approach to elaborating a protein network is to perform reiterative interactor hunts (e.g., [23]). Such hunts can start with one or several different baits known to be involved in the biology under study. Subsequent interactor hunts can then be performed using the newly isolated proteins as baits. Such a protein “interaction walk” can be performed using standard protocols like those described earlier in this chapter. If the goal is to elaborate a large protein network that may require many interactor hunts, the rate-limiting step can become subcloning and characterizing new baits. This process can be streamlined by making new bait-expressing plasmids from newly isolated library clones by PCR and *in vivo* recombination. In this approach, a single set of primers is used that have homology to the library vector immediately upstream and downstream of the cDNA insertion site and that also have 5' tails that are homologous to the bait vector. Amplification of library clones with these primers results in a product that can be co-transformed along with a linearized bait vector. The PCR product will be recombined into the bait vector by homologous recombination in the yeast, as described in **Subheading 3.5.1.** The resulting yeast strain can be used directly in an interactor hunt by mating with an aliquot of frozen pretransformed library strain, as described in **Subheading 3.3.3.**

Before the new bait strain is used, it should be tested to ensure that it expresses the new bait. The most efficient way to accomplish this is to mate the new bait strain (expressing bait B) with a strain expressing the original protein A (previously used as a bait) as an activation domain fusion (AD-A). This will require subcloning the original bait into the library plasmid. Successful interaction can streamline the approach by dispensing with some of the bait characterization steps, e.g., Western assay. For each subsequent iteration of the protein walk, the new bait strain can be tested by mating it with strains expressing the previously isolated library clones, as shown in **Fig. 5**.

An added benefit to this approach is that each interaction will get tested twice, each time with the two proteins expressed with different fusion moieties. However, there are many documented cases in which a two-hybrid interaction is not detectable when the DNA-binding domain (DBD) and AD are swapped, especially when full-length proteins are used.

3.5.1. Making and Testing New Bait Plasmids by Recombination Cloning

1. Construct a pJG4-5 derivative expressing an activation domain fused version of a protein that will interact with the new bait to be constructed. For example, if protein A was used as a bait in an interactor hunt to isolate the cDNA for protein B (which is now going to be made into a bait), construct a plasmid for expressing an AD fused version of (Note that this is necessary only for the first step in a protein interaction walk; for subsequent steps, the AD version of the previous bait will already be available from the previous hunt [see **Fig. 5**.])
2. Transform yeast strain SKY473 with the new AD fusion vector, and separately with pJG4-5. Select transformants on Glu/CM-Trp.
3. Perform PCR amplification using 1–2 μ L of yeast DNA prep (**Subheading 3.4.1.1, step 1**). Use a high-fidelity heat-stable polymerase, such as Vent (New England Biolabs)

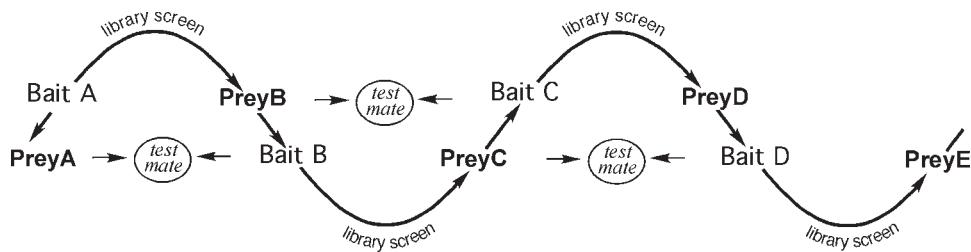


Fig. 5. Flowchart for reiterative screening (modified from <http://www.fccc.edu/research/labs/golemis/interactiontrapwork.html>). See text for details.

or Pfu (Stratagene), and the primers designed to transfer insert from pJG4-5 into pEG202 (Subheading 2.9.) (see Note 36).

4. Digest pMW103 to completion with EcoRI and XhoI. Dilute to 20 ng/μL.
5. Transform SKY48 containing one of the *URA3 lacZ* reporter plasmids (e.g., pMW111, pMW109, or pMW112) with 200 ng of the cut bait vector along with 20–200 ng of the PCR product from step 1. Set up separate transformations with an equal amount of the cut vector alone (with no PCR product), and with uncut vector as controls (see Subheading 3.4.1.2., step 3, and Notes 31 and 32). Select transformants on Glu/CM-Ura-His plates.
6. To assure compatibility with the 96-well plate format, put the YPD plate on a 6 × 8 grid (e.g., tape to the bottom one-half of the 200 mL tips insert).
7. Resuspend up to seven bait colonies in approx 30 mL of sterile water. From one suspension, put a small drop (approx 3 mL) on the medium surface over each of the spots in a six-spot row. Repeat for the remaining suspensions. A 7 × 6 rectangle is formed (see Note 37).
8. Resuspend up to five prey colonies (from step 2 above) in approx 30 mL of sterile water. From one suspension, put a small drop (approx 3 mL) on the medium surface over each of the spots in an eight-spot row (perpendicular to that in step 2 above). Of these, one drop will be placed on a bare surface, and the remaining seven will overlap drops of the bait strains. Repeat for the remaining suspensions. At this point, an 8 × 6 rectangle is formed (with one corner spot empty) (see Note 37). Incubate for 12–15 h to allow the two strains to mate.
9. Using replica plate technique, test for β-galactosidase activity and for leucine requirement exactly as described in Subheading 3.3.5. Only the mated diploid yeast at the intersections of the two sets will grow on these plates (a 7 × 5 rectangle should be formed on Glu/CM-Ura-His-Trp plates). Most of the clones of the new bait strain should produce an interaction phenotype (galactose-dependent Leu+ lacZ+) when crossed with the strain expressing the interacting AD fusion protein, but not with the strain containing the library vector alone (see Note 38). Bait-strain colonies that result in the correct positive interaction phenotype can also be characterized as described in Subheading 3.2. and can be used in subsequent interactor hunts by mating with frozen pretransformed library as described in Subheading 3.3.3.

3.6. Array Screening

Small arrays have been introduced in Subheading 3.3.5. for the confirmation of screening positives. However, arrays can also be used as a general tool for new screening-defined subsets of proteins. When combined with bioinformatic selection of candidate interactors, arrays are powerful tools to identify new interactions. For example, we know that SH3 domains usually bind to proline-rich sequences. That is, we can

predict all proline-rich candidate interactors for SH3 domains from complete genome sequences. These open reading frames can be amplified from cDNA libraries or genome sequences and cloned into prey vectors by recombination. Instead of screening a complex library, it may be sufficient to screen such a defined subset of preys. Prey sets can be selected based on a variety of criteria, e.g., sequence/homology, subcellular location, known or presumed function, and so on. Ultimately, whole genomes can be cloned into prey vectors, arrayed, and screened as has been done with the yeast and *Drosophila* genomes (14–16).

The main advantage of an array screen is its parallel nature, which allows one to compare preys as well as baits. Different preys have different affinities to any given bait; this will be indicated by different colony sizes on selective plates, or different color intensities when lacZ assays are performed. Furthermore, different baits can be compared when they are screened against the same prey array: some will generate only a few or no positives, while some will cause even activation. A certain group of baits will cause “random” activation, i.e., a relatively large number of positives with no or almost no background activation. In a “conventional” pretest for a library screen as described in **Subheading 3.2.1.**, such baits may not stand out because they are not tested against enough preys.

3.6.1. Array Screening Protocol

Array screening is simply an organized way of two-hybrid testing as described in **Subheading 3.3.5.**, although it may involve prey clones that have never been tested before. The general procedure and the reagents are the same as described above; hence, we provide only an abbreviated protocol below. More detailed protocols can be found in (13,24).

3.6.1.1. PREPARE PREY ARRAY

1. Clone a set of preys as described in **Subheading 3.1.** (cloning bait- and prey-encoding DNA fragments can be done in identical ways).
2. Test them for activation and/or expression as described in **Subheading 3.2.**
3. Place yeast colonies on microtiter-plate-sized agar plates (e.g., Nunc omnitrays) in a format corresponding to the 96 wells. Include at least one colony with empty prey vector (e.g., pJG4-5) (see **Note 39**).

3.6.1.2. PREPARE BAIT ARRAY AND SCREEN PREY ARRAY

1. Fill a sterile omnitrays with 10 mL of bait yeast suspension (e.g., pMW103 in overnight SKY48 culture).
2. Use self-made or commercial replica plater (frogger) to transfer 96 samples of bait suspension to omnitrays filled with YPD agar (you may place a 96-well plate underneath it for proper alignment). For larger numbers of assays, a commercial replicator or robotic device is recommended (e.g., Beckman Biomek, Genetix Q-bot).
3. Use sterile replica plater to transfer a small amount of prey yeast from the prey array on top of the bait spots, i.e., each bait spot should receive a different prey clone.
4. Mate bait and prey yeast overnight, or up to 2 d.
5. Transfer mated yeast with sterile replica plater to new omnitrays filled with –His, –Trp agar. Grow yeast for 2 d at 30°C. The bait plasmid pMW103 is selected on –His medium, the prey plasmid pJG4-5 on –Trp. This step selects for diploid cells and ensures that all colonies have mated properly. In the case of imprecise replicating, it may happen that not enough cells are transferred, resulting in inefficient mating.

6. Transfer successfully mated diploid to omnitrays filled with –His, –Trp, –Leu agar, using the sterile replica plater. Grow yeast at 30°C for at least 3 and up to 9 d. When preys of known identity are used, isolation of prey plasmids is not necessary.
7. (Optional) Proceed to the second confirmation of positives, as described in **Subheading 3.4.2.**

If the array is screened with many baits, specificity of interaction should be not a concern at this step. Most importantly, initial positives must show reproducible phenotype (i.e., activation of at least one reporter should be again shown).

3.7. Data Analysis and Bioinformatics Aspects

With the advent of large-scale interaction screening, systematic collection of interaction data in a database became essential. In examining such databases, besides the obvious goal of identifying which other interactors are known for the protein of interest, a number of other questions generally arise. Do the interacting proteins for the protein of interest have homologs in other species, and do these homologs interact as well? With what additional proteins do the interactors interact? Is an interaction chain containing several connected proteins part of a single protein complex, or an ordered signaling pathway?

Integration and visualization of interactions from different sources becomes increasingly challenging. At the time of this writing, hardly any software deals with these issues satisfactorily, although a number of academic groups and companies are actively working on solutions (25). We recommend monitoring of the Web sites and databases that are dedicated to these issues (such as www.two-hybrid.com; *see also* Chapter 78).

Interestingly, the increasing numbers of interactions recorded in the public domain not only has caused a tremendous need for visualization tools, but has also induced a surprisingly large number of studies aiming at the analysis of interaction networks. For example, it has been shown that protein interaction networks appear to have a scale-free topology, i.e., they have a few hubs of highly connected proteins and many less-well-connected proteins (26). Hubs in such networks are indeed more essential to a cell than peripheral proteins when mutated, as would be predicted by their central position (27). Interaction networks can be used to predict the function of previously uncharacterized proteins (Chapter 54 and **ref. 28**), because interacting proteins usually have related activities. Similarly, interactions can be used to predict protein complexes based on local clusters (“cliques”) of interactions (29). Two-hybrid interactions can also be used to predict interaction domains. When random libraries are used for screening, overlapping cDNA or genomics DNA fragments may inherently narrow down interacting fragments. However, even large-scale screens that used full-length clones contain enough information to select pairs of proteins that share common domains, subsequently allowing domain identification by sequence comparison (30). Finally, with a growing database of interactions, comparative “interactomics” becomes possible. We can now analyze the evolution of protein interactions and even of whole interaction networks (31). Combined with data from structural proteomics, this should allow us to understand which amino acids are important for the structure and function of protein complexes. Eventually, this may even be exploited for the prediction of inhibitors or other compounds that affect protein function.

However, we are not quite there yet. We need more data, better software and databases, and most importantly, better integration of various data sources. Eventually, this

will lead to a merger of many different areas of molecular biology into what is now called systems biology.

4. Notes

1. Standard molecular biology techniques or alternative cloning strategies (e.g., *in vivo* recombination [32,33]) can be used. In any case, it is a good idea to include a translational stop sequence at the carboxy-terminal end of the bait sequence. It is important to keep in mind that the assay depends on the ability of the bait to enter the nucleus, and requires the bait to be a transcriptional nonactivator. Hence, obvious membrane localization motifs or transcriptional activation domains should be removed. Using two-hybrid screens to find associating partners for proteins that are normally extracellular, even though they have worked in a few cases, should be regarded as extremely high risk.
2. A number of modified versions of the plasmid exist which contain additional cloning sites, altered reading frame, and alternate antibiotic resistance markers (see <http://www.fccc.edu/research/labs/golemis/interactiontrapinwork.html> for details).
3. It is important to use a fresh (thawed from -70°C and streaked to single colony less than approx 7 d previously) colony and maintain sterile conditions throughout all subsequent procedures.
4. A number of commercially available yeast transformation kits (e.g., BD Biosciences Clontech, #K1606-1; Zymo Research #T2001) or a simple chemical protocol (e.g., as in [19]) can be used.
5. An efficient transformation would yield approx 10^4 transformants per microgram of DNA (when two plasmids are being simultaneously transformed). Therefore, this experiment also provides a good chance to assess transformation efficiency, which will be of much higher importance by the time of library transformation. Thus, if only a very small number of colonies are obtained, or colonies are not apparent within 3 to 4 d, it would imply that transformation is for some reasons very inefficient, and results obtained in characterization experiments may not be typical. In this case, all solutions, media, and conditions must be double-checked or prepared fresh, and transformation be repeated. If very few transformants containing the bait plasmid appear (compared to the controls), or yeast expressing the bait protein grow noticeably more poorly than control yeast, or if the colony population appears much more heterogeneous than a control (e.g., presents a mix of large and small colonies), this would suggest that the bait protein is somewhat toxic to the yeast (see **Table 1** for suggested modifications).
6. This is important, because for some baits, protein expression level is heterogeneous between independent colonies, with accompanying heterogeneity of apparent ability to activate transcription of the two reporters.
7. A replicator (frogger) for the transfer of multiple colonies can be purchased or easily homemade; it is important that all of the spokes have a flat surface, and that spoke ends are level. A metal frogger can be sterilized by autoclaving or by alcohol/flaming. A plastic replicator must be cut in half to fit to a standard 90-mm Petri plate; it can be sterilized by autoclaving or rinsing with alcohol. The replicator should have 48 spokes in a 6×8 configuration. When making prints on a plate, dip the replicator in the wells of the microtiter plate, then put it on the surface of the solidified medium. Tilt slightly in circular movement, then lift replicator and put it back in the microtiter plate (keep the correct orientation!). Make sure all the drops left on the surface are of approximately the same size. If only one or two drops are missing, it is easy to correct by dropping approx 3 μL of yeast suspension on the missing spots from the corresponding wells. If many drops are missing, make sure that all the spokes of the replicator are in good contact with liquid in the

- microtiter plate (it may be necessary to cut off the side protrusions on the edge spokes of the plastic replicator) and redo the whole plate. Continue replicating by shuttling back and forth between microtiter and media plates. Let the liquid absorb into the agar before putting the plates upside down in the incubator.
- 8. The technique described here is much more sensitive than a standard X-Gal plate assay, can be done within 24 h of plating on appropriate medium, and is generally preferred in high-throughput analysis.
 - 9. Either of the bait activation controls can be used as a positive control for bait protein expression.
 - 10. Many fusion proteins exhibit sharp decreases in detectable levels of protein with the onset of stationary phase. Therefore, use of the saturated cultures is not recommended for this assay.
 - 11. Samples can be frozen at -70°C for subsequent use; such samples are stable for at least 4–6 mo.
 - 12. In addition to the simplified technique described here, a number of more elaborate (and time-consuming) protocols are available (e.g., see Clontech's Yeast Protocols Handbook, available at <http://www.clontech.com/clontech/Manuals/PDF/PT3024-1.pdf>).
 - 13. A good library transformation efficiency should be approx 10⁵ transformants per microgram of library DNA (for a single transformation). Transformation of yeast in multiple small aliquots in parallel helps reduces the likelihood of contamination; further, it frequently results in significantly better transformation efficiency than that obtained by using larger volumes in a smaller number of tubes. Finally, do not use excess transforming library DNA per aliquot of competent yeast cells (more than approx 1 µg) because cells may take up multiple library plasmids, complicating subsequent analysis. Under the conditions described here, less than 10% of yeast will contain two or more library plasmids.
 - 14. Keeping the glass beads on the lids while incubating the plates would facilitate the subsequent harvest of the library transformants (**Subheading 3.3.2**). Contamination is much less likely to occur on the glass beads than on the plates themselves.
 - 15. Thoroughly inspect the plates visually prior to the collecting transformants. If visible molds or other contaminants are observed on the plates, carefully excise them and a region around them using a sterile razor blade prior to beginning harvest of library transformants.
 - 16. This technique also minimizes the time the plates are open, and thus helps avoid contamination from airborne molds and bacteria.
 - 17. It is more important to ensure the same wash-off rate for all plates, than to collect as much yeast as possible (about one-third of the yeast slurry will be left on the plates). However, a second wash can greatly improve the yield (after the first wash, add 10 more mL of water, shake again, and transfer the slurry to the next unwashed plate; be careful to avoid contamination).
 - 18. Optionally, the 24 × 24 cm plates can be reused many times after removing the remaining agar, washing, and alcohol/ultraviolet sterilization. As these plates are quite expensive, this is a useful point of economy.
 - 19. The bait and reporter plasmids should have been transformed into the yeast less than approx 7–10 d prior to mating with pretransformed library or transforming with library plasmids.
 - 20. In general, for yeast frozen for less than 1 yr, viability will be greater than 90%. Refreezing a thawed aliquot results in the loss of viability.
 - 21. Titering can also be done later, in parallel with **Subheading 3.3.4**.
 - 22. If contamination occurred at an earlier step and results in the growth of many (greater than 500) colonies per plate, this will interfere with screening. In the event of bacterial contamination, the hunt may be salvaged by adding tetracycline (15 µg/mL) to the selective

- plates and repeating library induction/plating. If contamination is fungal, there is little to be done; mating (or even library transformation) must be repeated.
23. Compare selection plates seeded with lower and higher densities. The number of colonies should be roughly proportional to the seeding density, and there should be no background growth. If disproportionately more colonies (or a lawn) appear on the more densely seeded plates, this is background due to crossfeeding. In this case, a higher number of plates seeded at lower density should be used. Calculate how many plates at acceptable cell density are necessary for full representation of the desired number of diploids, and if needed, repeat induction and plating from another frozen mating aliquot.
 24. If colonies do not arise within the first week after plating, colonies appearing at later time points are not likely to represent bona fide positives. True interactors tend to come up in a window of time specific for a given bait, with false positives clustering at a different time point: hence, pregrouping by date of growth facilitates the decision of which clones to analyze first.
 25. The number of candidate colonies to pick and characterize should be based on the number of cDNA-independent false positives that arise on the same selection plates for the control mating. The higher the frequency of false positives, the more colonies should be picked to find rare true positives. Since the frequency of true positives will be unknown at this step, the goal will be to pick through all of the false positives that are expected in the number of library transformants being screened. For example, if the number of library transformants was 10^6 , the goal will be to pick through the number of false positives expected in 10^6 diploids. If the cDNA-independent false-positive frequency is 1 Leu+ colony in 10^4 cfu plated, it will be necessary to pick at least 100 Leu+ colonies to find a true positive that exists at a frequency of 1 in 10^6 .
 26. In general, test plates for auxotrophic reporter characterization lacking only leucine would automatically keep selective pressure for the presence of the prey and the corresponding bait plasmids. Using plates with fewer dropped-out components would slightly accelerate the growth, and the potential loss of other plasmids would not influence the results of the assay on these plates. At the investigator's discretion, however, -Leu plates can be substituted for -Ura-His-Trp-Leu.
 27. In some cases no positives are obtained from library screens. Reasons for this might include inappropriate library source; an inadequate number of screened colonies (<500,000); a bait that in spite of production at high levels is nevertheless incorrectly folded or post-translationally modified; or alternatively, a bait that does not interact with its partners with a sufficiently high affinity to be detected. Also, be aware of such simple explanations as a wrong batch of plates. In such cases, it may be worth trying screens again with a different variant of bait, screening strain, and/or library, although success is not guaranteed. It is rarely, if ever, profitable to continue to rescreen the same bait/strain/library combination through >3–5 million primary transformants.
 28. Modified versions of this protocol with extended elongation times were also found to work; the variant given above has amplified fragments of as much as 1.8 kb in pretty fair quantity.
 29. Sometimes, a single yeast cell will contain two or more different library plasmids. If this happens, it will be immediately revealed by PCR. In this case, two bands can be separately isolated from the gel, and reamplified. Also, after bacterial transformation, an increased number of clones should be checked to avoid the loss of the “real” interactor.
 30. Note, only the forward primer, FP1, works well in sequencing of PCR fragments; the reverse primer will work in sequencing only from purified plasmid. In general, the TA-rich nature of the ADH terminator sequences downstream of the polylinker in the pJG4-5 vector makes it difficult to design high-quality primers in this region.

31. Gel analysis produces little information on the completeness of digestion, since it is not possible to distinguish between plasmid species cut by one and two enzymes. Purification of the digested plasmid is not necessary.
32. This control experiment is an indicator of the degree of digestion of the library plasmid. The background level of colonies transformed with undigested empty library plasmid (a) should be minimal. In case the background is high, make sure that the digestion of the empty library plasmid is full and not partial by increasing the digestion incubation time or the restriction enzyme concentration.
33. Clones obtained in (b) should be positive. The proportion of clones that are correct can be assessed by replica-plating 12–24 clones to confirm their phenotype (as in **Subheading 3.3.5.**). Normally, it should be between 85 and 95%.
34. If using specialized LexA-fusion bait plasmid (they are all Ap^R) in combination with Ap^R library plasmid, it will be necessary to select specifically for transformants containing a library plasmid by the ability of the yeast *TRP1* gene to complement the *E. coli* *trpC* mutation (**19**).
35. Characterization of these DNA preps by restriction digest/PCR analysis may reduce the number of yeast transformations in the subsequent steps.
36. Primers to transfer a cDNA insert to alternative bait plasmids can also be designed (e.g., see **ref. 19** for primers to transfer cDNA from pJG4-5 to pNLex).
37. To archive selected clones, streak the rest of the resuspended colonies onto corresponding master plates (a Glu/CM–Ura–His for baits, a Glu/CM –Trp for preys). Keep track of which rows came from which colonies so that they can be used after testing.
38. About 10% of the new bait strain colonies may fail to produce the expected interaction phenotype. Some of these result from bait vectors that did not receive an insert, and others result from incorrect recombination events—for example, between the lacZ vector and the bait vector; the former class will result in Leu–lacZ– phenotype when mated, whereas the later class can lead to galactose-independent lacZ+ yeast.
39. This protocol was originally developed with a different combination of strains, plasmids, and reporters, namely the bait vector pOBD2 (compatible with pAS2), the prey vector pOAD (compatible with pACTII), and the strain PJ694-a/α (**24,34**). It should be adaptable to any other protocol that allows the selection of bait and prey plasmids in some yeast strain that carries appropriate marker and reporter genes. Time of mating and auxotrophic selection may need to be adjusted with different components.

References

1. Johnsson, N. and Varshavsky, A. (1994) Ubiquitin-assisted dissection of protein transport across membranes. *EMBO J.* **13**, 2686–2698.
2. Drees, B. L. (1999) Progress and variations in two-hybrid and three-hybrid technologies. *Curr. Opin. Chem. Biol.* **3**, 64–70.
3. Frederickson, R. M. (1998) Macromolecular matchmaking: advances in two-hybrid and related technologies. *Curr. Opin. Biotechnol.* **9**, 90–96.
4. Golemis, E. A. and Khazak, V. (1997) Alternative yeast two-hybrid systems. The interaction trap and interaction mating. *Methods Mol. Biol.* **63**, 197–218.
5. Fields, S. and Song, O. (1989) A novel genetic system to detect protein-protein interactions. *Nature* **340**, 245–246.
6. Schwartz, H., Alvares, C. P., White, M. B., and Fields, S. (1998) Mutation detection by a two-hybrid assay. *Hum. Mol. Genet.* **7**, 1029–1032.
7. Vidal, M. and Legrain, P. (1999) Yeast forward and reverse 'n'-hybrid systems. *Nucleic Acids Res.* **27**, 919–929.

8. Vidal, M. and Endoh, H. (1999) Prospects for drug screening using the reverse two-hybrid system. *Trends Biotechnol.* **17**, 374–381.
9. SenGupta, D. J., Zhang, B., Kraemer, B., Pochart, P., Fields, S., and Wickens, M. (1996) A three-hybrid system to detect RNA-protein interactions in vivo. *Proc. Natl. Acad. Sci. USA* **94**, 8496–8501.
10. Estojak, J., Brent, R., and Golemis, E. A. (1995) Correlation of two-hybrid affinity data with in vitro measurements. *Mol. Cell Biol.* **15**, 5820–5829.
11. Rain, J. C., Selig, L., De Reuse, H., et al. (2001) The protein-protein interaction map of *Helicobacter pylori*. *Nature* **409**, 211–215.
12. Raquet, X., Eckert, J. H., Muller, S., and Johnsson, N. (2001) Detection of altered protein conformations in living cells. *J. Mol. Biol.* **305**, 927–938.
13. Cagney, G., Uetz, P., and Fields, S. (2001) Two-hybrid analysis of the *Saccharomyces cerevisiae* 26S proteasome. *Physiol. Genomics* **7**, 27–34.
14. Giot, L., Bader, J. S., Brouwer, C., et al. (2003) A protein interaction map of *Drosophila melanogaster*. *Science* **302**, 1727–1736.
15. Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA* **98**, 4569–4574.
16. Uetz, P., Giot, L., Cagney, G., et al. (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623–627.
17. Stagljar, I. and Fields, S. (2002) Analysis of membrane protein interactions using yeast-based technologies. *Trends Biochem. Sci.* **27**, 559–563.
18. Serebriiskii, I. G., Mitina, O., Pugacheva, E. N., et al. (2002) Detection of peptides, proteins, and drugs that selectively interact with protein targets. *Genome Res.* **12**, 1785–1791.
19. Serebriiskii, I., Toby, G., Finley, R. L., and Golemis, E. A. (2001) Genomic analysis utilizing the yeast two-hybrid system. In: (Starkey, M., ed.) *Chimeric Genes and Proteins*, , Humana, Totowa, NJ: 415–454.
20. Duttweiler, H. M. (1996) A highly sensitive and non-lethal beta-galactosidase plate assay for yeast. *Trends Genet.* **12**, 340–341.
21. Finley, R. and Brent, R. (1994) Interaction mating reveals binary and ternary connections between *Drosophila* cell cycle regulators. *Proc. Natl. Acad. Sci. USA* **91**, 12,980–12,984.
22. Petermann, R., Mossier, B. M., Aryee, D. N., and Kovar, H. (1998) A recombination based method to rapidly assess specificity of two- hybrid clones in yeast. *Nucleic Acids Res.* **26**, 2252–2253.
23. Fromont-Racine, M., Mayes, A. E., Brunet-Simon, A., et al. (2000) Genome-wide protein interaction screens reveal functional networks involving Sm-like proteins. *Yeast* **17**, 95–110.
24. Cagney, G., Uetz, P., and Fields, S. (2000) High-throughput screening for protein-protein interactions using two-hybrid assay. *Methods Enzymol.* **328**, 3–14.
25. Uetz, P., Ideker, T., and Schwikowski, B. (2002) Visualization and integration of protein-protein interactions. In: (Golemis, E., ed.) *Protein-Protein Interactions—A Molecular Cloning Manual*, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY: 682.
26. Goh, K. I., Oh, E., Jeong, H., Kahng, B., and Kim, D. (2002) Classification of scale-free networks. *Proc. Natl. Acad. Sci. USA* **99**, 12,583–12,588.
27. Jeong, H., Mason, S. P., Barabasi, A. L., and Oltvai, Z. N. (2001) Lethality and centrality in protein networks. *Nature* **411**, 41–42.
28. Schwikowski, B., Uetz, P., and Fields, S. (2000) A network of protein-protein interactions in yeast. *Nat. Biotechnol.* **18**, 1257–1261.
29. Bader, G. D. and Hogue, C. W. (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* **4**, 2.

30. Sprinzak, E. and Margalit, H. (2001) Correlated sequence-signatures as markers of protein-protein interaction. *J. Mol. Biol.*, **311**, 681–692.
31. Kelley, B. P., Sharan, R., Karp, R. M., et al. (2003) Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proc. Natl. Acad. Sci. USA* **100**, 11,394–11,399.
32. DeMarini, D. J., Creasy, C. L., Lu, Q., et al. (2001) Oligonucleotide-mediated, PCR-independent cloning by homologous recombination. *Biotechniques* **30**, 520–523.
33. Oldenburg, K. R., Vo, K. T., Michaelis, S., and Paddon, C. (1997) Recombination-mediated PCR-directed plasmid construction in vivo in yeast. *Nucleic Acids Res.* **25**, 451–452.
34. James, P., Halladay, J., and Craig, E. A. (1996) Genomic libraries and a host strain designed for highly efficient two-hybrid selection in yeast. *Genetics* **144**, 1425–1436.

Antibody-Affinity Purification to Detect Interacting Proteins

Sonia Navarro and Lucio Comai

1. Introduction

Affinity purification is a procedure based on the specific binding interactions between a ligand chemically coupled to a resin and a target molecule. A common application is the use of antibody as immobilized ligands. The purification of antigens by antibody-affinity chromatography is widely used to detect factors interacting with a protein of interest, and when coupled to mass spectrometry, it is a powerful approach for the identification of associated factors. In general, interacting proteins are isolated from cells expressing an epitope-tagged version of the protein of interest (1). Short peptides (tags) engineered at the amino- or carboxy-terminal end of proteins are quite useful for the purification of protein complexes through the use of tag-specific antibody covalently coupled to a solid matrix. The use of tagged proteins is particularly effective for the isolation of native protein complexes from yeast cells, where these proteins can be expressed at their natural level (2–4). On the other hand, it is difficult to express engineered proteins at physiological levels in mammalian cells. Because overexpression may alter the network of protein interactions, the purification by affinity to an engineered epitope tag may not provide a faithful representation of the factors that are associated with the endogenous protein. To overcome this potential problem, many investigators choose to employ an affinity purification procedure that uses an antibody that recognizes the endogenous protein. This approach has been valuable in the identification of protein complexes involved in gene transcription (5–6) and DNA repair (7).

The overall strategy involves the capture of an antigen and the associated proteins from a cell extract by a specific antibody covalently linked to a solid matrix, such as agarose or sepharose. The most critical parameters for a successful isolation of a protein complex by antibody affinity chromatography are: (1) the antibody must be able to recognize the antigen within its native complex; and (2) the experimental conditions used to lyse the cell must preserve the interaction between the protein under investigation (antigen) and the associated factors. The preparation of nuclear extract using mild conditions may not efficiently extract the proteins under examination, while conditions that are too stringent may permanently disrupt the molecular interactions between the antigen and the associated factors. This chapter describes a basic antibody-affinity

purification procedure for the identification of factors associated with a nuclear protein in mammalian cells. The final identification of associated proteins is achieved by mass spectrometry of proteolytically cleaved peptides. Ultimately, the interaction between the newly identified factor and the antigen must be confirmed by reciprocal co-immunoprecipitations and other biochemical assays (8,9).

2. Materials

2.1. Crosslinking of Antibody to Protein A-Agarose Resin

1. Poly-Prep Chromatography Column (BIORAD Cat. No. 731-1550).
2. Phosphate-buffered saline (PBS): 1.37 M NaCl, 27 mM KCl, 43 mM Na₂HPO₄, 14 mM KH₂PO₄, 5 mM MgCl₂.
3. Protein A-Agarose Gel (BIORAD Cat. No. 153-6153). Store as a 50% slurry.
4. 0.2 M Borate-NaOH (pH 8.6).
5. Crosslinking coupling buffer: 0.2 M triethylamine (pH 8.3).
6. Crosslinker: dimethylpimelimidate (DMP) powder (stored at -20°C). Bring to room temperature before using.
7. Crosslinking termination buffer: 0.2 M ethanolamine-HCl (pH 8.2).
8. Purified antibody (see Note 1).

2.2. Preparation of the Nuclear Extract

1. Monolayer- or suspension-cultured cells.
2. Glass Dounce homogenizer with type "B" pestle.
3. PBS: 1.37 M NaCl, 27 mM KCl, 43 mM Na₂HPO₄, 14 mM KH₂PO₄, 5 mM MgCl₂.
4. Buffer H: 10 mM Tris-HCl (pH 8.0), 10 mM KCl, 1.5 mM MgCl₂.
5. Buffer D*: 50 mM Tris-HCl (pH 8.0), 420 mM KCl, 5 mM MgCl₂, 100 mM EDTA (pH 8.0), 20% glycerol, 10% sucrose.
6. Protease inhibitors (PI): 1 mM dithiothreitol (DTT), 0.2 mM phenylmethyl-sulfonyl fluoride (PMSF), 2 mM sodium metabisulfite (MBS), 5 μ g/mL pepstatin A, 5 μ g/mL leupeptin, and 5 μ g/mL aprotinin. Store at 4°C or -20°C, according to manufacturer's specifications.

2.3. Purification of Antigen and Associated Factors by Antibody-Affinity Chromatography

1. Column 1: Protein A agarose. Store as a 50% slurry in buffer B plus 0.01 mM NaN₃.
2. Column 2: crosslinked antibody/Protein A-agarose resin. Store as a 50% slurry in buffer B containing 0.01 mM NaN₃.
3. Equipment for sodium dodecyl sulfate (SDS)-polyacrylamide gel electrophoresis (PAGE).
4. Microdialyzer (optional).
5. Buffer D*: 50 mM Tris-HCl (pH 8.0), 420 mM KCl, 5 mM MgCl₂, 100 mM EDTA (pH 8.0), 20% glycerol, and 10% sucrose.
6. Buffer B: 50 mM Tris-HCl (pH 8.0), 150 mM NaCl, 1 mM EDTA (pH 8.0), 10% glycerol, and 1 mM PMSF.
7. Buffer C: 50 mM Tris-HCl (pH 8.0), 500 mM LiCl, 1 mM EDTA (pH 8.0), 10% glycerol, 1 mM PMSF.
8. Buffer D: 10 mM PIPES (pH 7.4), 5 mM NaCl, 1 mM EDTA (pH 8.0), 10% glycerol, 1 mM PMSF.
9. Buffer E: 20 mM triethylamine and 10% glycerol.
10. Buffer F: 10 mM PIPES (pH 7.0), 5 mM NaCl, 0.1 mM EDTA (pH 8.0), 10% glycerol, 0.5 mM PMSF, and 1 mM DTT.
11. 2X protein sample buffer (2X SB): 62.5 mM Tris-HCl (pH 6.8), 10% glycerol, 3% SDS, 5% β -mercaptoethanol. This solution can be stored at room temperature.

3. Methods

The following protocols describe: (1) adsorption and chemical crosslinking of the antibody to the protein A-agarose resin; (2) preparation of the nuclear extract under conditions that do not disrupt the protein–protein interactions; (3) purification of antigen and associated factors by affinity chromatography on an antibody resin.

3.1. Crosslinking of Antibody to Protein A-Agarose Resin

1. To an empty poly prep chromatography column add 1X PBS and add 400 μ L protein A-agarose gel slurry. Wash column twice with 1X PBS.
2. Add purified antibody.
3. Incubate for 15 min at 4°C on nutator.
4. Collect the flow-through on ice and repeat **steps 2 and 3**.
5. Collect and quick-freeze the flow-through.
6. Wash column four times with 1X PBS.
7. Pre-equilibrate column three times with 0.2 M borate-NaOH (pH 8.6).
8. Leave approx 1 mL of 0.2 M borate-NaOH (pH 8.6) above resin.
9. In a separate tube, prepare fresh crosslinking reagent: to 2.0 mL 0.2 M triethylamine (pH 8.3) add 15.5 mg of dimethylpimelimidate and vortex to mix well. Immediately add crosslinking reagent to column containing 1 mL of 0.2 M borate-NaOH (pH 8.6). Check pH, since coupling with DMP must be performed at pH above 8.3.
10. Incubate 30 min at room temperature.
11. Drain the column and terminate reaction with 10 mL 0.2 M ethanolamine-HCl (pH 8.2). Incubate for 5 min at room temperature.
12. Drain the column and add 10 mL 0.2 M ethanolamine-HCl (pH 8.2). Incubate for 1 h at room temperature.
13. Wash column twice with 1X PBS and store column in 1X PBS.
14. To check the effectiveness of crosslinking, add 2X SB to an aliquot of slurry before and after crosslinking and run on a 10% SDS-PAGE.

3.2. Preparation of the Nuclear Extract

During this procedure, it is critical to minimize proteolytic activity by keeping the cell lysate cold (<8°C) at all times. A cocktail of protease inhibitors is always included in the cell lysis buffer to minimize proteolysis.

1. Place tissue-culture dishes on ice. Remove the culture media.
2. Add 5 mL of 1X PBS/5 mM MgCl₂ to dish to collect cells by gently scraping with disposable cell scraper into ice-cold Falcon tube.
3. Centrifuge at 750g for 5 min at 4°C.
4. Repeat wash twice with 1X PBS/5 mM MgCl₂.
5. To the pelleted cells, add five packed-cell volumes (PCV) of buffer H supplemented with PI.
6. Resuspend cells by gently pipeting up and down and transfer to glass Dounce homogenizer.
7. Incubate on ice for 20 min, then homogenize with twenty up-and-down strokes using a type “B” pestle.
8. Transfer the homogenate to an ice-cold 15-mL Falcon tube and centrifuge at 750g for 5 min at 4°C.
9. Carefully transfer the supernatant to an ice-cold 12-mL Falcon tube and quick-freeze post-nuclear supernatant in liquid nitrogen.
10. To the pelleted nuclei, add 5 mL buffer H supplemented with PI. Resuspend nuclei by gently inverting to prevent breakage of nuclei.

11. Centrifuge at 750g for 5 min at 4°C and discard supernatant.
12. Add five packed-cell volumes (PCV) of buffer D* supplemented with PI and resuspend nuclei by gently pipeting up and down (*see Note 2*).
13. Transfer the extract to an ice-cold tube and centrifuge at 100,000g for 45 min at 4°C (*see Note 3*).
14. Transfer supernatant to a fresh tube and add 0.1% NP-40. Quick-freeze in liquid nitrogen and store at -80°C.

3.3. Purification of Antigen and Associated Factors by Antibody-Affinity Chromatography

Important: buffers and nuclear extract must be kept cold and the experiment must be performed in a cold room.

1. Pre-equilibrate columns with buffer D* plus 0.1% NP-40 before adding the nuclear extract.
2. Run nuclear extract through protein A column (column 1) to remove nonspecific binding of nuclear proteins. Collect the flow-through on ice and save it for next step.
3. Add flow-through to preequilibrated antibody column (column 2) and incubate at 4°C for 1 h on nutator (*see Note 4*).
4. Collect the flow-through and quick-freeze.
5. Wash antibody column with 10 mL of buffer C followed by 5 mL of buffer D (*see Note 5*).
6. Sequentially pass 5 to 10 samples of buffer E (approx 200 µL each) through the antibody column to elute bound proteins. Collect each fraction into separate microcentrifuge tubes containing 1/10 volume of neutralizing buffer (0.5 M PIPES [pH 7.0]). Flick collection tubes to mix solutions well.
7. Microdialyze samples for 3 h in 1 L of buffer F supplemented with PI.
8. Collect samples into fresh ice-cold tubes.
9. To examine the profile of the eluted proteins, an aliquot of each fraction is mixed with an equal volume of 2X SB, boiled for 2 min and loaded on a 10% SDS-polyacrylamide gel. Proteins are visualized by staining the gel with Coomassie blue or silver using standard protocols.
10. To identify the proteins associated to the protein of interest, fractions containing antigen and associated proteins are pooled, and proteins are concentrated by precipitation with trichloroacetic acid (*see Note 6*). Proteins are resolved on a 10% SDS-polyacrylamide gel, stained with Coomassie blue or silver. Bands of interest are excised and submitted to a mass-spectroscopy facility for further analysis (**9**) (*see Note 7*).
11. To restore columns, wash with 1mL buffer E, then equilibrate and store columns in buffer B containing 0.01 mM NaN₃.

4. Notes

1. The antibody used for the preparation of the affinity resin needs to be purified. Ammonium sulfate precipitation is a common method used for the purification of antibody from tissue-culture supernatant serum-free media. Polyclonal antibodies are generally purified from animal serum by immunoaffinity purification on a column containing the recombinant antigen. These and a variety of other methods used to purify antibodies are described in **ref. 10**.
2. The composition of the buffer used for the extraction of proteins from the nuclei is important for efficient purification of the protein complex. Ideally, the buffer has to release the complex from the nucleus and preserve the association between its components. The radioimmune precipitation assay (RIPA) is the buffer of choice for standard immunopre-

cipitation assays. However, it is not suitable for the affinity purification because it denatures some antigens and can disrupt the protein complex. Buffers such as buffer D*, which contain 420 mM KCl, are commonly used for the preparation of nuclear extracts from HeLa cells. The preparation of nuclear extracts from other cell lines may require the optimization of the extraction conditions by varying the concentrations of salts or nonionic detergents. As a general rule, it is critical to avoid nuclear lysis, which results in a viscous solution and leads to a high level of nonspecific binding of proteins to the affinity column.

3. In many instances, particularly when problems with high background are present, it is suggested to preclear the nuclear extract by centrifugation at 100,000g for 30 to 60 min at 4°C. Under these conditions, aggregates of denatured proteins, which tend to cause an increase in background are efficiently cleared from the lysate.
4. As an alternative to the nutation step, the nuclear extract can be loaded on the column several times before the washing step to maximize the binding of the antigen to the antibody resin.
5. Try to define the washing conditions empirically. Milder buffers, such as PBS, may yield higher background, while high-salts buffers may dissociate the interacting proteins from the antigen-antibody complex. In general, it is suggested to start from the same buffer used to prepare the nuclear extract and adjust conditions, such as salt and detergent concentrations, according to the results of the SDS-PAGE.
6. To concentrate the protein solution, it may be necessary to precipitate the proteins with trichloroacetic acid (TCA). For this purpose, add 1/4 vol of ice-cold TCA solution (100% w/v TCA plus 4 mg/mL sodium deoxycholate) to the protein mixture. Vortex tube for 10 s and incubate on ice for 20 min. Centrifuge at 16,000g for 10 min at 4°C. Carefully aspirate supernatant and avoid disturbing the pellet. The pellet may not be easily visible. Add 100% acetone (approx 700 µL) to wash off the residual TCA, vortex, and incubate on ice for an additional 10 min. Centrifuge at 18,000g for 10 min at 4°C. Aspirate the supernatant and air-dry the pellet. Resuspend the pellet in 15 µL of 2X SB. If the color turns yellow, immediately add 1 µL of 2 M Tris (pH 8.0). Heat at 95°C for 3 min prior to loading on a SDS-polyacrylamide gel.
7. It is generally useful, before proceeding with the identification of the associated factors by mass spectrometry, to confirm or exclude a specific protein by performing a parallel mock purification using a resin crosslinked to immunoglobulin G or to antibody against an unrelated protein.

References

1. Phizicky, E. M. and Fields, S. (1995) Protein-protein interactions: methods for detection and analysis. *Microbiol. Rev.* **59**, 94–123.
2. Puig, O., Caspary, F., Rigaut, G., et al. (2001) The tandem affinity purification (TAP) method: a general procedure of protein complex purification. *Methods* **24**, 218–229.
3. Honey, S., Schneider, B. L., Schieltz, D. M., Yates, J. R., and Futcher, B. (2001) A novel multiple affinity purification tag and its use in identification of proteins associated with a cyclin-CDK complex. *Nucleic Acids Res.* **29**, E24.
4. Reisdorf, P., Maarse, A. C., and Daignan-Fornier, B. (1993) Epitope-tagging vectors designed for yeast. *Curr. Genet.* **23**, 181–183.
5. Dynlacht, B. D., Hoey, T., and Tjian, R. (1991) Isolation of coactivators associated with the TATA-binding protein that mediate transcriptional activation. *Cell* **66**, 563–576.
6. Comai, L., Tanese, N., and Tjian, R. (1992) The TATA-binding protein and associated factors are integral components of the RNA polymerase I transcription factor, SL1. *Cell* **68**, 965–976.

7. Wang, Y., Cortez, D., Yazdi, P., Neff, N., Elledge, S. J., and Qin, J. (2000) BASC, a super complex of BRCA1-associated proteins involved in the recognition and repair of aberrant DNA structures. *Genes Dev.* **14**, 927–939.
8. Comai, L., Zomerdijk, J. C., Beckmann, H., Zhou, S., Admon, A., and Tjian, R. (1994) Reconstitution of transcription factor SL1: exclusive binding of TBP by SL1 or TFIID subunits. *Science* **266**, 1966–1972.
9. Pandey, A., Andersen, J. S., and Mann, M. (2000) Use of mass spectrometry to study signaling pathways. *Sci. STKE* **37**, PL1.
10. Harlow, E., and Lane, D. (1999). Using Antibodies: A Laboratory Manual. (*2nd edition*), *Cold Spring Harbor, Cold Spring Harbor, NY*.

Biomolecular Interaction Analysis Coupled With Mass Spectrometry to Detect Interacting Proteins

Setsuko Hashimoto, Toshiaki Isobe, and Tohru Natsume

1. Introduction

With the completion of the human genome sequencing, the objectives of life science have shifted to understanding the functions of proteins. One of the experimental processes of functional proteomics is the analysis of protein interaction. Surface plasmon resonance (SPR) sensors have become popular technology for the interaction analysis of proteins. The sensors can monitor protein interactions in real-time without labeling molecules. An SPR sensor is unique that it can provide kinetic information on interactions, such as association and dissociation rate constants, which often provide clues to evaluate the molecular interaction in terms of protein functions and biological mechanisms (1,2). With these features, SPR sensors have been intensively used for “ligand fishing” experiments to identify the binding partner proteins to a specific bait protein (3–8). Although a binding partner is found in a biological mixture, the task of identifying the ligand remains daunting. Purification of the binding partners can be time consuming and labor intensive, often requiring case-by-case strategies.

To facilitate ligand identification, the combination of SPR sensor and mass spectrometry (MS) has been developed (Fig. 1). MS is one of the most sensitive and specific techniques available for identification and characterization of biomolecules. Two ionization mechanisms are commonly used for MS:

- a. Matrix-assisted laser desorption/ionization (MALDI), in which proteins, crystallized on a sample support, are ionized by laser irradiation.
- b. Electrospray ionization (ESI), in which ions are generated directly from solution.

Identification of binding-partner molecules after affinity purification with SPR sensors has been demonstrated using both ESI-MS (9,10) and MALDI-MS (11–14).

The Biacore 3000 instrument, a most popular commercial SPR sensor, is particularly useful for ligand fishing experiments, in which it can be used to monitor the immobilization of a bait protein, to capture the target protein, and finally to prepare the recovered material automatically for MS analysis.

The major difficulty in combining Biacore with MS is the transfer of the sample from a Biacore instrument to the ion source of the mass spectrometer. Small quantities of protein samples are easily lost just by transferring from one vial to another or pipetting. To overcome this difficulty, we have developed a novel “on-chip” method to digest the bound proteins to minimize protein handling (Fig. 2). As a modification of

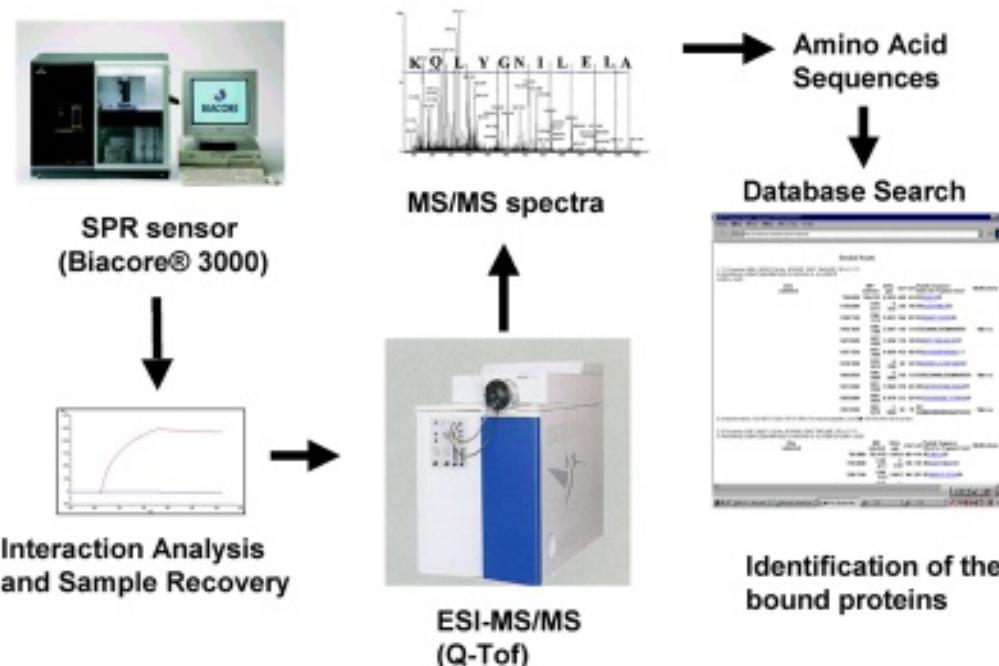


Fig. 1. Combination of Biacore and mass spectrometry (MS). The interaction with a bait protein immobilized on the sensor chip and prey proteins can be monitored by surface plasmon resonance sensor. The bound proteins were digested on-chip and the peptides were recovered for MS analysis. With the amino acid sequences of the recovered peptides, bound proteins can be identified unambiguously by a database search.

on-chip digestion, we have also demonstrated “on-chip phosphorylation” of the immobilized proteins and confirmed phosphorylation-dependent protein interactions. This will open up a new technical platform to study regulation of protein function with posttranslational modifications. Following on-chip digestion, the resulting peptide mixture is then trapped and concentrated in a minicolumn made by GELoader tip for subsequent tandem MS (MS/MS) analysis to obtain the amino acid sequences for the unambiguous identification of the bound proteins.

2. Materials

2.1. Instruments

1. Biacore® 3000 (Biacore AB, Sweden).
2. Q-TOF ESI-MS/MS (Micromass, UK).

2.2. Software

1. Mascot program (Matrixscience).

2.3. Reagents

1. Sensor chip NTA (Biacore AB, BR-1000-34): sensor chip with the surface of nitrilotriacetic acid (NTA) for capturing histidine-tagged fusion proteins via metal chelation.

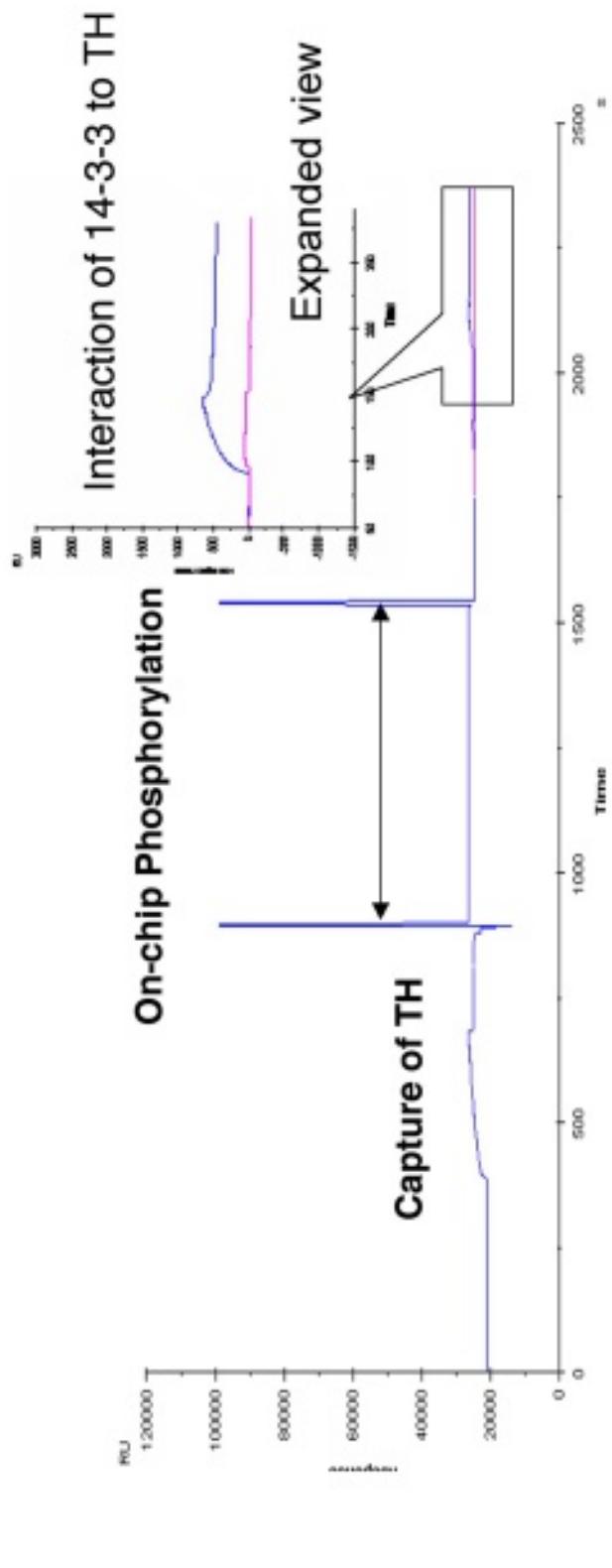


Fig. 2. Experimental design. Histidine-tagged tyrosine hydroxylase (TH) was immobilized on nitrilotriacetic acid (NTA) sensor chip by nickel chelation. On-chip phosphorylation of TH was performed by injecting CaM kinase on flow cell 2. Glutathione s-transferase-14-3-3 was then injected over the flow cells. The binding of 14-3-3 was observed only on flow cell 2.

2. Human tyrosine hydroxylase (TH): histidine-tagged TH was expressed with baculovirus expression system in Sf21 cells and purified with Ni^{2+} -agarose beads (15).
3. Bovine 14-3-3: glutathione s-transferase (GST)-fused 14-3-3 was expressed in *E. coli* and purified with glutathione-agarose beads (16).
4. Calmodulin (CaM) and CaM kinase II were purified from bovine and rat brains, respectively (17).
5. Lysyl endopeptidase (LEP).
6. POROS R2 bead (particle size 50 μm , Applied Biosystems).
7. GELoader tip (Eppendorf, Germany).

2.4. Running Buffer for Biacore

1. For immobilization of the histidine-tagged protein: 100 μM NiCl_2 , 50 mM HEPES, 150 mM NaCl, 50 μM ethylenediaminetetraacetic acid (EDTA), pH 7.4.
2. For interaction analysis: 50 mM HEPES, 150 mM NaCl, 50 μM EDTA, 0.005% *n*-octylglucopyranoside, pH 7.4.
3. For on-chip phosphorylation: 50 mM HEPES (pH 7.6), 0.5 mM ATP, 5 mM $\text{Mg}(\text{CH}_3\text{CO}_2)_2$, 0.1 mM CaCl_2 , 40 $\mu\text{g}/\text{mL}$ CaM, 3 $\mu\text{g}/\text{mL}$ CaM kinase II.
4. For on-chip digestion: 200 mM 2-amino-2-hydroxymethyl-1,3,-propanediol (pH 8.0), 4 M urea, 0.005% *n*-octylglucopyranoside.
5. Peptide recovery: A solvent—2% acetonitrile, 0.1% formic acid; B solvent—50% acetonitrile, 0.1% formic acid.

3. Methods

We describe the experiment to monitor the interaction of 14-3-3 to phosphorylated tyrosine hydroxylase (TH) followed by the identification of the proteins by MS/MS analysis as a case study. TH is known to be activated through phosphorylation by calmodulin (CaM) kinase II in response to a variety of stimuli that increase intracellular Ca^{2+} . The 14-3-3 protein family plays an important role in cell signaling processes, including monoamine synthesis, exocytosis, and cell-cycle regulations. It has been demonstrated that phosphorylated TH binds to 14-3-3 isoform. We outline (1) the immobilization of histidine-tagged TH to the sensor chip NTA; (2) interaction analysis of 14-3-3 to TH with and without in vitro phosphorylation by Cam kinase II; (3) on-chip digestion of the interaction complex; (4) recovery of the digested peptides; and (5) MS/MS analysis for protein identification.

3.1. Immobilization of Histidine-Tagged TH to the Sensor Chip NTA

TH expressed in insect cells with a hexa-histidine tag was affinity captured on the sensor chip NTA with the Biacore instrument (see Note 1).

1. Insert Sensor Chip NTA into Biacore 3000 instrument. “Prime” the instrument with a running buffer of 50 mM HEPES, 150 mM NaCl, 50 μM EDTA, 0.005% *n*-octylglucopyranoside (pH 7.4). The flow rate was set to 5 $\mu\text{L}/\text{min}$ and the reaction temperature to 25°C.
2. Inject 10–20 μL of 100 μM NiCl_2 , 50 mM HEPES, 150 mM NaCl, 50 μM EDTA (pH 7.4).
3. Inject 40 μL of 10 $\mu\text{g}/\text{mL}$ histidine-tagged TH over the Ni-chelated sensor surface. The differences in the response signals (RU) before and after the injection of TH is the amount of TH captured on the NTA surface. 1000 RU corresponds to 1 ng of the protein. Try to capture as much TH as possible.

3.2. Interaction Analysis of 14-3-3 to TH

The interaction of 14-3-3 with the phosphorylated TH was confirmed by the Biacore analysis.

1. Inject 100 μ L of CaM kinase II phosphorylation buffer at a flow rate of 5 μ L/min into flow cell 2 for 10 min (*see Note 2*).
2. Inject 120 μ g/mL GST-14-3-3 over flow cells 1 and 2. The interaction of 14-3-3 is observed only in flow cell 2, where TH is phosphorylated by CaM kinase II (**Fig. 2**).

3.3. On-Chip Digestion of the Interaction Complex

The interaction complex of TH and 14-3-3 was subjected to the peptidase digestion for MS analysis.

1. The autosampler and the microfluidics but not the flow cells were washed with 0.1% formic acid using the MS_WASH command (a bypass wash command of the Biacore 3000 control software).
2. 2 μ L of (25 ng/mL) lysyl endopeptidase was injected into the flow cells using the MS_RECOVER command (*see Note 3*). Once the enzyme solution reached the sensor surface, which can be monitored by the spike signal from the air in the flow, the flow of the instrument was halted, allowing the enzyme reaction to proceed on the sensor surface for approx 2 h. The progress of the enzyme reaction can be monitored by the decrease in the SPR signals (**Fig. 3**) (*see Notes 4 and 5*).

3.4. Recovery of the Digested Peptides

1. Once the decrease in the SPR signal stopped, suggesting the completion of the enzyme reaction, the flow of the Biacore instrument was restarted in the reverse direction.
2. The recovered solution was applied to a minicolumn made with the GELoader tip which was packed with POROS (18).
3. The column was washed with solvent A and the digested peptides were eluted in 2 μ L of solvent B into a nanospray tip directly.

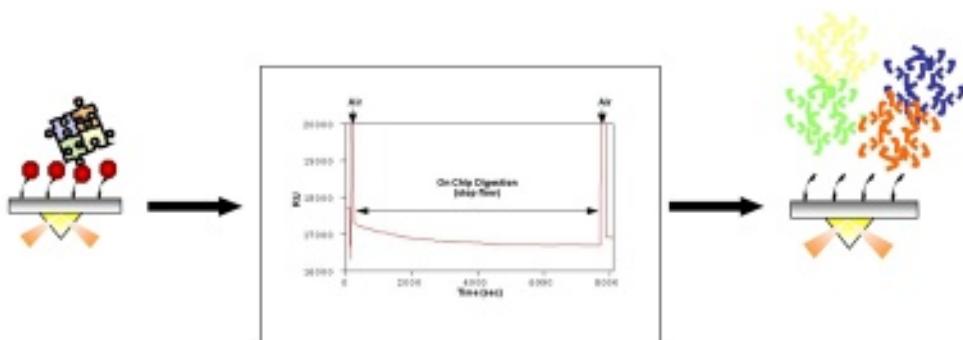
3.5. MS/MS Analysis for Protein Identification

1. The nanospray was analyzed by a quadrupole time-of-flight mass spectrometer (Q-TOF). If peptide ions were present in the MS scan, these precursor ions were selected for MS/MS analysis so that internal amino acid sequence information of the peptides was obtained for unambiguous identification (**Figs. 4 and 5**).

Proteins were identified using the Mascot program. The database maintained at the National Center for Biotechnology Information (NIH) was used for protein assignment.

4. Notes

1. Purified proteins were used as a model system in this case study. Crude samples can also be applied to identify novel binding partners.
2. Four flow cells are arranged parallel on the surface of the sensor chip. The samples and the buffer solutions can be sent over the flow cells in tandem or separately. In this experiment, TH was captured on flow cells 1 and 2. The CaM kinase solution was injected only over flow cell 2. When 14-3-3 was injected, the interaction was observed only on flow cell 2. No significant binding was detected on flow cell 1, where TH was not phosphorylated.



Real-time analysis of on-chip digestion

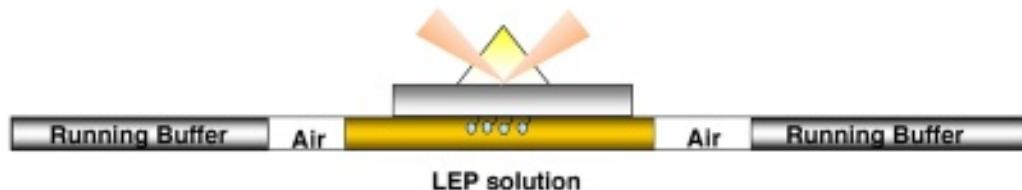


Fig. 3. On-chip digestion. Proteolytic enzyme solution was injected over the flow cells. The enzyme solution was sandwiched by the two air plugs at both ends to avoid dilution and dispersion. The flow was stopped when the first air spike was detected, followed by peptide digestion for 2 h. The digestion process can be monitored in the changes in the sensorgram.

3. The enzyme solution was partitioned from the running buffer by air plugs at both ends to prevent the enzyme from being diluted during the incubation and minimize dispersion of the digested peptides in the microfluidics.
4. The Biacore instrument must be kept as clean as possible to avoid any contamination in the flow systems, which would lead to noise in the MS signals. Run the recommended maintenance procedures routinely. For capturing of His-tagged proteins, special care must be taken to eliminate metal ions in the buffers and the instrument to achieve the efficient capturing of His-tag proteins to the NTA surface. Wash the instrument with 0.35 M EDTA and equilibrate the instrument with a continuous flow of the running buffer until the start of the experiment. Try to immobilize as much histidine-tagged TH as possible (ca. ≥ 5000 RU) for better detection of the MS signals.
5. The detergent surfactant P20 is a standard component of the running buffer for Biacore. However, it has been shown to give multiple intense ions in MS analysis. The recovery of the peptide decreases drastically in the absence of the detergent. Hence, a detergent that is compatible for use in mass spectrometry, such as *n*-octylglycopyranoside, was selected.

Acknowledgments

The authors thank Dr. Tohru Ichimura for providing TH and 14-3-3 proteins. We also thank Kazunobu Asano for performing Biacore and MS experiments and Dr. Hiroshi Nakayama for valuable advice for MS analysis. This work was supported in part by Grants for the Integrated Proteomics System Project, Pioneer Research on Genome the Frontier from MEXT of Japan.

A**BIA-MS TH-P-GST14-3-3 on chip digestion off-line rec nano spray by asano**

K16_01 173 (15.444) Cm (41:226)

677.044

TOF MS ES⁺
2.83e4

739.671

679.042

21(3+)

23 (3+)

3(3+)

2(3+)

31 (2+)

Expanded in Fig. 4B

776.448

776.119

826.165

825.835

826.495

838.007

787.138

839.004

839.498

787.469

741.670

839.634

836.244

635.074

637.357

673.534

654.389

674.314

695

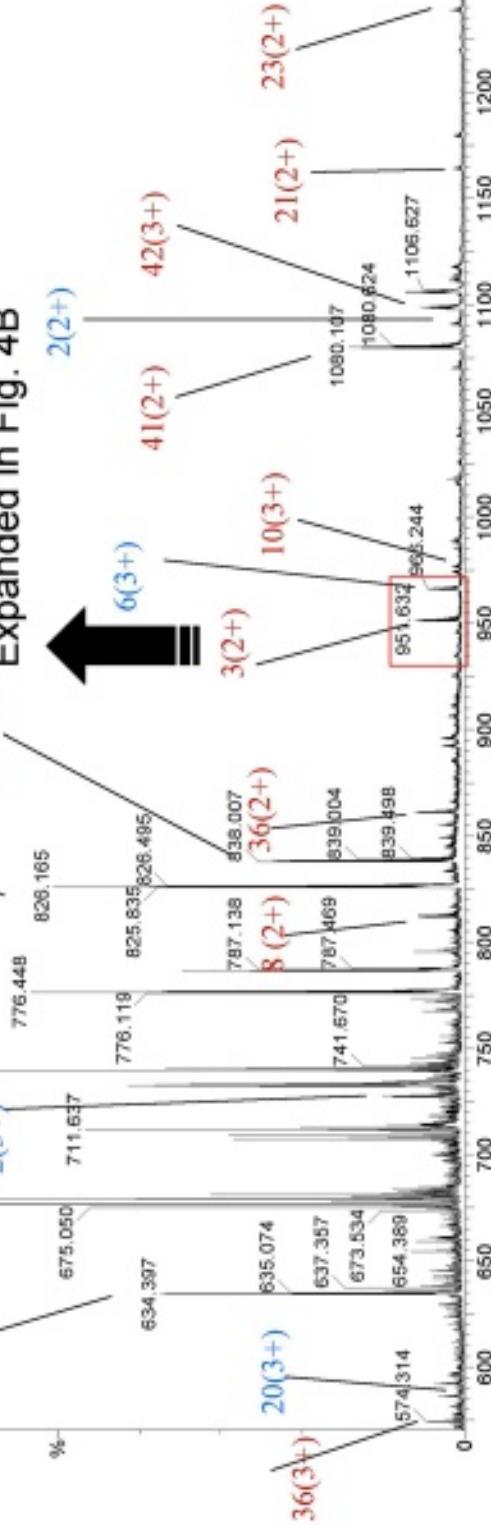


Fig. 4. (A) Mass spectrometry (MS) spectrum of on-chip digested tyrosine hydroxylase (TH)-14-3-3 complex. The recovered peptide sample was applied to electrospray ionization-tandem MS. Typical MS spectra are shown.

B**BIA-MS TH-P-GST14.3.3 on chip digestion off-line rec nano spray by asano**

K16_01 173 (15.444) Cm (21:226)

TOF MS ES+
1.96e3

100

%

GST14.3.3 < P3(2+)>**TH <P6(3+)>**

966.244

966.000

965.907

965.916

967.233

967.590

968.549

968.653

966.960

966.960

966.960

966.960

966.960

966.960

966.960

966.960

966.960

966.960

966.960

966.960

966.960

966.960

966.960

696

Fig. 4. (continued) (B) MS spectrum of on-chip digested TH-14-3-3 complex. MS spectra showing the peptides derived from 14-3-3 and TH are expanded in (B).



TOF MSMS 965.90ES+
244

$m/z = 965.82 +3$

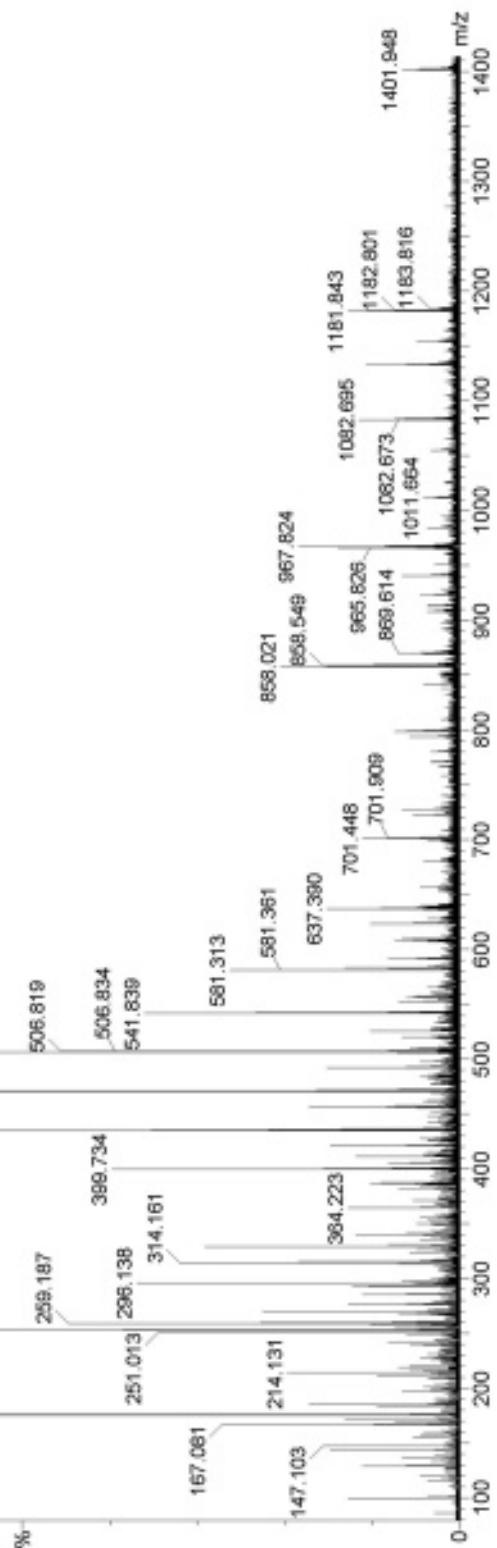


Fig. 5. Tandem mass spectrometry (MS/MS) spectrum and amino acid assignment. MS spectrum at $m/z = 965.82(3+)$ was fragmented and sequenced. The amino acid sequence was determined and it was assigned as the lysyl endopeptidase-digested peptide derived from tyrosine hydroxylase (EREAAVAAAAAVPSEPGDPLEAVAFEEK).

References

1. Jönsson, U., Fagerstam, L., Ivarsson, B., et al. (1991) Real-time biospecific interaction analysis using surface plasmon resonance and a sensor chip technology. *Biotechniques* **11**, 620–627.
2. Nagata K. and Handa, H. (eds) (2000) *Real-Time Analysis of Biomolecular Interactions*. Springer Verlag, Tokyo, Japan.
3. Lackmann, M., Bucci, T., Mann, R.J., et al. 1996 Purification of a ligand for the EPH-like receptor HEK using a biosensor-based affinity detection approach. *Proc. Natl. Acad. Sci. USA* **93**, 2523–2527
4. Sakano, S., Serizawa, R., Inada, T., et al. (1996) Characterization of a ligand for receptor protein-tyrosine kinase HTK expressed in immature hematopoietic cells. *Oncogene* **13**, 813–822.
5. Seok, Y. J., Sondej, M., Badawi, P., et al. (1997) High affinity binding and allosteric regulation of *Escherichia coli* glycogen phosphorylase by the histidine phosphocarrier protein, HPr. *J. Biol. Chem.* **272**, 26,511–26,521.
6. Markgren, P. O., Hamalainen, M., and Danielson, U. H. (1998) Screening of compounds interacting with HIV-1 proteinase using optical biosensor technology. *Anal. Biochem.* **265**, 340–350.
7. Iemura, S., Yamamoto, T. S., Takagi, C., Kobayashi, H., and Ueno, N. (1999) Isolation and characterization of bone morphogenetic protein-binding proteins from the early *Xenopus* embryo. *J. Biol. Chem.* **274**, 26,843–26,849.
8. Williams, C. and Addona, T. A. (2000) The integration of SPR biosensors with mass spectrometry: Possible applications for proteome analysis. *Trends Biotechnol.* **18**, 45–48.
9. Natsume, T., Nakayama, H., Jansson, Ö., Isobe, T., Takio, K., and Mikoshiba, K. (2000) Combination of biomolecular interaction analysis and mass spectrometric amino acid sequencing. *Anal. Chem.* **72**, 4193–4198.
10. Natsume, T., Nakayama H. and Isobe T. 2001 BIA-MS-MS: biomolecular interaction analysis for functional proteomics. *Trends Biotech.* **19** s28–s33.
11. Sönksen, C. P., Nordhoff, E., Jansson, Ö., Malmqvist, M., and Roepstorff, P. (1998) Combining MALDI mass spectrometry and biomolecular interaction analysis using a biomolecular interaction analysis instrument *Anal. Chem.* **70**, 2731–2736.
12. Nelson, R. W. and Krone, J. R. (1999) Advances in surface plasmon resonance biomolecular interaction analysis mass spectrometry (BIA/MS) *J. Mol. Recog.* **12**, 77–93.
13. Williams, C. and Addona, T. A. (2000) The integration of SPR biosensors with mass spectrometry: Possible applications for proteome analysis. *Trends Biotechnol.* **18**, 45–48.
14. Natsume, T., Taoka, M., Manki, H., Kume, S., Isobe, T., and Mikoshiba, K. (2002) Rapid analysis of protein interactions: on-chip micropurification of recombinant protein expressed in *Escherichia coli*. *Proteomics* **2**, 1247–1453.
15. Itagaki, C., Isobe, T., Taoka, M., et al. (1999) Stimulus-coupled interaction of tyrosine hydroxylase with 14-3-3 proteins. *Biochemistry* **38(47)**, 15,673–15,680.
16. Ogihara, T., Isobe, T., Ichimura, T., et al. (1997) 14-3-3 protein binds to insulin receptor substrate-1, one of the binding sites of which is in the phosphotyrosine binding domain. *J. Biol. Chem.* **272(40)**, 25,267–25,274.
17. Ichimura, T., Uchiyama, J., Kunihiro, O., et al. (1995) Identification of the site of interaction of the 14-3-3 protein with phosphorylated tryptophan hydroxylase. *J. Biol. Chem.* **270(48)**, 28,515–28,518.
18. Gobom, J., Nordhoff, E., Mirgorodskaya, E., Ekman, R., and Roepstorff, P. (1999) Sample purification and preparation technique based on nano-scale reversed-phase columns for the sensitive analysis of complex peptide mixtures by matrix-assisted laser desorption/ionization mass spectrometry. *J. Mass Spectrom.* **34**, 105–116.

Assessment of Antibody–Antigen Interaction Using SELDI Technology

Li-Shan Hsieh, Ramy Moharram, Emilia Caputo, and Brian M. Martin

1. Introduction

The aim of this chapter is to discuss/describe both the theoretical aspects of the study of antibody–antigen interaction employing ProteinChip Array® technology, also known as surface-enhanced laser desorption/ionization-time-of-flight mass spectrometry (SELDI-TOF™ MS), and to provide proven procedures for their use.

SELDI-TOF MS, first introduced in 1993 by Hutchens and Yip (1), is a novel approach based on two powerful techniques, chromatography and mass spectrometry. It consists of selective protein/peptide extraction and retention on chromatographic chip arrays and their subsequent analysis using a simple laser desorption/ionization mass spectrometer (2,3). The ProteinChip arrays have chemically derivatized surfaces utilizing classical chromatographic separation characteristics such as reverse phase, ion exchange, silica, immobilized metal affinity capture, and preactivated capture. The latter surface allows for covalent attachment of various molecules, such as antibodies, receptors, DNA, small molecules, and ligands. Bio-active proteins/peptides can thus be captured on these surfaces and/or identified through the recognition of their corresponding antibodies. SELDI-TOF MS technology has thus far been successful in various applications ranging from protein profiling of complex biological mixtures (4) to identification and characterization of biomolecules (5–7).

In this chapter, we illustrate the capabilities of this technology in reference to protein identification via antibody–antigen interactions and epitope determination. As a result of the diverse nature of antibodies and antigens, a number of methods to assess antibody–antigen interactions using SELDI-TOF MS are outlined. Examples include capturing allergens and determining epitopes using low-titer antibody. Experimental results have shown that these methods are as sensitive as Western blots.

2. Materials

1. Instrument and reagents: all mass spectra were recorded in the positive ion mode using a Ciphergen Biosystems, Inc. (Fremont, CA), PBS II ProteinChip Array reader and a linear laser desorption/ionization-time-of-flight mass spectrometer with time lag focusing (2). The instrument was operated in positive ion mode with a source and detector range of 2.0 and 2.2 kV, a digitizer rate of 250 MHz time focusing, pulse voltage (3000 V), pulse lag time (600 ns), and nitrogen laser (337 nm) with 175 μ J maximum energy/4 ns

- pulse and 20 Hz maximum pulse rate. Phosphate-buffered saline (PBS) and Tween-20 were purchased commercially.
2. ProteinChips: NP, H4, PS1, and PS2 ProteinChip Arrays were purchased from Ciphergen.
Normal phase (NP1 and NP2)—The active spots contain silicon oxide, which bind proteins via serine, threonine, arginine, or lysine residues. Their use is recommended for the analysis of hydrophilic proteins.
Hydrophobic surface (H4)—The active spots contain a 16-methylene group (C-16) that bind proteins abundant in alanine, valine, leucine, isoleucine, phenylalanine, tryptophan, or tyrosine residues in a manner analogous to reverse-phase chromatography. Their use is recommended for the analysis of hydrophobic proteins.
Preactivated surfaces (PS1 and PS2)—PS1 and PS2 chips (now replaced by PS10 and PS20) have a reactive carbonyl diimidazole moiety and an epoxy functional group on active spots, respectively. The PS20 chips also have a hydrophobic coating that restricts the sample to the spot. Because of differences in surface properties, the PS20 surface is especially recommended for sensitive detection, low nonspecific binding, and target protein less than 1% of total protein.
 3. Matrix: Sinapinic acid (SPA) and α -cyano-4-hydroxy-cinnamic acid (CHCA) are used as matrix for the analysis of proteins and peptides, respectively. The matrix is prepared as a saturated solution in 50% acetonitrile water containing 0.1% trifluoroacetic acid (TFA).
 4. Buffers:
 - a. *Antibody buffer*—Antibodies (see **Notes 1 and 2**) can be coupled directly to the surface of a preactivated ProteinChip array. To this end, buffer for the covalent coupling reaction must be free of nucleophilic agents (e.g., Trizma base, glycerin, and so on). We recommend PBS as the buffer. In addition, many antibodies are stored in solutions containing azide as a preservative. In this case, the azide will interfere with antibody coupling to the PS1/2 ProteinChip surface; therefore, it may be necessary to remove the azide (by dialysis) prior to coupling.
 - b. *ProteinChip buffers*—PBS acts as the buffer for the PS1/2 surfaces. Its composition is 2.7 mM potassium chloride, 120 mM sodium chloride, and 10 mM phosphate buffer maintained at a pH of 7.0.
 5. A detergent can also be added to a PBS solution in order to increase the stringency of the wash. In our case, we used a 0.05% concentration of Tween-20. Furthermore, prior to attaching an antibody to the ProteinChip surface or to protein G, it is necessary to block any remaining reactive sites that might distort mass spectrometry results. The most commonly used blocking solution is 1 M ethanolamine in PBS.
 6. Materials for Example 1: Twenty grams of banana fruit and 25 mL of 50 mM sodium bicarbonate were combined and homogenized; the volume was then adjusted to 50 mL with the same bicarbonate buffer. This mixture was centrifuged at 20,000g for 30 min. The supernatant was collected and used throughout. Sera were generously provided by the Department of Allergy, National Children's Hospital, Tokyo, Japan. A total of nine sera from patients allergic to banana and/or latex with positive case histories and characteristic type I allergic reactions were used for this study. Patient clinical characteristics are described in **Table 1**. For example, patient 9, who had low levels of banana immunoglobulin (Ig)E, was used as a negative control. A phosphatase-labeled goat antihuman IgE antibody (Kirkegaard and Perry Laboratories, Gaithersburg, MD) diluted 1:2500 in PBS was used. For comparison, sera were also analyzed for their IgE activities using banana extract by Western blot. Banana extract was analyzed by sodium dodecyl sulfate (SDS)-polyacrylamide gel electrophoresis (PAGE) and blotted to a polyvinylidene difluoride (PVDF) membrane. Membrane strips were incubated with individual patient's serum. The bound IgE was detected using peroxidase labeled antihuman IgE goat serum and assayed according to the manufacturer's instructions shown in **Fig. 1**.

Table 1
Clinical Manifestations and Tested IgEs of Nine Patients

Patient no.	Age	Sex	Allergic disease	Symptoms by banana	Total IgE U/mL	Banana IgE class	Latex IgE U/mL
1	18	M	AD	Skin rash	3425	1	9.55
2	7	F	BA, Lx	Skin rash	166	2	31.7
3	15	M	AD, BA	Skin rash	11,378	2	1.74
4	15	M	UR	UR	2438	2	3.43
5	9	M	DA, AR	Itching	1808	2	0.88
6	5	F	AD, FA, BA	None	9595	3	12.3
7	7	M	BA, FA	UR	998	3	5.82
8	9	M	Lx, AR, AC. OAS	OAS	119	3	6.38
9	1	F	AD, FA	UR	18.9	1	ND

Ig, immunoglobulin; AD, atopic dermatitis; BA, bronchial asthma; Lx, latex allergy; FA, food allergy; AC, allergic conjunctivitis; AR, allergic rhinitis; DA, drug allergy; OAS, oral allergy syndrome; UR, urticaria; ND, not detectable.

Banana IgE class: 1 (0.35–0.69 U/mL); 2 (0.75–3.4 U/mL); 3 (3.5–17.4 U/mL); 4 (17.5–49.9 U/mL); 5 (50–99 U/mL); 6 (>100 U/mL); IgE activity was determined by Pharmacia CAP system.

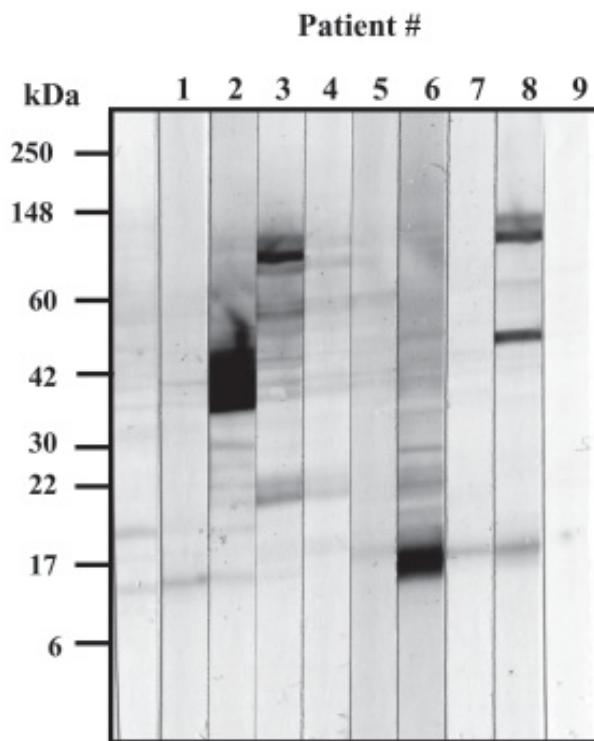


Fig. 1. Analysis of allergens by immunoblotting. Banana extract was separated under reducing conditions on 16% sodium dodecyl sulfate-polyacrylamide gel electrophoresis and then electrotransferred to a nylon membrane. The membrane was cut into 10 strips. Lane 0 shows the proteins stained with Coomassie blue from banana extract. The other 9 strips were separately incubated with individual sera from nine patients allergic to banana, indicated with the numbers from 1 to 9.

7. Materials for Example 2: Twenty residue peptides derived from Hev b 5, a major acidic latex allergen, with overlaps of nine amino acids, were synthesized by Chiron Corp. and were provided by the Center for Biologics Evaluation and Research, FDA (8–10). Peptides 20 and 21 were synthesized by CBER Core facility, Food and Drug Administration (FDA). Prior to the experiment, all peptides were examined on H4 ProteinChips and were used as reference controls for epitopes captured by individual sera. Pooled latex antisera S2 and S3 were obtained from CBER, FDA. Latex-allergic patient sera 1–3 as well as a pool of combined sera from patients 1–3 were used for capture of peptides. Anti-IgE antibody used was described above.

3. Methods

A brief description of the covalent attachment of proteins to ProteinChip arrays (**Subheading 3.1.**) and in particular the attachment of protein A or G (**Subheading 3.2.**) are outlined. Specific examples are described in **Subheadings 3.3.1.** and **3.4.1.**

3.1. Covalent Attachment of Proteins to the ProteinChip

1. Initially, rinse the ProteinChip surface with 50% acetonitrile water from a squeeze bottle by gentle spraying. Let the surface dry and add 1 to 3 μ L of PBS for several minutes. (Note: A hydrophobic pen may be used to circle the spot if the chip is not precoated.)
2. Remove the PBS (with a Kimwipe or a pipet) and add 1–5 μ L of protein to be coupled. Incubate the chip in a humid chamber for 1 to 4 h (or overnight at 4°C).
3. Remove the chip and wash with PBS containing 0.05% Tween-20 three times (10 μ L each time) and finally with PBS.
4. Add 2 μ L of ethanolamine (1 M in PBS) to block any remaining reactive sites on the surface. Incubate 30 min in the humid chamber and then wash extensively with PBS. (Note: The chip should be kept damp until the next step or for future use.) If necessary, the chip may be stored in a sealed container filled with PBS for up to 1 wk.

3.2. Attachment of Protein A or G for Amplification of Antibody Detection

In cases where it is necessary to use protein A or G for capture of the antibody, one should check the integrity of the protein A or G using an H4 chip as described above, due to its tendency to break down over time when stored in solution at 4°C. In the past, there have been some problems with the quality of commercially available protein A; therefore, we prefer protein G (*see Note 2*).

1. Incubate the ProteinChip with 3 μ L of PBS for 5 min and then remove the PBS (with a Kimwipe or pipet).
2. Apply 1.0 μ L of protein A or G (1 mg/mL in PBS) to the spot, then incubate in a humid chamber for 1 h at room temperature.
3. Wash the ProteinChip with buffer (PBS with 0.05% Tween), three times. (Alternatively, one may leave the chip in a tightly closed tube, filled with wash buffer, and place on a rocker or shaker for 1–2 min.)
4. Wash with PBS, one time (without Tween, to prevent the collapse of drops on the chip surface).
5. Incubate with ethanolamine as above and repeat washes. (Note: do not let the chip dry.)

3.3. Example 1. Detection of Allergens Using Patients' Serum

3.3.1. Preparation of IgE-Conjugated ProteinChip

Serum IgE was captured by affinity purified goat anti-human IgE antibody that was covalently attached to a PS1 ProteinChip Array surface.

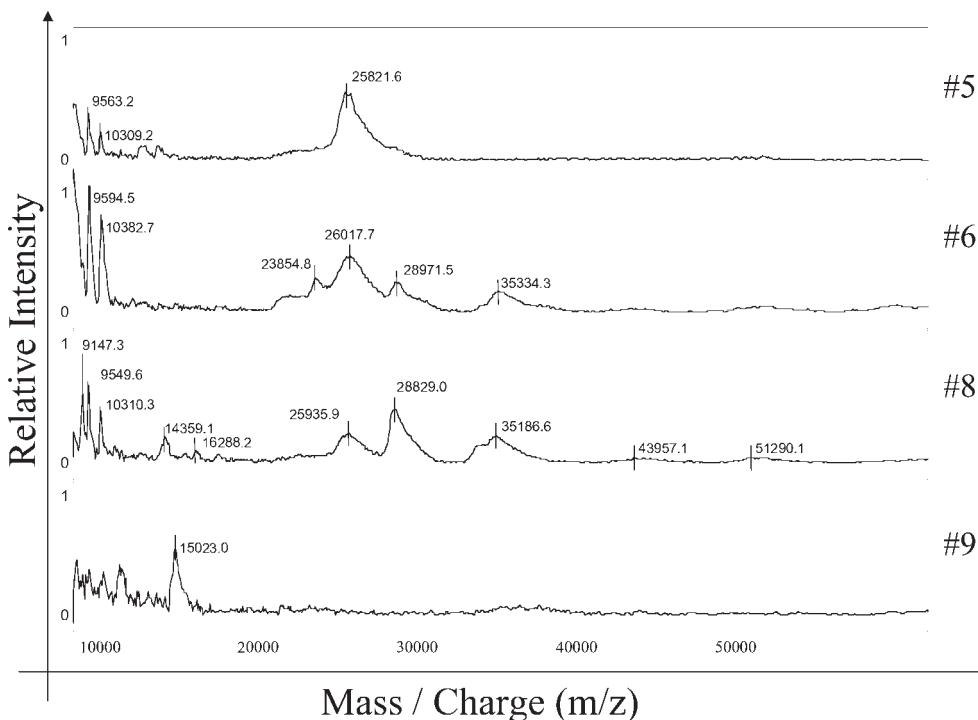


Fig. 2. Surface-enhanced laser desorption/ionization-time-of-flight mass spectrometry spectra of allergens captured by immunoglobulin E-conjugated ProteinChip®. Banana allergens captured by allergenic sera are shown. The mass/charge (m/z) values of each detected species are reported.

1. Incubate the ProteinChip with 3 μ L of PBS for 5 min and then remove the PBS.
2. Apply 2.0 μ L of goat antihuman IgE (1:2500 dilution in PBS) to the spot, then incubate in a humid chamber for 1 h at room temperature.
3. Wash the ProteinChip with buffer (PBS with 0.05% Tween) three times. Incubate with ethanolamine as above and repeat washes. (Note: do not let the chip dry.)
4. Apply an aliquot of 3 μ L of 10X diluted patient serum to the spot with anti-IgE antibody conjugated, and incubate in a humid chamber for 1 h.
5. Wash the ProteinChip with buffer (PBS with 0.05% Tween) three times, followed by a final PBS wash.

3.3.2. Capture of Allergens on Chip Surface

The following outline presents the steps for capturing allergens from crude extract by IgE-conjugated ProteinChip.

1. Apply the 3 μ L of 10X diluted banana extract to the IgE-captured ProteinChip and then incubate in a humid chamber for 1 h at room temperature.
2. Wash the ProteinChip with buffer (PBS with 0.05% Tween) three times and PBS one time.
3. Wash the ProteinChip with deionized water and allow to air dry.
4. Add SPA as a matrix and read on the mass spectrometer.

3.3.3. Detection of Captured Allergens or Peptides

1. Apply 0.5 μ L SPA (CHCA for peptides) to the chip surfaces and allow to air dry.
2. Read the chip on the PBS IIC. Raw data were analyzed using the computer software provided by the manufacturer and are shown in **Fig. 2**. It is important to consider the

differences observed between the data obtained from the SELDI-TOF MS and the immuno-blotting analysis of the allergens resulting from different techniques and experimental conditions used.

3.4. Example 2: Epitope Mapping of Latex Allergen *Hev b.5*

3.4.1. Preparation of IgE-Conjugated ProteinChip

Steps are the same as those outlined in Example 1, first coupling the anti-IgE and then capturing the IgE from the latex antisera or patient samples.

3.4.2. Capturing of Peptide (Epitope) on Chip Surface

The IgE-conjugated ProteinChip then can be used for capturing peptides that contain epitopes.

1. Apply the 3 μ L of peptide solution (diluted to working concentration approx 10 μ g/mL) to the IgE ProteinChip and then incubate in a humid chamber for 1 h at room temperature.
2. Wash the ProteinChip with buffer (PBS with 0.05% Tween) three times and PBS one time.
3. Wash the ProteinChip with deionized water and allow to air dry.

3.4.3. Detection of Peptides

Steps are the same as those outlined elsewhere in this chapter, using CHCA as matrix. This method yields a quick and unambiguous detection of peptides captured by IgE, as shown in **Fig. 3**. The summary of re-activity of all the sera and peptides captured are shown in **Table 2**. We were able to determine both weak and strong serum-binding sequences and identify the resulting epitopes (**Fig. 4**).

4. Notes

1. Antibodies belong to the Ig supergene family and are classified into five major classes (IgG, IgM, IgA, IgD, IgE) with respect to their physical, chemical, and biological properties (11). Antibodies are typically raised against proteins/peptides called antigens. The antigenic region within the protein/peptide to which the antibody binds is known as the epitope; this may be a continuous linear sequence of amino acid residues, or may be conformation dependent. Therefore, some antibodies recognize the epitope on native proteins, others recognize the epitope on the denatured protein, while others recognize epitopes in both the native and the denatured forms.

Among the five different types of immunoglobulins in serum, IgG is the most abundant (more than 75%), is synthesized at a high rate and with high-affinity binding. On the contrary, IgE antibodies are present in serum at the lowest concentration (less than 1%) and are used to identify proteins/peptides that cause allergic reactions (12).

Furthermore, there are two basic categories of antibodies: polyclonal and monoclonal. Polyclonal antibodies recognize multiple sites on the target protein. They may be used directly from the serum or purified over protein A/protein G affinity columns (standard polyclonal preparations have a high amount of contaminating proteins and antibodies). Monoclonal antibodies are derived from a cell line that has been isolated based on its ability to produce one antigen-specific antibody. Because of their purity and specificity, monoclonal antibodies are not usually put through an addition purification step and are preferred for ProteinChip studies. Alternatively, protein A or protein G can be bound to the array first to non-covalently enrich the antibody on the surface.

Prior to any experiment, it is imperative that all component parts of the reaction are validated. Therefore, the first step is examination of the antigen and antibody using either an

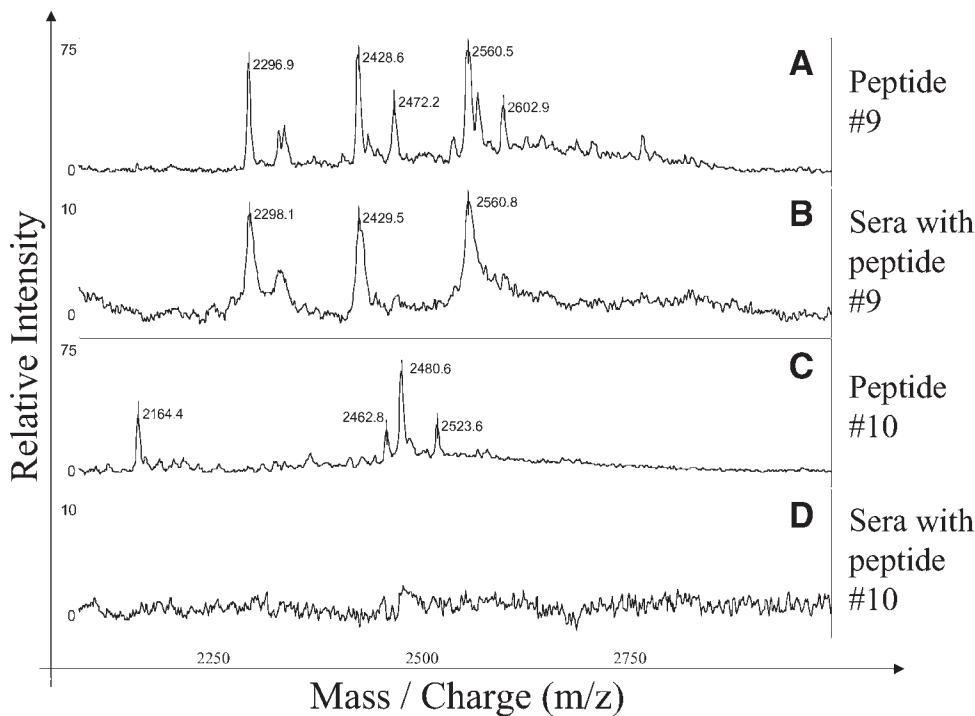


Fig. 3. Surface-enhanced laser desorption/ionization-time-of-flight mass spectrometry spectra of peptides captured by immunoglobulin (Ig)E-conjugated ProteinChip®. The peptides indicated as #9 and #10 deriving from Hev b 5 allergen were spotted on H4 ProteinChip and are shown in panels a and c. The same peptides were separately incubated on the IgE ProteinChip prepared as described in the text (panels B and D). The mass/charge (m/z) values of each detected species are reported.

Table 2
Sera Reactivity Against Peptides Deriving From the Latex Allergen Hev b 5

PS2	Peptide number																		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	20	21	
Serum	—	—	—	—	+/-	—	—	—	+	+/-	—	—	—	—	—	+	+/-	+	
A_Apo01	—	—	—	—	+/-	—	—	—	—	—	—	—	—	—	—	—	+/-	+	
Patient 1	+	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	
Patient 2	—	—	—	—	—	—	—	—	+	—	—	—	—	—	—	—	+	—	
Patient 3	—	—	—	—	+/-	—	—	+/-	—	+/-	—	—	—	—	—	—	+/-	+	
CBER S2	+	—	—	—	+/-	+/-	—	+	+	—	—	—	—	—	—	—	—	+	
CBER S3	+	—	—	—	—	—	—	+	+	—	—	—	—	—	—	+	+	+	

PS1																			
A_Apo01	—	—	—	—	+/-	—	—	—	+	—	—	—	—	—	—	—	—	+	
Patient 3	+	—	—	—	+/-	—	—	+	+	—	—	—	—	—	—	+/-	+/-	+	
CBER S2	+	—	—	—	+/-	—	—	+/-	+	—	—	—	—	—	—	+/-	+/-	—	
CBER S3	+	—	—	—	+/-	—	—	+	+/-	—	—	—	—	—	—	+	+	+	

The symbols +, +/- and – indicate a strong, weak, and minimal reactivity of sera against the peptides determined by the signal intensity observed by surface-enhanced laser desorption/ionization-time-of-flight mass spectrometry (SELDI-TOF MS) analysis.

A curve of intensity signal obtained from each peptide by SELDI-TOF MS was performed at different peptide concentrations.

IgE strong binding region

	21	KTEEPAPEADQTTPEEKPAE
8	PETEKAAEVEKIEKTEEPAP	
	9	EKIEKTEEPAPEADQTTPEE
	10	APEADQTTPEEKPAEPEPVA
15	AEEKPITEAAETATTEVPV	
16	PITEAAETATTEVPVEKTEE	
1	MASVEVESAAATALPKNETPE	
2	ATALPKNETPEVTKAETKT	

IgE weak binding region

5	ASEQETADATPEKEEPTAAP	
20	TPEKEEPTAAPAEPEAPAP	
6	TPEKEEPTAAPAEPEAPAPE	

Fig. 4. Sequences of weak and strong epitopes from Hev b 5 allergen. The sequences of strong and weak immunoglobulin (Ig)E-binding peptides determined by surface-enhanced laser desorption/ionization-time-of-flight mass spectrometry are listed. The IgE strong binding regions are enclosed in a shaded box, and the IgE weak binding regions are in gray.

H4 or NP chip (in some cases, both may be necessary) in order to ascertain the quality of the components. This is accomplished by spotting the material directly on the ProteinChip surface, letting it dry, and then washing the surface with 5% acetonitrile or water for H4 or NP surfaces, respectively. Finally, sinapinic acid is added and the spot analyzed on the mass spectrometer. For example, if there appear to be a large number of molecular-weight species not representative of IgG or if a 45-kDa species, indicative of IgG breakdown, is present in large amounts, it is absolutely necessary to purify the antibody solution either by protein A or G affinity chromatography or directly on the ProteinChip (discussed later).

2. There are several points to raise that may increase the likelihood of success in the methods described. If the antibody is captured directly on a preactivated surface, it may be advisable to try both PS1 and PS2 surfaces to determine which binds optimally. In addition, more stringent washing of the chip surface with higher concentrations of either detergent and/or salt can increase the specificity of antigen binding. Lastly, there may be an advantage to employing protein A for capture of antibody in certain circumstances, although we normally do not use it.
3. The ProteinChip immunoassay has the potential to monitor the quantity of antigen present in the sample analyzed. Furthermore, information about its biologically active form can be visualized, and one can potentially monitor fragments that could be useful in following the biological regulation of these molecules *in vivo*. An example of this potential application is the identification of biomarkers in cerebrospinal fluid. Cystatin C has previously been shown to be a biomarker for pain, using a polyclonal antibody to distinguish expression levels in patients vs controls (13). The development of a cytokine assay using ProteinChip technology represents another example of its application. The monitoring of cytokines from secretory fluids through the use of cytokine capture surfaces

has the potential to facilitate the diagnosis and/or the treatment of several disease states (14). Furthermore, this approach has been helpful in the study of protein–protein interactions (6).

4. These approaches are known to be technically limited by the concentration of the antibody that can be immobilized on the ProteinChip surface and by the volume of antigen containing sample that can be analyzed. In addition, the analysis of multiple samples on a single chip could result in cross-contamination between sample spots, even if a bioprocessor is used (15).

Acknowledgments

The authors thank Dr. Sandy Markey for his continued support.

References

1. Hutchens, T. W. and Yip, T.-T. (1993) New desorption strategies for the mass spectrometric analysis of macromolecules. *Rapid Commun. Mass Spectrom.* **7**, 576–580.
2. Merchant, M. and Weinberger, S. (2000) Recent advancements in surface enhanced laser desorption/ionization time of flight mass spectrometry. *Electrophoresis* **21**, 1164–1167.
3. Weinberger, S. R., Morris, T. S., and Pawlak, M. (2000) Recent trends in protein biochip technology. *Pharmacogenomic* **1**, 395–416.
4. Issaq, H. J., Veenstra, T. D., Conrads, T. P., and Felschow, D. (2002) The SELDI-TOF MS approach to proteomics: protein profiling and biomarker identification. *Biochem. Biophys. Res. Comm.* **292**, 587–592.
5. Carter, D., Douglass, J. F., Cornellison, C. D., et al. (2002) Purification and characterization of the gammaglobulin/lipophilin B complex, a promising diagnostic marker for breast cancer. *Biochemistry* **41**, 6714–6722.
6. Caputo, E., Camarca, A., Moharram, R., et al. (2003) Structural study of GCDFP-15/gp17 in disease *versus* physiological conditions using a proteomic approach. *Biochemistry*, **42**, 6169–6178.
7. Caputo, E., Moharram, R., and Martin, B. M. (2003) Methods for on-chip protein analysis. *Anal. Biochem.* **321** (1), 116–124.
8. Slater, J. E., Paupore, E. J., and O’Hehir, R. E. (1999) Murine B-cell and T-cell epitopes of the allergen Hev b 5 from natural rubber latex. *Mol. Immunol.* **36**, 135–143.
9. Askasaw, A., Hsieh, L. S., Martin, B. M., and Lin, Y. (1996) A novel acidic allergen, Hev b 5, in latex. Purification, cloning and characterization. *J. Biol. Chem.* **271**, 25,389–25,393.
10. Slater, J. E., Vedick, T., Arthur-Smith, A., Trybul, D. E., and Kekwick, R. G. (1996) Identification, cloning and sequence of a major allergen (Hev b 5) from natural rubber latex (*Hevea brasiliensis*). *J. Biol. Chem.* **271**, 25,394–25,399.
11. Frazer, J. K. and Capra, J. D. (1999) Immunoglobulins: structure and function. In Paul, W. E. (ed), *Fundamental Immunology*, chapter 3, Lippincott-Raven Publishers, Philadelphia.
12. Berzofsky, J. A., Berkower, I. J., and Epstein, S. L. (1999) Antigen-antibody interactions and monoclonal antibodies. In: (Paul, W. E., ed.) *Fundamental Immunology*, Lippincott-Raven, Philadelphia.
13. Mannes, A. J., Martin, B. M., Yang, H. Y., et al. (2003) Cystatin C as a cerebrospinal fluid biomarker for pain in humans. *Pain* **102**, 251–256.
14. Bai, H., Buller, R. M. L., Chen, N., and Boyle, M. D. P. (2002) Viral binding proteins as antibody surrogates in immunoassays of cytokine. *BioTechniques* **32**, 160–171.
15. Saouda, M., Romer, T., and Boyle, M. D. P. (2002) Application of immuno-mass spectrometry to analysis of a bacterial virulence factor. *BioTechniques* **32**, 916–923.

Protein and Peptide Microarray-Based Assay Technology

Scott T. Clarke

1. Introduction

The vast majority of the current technology in the field of microarrays focuses on the field of genomics, with gene expression being the area of greatest commitment, leaving the field of protein microarrays still in its infancy. In fact, a recently published book on microarray analysis dedicates only two pages under the chapter of novel microarray technologies to the topic of protein microarrays (1). The use of antibody-based microarrays for the detection of both research and clinically relevant markers has increased in popularity over the last several years, as protein-compatible surfaces have become commercially available. As with all new technologies, there are multiple steps in the process of producing protein microarrays that still need to be proven. An overview of the process, from the printing of the microarrays through the testing and validation of a system follows:

Printing of microarrays can be described as the deposition of a known volume of solution onto a solid substrate. Printers can be of two basic classes; contact or contactless. By far the largest number of printers available on the market are contact printers.

The basic process for contact printing uses small stainless steel pins that are dipped into the protein solution and tapped on the slide substrate at specific addressable locations. These pins can be either split or solid, and function much like a quill ink pen through the release of a discrete volume of liquid as a function of the tapping force and the contact time with the substrate. Multiple pins arranged in blocks compatible with 96/384-well plates allow for rapid printing on the substrate. Considerations for contact printers include versatility in programming, environmental control, robot precision, efficiency of pin wash stations, and air particulate contamination control. Protein printing requires refrigeration on the deck where the 96/384-well source plate and substrates are held, and air temperature and humidity control. The biggest influences on the quality of contact dispensing are the choice of printing buffer and the drying time of the dispensed spot. There are many commercially available printing buffers that have been optimized for specific substrate attachment, and these will be discussed briefly in the following section. Evaporation rates resulting from high temperature and low humidity can become a printing variable on long printing runs, and result in poor spot morphology and unknown dispense volume. Additional factors that can affect the print quality are the methods by which the substrate is held to the deck. Vacuum manifolds

that hold the glass fixed to the deck tend to give more consistent array placement than clip assemblies. The number of times that a pin needs to be blotted, and the number of spots that can be dispensed before a pin is re-dipped, are variables that need to be optimized for each machine, the pin type, and printing buffer. Split pins can print approx 80–100 spots before wicking of the solution begins to significantly alter spot morphology (Scott Clarke, unpublished data). When using multiple pins, the pins must be matched for delivery volume consistency. Once matched, care should be taken to maintain the integrity of the set. Thorough washing, sonication, drying, and periodic microscopic examination of the pin tip are required to maintain optimum performance. Wash programs and buffers must be optimized to reduce sample carryover to an acceptable level. Typically, wash buffers that work best do not alter the wicking and dispensing characteristics of the pins. Detergent and surfactants tend to have adverse effects on dispensing. Some manufacturers recommend pin conditioning by printing blocking protein, e.g., bovine serum albumin (BSA) prior to printing a protein array. Stable vibration-free work surfaces are usually necessary for good spot morphology.

Ink-jet and piezo-electric dispensing are the two main forms of contactless printing. A falling drop does not damage the gel-coated substrates that may be more suitable for use in protein microarrays. Currently, ink-jet printing makes use of fluidic handling where a syringe pump builds pressure behind a microsolenoid, which then opens and delivers a drop. This technology is currently best adapted to delivering many replicates of a single sample in the liquid-handling mode. One of the drawbacks of this method of dispensing is higher sample volume requirements. Also, since samples flow through much of the tubing, extensive line washing is required to prevent sample carryover. Piezo-electric employs a glass capillary with a ceramic piezo-collar. When voltage is applied, the piezo device exerts a force on the capillary and a drop is expelled. The drop volume is controlled by the pressure and is typically adjusted to deliver 333 picoliter on a commercially available system. Dispensing volumes tend to be very precise, since it is not dependent upon a wicking effect but rather on applying pressure to a fluidics line. These systems are well suited for microarraying applications, which demand high accuracy and precision in the spot volume and placement. Since the sample is drawn up into the capillary tip, volume requirements are very low. Sample carryover after standard washing is very low. Typically, 0.9 μ L is drawn up into the tip and is sufficient to dispense hundreds of spots. Spot morphology of a piezo-dispensed microarray is likely the best of all printing methods when environmental conditions are controlled. The main disadvantage of piezo dispensing is the requirement for a clean-room environment to achieve trouble-free dispensing. Additionally, all samples and wash liquids must be filtered to prevent fouling of the piezo-tip. Printing rates of commercially available piezo-electric printers are typically similar to single-pin contact printing and are limited by the low number of tips on the printer. Tip replacement and repairs take much more time and expense than on contact printers.

A number of attachment chemistries on glass or plastic have been taken from DNA microarrays and adapted to attaching proteins. Nonspecific as well as covalent oriented attachment schemes have been employed for protein arrays, with varying degrees of success (2). Aldehyde, epoxy, and *N*-hydroxysuccinimidyl (NHS) ester coupling all make use of reactive groups on the microarray substrate for covalent attachment of protein to the glass surface. Streptavidin, protein A, and protein G substrates have been

tested for affinity binding of biotinylated proteins or antibodies. His-tagged proteins have been attached to substrates containing nitrilotriacetic acid (NTA) functional groups. The advent of gel-coated microarrays on the market has permitted increased stability and sensitivity for use of proteins as a detection platform. A second generation of gel-coated surfaces including poly(ethylene glycol)-epoxy and brushy dendrimeric surfaces have been specifically designed to stabilize proteins (3). Efficiency of attachment and performance is typically measured by either sandwich type of immunoassay or by direct detection of dye-conjugated proteins.

Antibody attachment through the epsilon amino of lysine is achieved using dispensing buffer with the pH adjusted to between 8.0 and 9.3 (typically 50 mM sodium carbonate buffer) on aldehyde, or NHS ester slide chemistry. Neutral pH buffers like phosphate-buffered saline (PBS) can be successfully used on gel-coated and plastic microarrays that bind through electrostatic interactions and do not have pH-dependent attachment chemistry. Dispensing buffers or samples used for printing should not contain free amines (e.g., Tris). Protein samples that contain Tris buffer need to be dialyzed against PBS or another compatible buffer. After printing, the slides are usually treated to an overnight incubation step in a 70% humidity chamber to complete the chemical coupling reaction. The humidity is achieved using a water-saturated solution of NaCl in a closed environment. After completion of the chemical coupling step, the slides are blocked prior to use so unreacted functional groups will not react with sample in the subsequent steps.

Peptide arrays are less common than protein arrays. Attachment of peptides to microarray surfaces presents greater problems, since the functional groups used for attaching proteins to the substrate can often be missing from the peptide sequence. Custom peptides synthesized to include cysteine, lysine, or biotin have been used for attachment to a variety of surfaces. Cysteine-modified peptides can be attached to thiol-reactive species of self-assembled monolayers (SAMs) on gold surfaces (4,5). Peptides with additional lysines can be attached using NHS chemistry, described in the section on protein attachment. Nonspecific adsorption by electrostatic interaction can be effective for attaching peptides without needing a tag-specific binding sequence (6). Phosphorylated amino acid dispensed on HydroGel microarrays have been successfully detected with labeled antibodies (Pastula, C., and Johnson, I., unpublished data). Attachment efficiency needs to be validated if quantitative data are going to be achieved. Dispensing of Alexa Fluor® 488-biotin conjugate (cat. no. P-12924, Molecular Probes) and performing pre- and post-wash scans is a simple method for determining binding efficiency of small molecules.

Several methods of incubating the target protein on the microarray are commercially available. HybriWell™ hybridization sealing systems are available in a variety of sizes and have a water-tight adhesive coating around the perimeter of the chamber. Access ports allow for easy filling of the array with the target solution. Adhesive seal tabs are used to close the chamber during the incubation and prevent evaporation of the solutions on the array.

Several devices have been designed that can clamp onto a slide to divide the slide into multiple wells. These have the advantage of allowing for multiple reactions on the same slide. There are also a limited number of array surfaces that are compatible with the frame of a 96/384-well plate. They attach to the frames with adhesive or clamps.

The disadvantage of this larger format is that they cannot be read on standard microarray scanners, but require plate scanners, which have lower resolution or issues with stitched images.

2. Materials

2.1. Printing of Protein or Peptide Microarrays

1. 384-Well Proxy plate (cat. no. 6006280, Perkin Elmer).
2. Uniseal sealant tape (cat. no. 7704-0009, Whatman).
3. HydroGel slides (cat. no. 605001, Perkin Elmer).
4. Phosphate-buffered saline adjusted to pH 8.3 with NaOH (see Note 1).
5. 18 Mega-ohm water.
6. Antibody or peptide stock solutions.

2.2. Antibody-Based Assay on PARAGON™ Printed Microarray Using Multiwell ProPlate Device

1. 30-mL slide-holder tubes (cat. no. 240-5420 G8K, Evergreen).
2. Blocking buffer: 0.05 M Tris-HCl (pH 7.5), 0.15 M NaCl, 0.25% Mowiol® 4-88, 0.2% Tween-20, 0.5% bovine serum albumin (see Note 1).
3. Washing buffer: 0.05 M Tris-HCl (pH 7.5), 0.15 M NaCl, 0.25% Mowiol® 4-88.
4. ProPlate™ multiarray slide module (cat. no. 204862, Grace BioLabs).
5. Alexa Fluor 488 goat antimouse immunoglobulin (IgG (H+L) conjugate (cat. no. A-11001, Molecular Probes) (see Note 2).
6. Alexa Fluor 488 goat antimouse IgM conjugate (cat. no. A-21042, Molecular Probes).

2.3. Labeled-Antibody Assay on Two-Pad HydroGel

1. 30-mL slide-holder tubes (cat. no. 240-5420 G8K, Evergreen).
2. 22 × 22 mm HyrbiWell (cat. no. H-24723, Molecular Probes).
3. Blocking buffer: 0.05 M Tris-HCl (pH 7.5), 0.15 M NaCl, 0.25% Mowiol® 4-88, 0.2% Tween-20, 0.5% bovine serum albumin.
4. Washing buffer: 0.05 M Tris-HCl (pH 7.5), 0.15 M NaCl, 0.25% Mowiol 4-88.

2.4. Printing of Protein Microarrays

1. BioChip Arrayer® Piezo Dispenser.
2. Clean-room gloves.
3. High-efficiency particulate air (HEPA) positive-pressure hood.
4. 16-Tip multichannel pipettor.
5. Pipet tips.
6. Enzyme-linked immunosorbent assay (ELISA) reagent trough.
7. 0.2-μm Filters.
8. 1-mL Disposable syringes.
9. Microfuge.
10. Vortex.
11. Lint-free alcohol wipes.
12. Compressed air.
13. 0.5-mL Microfuge tubes.

2.5. Antibody-Based Assay on PARAGON Printed Microarray Using Multiwell ProPlate Device

1. Forceps and gloves for handling slides.
2. Rotary or reciprocal rocker.

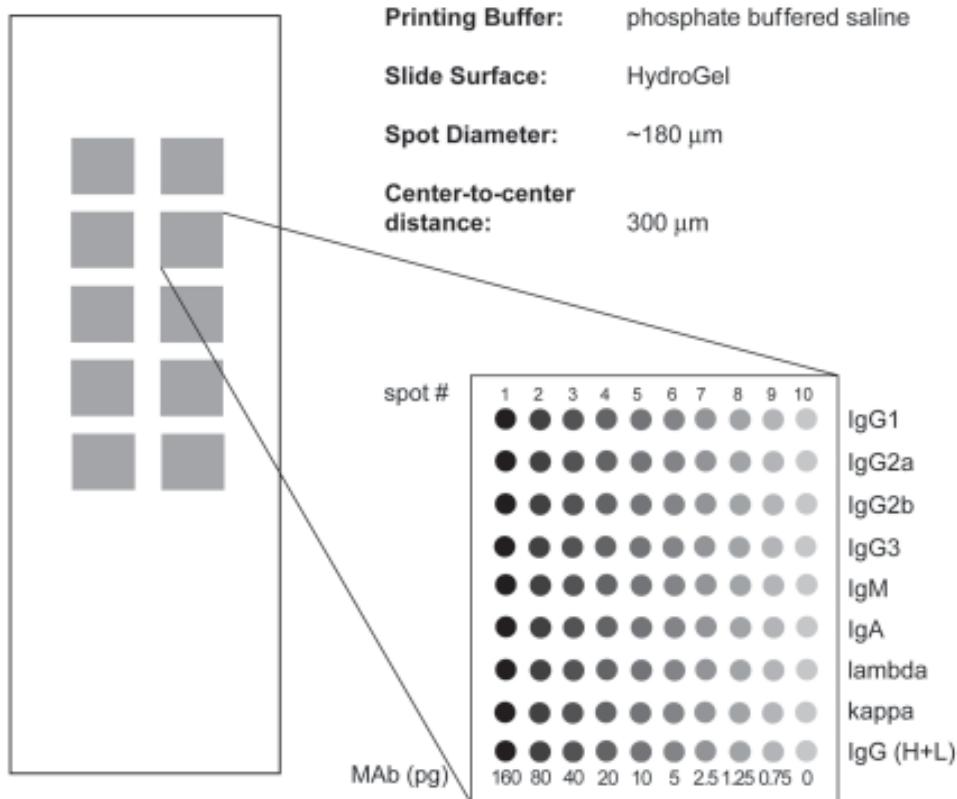


Fig. 1. Layout of a microarray printed in 10 blocks designed to fit the ProPlate device. Blocks of goat anti-mouse isotype-specific antibodies are printed in rows of serial one-half dilutions on HydroGel slides.

3. Centrifuge with a slide-holder adapter (e.g., Telechem part no. HTWSR) or a microfuge with a slide holder (Laboratory Products Sales C 1303-T).
4. Microarray scanner.
5. Microarray data analysis software.

2.6. Labeled-Antibody Assay on Two-Pad HydroGel

1. Forceps and gloves for handling slides.
2. Rotary or reciprocal rocker.
3. Centrifuge with a slide-holder adapter or a microfuge with slide holder.
4. 500-mL squeeze bottle.
5. Microarray scanner.
6. Microrray data analysis software.

3. Methods

3.1. Printing of Protein Microarrays

The following method has been optimized for the preparation of antibody-based microarray dispensed on HydroGel slides using a piezo-electric type of arraying (Biochip Arrayer). This process is performed in a clean-room environment with particular attention given to controlling the environmental factors which can affect spot morphology. This method prepares microarrays with ten blocks of twofold serial dilutions of antibody in ten-block format, which is compatible with the ProPlate device (Fig. 1).

1. Assemble all the materials needed for creating a source plate in a positive-flow hood.
2. Using the compressed air, blow out all containers to be used for making the source plate, including microfuge tubes and source plates.
3. Wipe off all surfaces and equipment in the hood with alcohol-based lint-free wipes.
4. Filter a stock solution of pH-adjusted PBS for diluent and pour into ELISA trough.
5. Centrifuge all stock solutions of antibodies at 1000g for 10 min prior to diluting.
6. Prepare 50 μ L stock solution of each antibody to be used by diluting down to 0.5 mg/mL with PBS (pH 8.3).
7. Filter the antibody stock solution through 0.2- μ m syringe-tip filter (*see Note 4*).
8. Add 15 μ L of PBS diluent to each of the wells to be used for diluted antibody solution (*see Note 5*).
9. Add 30 μ L of 0.5 mg/mL antibody to first column of the 384-well plate.
10. Serially dilute one-half using the multichannel pipettor through the series of samples.
11. Apply the sealing tape to the 384-well plate.
12. Centrifuge the plate immediately prior to printing at low speed for 2 min to remove air bubbles from the bottom of the wells and condensate from the sides and sealing tape.
13. Set the source plate on the deck of the arrayer.
14. Perform the necessary print verification checks required for daily maintenance of the arrayer.
15. Adjust humidity of the printing environment to between 40 and 50%.
16. Set the temperature of the deck refrigeration unit to two degrees above the dew point of the local environment (*see Note 6*).
17. Load the slides onto the arrayer deck, after first blowing off any glass fragments using compressed air. Pay particular attention to aligning the glass substrate and checking that the slides are flat on the deck.
18. Connect the vacuum line and coolant lines.
19. Remove the seal tape from the source plate.
20. Load and run the printing program.
21. Inspect the arrays for missed spots after the printing is completed. For examination of unlabeled proteins, the best method to look for spotting errors is to scan the slide in a reflective mode. The image is similar to those seen with differential interference contrast (DIC) microscopy. Set the excitation wavelength of the scanner to a longer wavelength (e.g., 488 nm) than the emission wavelength (532 nm) and boost the laser setting to 100% and the photomultiplier (PMT) to 50%. By adjusting the brightness and contrast settings, the arrayed pattern can be observed.
22. Create and run a recovery print file, if necessary. The recovery file is a unique print program, which will re-dispense any missed spots on the array.
23. Turn off vacuum lines and place the printed slides into a clean microscope box.
24. Place the microscope slide box in a 70% humidity chamber overnight.
25. Remove the slides from the humidity chamber and store at 4°C until used.

3.2. Printing of Peptide Microarrays

Printing of peptide arrays on HydroGels is quite straightforward and follows the same steps described in **Subheading 3.1.** with the following exceptions:

1. Dilute peptide to 2.0 mg/mL in filtered 18-mega-ohm water (*see Note 7*).
2. Stock solutions of peptides should be prepared fresh the day of printing.
3. Print as described in **Subheading 3.1.**
4. Place the microscope slide box in a 70% humidity chamber overnight.
5. Store printed slides in humidity chamber at room temperature until used.

3.3. Antibody-Based Assay on PARAGON Printed Microarray Using Multiwell ProPlate Device

The method of treatment of an antibody-based microarray is similar to those used in sandwich-styled ELISA techniques (8). After initially blocking and rinsing, the primary target is incubated on the arrays, followed by a wash and incubation with a secondary reporter antibody-dye conjugate or biotinylated antibody followed by a streptavidin-dye conjugate incubation step. After sufficient incubation with reporter antibody, the array is washed, rinsed, spun dry, and scanned. The following method describes the use of a ProPlate device on a 10-block array printed with dilutions of goat anti-mouse antibodies.

1. Remove the sealing tape from the top of the wells of the microarray attached to the ProPlate device. Add 200 μ L of blocking buffer to each of the 10 wells. Reseal the fixture with sealing tape and rock gently on a reciprocal or rotary shaker at room temperature for 2 h or overnight at 4°C.
2. After blocking, the sealing tape is removed and the wells are rinsed either manually with a squeeze bottle filled with wash buffer or using an automated plate washer. To rinse the device manually, empty the wells by flicking the device over the sink. Blot vigorously several times on clean absorbent paper. Refill the wells with M/TBS and repeat the rinsing process several times to ensure that any unbound capture antibody is removed from the slide surface. If using an automatic plate washer, adjust the tip positions to contact the outside edge of the wells. This avoids damaging the arrayed surface printed along the inside edge.
3. Add 75 μ L of appropriately diluted mouse antibody sample or control sample to each of the wells (see Note 8). A typical dilution of 1/100 hybridoma cellular supernatant in blocking buffer gives good results. Reseal the wells with sealing tape (component B). Place the fixture on a reciprocating rocker or rotation platform for 1.5 h. For mouse antibody concentrations of 1 μ g/mL, a significant amount of antibody binds within the first 10 min of addition (see Fig. 2).
4. Wash out the wells to remove the unbound target using multiple rinses with M/TBS. Blot the multiwell chamber onto clean absorbent paper each time it is emptied. Using the normal amount of precaution, cross-contamination between wells will not occur during the washing steps. Thoroughness of washing at this step will help to reduce nonspecific background. Rinse and blot a minimum of six times.
5. Add 75 μ L of a mixture containing 200 ng/mL of Alexa Fluor 488 dye conjugated goat antimouse IgG (H+L) plus 200 ng/mL Alexa Fluor dye conjugate goat antimouse IgM diluted antibody, diluted in blocking buffer, to each well, and incubate for 1 h on a rocking platform. Protect the fixture from light during this and all subsequent steps.
6. After incubation with the secondary antibody, the wells are rinsed six times with washing buffer using the same flicking and blotting method described above. Fill a wash chamber with 30 mL of washing buffer. Remove the slide from the device and insert into the slideholder tube (see Note 9).
7. Place slide on a reciprocal rocker platform protected from light and rock for 30 min. Wash slide by pouring off the blocking solution and refilling the wash chamber three times for 5 min each with washing buffer followed by three de-ionized water rinses to remove salts from the glass. Using forceps, remove the slide directly from the water of the wash chamber and spin dry for 5 min at 1200 rpm in a tabletop centrifuge or in a microfuge fitted with a slide adapter (see Note 10).

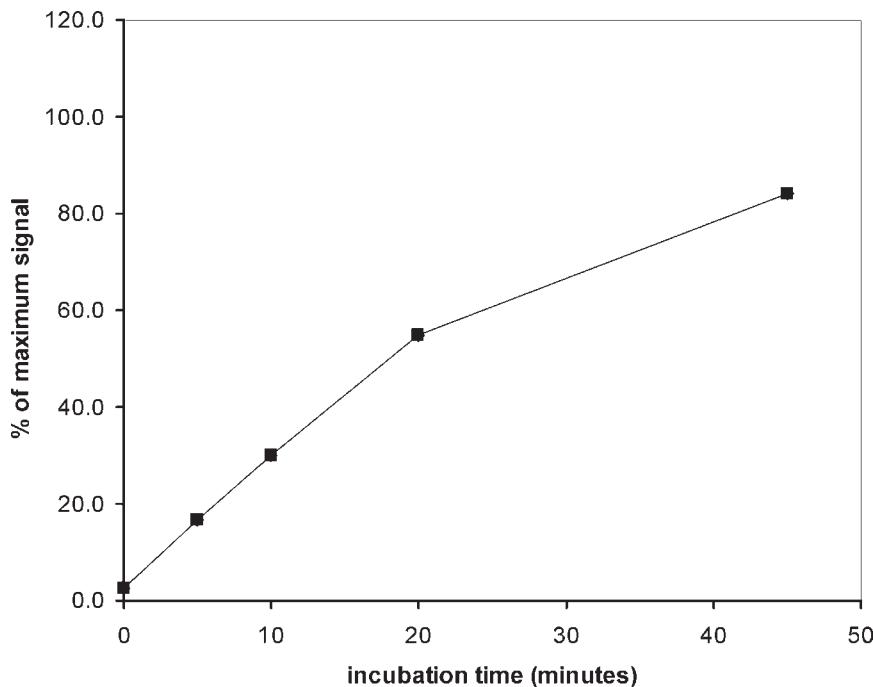


Fig. 2. Kinetics of binding of target protein to antibody microarray. A time course of addition of 100 ng/mL mouse antibody followed by incubation with Alexa Fluor® 488 goat antimouse IgG (H+L)-labeled reporter antibody (Cat. no. A-11001, Molecular Probes) shows that binding can be detected within 5 min after addition of the target protein.

3.3.1. Detection of Labeled Antibody on PARAGON Two-Pad HydroGel Array

An array printed with a series of primary isotypes can be used to screen the selectivity and cross-reactivity of labeled secondary antibody (see Fig. 3). Biotin-labeled secondary antibody can be imaged by adding an additional incubation step with labeled streptavidin. Enzyme-labeled antibody (e.g., horseradish peroxidase [HRP]) can be imaged using precipitating substrates like tyramide signal amplification (TSA) (cat. no. T-20912, Molecular Probes). The arrayed pad is covered with a HybriWell and injected with a labeled antibody. After incubation, the HybriWell is removed and the slide is washed and prepared for imaging.

1. Block the slide in 30 mL of blocking solution in a slide-holder tube. Fill the tube with blocking solution, then insert the slide with the barcode towards the top of the tube. Replace the cap and rock for 2 h at room temperature.
2. Decant the blocking solution, holding the cap over the top of the tube to prevent the slide from falling out.
3. Rinse the slide with 3 × 30 mL of washing solution, capping and inverting several times for each rinse.
4. Rinse the slide with 3 × 30 mL of de-ionized water, capping and inverting several times for each rinse. After the last rinse, fill the tube with de-ionized water.
5. Remove the slide from the filled tube and place in a centrifuge or microfuge adapted for slides. Spin at 1200 rpm for 5 min or until the pads are no longer wet. Do not let the slide dry out prior to spinning.

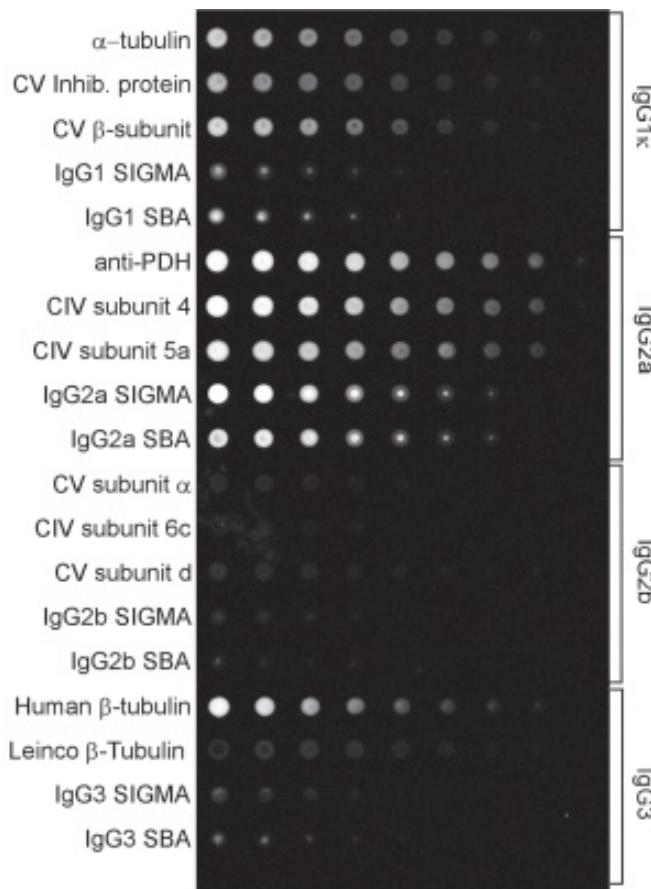


Fig. 3. Antibody microarray showing signal dilution serials response of Alexa Fluor® 488 goat antimouse immunoglobulin (Ig)G2a conjugate to a panel of primary mouse antibodies. Serial dilutions of 80 to 0.75 pg/spot of a panel of isotype-specific mouse antibodies are arrayed on a HydroGel surface. Antibodies are arrayed in blocks by isotypes. The binding selectivity of the labeled GAM shows a wide response to the isotype-specific array. Inconsistencies in the predicted response can be visually observed using antibody arrays and help to explain anomalous immunohistochemical results. Human β -tubulin, a mouse IgG3, shows strong binding of the GAM IgG2a.

6. Remove the slide from the centrifuge and place on a clean surface of Parafilm.
7. Apply the HybriWells to each of the arrayed pads by first removing the protective film from the HybriWell and carefully centering the HybriWells over the pads prior to contact with the glass.
8. Seal the outside adhesive edges of the HybriWells with the back end of the forceps or other blunt instrument.
9. Dilute 200 μ L labeled secondary antibody to 10–1000 ng/mL in blocking buffer.
10. Slowly inject 100 μ L of diluted antibody into one of the ports until the HybriWell is filled (see **Note 11**).
11. Cover the chamber with the seal ports.
12. Place slide on rotating or reciprocal device for 1.5 h. Protect the slide from light during this and all other incubation steps.

13. Remove the HybriWell from each of the pads using a pair of strong forceps. Prepare a blotting surface using lint-free wipes on a fresh piece of Parafilm. Slowly lift the HybriWell, allowing the antibody solution to run off the slide and be absorbed on the blotting surface.
14. Briefly rinse the pads with a squeeze bottle containing blocking solution, taking care not to mix solutions of the two pads.
15. Place the slide in a wash tube containing 30 mL of blocking buffer.
16. Rinse twice with blocking buffer after capping, and invert several times.
17. Replace with 30 mL of blocking buffer and incubate while rocking for 30 min. Protect the sample from light.
18. Decant the blocking buffer and rinse three times with wash buffer, capping and inverting between rinses.
19. Rinse the slide with 3×30 mL of de-ionized water, capping and inverting several times for each rinse. After the last rinse, fill the tube with de-ionized water.
20. Using forceps, remove the slide directly from the water of the wash chamber and spin dry for 5 min at 500g in a tabletop centrifuge or in a microfuge fitted with a slide adapter.

3.3.2. *Detection of Kinase Activity on PARAGON Kinase Array*

Detection of kinase activity on peptide arrays is described in detail in Chapter 45, with applications of Pro-Q® Diamond phosphoprotein microarray stain for visualizing kinase modified peptides.

3.4. *Microarray Analysis and Data Quantitation*

Microarray slides are scanned dry using a laser scanner, such as the Axon GenePix 4200A scanner, Fuji FLA-8000, or Perkin Elmer ScanArray 5000 set for the appropriate dye excitation and emission, according to manufacturer's specifications. Scanners are now available that have three and four color lasers, so a number of dyes can be used for labeling. For slides labeled with Alexa Fluor 488, use excitation wavelength = 488 nm and emission wavelength = 522 nm. For slides labeled with GAM Alexa Fluor 555, use excitation wavelength = 532 nm and emission wavelength = 570 nm. For slides labeled with Alexa Fluor 647, use excitation wavelength = 633 nm and emission wavelength = 660 nm.

1. Turn on scanner and ignite the lasers. The ScanArray 5000 has three lasers that require a minimum of 15 min warmup prior to scanning for the green, yellow, and red HeNe lasers. The blue Argon laser only requires 5 min warmup. The solid-state lasers on the Axon scanner ignite automatically when the machine is switched on. The Fuji lasers need to be warmed up 15 min.
2. Create a scan protocol selecting the lasers that correspond to the fluorescent dyes on the arrays. Set laser and PMT corresponding to 75% of the maximum signal on the array.
3. Select the area of interest to scan and the scan resolution. 10 μm resolution is usually sufficient for feature diameters of 180 μm or greater.
4. Insert slide into scanner. For the Axon scanner, the microarray surface is face down. For the ScanArrayer 5000 the arrayed surface is placed into the machine face up.
5. Start the scan and check that the signal is high enough without being saturated. For these scanners, the image is collected in a 16-bit format, and signal saturation is 65,535 relative fluorescent units. Reset the laser and PMT as necessary if the signal is saturated.
6. Save the image as either a single TIFF file, multiple TIFF, or Bitmap image. The single-image TIFF file and Bitmap images are the most useful for opening the image in other applications.

7. Using GenePix Pro software, open the image and define new blocks, which match the printed array pattern. A pitch of between 300 and 500 μm spacing is commonly used for low-density arrays. Piezo-printed feature diameter is generally about 180 μm . Save the block definitions with “Alt H” and chose the GAL file.
8. Open the GAL file in Excel and fill in the sample name and concentrations in the name and id columns. Save and close the Excel file and return to GenePix; apply the GAL file to the array with “Alt Y.” Sample name and concentration will automatically be assigned to each feature of the array.
9. Align image tool when two or more colored images are opened in the same window. GenePix Pro 5.0 permits four images to be opened simultaneously and ratiometrically compared with each other.
10. Find the features of the array using “Alt F5” to align all blocks to the features. Visually inspect the image to see that the block alignment fits the features.
11. Apply the automatic analysis function (Alt A) to the defined blocks. Ratiometric and statistically relevant data will be calculated and assembled into a spreadsheet.
12. Data can be exported into Excel for further analysis and graphing (see **Fig. 4** for graphic rendition of image in **Fig. 3**).

4. Notes

1. Other protein components can be added to the blocking solution, or left out all together. Other suitable nonionic detergents can be used.
2. Alexa Fluor goat anti-mouse conjugates are available in many other wavelengths, including Alexa Fluor 555 and Alexa Fluor 647 conjugates, which are compatible with two laser scanner systems.
3. 100 mM sodium carbonate buffer (pH 8.4) can be used in place of the pH-adjusted PBS as a printing buffer.
4. Some loss due to hold-up volume in the syringe will occur. For preparing the source plate, only 30 μL of stock is needed. To minimize the volume loss, pipet the unfiltered stock into the base of the filter, then attach the 1-mL syringe.
5. The piezo-electric printer uses four alternate rows (e.g., A, C, E, and G) in a 384-well plate to pick up sample for dispensing. Plate layout should be matched to the print program of the arrayer.
6. The dewpoint can be determined using online dewpoint calculators if the temperature and the humidity are known.
7. Small volumes of sodium hydroxide may be needed to solubilize some peptides prior to diluting in water.
8. Volumes can be successfully reduced to 35 μL .
9. To remove the slide from the fixture, unclip the rails from each side by peeling off the bottom edge of each rail away from the glass. After both rails are removed, grasp the glass by the edges and separate it from the silicon gasket. Immediately transfer the slide to the blocking buffer in the slideholder tube and close chamber.
10. To reduce background and streaking, do not pour off the water in the wash chamber, but remove the slide from the water.
11. To prevent bubbles from being trapped in the HybriWell, carefully tap the top of the well while injecting the sample into the chamber. This will help to distribute the liquid over the whole pad.

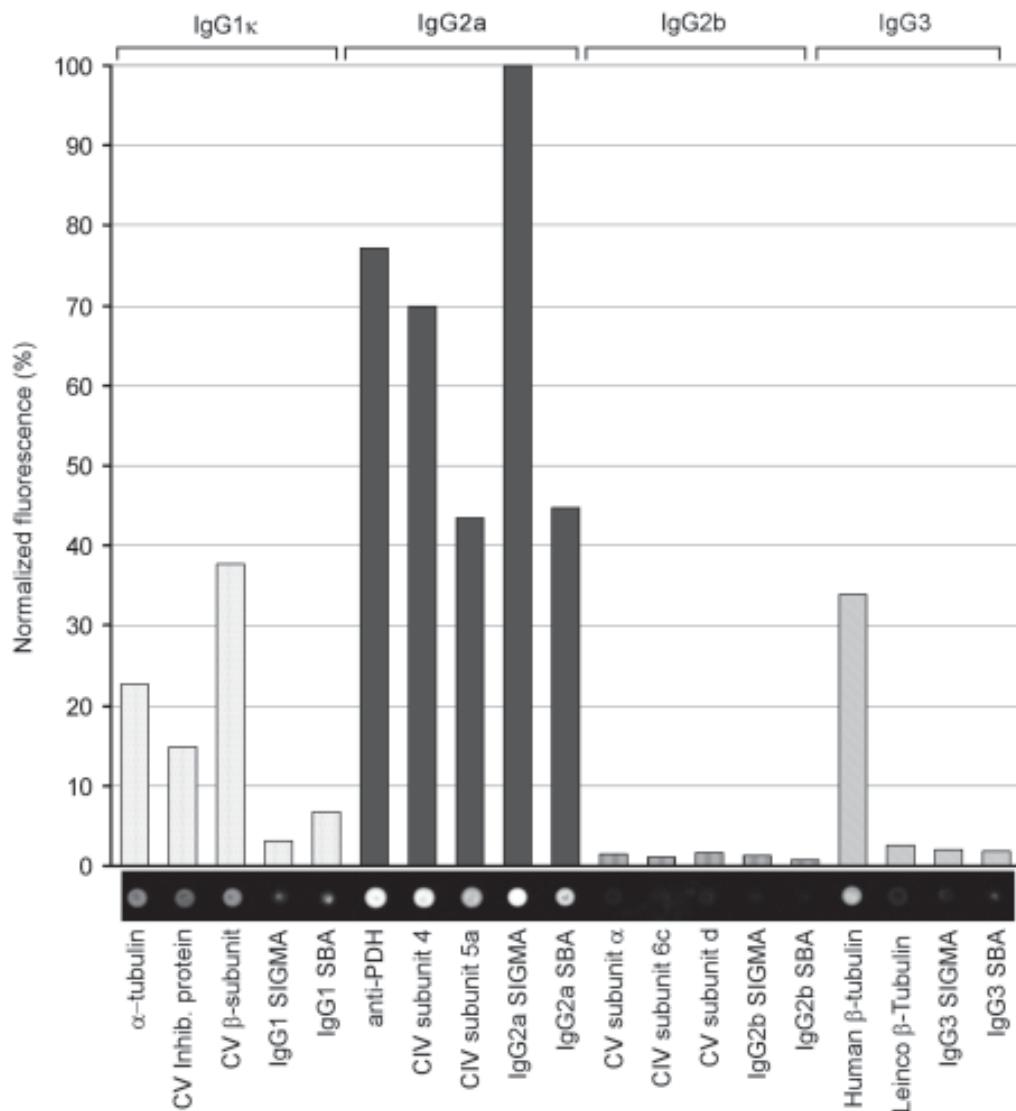


Fig. 4. Quantitation of fluorescence signal from GAM immunoglobulin (Ig)G2a binding to the isotype-specific array. Using Axon GenePix Pro 5.0 quantitation software, the signal intensities of the 40-pg/spot series of column two are calculated, and then plotted in Excel. Spot intensities are normalized to the isotype-specific IgG2a Sigma antibody.

References

1. Schena, M. (ed.), (2003) *Microarray Analysis*, John Wiley & Sons Inc., Hoboken, NJ.
2. Angenendt, P., Glokler, J., Murphy, D., Lehrach, H., and Cahill, D. (2002) Toward optimized antibody microarrays: a comparison of current microarray support materials. *Anal. Biochem.* **309**, 253–260.
3. Angenendt, P., Glokler, J., Sobek, J., Lehrach, H., and Cahill, D. J. (2003) Next generation of protein microarray support materials: evaluation for protein and antibody microarray applications. *J. Chromatogr. A* **1009**, 97–104.

4. Reimer, U., Reineke, U., and Schneider-Mergener, J. (2002) Peptide arrays: from macro to micro. *Curr. Opin. Biotechnol.* **4**, 315–320.
5. Lesaicherre, M. L., Uttamchandani, M., Chen, G. Y., and Yao, S. Q. (2002) Developing site-specific immobilization strategies of peptides in a microarray. *Bioorg. Med. Chem. Lett.* **12**, 2079–2083.
6. Martin, K., Steinberg, T. H., Cooley, L. A., Gee, K.R., Beechem, J. M., and Patton, W. F. (2003) Quantitative analysis of protein phosphorylation status and protein kinase activity on microarrays using a novel fluorescent phosphorylation sensor dye. *Proteomics* **7**, 244–255.
7. Ausubel, F. M. (ed.) (1995) *Short Protocols in Molecular Biology*, John Wiley & Sons, Inc., Hoboken, NJ.
8. Harlow, E., and Lane, D. (eds) (1988) *Antibodies—A Laboratory Manual*, Cold Springs Harbor Publication, Cold Springs Harbor, NY.

Production of Protein Microarrays Using Robotic Pin Printing Technologies

Ye Fang, Ann M. Ferrie, and Fang Lai

1. Introduction

The increased numbers of potential drug targets uncovered through genomics-based approaches have created a demand for screening technologies that enable robust and parallel analysis of many targets, given that it is costly to sort out the targets one by one. Array-based expression analysis (1) and mutation mapping (2) of many genes have made a major impact on biology and on drug discovery and development. Whereas genes contain the information of life, their encoded proteins perform nearly all the functions in the cell. Because of this, and the fact that proteins are the targets against which most drugs are designed, it therefore becomes obvious that nothing is more important than deciphering the functions of proteins. Together with genomics, advanced chemical technologies, and high-throughput screening, protein microarray technology has the potential to aid in understanding biological systems or system biology, as well as in developing tomorrow's new medicines.

Protein microarrays can be classified into protein-detecting microarrays (e.g., antibody microarrays) and functional protein microarrays (e.g., enzyme substrate microarrays and G protein-coupled receptor [GPCR] microarrays) (3–5). A protein-detecting microarray uses protein capture reagents as probes arrayed on the surface of a solid substrate. The protein capture reagents are capable of recognizing and interacting with their target protein(s) in a biological sample (e.g., a bio-fluid or a cell lysate). The capture reagents could be antibodies (6,7), antigens (8), ligands (9), carbohydrates (10,11), or glycolipids (12). This type of array could be used to profile protein abundance and/or modification, and to identify biomarkers of diseases. Functional protein microarrays, in contrast, use native proteins as probes arrayed on the surface. Arrays of this type are useful for parallel studies of the activity of native proteins, such as protein–protein and protein–small molecule interactions (13–15).

This chapter describes protocols for fabricating two distinct types of protein microarrays: antibody microarrays and GPCR microarrays, on γ -aminopropylsilane (GAPS)-modified surfaces using robotic pin printing technology.

2. Materials

2.1. Chemicals

1. C-reactive protein (CRP) (GenWay Biotech Inc., San Diego, CA).
2. Human growth hormone (hGH) (GenWay Biotech Inc.).
3. Recombinant human insulin (rHI) (GenWay Biotech Inc.).
4. Egg yolk IgY anti-CRP (IgY-CRP) (GenWay Biotech Inc.).
5. Egg yolk immunoglobulin (IgY) anti-hGH (IgY-hGH) (GenWay Biotech Inc.).
6. Egg yolk IgY anti-rHI (IgY-rHI) (GenWay Biotech Inc.).
7. Mouse monoclonal anti-CRP (mAb-CRP) (Fitzgerald Industries International Inc., Concord, MA).
8. Mouse monoclonal anti-hGH (mAb-hGH) (Fitzgerald Industries International Inc.).
9. Mouse monoclonal anti-rHI (mAb-rHI) (Fitzgerald Industries International Inc.).
10. Cy5-goat anti-mouse IgG (Cy5-IgG) (Amersham, Pharmacia Biotech, Piscataway, NJ).
11. Human β -adrenergic receptor subtype 1 (β 1) membrane preparation (Biosignal, Montreal, Canada).
12. Human β -adrenergic receptor subtype 2 (β 2) membrane preparation (Biosignal).
13. Human α -adrenergic receptor subtype 2A (α 2A) membrane preparation (Biosignal).
14. Human neurotensin receptor subtype 1 (NTR1) membrane preparation (PerkinElmer Life Sciences, Boston, MA).
15. Bodipy-TMR-CGP12177 (BT-CGP) (Molecular Probes, Eugene, OR).
16. Bodipy-TMR-neurotensin (BT-NT) (PerkinElmer Life Sciences).
17. Deionized water ($>18\text{ M}\Omega$; MilliQ-UV, Millipore, Bedford, MA).

2.2. Solutions

1. 1X phosphate-buffered saline (PBS): 10 mM sodium phosphate, 150 mM sodium chloride, pH 7.4 (Bio-Rad, Hercules, CA).
2. Blocking solution: 3% dry milk in 1X PBS.
3. Washing solution: 0.05% Tween-20 and 0.3% dry milk in 1X PBS.
4. Rinsing solution: 3% bovine serum albumin (BSA) in 1X PBS.
5. Cy5-antimouse IgG binding solution: 2 $\mu\text{g/mL}$ Cy5-antimouse IgG diluted in 1X PBS/3% BSA.
6. Capture antibody solution for printing: freshly prepared IgY-CRP, IgY-hGH, or IgY-rHI solution (100 $\mu\text{g/mL}$) in 60% (v/v) 1X PBS and 40% glycerol.
7. GPCR solution for printing: commercially available GPCR preparation reformulated in a buffer solution containing 10% sucrose and 10% glycerol to a final total protein concentration of 0.5–5 mg/mL (see Note 1).
8. GPCR binding buffer: 50 mM Tris-HCl (pH 7.4), 10 mM MgCl₂, 1 mM ethylenediaminetetraacetic acid (EDTA), and 0.1% BSA. This should be used within the same day.
9. Labeled ligand storage solution: 2 μM Bodipy-TMR-CGP 12177 or Bodipy-TMR-neurotensin in dimethylsulfoxide (DMSO). These solutions can be stored at -80°C for up to 6 mo (see Note 2).
10. Compound storage solution: 1 mM CGP 12177, ICI 118551, neuromedin N, or neurotensin in DMSO. These solutions can be stored at -20°C for up to 6 mo.

2.3. Instruments and Consumables

1. Genipix 4000B scanner (Axon Instruments, Union City, CA).
2. Pixsys 5500C Arrayer (Cartesian Technologies, Irvine, CA).
3. Quill pin CMP3 (Telechem, Atlanta, GA).
4. Pipettors and pipet tips.

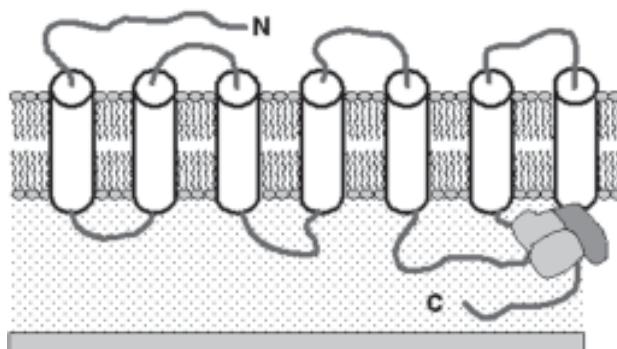


Fig. 1. Ideal representation of a G protein-coupled receptor (GPCR) immobilized on a surface in a microarray. The GPCR is associated with lipid bilayer membranes and with a heterotrimeric G protein. The receptor is immobilized on the surface such that the extracellular side faces the solution and the receptor is offset from the surface by the hydration layer between the membrane and the surface.

5. Vortexer.
6. White low-volume 384-well microplate (cat. no. 3674, Corning Inc., Acton, MA).
7. GAPS slide (made in-house; unavailable commercially; for use in GPCR arrays).
8. FlexiPERM micro 12 chambers (Vivascience, Goettingen Germany).
9. Humidified box.
10. CultureWell Isolater (Grace Biolabs, Bend, OR).
11. Eppendorf Centrifuge 5804R (Eppendorf, Westbury, NY).

3. Methods

3.1. General Considerations

3.1.1. Surface Chemistries

The interaction of proteins with a surface complicates the preparation of protein microarrays. This is because (1) proteins could denature at the interface between an aqueous solution and a solid surface, and (2) random immobilization of proteins on a surface may cause the active site(s) of the proteins to be inaccessible for binding. Fabricating GPCR microarrays is particularly challenging, mainly because GPCRs require association with lipid membrane to retain their correctly folded conformation and function (Fig. 1). Covalent immobilization of the entire membrane is not desirable because lateral mobility is an intrinsic and physiologically important property of biological membranes. In addition, the GPCR–G protein complex should be preserved after being arrayed onto a surface because the correct configuration of the receptor and G protein is a prerequisite for the binding of agonists to the receptor with physiological binding affinity (16). The surface could have a significant impact on the structure and functionality of the receptors, but it also plays a critical role in the structure and mechanical stability of the immobilized lipid membranes.

To achieve maximum binding capacity and desired stability of proteins on a surface with largely preserved structure and activity, the surface of solid supports generally need to be re-engineered. Examples include “deformable” polymer-grafted surfaces for immobilization of proteins (17) (e.g., HydroGel-coated slides, PerkinElmer Life

Science, Boston, MA), amine- or thiol-reactive surfaces for covalent coupling of proteins (6), or functional group-presenting surfaces for specific binding of proteins. Functional group-presenting surfaces include avidin-coated surfaces for biotinylated proteins (18), Ni^{2+} -chelating surfaces for histidine-tagged proteins (13), or antibody-modified surfaces for native proteins (19).

We have demonstrated that model lipid membranes can be immobilized onto amine-presenting surfaces (e.g., GAPS surfaces) with rapid kinetics, desired structures, preserved lateral fluidity, and significant mechanical stability (15; data not shown). Ligand binding to GPCR microarrays on these surfaces is specific; binding affinities are similar to those obtained using traditional methods (15,20,21). This chapter focuses on the use of GAPS surfaces (*see Note 3*) for producing both antibody microarrays and GPCR microarrays.

3.1.2. Printing Technologies

With the advent of DNA microarray technology, a number of printing technologies have become amenable to production-scale fabrication of protein microarrays. The most popular ones are contact pin printing and non-contact ink-jet printing (3–5,22). Although ink-jet arrayers are less restricted as to surface structure, they may adversely affect the activity of proteins, as a result of the shearing force and/or thermal effect during drop formation. Contact pin arrayers generally deliver subnanoliter volumes of protein solution directly to a surface using tiny pins with or without capillary slots (“quill pin” vs “solid pin,” respectively). The use of quill-pin printers is more suitable for large-scale production of protein microarrays, because one sample pickup can produce tens or even hundreds of reproducible and consistent microspots. However, considerable optimization is required to prepare high-quality microarrays. For instance, the proteins should be buffered in a solution that leads to optimal printing reproducibility with desired spot morphology. In addition, a pin-cleaning protocol may also be included to avoid clogging of the pins and cross-contamination between samples. To keep proteins in a wet environment as well as prevent the tiny solution inside the pin from evaporating during printing, a high percentage of glycerol (30–50%) is generally used in the sample buffer, and the printing is carried out in a humidity-controlled environment.

The preparation of protein microarrays could be significantly slower when a greater number of elements are arrayed or numerous microarrays are produced. Proteins, in particular membrane proteins, are susceptible to environmental changes, and may become aggregated or damaged during printing. Therefore, the stability of proteins during the printing should be also considered.

Printed protein arrays generally require sequential post-printing processes to achieve maximum immobilization or attachment. For example, printed GPCR arrays should undergo a 1-h incubation under controlled humidity before use or storage, because the immobilization of biological membranes to achieve stable association with a surface is relatively slow.

3.1.3. Reformulation of Proteins for Printing

Once the appropriate surface chemistry and printing instrument are selected, the next crucial step is to develop a printing “ink.” Proteins differ greatly in their amino acid sequence, shape, size, structure, function, stability, and solubility. Some are soluble in aqueous solutions, whereas membrane proteins have to associate with lipid mem-

branes, and are thus able to be only suspended in aqueous solution. Such individuality and diversity have created problems for the production of functional protein microarrays. Thus, for a given set of proteins, a universal ink should be developed and optimized. The ideal printing ink should allow consistent sample volume to be delivered onto the surface during pin contact. It should also enhance the immobilization of probe proteins to the surface while preserving protein activity.

3.2. Antibody Microarrays

The protocol below is related to the fabrication and use of microarrays of three capture antibodies: IgY-CRP, IgY-hGH, and IgY-rhI.

3.2.1. Fabrication of Antibody Microarrays

Antibodies, unlike intracellular or membrane-bound proteins, are quite stable. Antibody microarrays with acceptable consistency and desired morphology can be fabricated using PBS buffer containing 40% glycerol on GAPS slides (6) (an example after assays is shown in **Fig. 2**). The concentration of capture antibodies used is generally in the range of 1–1000 µg/mL, preferably 50–200 µg/mL.

For array fabrication, 5–10 µL of a formulated capture antibody solution is loaded into one well of a 384-well low-volume microplate. A single insertion of quill pin (CMP3) into the solution is used to pick up a sample and print continuously up to a couple of hundred almost identical spots with small printing variability. Multiple submicroarrays are made on a single GAPS slide (2 × 8 subarrays, each array has 5 × 3 microspots, with each capture antibody having five replicates for statistic purposes). After printing, the arrays are incubated in a humid chamber at room temperature for at least 30 min before use.

1. Set up a program of the arrayer to print desired configuration of microarrays.
2. Load 7 µL of each the freshly prepared capture antibody solution into a well of a 384-well low-volume microplate.
3. Clean a CMP3 quill pin in an ultrasonic water bath.
4. Insert the pin into the capture antibody solution and load up the solution.
5. Preprint 20–50 microspots on a spare slide.
6. Continuously print a given number of microspots on GAPS slides in the given array format.
7. Repeat **steps 2–6** to fabricate microarrays of multiple capture antibodies.
8. Transfer slides with the printed arrays to a humidity chamber.
9. Incubate for at least 30 min before use.

3.2.2. Profiling Target Proteins in a Sample

Several assays have been developed for the use of antibody microarrays to profile target proteins in a sample. One popular assay is the “sandwich” assay (23). This assay involves two antibodies that bind to the same target protein: a capture antibody immobilized on surface, and a second antibody to detect the presence of bound target proteins. The second antibody could be labeled or unlabeled. If unlabeled, a third, labeled antibody (e.g., fluorescently labeled anti-mouse IgG) is used as a universal readout to detect the binding. The following protocol is related to the sandwich assay, and comprises five sequential steps: (1) preblocking of the array with a blocker (e.g., BSA or dry milk), (2) binding of target protein(s) in a sample, (3) binding of a target-detecting antibody, (4) binding of labeled “readout” antibody, and (5) data acquisition and analysis.

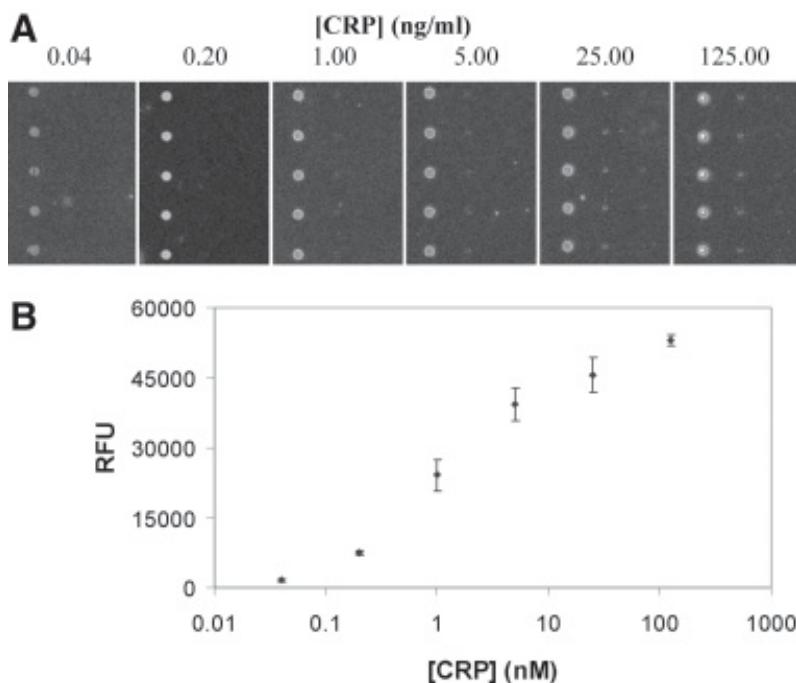


Fig. 2. Demonstration of antibody microarrays to profile protein abundance in a sample. The binding of C-reactive protein (CRP) to microarrays of antibodies (IgY-CRP, IgY-hGH, and IgY-rhI) as a function of CRP concentration is examined using a “sandwich” assay. In this assay, a second anti-CRP is added to interact with the bound CRP, and a Cy5-antimouse IgG is used to detect the binding of the second anti-CRP in a sequential reaction. (A) Fluorescence images of the capture antibody microarrays after sequential incubation with (i) CRP at different concentrations, (ii) an anti-CRP, and (iii) a Cy5-labeled detection antibody (Cy5-antimouse IgG). (B) The graph shows an excellent linear-log correlation between fluorescence intensity and CRP concentration. The error bars represent the standard deviation of five replicates in a single array.

3.2.2.1. STEP A: PREBLOCKING

1. Place a CultureWell Isolator on the top of a slide to form virtual wells each of which has a subarray located in the center.
2. Load 30 μ L of the blocking solution into each well.
3. Incubate the arrays at room temperature with gentle shaking for 20 min.
4. Aspirate the solution.

3.2.2.2. STEP B: BINDING OF TARGET PROTEIN(S) IN A SAMPLE

1. Load 20 μ L of a sample solution into each well. The sample contains the target protein(s) to be detected.
2. Incubate at room temperature for 1 h.
3. Aspirate the solution.
4. Wash each well by adding 50 μ L of the washing solution and gently shaking the slides at room temperature for 15 min.
5. Repeat the wash with the same solution four times, 2 min each time, and aspirating the solution between washes.
6. Rinse each well twice with the rinsing solution.

3.2.2.3. STEP C: BINDING OF TARGET-DETECTING ANTIBODY

1. Add 20 μ L of a diluted solution of mAb (mAb at a final concentration of 0.2 μ g/mL in 1X PBS/3% BSA) to each well where its cognate protein was added in the previous step.
2. Incubate in a humidified box for 1 h at room temperature.
3. Aspirate the solution.
4. Submerge the slide in a box filled with approx 50 mL of the washing solution for 5 min at room temperature. Repeat this wash three times, 2 min each time with shaking.
5. Rinse the slide with the rinsing solution.

3.2.2.4. STEP D: BINDING OF LABELED “READOUT” ANTIBODY

1. Add 20 μ L of the Cy5-anti-mouse IgG binding solution to each well.
2. Incubate for 30 min with gentle shaking at room temperature.
3. Aspirate the solution.
4. Wash with 50 mL of the washing solution for 15 min with shaking at room temperature.
5. Remove the CultureWell Isolator.
6. Repeat the wash four times, 2 min each time.
7. Spin-dry the slide by centrifugation at 2000 rpm (425g) for 2 min.

3.2.2.5. STEP E: DATA ACQUISITION AND ANALYSIS

1. Image the slide with a GenePix 4000B scanner at proper photomultiplier tube (PMT) settings.
2. Analyze the images using GenPix Pro 3.0 software. The average relative fluorescent units (RFU) and the standard deviation are obtained from the five replicate spots within the same subarray. A result is summarized in **Fig. 2**.

3.3. GPCR Microarrays

The protocol below is related to the fabrication and use of microarrays of four GPCRs: β 1-, β 2-, and α 2A adrenergic receptors, and NTR1.

3.3.1. Fabrication and Storage of GPCR Microarrays

GPCR cell-membrane preparations can be directly used to prepare functional GPCR microarrays (see **Note 4**) with contact pin printing technology (15,20,21). Several specifications related to the GPCR membrane preparations have been found to play important roles in array printing quality, assay sensitivity, and robustness (see **Note 5**).

For array fabrication, in a typical print run, 5–10 μ L of each GPCR membrane preparation is added to a 384-well low-volume microplate. A single insertion of a quill pin into the solution enables the printing of up to 400 almost identical spots. The printing variability is quite small (less than 10%), as determined by binding-assay robustness (unpublished data). After printing, the arrays are incubated in a humid chamber at room temperature for 1 h before use. For longer-term storage, the arrays are stored at 4°C in a dessicator filled with nitrogen.

1. Load 7 μ L of each reformulated GPCR solution into a well of a 384-well microplate.
2. Clean a CMP3 quill pin in an ultrasonic water bath.
3. Insert the pin into the receptor solution and load up the solution.
4. Pre-print 25–50 microspots on a spare slide.
5. Continuously print a given number of microspots on GAPS slides in a given format.
6. Repeat **steps 1–5** to prepare microarrays of multiple GPCRs.
7. Transfer slides with the printed arrays to a humidity chamber.
8. Incubate for 1 h.
9. Store the slides in a dessicator filled with nitrogen at 4°C.

3.3.2. Profiling a Target Compound in a Sample

Generally, multiple microarrays of GPCRs are fabricated on a single GAPS slide. To perform binding assays, each array is incubated for 1 h with 10 μ L of a buffered solution containing fluorescently labeled ligands, or a mixture of fluorescently labeled ligands and unlabeled compounds for competitive binding assays. After incubation, the solution is carefully removed using a vacuum aspirator. The slides are rinsed briefly with water, dried under a stream of nitrogen, and imaged in a fluorescence scanner. Alternatively, a FlexiPERM micro 12-chamber gasket is attached to the slide so as to position each printed array in the center of each well. An assay solution (30 μ L) is added to each well. Following incubation, similar washing and drying protocols are used after the FlexiPERM gasket is removed.

Several assays have been developed for pharmacological profiling of target compounds. For example, saturation assays are used to examine binding affinity of a labeled ligand, and competitive binding assays are used to determine the relative potency of two target compounds against one receptor, or the selective potency of one target compound against multiple receptors. Compound screening can be carried out in many different formats and with different types of GPCRs. For example, compounds can be tested against an array consisting of one member of each GPCR family, or against an array consisting of all of the GPCRs within a family (e.g., the adrenergic receptors), or against a full index GPCR array that contains receptors from different families.

The following protocol is a competitive binding assay used for compound potency profiling and compound screening, in which microarrays of three adrenergic receptor family members, (β 1, β 2, and α 2A), are used as a model system. **Figure 3** shows the relative potency of neuropeptides against the binding of BT-NT (see Note 6) to NTR1 arrays; **Fig. 4** shows the results of compound selectivity screening using arrays of the adrenergic receptors.

1. Prepare a series of solutions of BT-CGP12177 at a fixed concentration (2 nM) in the presence of different unlabeled compounds at different concentrations using the binding buffer.
2. Apply 15 μ L each of the above solutions to an array of adrenergic receptors on a slide.
3. Incubate the slide with the solution for 1 h at room temperature.
4. Aspirate the solution, briefly rinse the slides with water, and dry immediately.
5. Scan the slide using a Genipix scanner.
6. Analyze the image to determine the total binding signal intensity.

4. Notes

1. The buffer composition for reformulating GPCRs should contain a pH buffer (pH 7.4–7.5), inorganic salt (e.g., $MgCl_2$), membrane stabilizer (e.g., sucrose), and glycerol.
2. Fluorescently labeled ligands should be protected from light in order to minimize photobleaching of the dye molecules. Solutions of all ligands and compounds can be aliquotted and stored at low temperature (generally $-80^{\circ}C$) and should be stable over 6 mo. Freeze/thaw cycles should be minimized (<15 cycles).
3. Not all GAPS-coated surfaces perform equivalently for antibody microarrays and GPCR microarrays; performance depends upon several factors, including the coating and post-coating process of the slides, as well as the surface properties.
4. A large number of GPCR membrane preparations are commercially available from PerkinElmer Life Sciences, Amersham Biotech, or Euroscreen (Brussels, Belgium). The GPCRs are prepared from a cell line that has a high expression level of either a cloned or

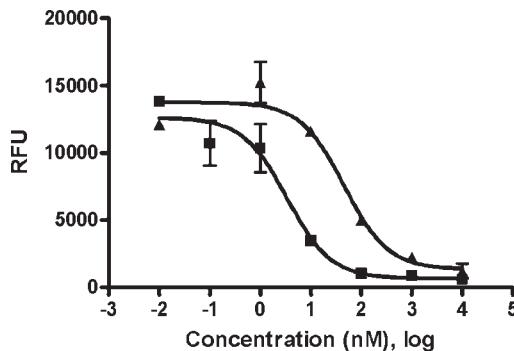


Fig. 3. Relative potency of neurotensin (NT) and neuromedin N (NN) against BT-NT when binding to arrays of NTR1, obtained using competitive binding assays. The binding of BT-NT to NTR1 microarrays as a function of NT (■) or NN (▲) concentration is examined. The IC_{50} is 3.4 nM for NT, 46 nM for NN; these values are close to those reported in the literature for solution-based assays. The error bars represent the standard deviation of four replicates in a single array.

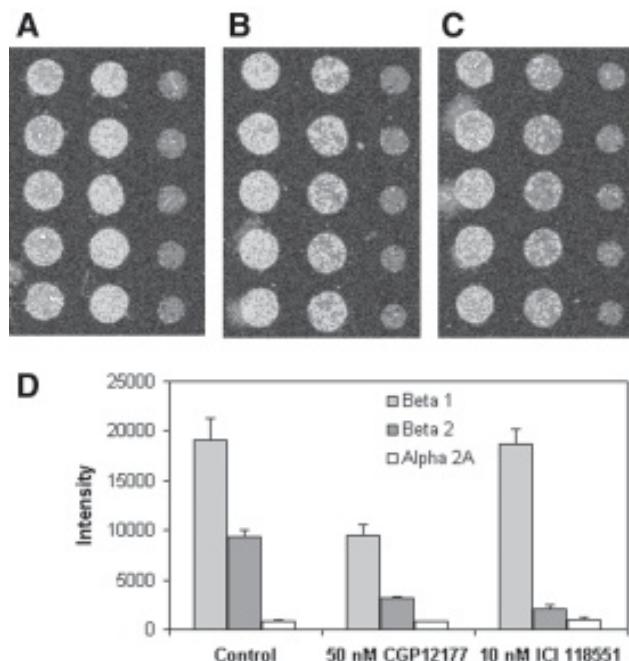


Fig. 4. Demonstration of selectivity screening using G protein-coupled receptor (GPCR) arrays. Three separate arrays of the $\beta 1$, $\beta 2$, and $\alpha 2A$ receptors were printed on a single γ -aminopropylsilane (GAPS)-coated gold slide. The arrays were incubated with a solution containing 2nM BT-CGP in the absence (positive control) and presence of either CGP 12177 (50 nM) or ICI 118551 (10 nM). Analysis of the histogram shows that the presence of 50 nM CGP 12177 significantly inhibits the binding of BT-CGP to microspots corresponding to the $\beta 1$ and $\beta 2$ receptors. This is consistent with the fact that CGP 12177 binds to $\beta 1$ and $\beta 2$ with similar affinity (K_i value is approx 0.6 nM for both receptors) (24). However, the presence of 10 nM ICI 118551 significantly inhibits the binding of BT-CGP only to microspots corresponding to the $\beta 2$ receptor. This is consistent with the fact that ICI 118551 binds to $\beta 2$ with much higher affinity (K_i is 1.2 nM) relative to $\beta 1$ (K_i is 120 nM) (26). The $\alpha 2A$ receptor is included in the array as a negative control (15).

- an endogenous receptor. These preparations contain cell-membrane fragments with embedded receptors.
5. Important parameters of a GPCR membrane preparation include: (1) active receptor concentration (B_{max}), (2) concentration of total membrane proteins, and (3) buffer composition; this information is provided by the vendors. The B_{max} value is the single most important parameter determining assay sensitivity. The total protein concentration, buffer composition, and homogeneity of the preparations can significantly affect spot morphology, printing quality, and array performance (21). Membrane preparations with greater homogeneity, higher active receptor, and appropriate total membrane protein concentration yield arrays with better performance.
 6. Bodipy-TMR-CGP12177 (BT-CGP), a fluorescently labeled CGP 12177 analog, is used as a probe for $\beta 1$ and $\beta 2$. CGP 12177 is a β -adrenergic receptor-selective antagonist that binds to $\beta 1$ and $\beta 2$ with similar affinity (24). Bodipy-TMR-neurotensin (BT-NT), a fluorescently labeled neurotensin, is used as a probe for the neurotensin receptor subtype I. Neurotensin is a natural agonist for NTR1 (25).

References

1. Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470.
2. Favis, R., Day, J. P., Gerry, N. P., Phelan, C., Narod, S., and Barany, F. (2002) Universal DNA array detection of small insertions and deletions in BRCA1 and BRCA2. *Nat. Biotechnol.* **18**, 561–564.
3. Lee, K. H. (2001) Proteomics: a technology-driven and technology-limited discovery science. *Trends Biotechnol.* **19**, 217–222.
4. Zhu, H. and Synder, M. (2001) Protein arrays and microarrays. *Nat. Biotechnol.* **5**, 40–45.
5. Mitchell, P. (2002) A perspective on protein microarrays. *Nat. Biotechnol.* **20**, 225–229.
6. MacBeath, G. and Schreiber, S. L. (2000) Printing proteins as microarrays for high-throughput function determination. *Science* **289**, 1760–1763.
7. Schweitzer, B., Wiltshire, S., Lambert, J., et al. (2000) Immunoassays with rolling circle DNA amplification: a versatile platform for ultrasensitive antigen detection. *Proc. Natl. Acad. Sci. USA* **97**, 10,113–10,119.
8. Wiltshire, S., O’Malley, S., Lambert, J., et al. (2000) Detection of multiple allergen-specific IgEs on microarrays by immunoassay with rolling circle amplification. *Clin. Chem.* **46**, 1990–1993.
9. MacBeath, G., Koehler, A. N., and Schreiber, S. L. (1999) Printing small molecules as microarrays and detecting protein-ligand interactions *en masse*. *J. Am. Chem. Soc.* **121**, 7967–7968.
10. Houseman, B. T. and Mrksich, M. (2002) Carbohydrate arrays for the evaluation of protein binding and enzymatic modification. *Chem. Biol.* **9**, 443–454.
11. Wang, D., Liu, S., Trummer, B. J., Deng, C., and Wang, A. (2002) Carbohydrate microarrays for the recognition of cross-reactive molecular markers of microbes and host cells. *Nat. Biotechnol.* **20**, 275–281.
12. Fang, Y., Frutos, A. G., and Lahiri, J. (2003) Ganglioside microarrays for toxin detection. *Langmuir* **19**, 1500–1505.
13. Zhu, H., Bilgin, M., Bangham, R., et al. (2001) Global analysis of protein activities using protein chips. *Science* **293**, 2101–2105.
14. Houseman, B. T., Huh, J. H., Kron, S. J., and Mrksich, M. (2002) Peptide chips for the quantitative evaluation of protein kinase activity. *Nat. Biotechnol.* **20**, 270–274.
15. Fang, Y., Frutos, A. G., and Lahiri, J. (2002) Membrane protein microarrays. *J. Am. Chem. Soc.* **124**, 2394–2395.

16. Haga, T. and Berstein, G. (eds) (1999) *G Protein-Coupled Receptors*. CRC Press, Boca Raton, FL.
17. Rubina, Y. A., Dementieva, E. I., Stomakhin, A. A., et al. (2003) Hydrogel-based protein microchips: manufacturing, properties, and applications. *BioTechniques* **34**, 1008–1012.
18. Rowe, C. A., Tender, L. M., Feldstein, M. J., et al. (1999) Array biosensor for simultaneous identification of bacterial, viral, and protein analytes. *Anal. Chem.* **71**, 3846–3852.
19. Vijayendran, R. A. and Leckband, D. E. (2001) A quantitative assessment of heterogeneity for surface-immobilized proteins. *Anal. Chem.* **73**, 471–480.
20. Fang, Y., Frutos, A. G., and Lahiri, J. (2002) G protein-coupled receptor microarrays. *ChemBioChem* **3**, 987–991.
21. Fang, Y., Lahiri, J., and Picard, L. (2003) G protein-coupled receptor microarrays for drug discovery. *Drug Discovery Today* **8**, 755–761.
22. Okamoto, T., Suzuki, T., and Yamamoto, N. (2000) Microarray fabrication with covalent attachment of DNA using bubble jet technology. *Nat. Biotechnol.* **18**, 438–441.
23. Kodadek, T. (2001) Protein microarrays: prospects and problems. *Chem. Biol.* **8**, 105–115.
24. Heithier, H., Hallmann, D., Boege, F., et al. (1994) Synthesis and properties of fluorescent beta-adrenoceptor ligands. *Biochemistry* **33**, 9126–9134.
25. Barroso, S., Francoise, R., Nicolas-Etheve, D., et al. (2000) Identification of residues involved in neurotensin binding and modeling of the agonist binding site in neurotensin receptor 1. *J. Biol. Chem.* **275**, 328–336.
26. Bilski, A., Dorries, S., Fitzgerald, J. D., Jessup, R., Tucker, H., and Wale, J. (1980) ICI 118551, a potent β 2 adrenoreceptor antagonist. *Br. J. Pharmacol.* **69**, 292–305.

PCR-Directed Protein *In Situ* Arrays

Joe Boutell and Mingyue He

1. Introduction

With the completion of the human genome sequence, the next priority is to identify the function of the thousands of proteins encoded within. One powerful technology that enables the high-throughput analysis of protein function is that of protein arrays. Protein arrays are usually produced by immobilizing many hundreds of individual proteins in a defined pattern onto a solid surface (1). Such arrays allow simultaneous screening of large numbers of proteins and permit parallel analysis of protein function. They can also be used to identify molecular interactions or, in the form of antibody arrays, to study protein expression profiling within patient samples (2). However, the main limitation to protein array technology currently is the production of the huge diversity of proteins that form the array elements. Many proteins, especially human proteins, are not expressed as functional molecules in heterologous hosts (3), and cloning of individual genes is also a time-consuming process. To overcome these problems, we have developed a cell-free protein array method, protein *in situ* arrays (PISA), which creates functional protein arrays directly from polymerase chain reaction (PCR) DNA by *in vitro* synthesis of individual tagged proteins on tag-binding surfaces, such that the tagged proteins are immobilized *in situ* as they are synthesized (4) (Fig. 1).

The PISA technology avoids cloning and *Escherichia coli* expression processes, providing a rapid tool for arraying proteins or domains of proteins, even if DNA clones for those proteins are not available. It is also particularly useful for proteins that cannot be functionally produced in heterologous hosts. With recent improvements in cell-free expression systems (5,6) and sensitive detection or readout technologies, this method has the potential to be adapted for high-throughput application and automation. We have used this technology to generate arrays of different proteins and protein fragments and demonstrated their use for rapid functional analysis (4,7). Here, we describe details of this method for general applicability.

2. Materials

2.1. Primers

2.1.1. Primers for Making PCR Constructs for the Rabbit Reticulocyte Lysate System

1. T7(Rb): 5'-GCAGCTAATACGACTCACTATAGGAACAGACCACCATG-3'—an upstream primer containing T7 promoter (italics) and Kozak sequence (underlined) for translation in rabbit reticulocyte lysate system. The start codon ATG is indicated in bold.

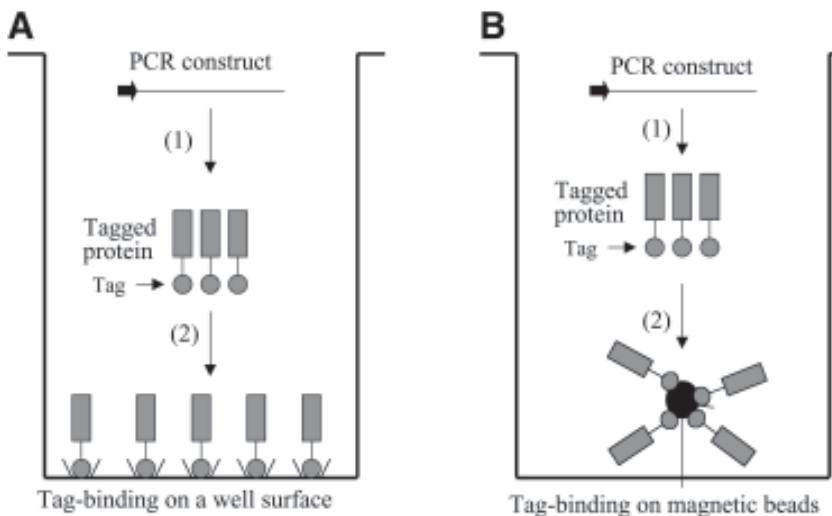


Fig. 1. Protein *in situ* array procedure showing cell-free synthesis of a tagged protein on the tag-binding surface and *in situ* immobilisation. (A) Arraying protein in the well of a microtiter plate; (B) arraying protein on magnetic beads. (1) Coupled cell-free transcription and translation. (2) *In situ* protein immobilization.

2. G/back (Rb): 5'-TAGGAACAGACCACCATG(N)₁₅₋₂₅-3'—an upstream primer for PCR amplification of the gene of interest. It contains a sequence overlapping with T7 (Rb) (underlined) and 15–25 nucleotides from the 5' sequence of the gene of interest. (N)₁₅₋₂₅ indicates the number of nucleotides.
3. G/for: 5'-CACCGCCTCTAGAGCG(N)₁₅₋₂₅-3'—a downstream primer for PCR amplification of the gene of interest. It contains a sequence (underlined) overlapping with a PCR fragment encoding a T-domain (see Subheading 2.2.) and 15–25 nucleotides complementary to the 3' region of the gene of interest.

2.1.2. Primers for Making PCR Constructs for *E. coli* S30 Extracts

1. T7(E): 5'-GAAATTAATACGACTCACTATAGGGAGACCACAACGTTCCCTCTAG AAATAATTTTGTTAACTTAAGAAGGAGATACCATG—an upstream primer containing T7 promoter (italics) and ribosome binding site (underlined) for translation in *E. coli* S30 extracts. The start codon ATG is indicated in bold.
2. G/back (E): 5'-CTTTAAGAAGGAGATACCATG(N)₁₅₋₂₅-3'—an upstream primer for PCR amplification of the gene of interest. It contains a sequence overlapping with T7 (E) (underlined) and 15–25 nucleotides from the 5' sequence of the gene of interest. (N)₁₅₋₂₅ indicates the number of nucleotides.
3. G/for: 5'-CACCGCCTCTAGAGCG(N)₁₅₋₂₅-3'—a downstream primer for PCR amplification of the gene of interest. It contains a sequence (underlined) overlapping with a PCR fragment encoding a T-domain (see Subheading 2.2.) and 15–25 nucleotides complementary to the 3' region of the gene of interest.

2.1.3. Primers for Making a T-Domain for PCR Assembly

1. Linker-tag/back: 5'-GCTCTAGAGGCGGTGGC-3'—an upstream primer for PCR generation of a T-domain fragment in combination with T-term/for (see next step).
2. T-term/for: 5'-TCCGGATATAGTTCCCTCC-3'—a downstream primer for PCR generation of either the T-domain fragment in combination with the Linker-tag/back or the full-

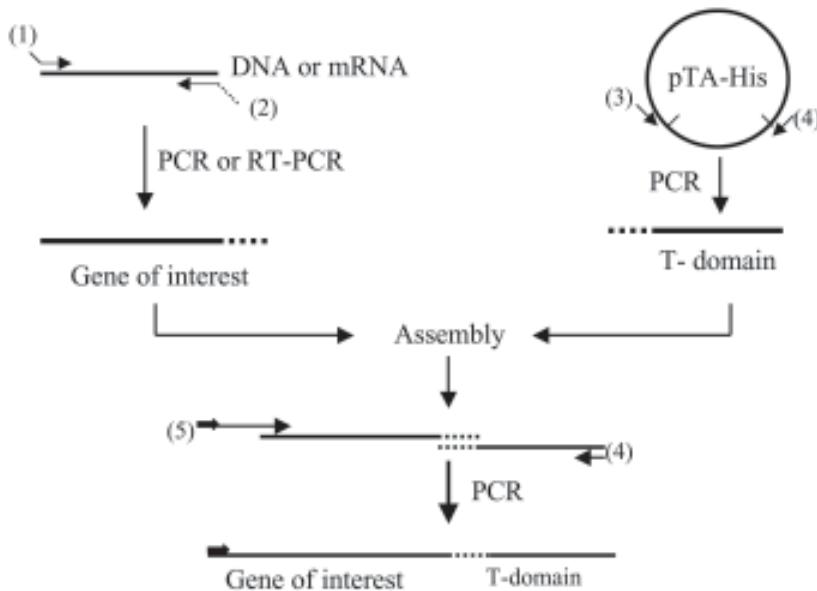


Fig. 2. A PCR construction strategy. The primers used are: (1) G/back (Rb) or G/back (E); (2) G/for; (3) Linker-tag/back; (4) T-term/for; (5) T7 (Rb) or T7 (E). The broken line indicates the linker.

length construct in combination with one of the T7 primers (see **Subheadings 2.1.1.** and **2.1.2.** and **Fig. 2**)

2.2. Plasmid Encoding a T-Domain

Plasmid pTA-His contains a DNA fragment encoding (in order) a flexible linker and a double (His)₆ tag, followed by two stop codons, a poly(A) tail, and a transcription termination region (4). The DNA fragment is called T-domain and the detailed sequence is: GCTCTAGAggcgggtggctctggggcggttcggccgggtggcaccgggtggcggtctggcggtggc AACGGGCTGATGCTGCCACATCACCATCACACTCTAGAGCTTGG CGTCACCCGCAGTCGGTGGTCACCACCACCACTAATAA(A)₂₈ CCGCTGAGCAATAACTAGCATAACCCTTGGGGCTCTAAACGGGTCTTGA GGGGTTTTGCTGAAAGGAGGAACTATCCGA-3'. The lower case indicates the linker encoding 19 amino acids (8); the double (His)₆ tag is underlined. Stop codons are in bold, and (A)₂₈ is a poly-A tail comprising 28 × A. The transcription termination region is shown in italics.

2.3. Cell-Free System and Molecular Biology Reagents and Kits

1. TNT T7 Quick for PCR DNA (Promega, UK).
2. RTS100 *E. coli* HY (Roche Molecular Biochemicals, UK).
3. Nucleotides (Sigma, UK).
4. Agarose (Sigma, UK).
5. Taq DNA polymerase (Qiagen, UK).
6. Gel elution kit QIAEX II (Qiagen, UK).
7. Ni-NTA-coated HisSorb strip/plates (Qiagen, UK).
8. Ni-NTA-coated magnetic agarose beads (Qiagen, UK).

9. TitanTM one tube reverse transcriptase (RT)-PCR system (Roche Molecular Biochemicals, UK).
10. Horseradish peroxidase (HRP)-linked Anti- κ antibody (The Binding Site, UK).
11. HRP-linked streptavidin (Amersham, UK).
12. 3,3',5,5' tetramethylbenzidine (TMB) liquid substrate system for ELISA (Sigma, UK).

2.4. Solutions

1. Biotin, 1 mg/mL in water (Pierce, UK).
2. Antigen solution (0.5–1 mg/mL in PBS).
3. Luciferase Assay Reagent (Promega, UK).
4. 100 mM Magnesium acetate.
5. Superblock (Pierce, UK).
6. Phosphate-buffered saline (PBS), pH 7.4.
7. Wash buffer 1: PBS containing 300 mM NaCl, 20 mM imidazole (pH 8.0).
8. Wash buffer 2: PBS containing 0.05% Tween-20.
9. Stripping buffer: 1 M $(\text{NH}_4)_2\text{SO}_4$, 1 M urea.

3. Methods

The methods described below outline (1) generation of a PCR construct, (2) protein arrays by PISA, and (3) functional assay of immobilized proteins.

3.1. Generation of a PCR Construct

A PCR template is required for protein synthesis in a cell-free system (the commonly used systems are rabbit reticulocyte lysate, wheat germ, and *E. coli* S30 extracts). The PCR construct must contain some essential elements, such as a promoter (T7 is the most commonly used promoter), translation initiation site, and transcription and translation termination regions. The translation initiation site used for eukaryotic systems is different in *E. coli* S30 extracts (Fig. 3). To promote protein expression, a poly-A tail is also included following the stop codon. An affinity tag sequence is also required at either the N- or C-terminus of the arrayed protein for protein immobilization (Fig. 3A,B) (see Note 1). To reduce any possible interference of the tag sequence on protein folding and also make the tag sequence accessible to the tag-binding reagent on the surface, a flexible linker is placed between the arrayed protein and the tag sequence (Fig. 3).

To simplify the process of PCR construction, all the common elements, such as the tag sequence, flexible linker, poly-A tail, and termination region, can be cloned into a plasmid, which is then used as a template to generate a PCR fragment to assemble with the protein to be arrayed (Fig. 2). This chapter describes the use of a cloned DNA fragment (termed *T-domain*), which encodes a flexible linker followed by a novel double His-tag sequence, stop codon, a poly-(A)₂₈ tail, and a transcription termination region (4). The T-domain fragment can be generated in large quantities by PCR and used for assembly as a C-terminal fusion with the protein of interest (see Fig. 3). The T7 promoter and translation initiation site can be chemically synthesized as a long oligonucleotide and used as a primer to introduce these elements into the construct. Figure 2 shows the PCR construction process, which contains two main steps: (1) PCR generation of the gene of interest and the T-domain separately, and (2) PCR assembly of the gene and the T-domain.

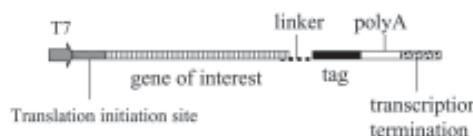
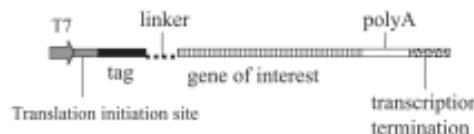
A Construct encoding a C-terminal tag**B** Construct encoding a N-terminal tag

Fig. 3. Polymerase chain reaction construct for protein *in situ* array. (A) Construct with a C-terminal tag. (b) Construct with an N-terminal tag. T7: T7 promoter sequence. Translation initiation site: Kozak consensus sequence (CAGACCACCC) should be used in rabbit reticulocyte system; Shine–Dalgarno (S/D) sequence (AAGGAG) is required for *Escherichia coli* system. tag: an affinity tag sequence. PolyA: poly (A)_n tail. linker: peptide flexible linker.

3.1.1. PCR Amplification of the Gene of Interest and the T-Domain

1. Generate target DNA by PCR (for a plasmid or cDNA template) or RT-PCR (for mRNA template) using the primers G/back and G/for (see Fig. 2). Standard PCR and RT-PCR procedure is performed using the enzymes according to manufacturer's instructions. Similarly, generate the T-domain by PCR using the plasmid pTA-His template and primers Linker-tag/back and T-term/for (see Fig. 2) (see Note 2).
2. Analyze the resultant PCR products by agarose gel electrophoresis and purify the fragments using the Qiagen gel extraction kit.

3.1.2. PCR Assembly of the Gene and the T-Domain

1. Set up a 25- μ L PCR solution by mixing the two fragments in equimolar ratios (total DNA 50–100 ng). Place the PCR mixture in a thermal cycler for eight cycles (94°C for 30 s, 54°C for 1 min, and 72°C for 1 min) to assemble the two fragments. Then, amplify the assembled product by transferring 2 μ L of the assembled product to a second PCR solution in a final volume of 50 μ L for a further 30 cycles (94°C for 30 s, 54°C for 1 min, and 72°C for 1 min) using one of the T7 primers and T-term/for.
2. Analyze the PCR product by agarose gel electrophoresis and purify the DNA if needed.
3. Confirm the construct identity by PCR mapping using primers at various positions along the desired sequence (see Note 3).

3.2. Protein Arrays by PISA

Protein *in situ* arrays are generated by cell-free protein synthesis and *in situ* protein immobilization on a surface. Cell-free expression is performed in a coupled cell-free system such as the rabbit reticulocyte lysate TNT system or the RTS100 *E. coli* HY system. *In situ* protein immobilization is carried out on a nickel-coated surface for His-tagged proteins. In our laboratory, both Ni-NTA-coated microtiter plates and Ni-NTA magnetic agarose beads have been successfully used (see Note 4).

3.2.1. Protein In Situ Arrays Using the Rabbit Reticulocyte Lysate TNT System

1. Set up TNT translation mixture as follows: (see Note 5): 20 μ L TNT T7 Quick for PCR DNA; 0.5 μ L 1 mM methionine (from the kit); 0.25 μ L 100mM magnesium acetate (see Note 6); 0.25–0.5 μ g PCR DNA; H₂O to 25 μ L.
2. Add the TNT mixture directly to either of following surfaces: (a) A single well of Ni-NTA-coated HisSorb strip or plate, or (b) 5–10 μ L Ni-NTA-coated magnetic beads. Incubate the mixture at 30°C for 2 h with gentle shaking.
3. Remove the mixture and wash three times with 100 μ L wash buffer 1 (see Note 7), followed by a final wash with 100 μ L PBS (pH 7.4). Immobilized proteins are then ready for functional assay (see Subheading 3.3.) or stored at 4°C (see Note 8)

3.2.2. Protein In Situ Arrays Using RTS100 *E. coli* HY System

1. Set up RTS100 *E. coli* HY mixture as follows (25 μ L) (see Note 9): 6 μ L *E. coli* lysate (from the kit); 5 μ L reaction mix (from the kit); 6 μ L amino acids (from the kit); 0.5 μ L methionine (from the kit); 2.5 μ L reconstitution buffer (from the kit); 0.2–0.5 μ g PCR DNA; H₂O to 25 μ L.
2. Add the mixture directly to either of following surfaces: (a) a single well of Ni-NTA-coated HisSorb strip or plate, or (ib) 5–10 μ L Ni-NTA-coated magnetic beads. Incubate the mixture at 30°C for 4 h with gentle shaking,
3. Remove the mixture and wash three times with 100 μ L wash buffer 1, followed by a final wash with 100 μ L PBS (pH 7.4). Immobilized proteins are then ready for functional assay (see Subheading 3.3.) or stored in PBS at 4°C.

3.3. Functional Assay of Immobilized Proteins

The immobilized proteins can be used for functional assays such as the detection of protein–protein interactions, ligand-binding, or enzymatic activity (4,7). The following sections describe procedures for the detection and functional analysis of single-chain antibodies and luciferase.

3.3.1. Detection of Immobilized Proteins

1. Add 25 μ L horseradish peroxidase (HRP)-linked antibody (appropriately diluted with Superblock) against the immobilized protein—e.g., use HRP-linked anti- κ antibody to detect single-chain V_H/K fragments (Fig. 4) (4).
2. Incubate the mixture at room temperature for 1 h.
3. Wash three times with 100 μ L wash buffer 2, then a final wash with PBS.
4. Develop HRP activity using 25 μ L of 3,3',5,5' tetramethylbenzidine (TMB) liquid substrate system for ELISA.

3.3.2. Analysis of Antigen Binding of Immobilized Antibody Fragments

1. Add 25 μ L of biotinylated antigen at a dilution determined experimentally (4,9).
2. Incubate the mixture for 1 h at room temperature.
3. Remove the mixture, wash three times with 100 μ L wash buffer 2, and add 25 μ L HRP-linked streptavidin diluted 1:1000 in Superblock.
4. Incubate the mixture at room temperature for 1 h.
5. Wash three times with 100 μ L wash buffer 2, then a final wash with PBS.
6. Develop the HRP activity as described in Subheading 3.3.1. (see Fig. 4).

3.3.3. Enzymatic Activity of Immobilized Luciferase

1. Add 100 μ L of Luciferase Assay Reagent to luciferase immobilized on magnetic beads (see Note 4).
2. Mix up and measure the light intensity immediately on a BioOrbit luminometer (see Fig. 5).

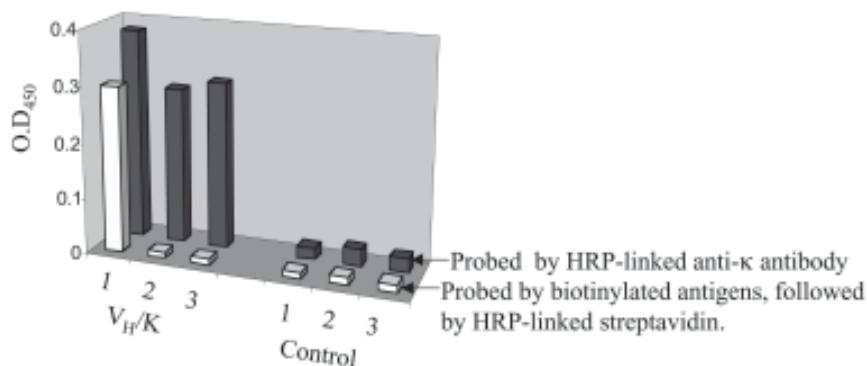


Fig. 4. Specificity analysis of a human anti-progesterone V_H/K fragment immobilized on wells of Ni-NTA-coated plates by protein *in situ* array. Front row: V_H/K detected by (1) biotinylated antigen progesterone-bovine serum albumin (BSA); (2) biotinylated nonantigen BSA; and (3) biotinylated nonantigen carcinoembryonic antigen, followed by horseradish peroxidase (HRP)-linked streptavidin. Back row: The same V_H/K array, after stripping, detected by HRP-linked anti- κ antibody. Control: rabbit reticulocyte lysate lacking polymerase chain reaction template.

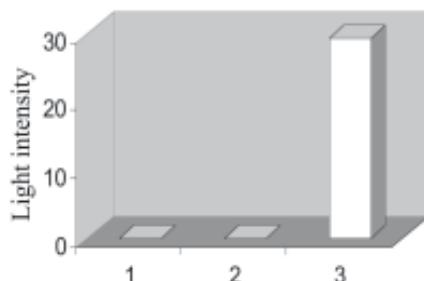


Fig. 5. Enzyme activity of luciferase after immobilisation on Ni-coated magnetic beads. (1) Control: rabbit reticulocyte lysate lacking polymerase chain reaction template; (2) Luciferase construct lacking the (His)₆ tag; (3) Luciferase construct with the (His)₆ tag.

3.4. Re-Use of the Arrays

After exposure to detection reagents, the arrays can be washed and re-used up to three times.

1. Remove the detection reagents and wash three times with 100 μ L wash buffer.
2. Incubate with 50 μ L freshly prepared stripping buffer at room temperature for 2 h.
3. Wash three times again with 100 μ L wash buffer 2 and one final wash with PBS.
4. Re-expose the arrays to detection reagents as described in Subheading 3.3. (Fig. 4).

4. Notes

1. It has been reported that the (His)₆ tag was not accessible when it was fused to the C-terminus of some proteins (10). In these cases, the tag sequence may be placed at the N-terminus. In some circumstances, the location of a tag sequence may affect the activity of some proteins.

2. The T-domain can be produced in large quantities by PCR and stored at -20°C for future use.
3. PCR mapping is carried out by using a combination of various primers annealing at different positions in the construct. If all PCR reactions give the expected size, it suggests the correct construction.
4. The use of Ni-coated magnetic beads to capture His-tagged proteins is useful, particularly when the immobilized protein needs to be tested in different tubes (e.g., to measure luciferase activity) or using different amounts.
5. The volume of TNT mixture used for cell-free expression can be scaled down to 10 μL .
6. Magnesium acetate added to TNT mixture during translation has been found to improve protein expression. We have shown that single-chain antibodies can be more efficiently produced with additional magnesium concentrations, ranging from 0.5 mM to 2 mM.
7. TNT lysate contains large amounts of hemoglobin, which sometimes sticks to Ni-coated magnetic beads. More washes are required to remove hemoglobin from the beads.
8. For most of the single-chain antibody fragments we tested, the arrays can be stored in 50 μL PBS at 4°C for 2 wks.
9. RTS100 *E. coli* HY can produce 3–25 μg of proteins in a 50 μL reaction. The reaction volume can be scaled down to 12.5 μL .

References

1. Pandey, A. and Mann, M. (2000) Proteomics to study genes and genomes. *Nature* **405**, 837–846.
2. Michaud, G. A. and Snyder, M. (2002) Review: proteomic approaches for the global analysis of proteins. *BioTechniques* **33**, 1308–1316.
3. Stevens, R. C. (2000). Design of high-throughput methods of protein production for structural biology. *Structure Fold. Des.* **8**, R177–185.
4. He, M. and Taussig, M. J. (2001). Single step generation of protein arrays from DNA by cell-free expression and *in situ* immobilization (PISA method). *Nucleic Acid. Res.* **29**, e73.
5. Chekulayeva, M. N., Kurnasov, O. V., Shirokov, V. A., and Spirin, A. S. (2001) Continuous-exchange cell-free protein-synthesizing system: synthesis of HIV-1 antigen Nef. *Biochem. Biophys. Res. Commun.* **280**, 914–917.
6. Sawasaki, T., Ogasawara, T., Morishita, R., and Endo, Y. (2002) A cell free protein synthesis system for high-throughput proteomics. *Proc. Natl. Acad. Sci. USA* **99**, 14,652–14,657.
7. He, M. and Taussig, M. J. (2003) DiscernArrayTM technology: a cell-free method for the generation of protein arrays from PCR DNA. *J. Immun. Methods* **274**, 265–270.
8. Robinson, C. R. and Sauer, R. T. (1998) Optimizing the stability of single-chain proteins by linker length and composition mutagenesis. *Proc. Natl. Acad. Sci. USA* **95**, 5929–5934.
9. Taussig, M. J., Groves, M. A., Menges, M., Liu, H., and He, M. (2000) ARM complexes for *in vitro* display and evolution of antibody combining sites. In: (Shepherd, P. S., Shepherd, P. S., Dean, C. J., eds.) *Monoclonal Antibodies: A Practical Approach*. Oxford University Press, Oxford, UK: 91–109.
10. Braun, P., Hu, Y., Shen, B., et al. (2002) Proteome-scale purification of human proteins from bacteria. *Proc. Natl. Acad. Sci. USA* **99**, 2654–2659.

Site-Specific Immobilization of Proteins in a Microarray

Yee-Peng R. Lue, Su-Yin D. Yeo, Lay-Pheng Tan, Grace Y. J. Chen, and Shao Q. Yao

1. Introduction

The completion of the Human Genome Project has generated enormous opportunities, as well as challenges, to protein scientists. The key issue now is to develop efficient strategies that allow high-throughput studies of many thousands of new proteins. The DNA microarray technology makes it possible for simultaneous expression profiling of thousands of genes from various biological sources. However, it is now well known that, at the cellular level, the relative abundance of messenger RNAs does not always correlate to their protein expression level (1). Therefore, it is essential to study the large number of proteins present in an organism in order to better understand its molecular functions. Protein microarray, which adopts the same spotting technology used to fabricate DNA microarray, has recently been developed (2–5). It promises to provide a means for high-throughput identification and quantification of proteins from different biological samples. In a protein microarray, tens of thousands of proteins may be immobilized on a solid surface, such as a glass slide (4,5), and the screening of protein activities could be carried out simultaneously. One of the main challenges in the fabrication of protein microarrays currently is the ability to immobilize proteins in their native conformation on surfaces, while preserving their active sites for functional studies (5,6). Several approaches have been developed (for review, *see ref. 2*). In most cases, however, these modes of protein attachment are unspecific, causing the molecules to be immobilized in the “wrong” orientation. A number of strategies have recently been reported, which allow site-specific immobilization of molecules in the microarray format (5–8). So far, there has been only one report of site-specific attachment of proteins on glass slides using His-tag (5). However, the binding between Ni-NTA and His-tag proteins is not very stable, and is often susceptible to interference by many commonly used chemicals and salts (9). A couple of new methods have recently been proposed that allow site-specific covalent immobilization on surfaces, but they remain to be experimentally demonstrated on the glass slide used in a microarray (10,11).

We have recently exploited the high affinity of avidin for biotin in protein microarray applications (6). By taking advantage of an intein-mediated protein expression system, we developed a method to purify and site-specifically biotinylate recombinant proteins

at their C-terminal end within a single column-purification step (**Fig. 1B**). With this approach, we were able to utilize intein-mediated expression of proteins to generate site-specifically biotinylated proteins with high efficiency (6). Subsequently, these biotin-labeled proteins can be immobilized onto an avidin-functionalized glass slide for array studies. Besides in vitro biotin tagging, we have also identified the usefulness of intein tag for in vivo biotinylation of proteins (**Fig. 1B**) (12). We now report (1) the detailed protocols of this strategy, its extension for in vivo biotinylation, as well as two other methods for site-specific immobilization of proteins in a microarray, namely (2) immobilization using biotinylated proteins generated by cell-free expression systems (**Fig. 1C**), and (3) immobilization using N-terminal cysteine proteins (**Fig. 1A**).

2. Materials

1. *Escherichia coli* strains DH5 α and ER2566 (NEB, Beverly, MA).
2. pTYB1 expression vector (NEB).
3. pTWIN expression vector (NEB).
4. Oligonucleotide primers.
5. dNTPs (Promega, Madison, WI).
6. HotStar *Taq* polymerase (Qiagen, Valencia, CA).
7. Restriction enzymes: *Pst*I and *Sap*I (NEB).
8. Qiaquick spin column (Qiagen).
9. T4 DNA ligase (NEB).
10. Shrimp alkaline phosphatase (Promega).
11. Agarose (BioRad, Hercules, CA).
12. Thermal cycler (MJ Research, Waltham, MA).
13. DNA gel electrophoresis equipment (BioRad).
14. DNA sequencing reagents and equipment (ABI, Foster City, CA).
15. Luria Bertani (LB) medium.
16. Ampicillin (100 mg/mL).
17. Isopropyl thiogalactosidase (IPTG), 1 M (Bio-Rad).
18. 2-mercaptoethanesulfonic acid (MESNA), 1 M (Bio-Rad).
19. Cysteine-biotin (prepared per procedures described in ref. 6).
20. Wash buffer: 20 mM Tris-HCl (pH 7.0–8.0), 500 mM NaCl, 1 mM ethylenediamine-tetraacetic acid (EDTA).
21. Lysis buffer: 20 mM Tris-HCl (pH 7.0–8.0), 500 mM NaCl, 1 mM EDTA, 0.1% Triton X-100, 1 mM PMSF.
22. Acid-washed beads (Bio-Rad).
23. Incubator shaker.
24. Orbital shaker.
25. Ultrasonic liquid processor.
26. Sodium dodecyl sulfate (SDS)-polyacrylamide gel electrophoresis reagents and equipment.
27. Polyvinylidene difluoride (PVDF) membrane (BioRad).
28. Whatman filter paper no. 1.
29. Transblot reagents and equipment.
30. Nonfat dry milk powder.
31. Phosphate-buffered saline (PBS), pH 7.4.
32. 0.1% Tween-20 in PBS (PBST).
33. Horseradish peroxidase (HRP)-conjugated anti-biotin antibody (NEB).
34. Enhanced ChemiLuminescent (ECL) kit (Amersham, UK).
35. Western blotting reagents and equipment.

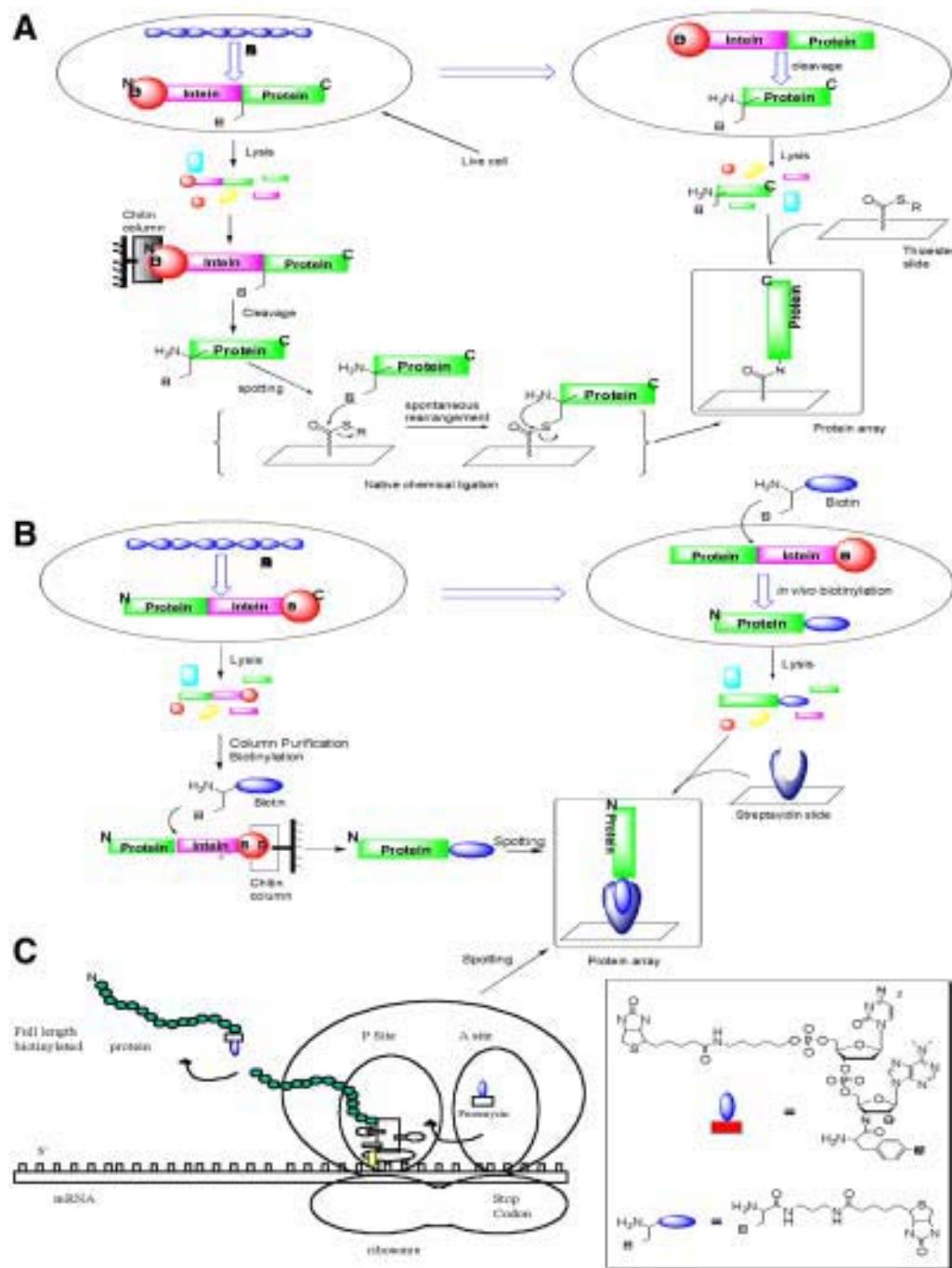


Fig. 1. Overview of three strategies for site-specific immobilization of proteins on a protein microarray.

36. Rapid translation system RTS 100 *E. coli* HY kit (Roche, Nutley, NJ).
37. 5'-Biotin-dC-Pmn (Dharmacon, Lafayette, CO).
38. QIAprep miniprep (Qiagen).

39. MicroSpin™ G-25 columns (Amersham, UK).
40. Microscope glass slides (Fisher or Sigma, USA) and slide tray.
41. Piranha solution ($H_2SO_4:H_2O_2$, 7:3).
42. 1% 3-Glycidyloxypropyltrimethoxysilane in 95% ethanol plus 16 mM acetic acid.
43. Cover slips.
44. 1 mg/mL Avidin in 10 mM $NaHCO_3$.
45. 2 mM Aspartic acid in a 0.5 M $NaHCO_3$ buffer, pH 9.0.
46. 10 mM Diamine-PEG in 0.1 M $NaHCO_3$, pH 9.0.
47. 180 mM Succinic anhydride in dimethylformamide (DMF), pH 9.0.
48. *N*-hydroxysuccinimide (NHS) solution: 100 mM *O*-benzotriazol-1-yl-*N,N,N',N'*-tetramethyluronium tetrafluoroborate (TBTU), 100 mM 1-hydroxybenzotriazole hydrate (HOEt), 200 mM diisopropylethylamine (DIEA), and 100 mM NHS in DMF.
49. Benzylmercaptan: 120 mM DIEA and 100 mM benzylmercaptan in DMF.
50. Arrayer (ESI SMA™, Toronto, Canada).
51. 384-Well polypropylene source plate.
52. Fluorescence-labeled monoclonal antibody against target protein.
53. ArrayWoRx™ microarray scanner (Applied Precision, Issaquah, WA).

3. Methods

The methods described below outline (1) the *in vivo* biotinylation of proteins, (2) the cell-free protein biotinylation, (3) the spotting of biotin-labeled proteins on avidin slides, and (4) the expression and immobilization of N-terminal cysteine proteins. Full accounts of the strategies have been reported elsewhere (12).

3.1. *In Vivo Biotinylation of Proteins (see Note 1)*

The steps below describe the *in vivo* biotinylation procedures for bacterial expressed proteins using a *Sce* intein tag. For *in vitro* biotinylation procedures, please refer to ref. 6 for details. The target gene has to be first cloned into a pTYB1 expression vector such that the C-terminus of the target protein is fused to the intein tag (6,12). The resulting T7-driven expression plasmid is then transformed into *E. coli* ER2566 host for protein expression using standard molecular biology techniques (13).

1. Inoculate 0.5 mL of freshly grown transformed *E. coli* cells into 50 mL of LB medium supplemented with 100 μ g/mL ampicillin.
2. Incubate the culture at 37°C in a 250-rpm incubator shaker to an OD_{600} of about 0.5 (about 3 h).
3. Add IPTG to a final concentration of 0.3–0.5 mM to induce fusion protein expression.
4. Incubate the culture overnight at room temperature on an orbital shaker.
5. Add MESNA and cysteine-biotin to the induced bacterial culture at a final concentration of 30 mM and 3 mM respectively.
6. Incubate the culture at 4°C for 24 h on an orbital shaker.
7. Harvest cells by centrifugation (6000g, 15 min, 4°C).
8. Wash cell pellet at least twice with wash buffer or PBS to remove excess cysteine-biotin.
9. Resuspend cell pellet in 1 mL lysis buffer for protein extraction. Cell pellet can also be stored at –20°C without any significant degradation of the biotinylated protein.
10. Lyse the cells by sonication on ice at 50% duty, 20% power in five treatments of 30 s each with 30 s cooling interval.
11. Remove cell debris by centrifugation (20,000g, 20 min, and 4°C).
12. Collect clarified cell lysate (supernatant) for spotting onto avidin slides.

13. The cell lysate can be spotted directly onto the avidin slides without any further treatment (see Note 2).
14. Confirm the presence of biotin label on the target protein by Western blot using HRP-conjugated antibiotin antibody.

3.2. Cell-Free Protein Biotinylation

This section outlines the steps for PCR-based in vitro biotinylation of proteins using puromycin. By utilizing suitable amounts of a puromycin-conjugated biotin in a cell-free protein expression reaction, we were able to efficiently incorporate the puromycin-biotin moiety to the carboxyl-terminal end of the protein (see Fig. 1C), which could be used directly for subsequent immobilization in a protein microarray.

1. Prepare reaction solution in one 0.2-mL PCR tube from the Rapid translation system RTS 100 *E. coli* HY kit (i.e., 12 µL *E. coli* lysate, 10 µL reaction mix, 12 µL amino acids, 1 µL of 1 mM methionine, 5 µL of reconstitution buffer) on ice.
2. Add DNA template: 0.5 µg of plasmid DNA, 0.5 µg of linear template generated via a standard PCR reaction or 0.1 µg of linear template generated via two-step PCR containing gene of interest and appropriate T7 regulatory regions (see Note 3).
3. Add 5'-biotin-dC-Pmn to a final concentration of 35 µM and RNase-free deionized water to a final reaction volume of 50 µL.
4. Start the reaction at 30°C for 6 h in a thermal cycler (see Note 4).
5. Remove reaction solution from the thermal cycler and store it at -20°C until further processing.
6. Use 5 µL of the reaction solution for Western blot to confirm presences of the biotin-labeled target protein.
7. For downstream microarray application, proteins generated via in vitro expression reaction need to be desalted. Prepare the MicroSpin™ G-25 column by resuspending the resin in the column (vortexing gently).
8. Loosen the cap one-fourth turn and snap off the bottom closure.
9. Place the column in a 1.5-mL screw-cap microcentrifuge tube for support. Alternatively, cut the cap from a flip-top tube and use this tube for support.
10. Pre-spin the column for 1 min at 735g.
11. Place the column in a new 1.5-mL tube and slowly apply the reaction solution to the center of the angled surface of the compacted resin bed, being careful not to disturb the resin. Careful application of the reaction solution to the center of the bed is essential for good separation. Do not allow any of the reaction solution to flow around the sides of the bed.
12. Spin the column for 2 min at 735g and collect the desalted reaction solution at the bottom of the support tube.
13. Discard the column, and the reaction solution is ready for spotting onto avidin slides.

3.3. Immobilization of Biotin-Labeled Proteins

Biotin-labeled proteins can be immobilized onto the avidin-functionalized slides, using the clarified cell lysate or reaction solution obtained from the procedures described in Subheadings 3.1. and 3.2., following the steps described below.

3.3.1. Preparation of Avidin-Functionalized Slide

1. Clean glass slides in piranha solution for at least 2 h (see Note 5).
2. Wash the slides copiously with deionized water, rinse with 95% ethanol, and finally dry the slides.
3. Soak the freshly clean slides in glycidyloxypropyltrimethoxysilane for 1 h.

4. Place the derivatized slides in a slide holder and wash two two three times with 95% ethanol.
5. Cure slides at 150°C for at least 2 h (overnight curing gives the same result). Rinse the slides with ethanol and dry.
6. Add 40–60 μ L of 1 mg/mL avidin onto the slides, cover with cover slip and incubate for 30 min (see Note 6).
7. Subsequently, wash the slides with deionized water in slide tray and dry the slides.
8. React the remaining epoxides by adding 2 mM aspartic acid onto the slides and covering with cover slip.
9. Finally, wash the slides with deionized water and dry them for spotting.

3.3.2. Immobilization of Biotin-Labeled Proteins onto Avidin-Functionalized Slides

1. Add 10 μ L of the clarified cell lysate or reaction solution into source plate.
2. Spot the cell lysate or reaction solution onto the avidin-functionalized slides using an ESI SMA™ arrayer.
3. Incubate the spotted slides at room temperature for approx 2–3 h.
4. Wash spotted slides with PBS for a few minutes before drying in air.
5. To visualize the immobilized proteins on the avidin slides, incubate the spotted slides with fluorescently labeled monoclonal antibody for 1 h (see Note 7).
6. Wash the slides twice with PBST on an orbital shaker (each time 15 min).
7. Finally, rinse the slides with distilled water to remove salt debris.
8. Dry slide and visualize spots with an ArrayWoRx microarray scanner.

3.4. Expression and Immobilization of N-Terminal Cysteine Proteins

In a separate but complementary method, the *Ssp* intein tag is used to generate N-terminal cysteine containing proteins for site-specific immobilization onto thioester-functionalized glass slides by means of a highly specific chemical reaction known as native chemical ligation (Yeo, S.Y.D, Yao, S.Q., unpublished results) (14,15). Only the N-terminal cysteine moiety would react with the thioester on the glass surface to form a stable native peptide bond (see Fig. 1A), while the presence of other reactive amino acid side chains, including internal cysteines, is tolerated (14,15). The expression and immobilization of N-terminal cysteine proteins are described in Subheadings 3.4.1.–3.4.3. The steps for (a) the cloning of target genes into pTWIN expression vector, (b) the expression of N-terminal cysteine proteins, and (c) the immobilization of N-terminal cysteine proteins onto thioester slides will be covered in this section.

3.4.1. Cloning of Target Genes into pTWIN Expression Vector

To generate an N-terminal cysteine protein, the target gene has to be first cloned into pTWIN expression vectors (see Note 8) using standard molecular biology techniques (13). The target gene fragment was first PCR amplified using an upstream primer (5'-GGT GGT TGC TCT TCC AAC TGC AGA GCC N_{15–18}-3') containing a *Sap*I site (underlined) and a cysteine residue (TGC) immediately after the last amino acid (AAC) of the intein with the sense strand sequence of the target gene. The downstream primer contains a *Pst*I site (5'-GGT GGT CTG CAG tta N_{15–18}-3') followed by the antisense strand sequence of a translation stop codon (TTA) and the C-terminal of the target gene. The PCR product was then double digested with *Sap*I and *Pst*I and gel purified using Qiaquick spin column. The gel-purified fragment was ligated to the *Sap*I-*Pst*I digested and dephosphorylated pTWIN vectors to yield the intein fusion construct, which was then transformed into *E. coli* DH5 α cells by the heat shock method. The

transformed cells were plated on LB plates containing 100 µg/mL of ampicillin and incubated overnight at 37°C. Positive clones carrying the desired expression plasmid were selected by colony PCR and grown overnight in LB broth with ampicillin. The final constructed expression plasmid, shown to be free of mutation by DNA sequencing, was then transformed into *E. coli* ER2566 host for protein expression.

3.4.2. Expression of N-Terminal Cysteine Proteins

1. Inoculate 2 mL of freshly grown transformed ER2566 cells into 200 mL of LB medium supplemented with 100 µg/mL ampicillin.
2. Incubate the culture at 37°C in a 250-rpm incubator shaker to an OD₆₀₀ of approx 0.5 (about 3 h).
3. Add IPTG to a final concentration of 0.3–0.5 mM to induce fusion protein expression.
4. Incubate the culture overnight at room temperature on an orbital shaker. For optimization of in vivo cleavage of fusion protein, incubate the culture for at least 18 h before harvesting (see Note 9).
5. Harvest cells by centrifugation (6000g, 15 min, 4°C).
6. Discard supernatant and resuspend cell pellet in 5 mL lysis buffer.
7. Lyse bacterial cells by sonication on ice at 50% duty, 20% power, in three treatments of 30 s each with 30-s cooling interval.
8. Centrifuge the cell lysate at 20,000g, 30 min, and 4°C.
9. Use clarified supernatant for direct spotting onto thioester glass slides.

3.4.3. Preparation of Thioester Slides and Immobilization of N-Terminal Cysteine Proteins

1. Incubate epoxy-derivatized slides with 10 mM diamine-PEG for 30 min. Details of the slide preparations had been previously described (7,8).
2. Wash slides with deionized water and place them in a solution of 180 mM succinic anhydride for 30 min and then in boiling water for 2 min.
3. Prepare NHS solution and incubate it with the slides for 3 h.
4. Rinse slides with deionized water and react overnight with a solution of benzylmercaptan.
5. Finally, wash the slides with deionized water and dry them for spotting.
6. Add 10 µL of the clarified cell lysate from Subheading 3.4.1. into source plate.
7. Spot the cell lysate onto the thioester-functionalized slides using an ESI SMA arrayer, and incubate for approx 10 min.
8. Wash spotted slides with PBS for a few minutes before drying in air.
9. Slides ready for detection by fluorescently labeled monoclonal antibody as described in Subheading 3.3.2. (see Note 7).

4. Notes

1. Our experimental results show that a substantial level of the target protein can be labeled with biotin in vivo using *Sce* intein. This is achieved simply with the addition of dithiothreitol (DTT) and cysteine-biotin to the induced bacterial culture. Cleavage and biotinylation of the target protein occurs within the bacterial cells at 4°C overnight. Though in vivo biotinylation of proteins using *Sce* intein is less efficient as compared to its in vitro system (6,12), nevertheless it provides an alternative method for researchers to biotinylate expressed proteins for subsequent spotting onto an avidin-functionalized slide.
2. No further treatment of the clarified cell lysate was needed prior to spotting, since trace amounts of the cysteine-biotin probe and endogenous biotinylated protein (acetyl-CoA carboxylase) in the *E. coli* lysate did not seem to interfere with binding of the target protein to the avidin slide.

3. Any vector or linear DNA to be used in combination with the Rapid Translation System must include the following elements and structural features: (1) target gene under control of T7 promoter located downstream of a ribosomal binding site (RBS) sequence, (2) distance between T7 promoter and start ATG should not exceed 100 base pairs, (3) distance between the RBS sequence and start ATG should be more than five to eight base pairs, and (4) T7 terminator sequence at the 3' end of the gene. A two-step PCR protocol is recommended for incorporation of the 5' and 3' T7 regulatory regions into the linear template. The purity ($OD_{260/280} = 1.7$) of plasmids obtained from commercially available DNA preparation kits is sufficient for the use as template in the Rapid Translation System.
4. Optimal temperature for most protein synthesis is 30°C. However, lower temperatures may be used for proteins that tend to aggregate. Protein synthesis can proceed for up to 6 h, but the synthesis reaction is usually 90% complete after 4 h.
5. When handling the slides, care must be taken to ensure that the slide is kept clean at all times, and that nothing comes into contact with spotting surface. Dust especially may result in extraneous fluorescence and may affect the fluorescent readout when the slide is scanned. Also gloves, if used, should be of the powder-free variety to ensure that the slides remain uncontaminated even after handling.
6. If there is sufficient reagent, it may be convenient to react both surfaces of the slides by placing them on slide racks in deep-well dishes. However, for expensive reagents, where it is preferable to utilize a conservative volume of the chemical, coverslips may be used. For a 22 × 60 mm coverslip, a 50-μL preparation is sufficient to allow for confluent coverage. Two methods may be used to apply the reagent on the surface. Either the reaction mix is first applied to the slide, and the coverslip is applied, or it could be applied to the coverslip and the slide may be inverted upon it. Both methods work equally well, but one ought to use the method that would allow production of a uniform spread of the reagent across the slide surface, without introducing any bubbles or voids between the coverslip and the slide (where the reagent does not come into contact with the slide surface). Coverslips may be slid off the slide once the reaction is complete, or be removed by vigorously shaking the slide in a water (or solvent) bath, until the coverslip slowly comes off.
7. Fluorescently labeled monoclonal antibody against target protein can be used to confirm successful immobilization of biotinylated proteins onto avidin slides. Some of these fluorescently labeled monoclonal antibodies may be commercially available, while others might require self labeling using fluorescent dye.
8. pTWIN vectors are cloning vectors designed for expression, labeling, and cyclization of recombinant proteins. Cloning of target gene within the multiple cloning site (MCS) allows fusion of an intein tag to the N-terminus, C-terminus, or both, of the cloned target protein. pTWIN1 is identical to pTWIN2 except that the *Mxe* RIR1 intein is substituted for the *Mth* GyrA intein. Both pTWIN vectors encode for the *Ssp* DnaB intein, upstream of the MCS, required for the expression of N-terminal cysteine proteins.
9. The expression level and amount of cleavage of fusion protein from the pTWIN vector is greatly influenced by the nature of the fusion protein and induction conditions such as induction temperature, duration of induction, and IPTG concentration. Different induction conditions (e.g., 30°C for 3 h, 20–25°C for 6–8 h, or 12–16°C overnight) should therefore be tested out for each type of fusion protein. Cleavage of fusion protein *in vivo* should be optimized for spotting of crude cell lysate onto thioester slides. In general, higher temperatures (room temperature and above) and longer induction duration (typically more than 18 h at room temperature) should be used to obtain maximum *in vivo* cleavage of fusion protein.

Acknowledgments

This work was supported by the National University of Singapore (NUS) and the Agency for Science, Technology and Research (A*STAR) of Singapore.

References

1. Hu, Y., Huan, X., Chen, G. Y. J., and Yao, S. Q. (2004) Recent advances in gel-based proteome profiling techniques. *Mol. Biotechnol.* **28**, 63–76.
2. Chen, G. Y. J., Uttamchandani, M., Lue, Y. P. R., Lesaicherre, M. L., and Yao, S. Q. (2003) Array-based technologies and their applications in proteomics. *Curr. Top. Med. Chem.* **3**, 705–724.
3. Schweitzer, B. and Kingsmore, S. (2002) Measuring proteins on microarray. *Curr. Opin. Biotechnol.* **13**, 14–19.
4. MacBeath, G. and Schreiber, S. L. (2000) Printing proteins as microarrays for high-throughput function determination. *Science* **289**, 1760–1763.
5. Zhu, H., Bilgin, M., Bangham, R., et al. (2001) Global analysis of protein activities using proteome chips. *Science* **293**, 2101–2105.
6. Lesaicherre, M. L., Lue, Y. P. R., Chen, G. Y. J., Zhu, Q., and Yao, S. Q. (2002) Intein-mediated biotinylation of proteins and its application in a protein microarray. *J. Am. Chem. Soc.* **124**, 8768–8769.
7. Lesaicherre, M. L., Uttamchandani, M., Chen, G. Y. J., and Yao, S. Q. (2002) Developing site-specific immobilization strategies of peptides in a microarray. *Bioorg. Med. Chem. Lett.* **12**, 2079–2083.
8. Lesaicherre, M. L., Uttamchandani, M., Chen, G. Y. J., and Yao, S. Q. (2002) Antibody-based fluorescence detection of kinase activity on a peptide array. *Bioorg. Med. Chem. Lett.* **12**, 2085–2088.
9. Paborsky, L. R., Dunn, K. E., Gibbs, C. S., and Dougherty, J. P. (1996) A nickel chelate microtiter plate assay for six histidine-containing proteins. *Anal. Biochem.* **234**, 60–65.
10. Hodneland, C. D., Lee, Y. S., Min, D. H., and Mrksich, M. (2002) Selective immobilization of proteins to self-assembled monolayers presenting active site-directed capture ligands. *Proc. Natl. Acad. Sci. USA* **99**, 5048–5052.
11. Kindermann, M., George, N., Johnsson, N., and Johnsson, K. (2003) Covalent and selective immobilization of fusion proteins. *J. Am. Chem. Soc.* **125**, 7810–7811.
12. Lue, R. Y., Chen, G. Y., Hu, Y., Zhu, Q., and Yao, S. Q. (2004) Versatile protein biotinylation strategies for potential high-throughput proteomics. *J. Am. Chem. Soc.* **126**, 1055–1062.
13. Sambrook, J., Fritsch, E. F., and Maniatis, T. (1989) *Molecular Cloning, A Laboratory Manual*, second ed. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
14. Dawson, P. E., Muir, T. W., Clark-Lewis, I., and Kent, S. B. H. (1994) Synthesis of proteins by native chemical ligation. *Science* **266**, 776–779.
15. Muir, T. W., Sondhi, D., and Cole, P. A. (1998) Expressed protein ligation: a general method for protein engineering. *Proc. Natl. Acad. Sci. USA* **95**, 6705–6710.
16. Uttamchandani, M., Chan, E. W. S., Chen, G. Y. J., and Yao, S. Q. (2003) Combinatorial peptide microarrays for the rapid determination of kinase specificity. *Bioorg. Med. Chem. Lett.* **13**, 2997–3000.

A Guide to Protein Interaction Databases

Tiffany B. Fischer, Melissa Paczkowski, Michael F. Zettel, and Jerry Tsai

1. Introduction

With the continual completion of new genomes, discovering the purpose of all these nucleotide sequences takes on greater implications. Proteins, the predominant products of gene sequences, have been a natural place to start, and projects have begun characterizing all the proteins from a genome (the proteome). One particular aspect of the proteome that has been amenable to large-scale bioinformatics studies is the identification of interacting proteins and the mapping of protein interaction networks (a complete set for a genome is commonly referred to as the *interactome*). Because studies of protein interaction produce large amounts of data, the challenge has become how to present such data sets in a meaningful and informative manner, so that they are a resource for the general biological community. The typical scientific medium to share information is publication in a journal article. As a physical medium of paper and text, publications are inadequate at presenting such large sets of information and are limited to an overview and some general conclusions about the data. Instead, presentation of these large data sets has taken advantage of the relational capabilities provided by computers and broad accessibility provided by the Internet. These protein interaction data sets are stored in a database, enabling simple implementations of search and browse functions, and are presented on the Internet with a World-Wide Web front end. As the amount of the protein interaction data has increased, so too has the number and variety of databases. This chapter is intended for the general research community as a guide to increase the visibility and accessibility of these information resources on protein interactions.

As pointed out above, these interaction databases are meant as proteomic resources. While they hold a wide range of information, one of the primary uses of protein interaction data is to construct signal transduction pathways. In particular, researchers can find a protein's role(s) in the interactome. Such mapping provides information about putative interacting proteins as well as alternative roles in cellular function. Additionally, the placement of a protein in the interactome hints at a protein's regulatory capacity: which network controls a protein's function and/or whether the protein acts as a control point. For a protein of unknown function, knowledge of a protein's interactive characteristics can help identify its functional role. As an information resource on protein interactions, these databases should help researchers make informed decisions about their protein of interest and potential directions for future research.

Effectively, these protein interaction databases are meta-servers of experimental data. Many of these experiments are high-throughput methods testing protein interactions (1) that provide direct data for these databases. An indirect method, nucleotide arrays have been widely used to identify co-expression and therefore potentially interacting proteins (2–4). More direct measures of protein interactions are the yeast two-hybrid system (5) and two-dimensional gel electrophoresis (6), which has evolved into column separation of protein complexes and identification using mass spectrometry (7). Another rich source of protein interaction information that is also used to validate the large-scale studies is in the primary scientific literature (8). As cataloged by Medline and PubMed, this literature is now the focus of data-mining efforts to find protein–protein interactions (9). Additionally, the Protein Data Bank (10) contains direct structural verification of interacting proteins. Depending on their focus, protein interaction databases serve up different slices of the complete interactome. Some attempt to be comprehensive and others are species specific or focus on a category of interactions.

Finding the appropriate databases and making sense of the different content that they provide can be somewhat overwhelming. An exhaustive search of protein-interaction databases was performed, which has been listed in the following tables. An honest attempt has been made to make this list as comprehensive as possible, and most of the major protein interaction databases are listed. After reviewing each database in the list individually for content and presentation, some databases were chosen for a more detailed review in the following sections. Selection of a database was based on its contents, ease of use (especially search functions), and clarity in the presentation of results. At the current time, if a database could not be accessed for review, it was not included. Information is as up to date as possible at the time of writing. Because of the current dynamic nature of this field, we would like to warn the reader that some of the following databases will have changed by the time of publication. Also, to simplify some discussions, acronyms for databases mentioned below are used many times without expansion to its full name. The full name and information on these databases, as well as the other databases not reviewed, are in **Tables 1–3**. The databases are loosely categorized by their coverage and contents. Comprehensive interaction databases (**Table 1**) are those that encompass most if not all interactions with proteins (protein–protein, protein–nucleic acid, and protein–ligand/substrate) and consist of the more established databases that produce protein network maps of interactomes. **Table 2** consists of all the other protein–protein interactions databases. This table is further subdivided into sections depending on the information presented by the database: organism specific, structure/mutational, and general. The last table describes the protein–nucleic acid and protein–ligand/substrate databases (**Table 3**). While primarily focused on the end-user, this chapter is also targeted to database designers as a guide to constructing more transparent interfaces for their protein interaction data.

2. Overview of Databases

2.1. Comprehensive Interaction Databases

With the boom of protein interaction information, several interaction databases have been developed to completely catalog the wide variety of protein interaction data: protein–protein, protein–nucleic acid, and protein–ligand. **Table 1** lists a group of such databases. The interaction data are usually from a number of sources: high-throughput

Table 1
Comprehensive Interaction Databases

Name	Abbr.	Entries*	Methods	URL
Database of Interacting Proteins	DIP	31,777	Various	dip.doe-mbi.ucla.edu
Biomolecular Interaction Network Database	BIND	41,776	Various	bind.ca
Molecular Interactions Database	MINT	15,068	Direct and indirect binding	cbm.bio.uniroma2.it/mint
Kinetic Data Bio-Molecular Interactions	KDBI	1231	Kinetic measurements	xin.cz3.nus.edu.sg/group/kdbi/kdbi.asp
Binding Database	BindingDB	438	ITC & enzyme inhibition	www.bindingdb.org

*Entries indicate the number of interactions found for that database as of November 1, 2003. The values shown act as a guide for comparison, as they are expected to change.
Abbr, abbreviation; ITC, isothermal titration calorimetry.

Table 2
Protein–Protein Interaction Databases

Name	Abbr.	Entries*	Organism(s)	Methods	URL
Yeast Interacting Protein Database	YPD	15,068	<i>S. cerevisiae</i>	Yeast two-hybrid	genome.c.kanazawa-u.ac.jp/Y2H/
Comprehensive Yeast Genome Database	CYGD	10,570	<i>S. cerevisiae</i>	Various	mips.gsf.de/genre/proj/yeast/index.jsp
Kinetic Data Bio-Molecular Interactions	MDS Proteomics	3617	<i>S. cerevisiae</i>	Mass spectrometry	www.mdsp.com/yeast/
Biomolecular Relations in Information Transmission and Expression	BRITE	N/A	<i>S. cerevisiae, H. pylori, E. coli</i>	Yeast two-hybrid	www.genome.ad.jp/brite/brite.html
Human Protein Reference Database	HPRD	13,194	Human	Various	www.hprd.org
Human Unidentified Gene-Encoded Protein–Protein Interaction Database	HUGE ppi	124	Human	Yeast two-hybrid	www.kazusa.or.jp/huge/ppi/
Cell Signaling Networks Database	CSNDB	N/A	Human	Literature search	geo.nihs.go.jp/csndb/
Subtilis Protein Interaction Database	SPiD	95	<i>B. subtilis</i>	Yeast two-hybrid	www.mig.jouy.inra.fr/bdsi/SPiD/
Drosophila Protein–Protein Interaction Map	FlyNet	20,405	<i>D. melanogaster</i>	Yeast two-hybrid	www.jubimed.org/perl/flynet.pl
Drosophila Protein Interaction Map Database	PIM	129	<i>D. melanogaster</i>	Yeast two-hybrid	proteome.wayne.edu/PIMdb.html

Protein–Protein Interactions Tables for Human herpesvirus 1	PPIHH1	62	Human herpesvirus 1	Various	www.stdgen.lanl.gov/cgi-bin/ pp.cgi?dbname=hhv1
Protein–Protein Interaction Panel using Full-length cDNAs	PPIP	145	Mouse	Yeast two-hybrid	genome.gsc.riken.jp/ppi/
<i>Structure/mutational</i>					
Surface Properties of Interfaces—Protein –Protein Interfaces	SPIN-PP	855	Many	Crystallography & NMR	trantor.bioc.Columbia.edu/cgi-bin/SPIN/
Thermodynamic Database for Proteins and Mutants	ProTherm	619	Many	Various	gihk26.bse.kyutech.ac.jp/jouhou/Protherm/ Protherm.html
Binding Interface Database	BID	308	Many	Mutagenesis	tsailab.org/BID/
Alanine Scanning Energetics Database	ASEdb	98	Many	Alanine scanning	www.asedb.org/
Database of Protein– Protein Complexes	COMBASE	N/A	Many	Crystallography & NMR	salilab.org/sub-pages/combase.html
<i>General</i>					
Kinase Pathway Domain Functional Association of Proteins in Complete Genomes	KPD Allfuse	78,029 66,406	Many Many	Abstract search Theoretical: gene fusion	kinasedb.ontology.ims.u-tokyo.ac.jp/ maine.ebi.ac.uk:800/services/allfuse/
Database of Interacting Domains	InterDom	15,058	Many	Theoretical and experimental	interdom.lit.org.sg/
General Repository for Interaction Datasets	GRID	21,236	Many	Various	biodata.mshri.on.ca/grid

Table 2 (Continued)
Protein–Protein Interaction Databases

Name	Abbr.	Entries*	Organism(s)	Methods	URL
Database of Oligomerization Domains from Lambda	Doodle	N/A	Many	Lambda repressor	oligomer.tamu.edu/
Genomic Knowledge Database	GKD	N/A	Many	Theoretical: gene fusion	big.gsc.riken.go.jp/GKDtext.htm
PDZ Domains and their binding partners	PDZ	N/A	Many	Various	mamba.bio.uci.edu/%7Epjibryant/lab/ PDZ_binders.html
Signaling pathway database for genetic information and signal transduction systems	SPAD	N/A	Many	Various	www.grt.kyushu-u.ac.jp/eny-doc/spad.html
Charting pathways of life	BioCarta	N/A	Many	Various	www.biocarta.com/genes/index.asp
Path database	PathDB	N/A	Many	N/A	www.ncgr.org/pathdb/
Protein to Protein Interaction Database	PPID	N/A	N/A	N/A	www.anc.ed.ac.uk/mscs/PPID/

*Entries indicate the number of unique interactions found for that database as of November 1, 2003. The values shown act as a guide for comparison, as they are expected to change.
 Abbr., abbreviation; NMR, nuclear magnetic resonance.

Table 3
Other Protein Interaction Databases

Name	Abbr.	Entries*	Contents	URL
<i>Protein–nucleic acid interactions</i>				
PathoDB	PathoDB	10,450		www.gene-regulation.com/cgi-bin/pub/databases/pathodbd/search.cgi
A Database on Transcriptional Regulation and Genome Organization	RegulonDB	179		www.cifn.unam.mx/Computational_Genomics/regulondb/
Kinetic Data Biomolecular Interactions	ProNIt	164		gibl26.bse.kyutech.ac.jp/jouhou/pronit/pronit.html
<i>Protein–ligand interactions</i>				
The Comprehensive Enzyme Information System	BRENDA	Many	Various	www.brenda.uni-koeln.de/
The University of Minnesota Biocatalysis/Biodegradation Database	UM-BBD	900	Enzymes	umbbd.ahc.umn.edu/
Protein ligand interaction data Enzymes and Metabolic pathways database	PLD EMP DB	273 30,000	Various Enzymes only	www.mitchel.ch.cam.ac.uk/pld/index.html emp.mcs.anl.gov/
GroEL Protein interaction database Protein Function and Biochemical Pathways	GELPI PFBP	40 N/A	Various Enzymes only	bioc02.uthscsa.edu/~seale/Chap/el_int.html www.ebi.ac.uk/research/pfmp/
MHC peptide interaction database Comprehensive Database of MHC Binding and Non-binding Peptides	MPID MHCDB	90 N/A	Many Protein/peptide	surya.bic.nus.edu.sg/mpid www.hgmp.mrc.ac.uk/Registered/Option/mhcdb.html
A database of MHC Binding Peptides (v1.3)	MHCPEP	N/A	Protein/peptide	wehi.wehi.edu.au/mhcpep/

*Entries indicate the number of unique interactions found for that database as of November 1, 2003. The values shown act as a guide for comparison, as they are expected to change.

Abbr., abbreviation; MHC, major histocompatibility complex.

experimental methods such as yeast two-hybrid (5) and lambda repressor (11) systems as well as data-mining the literature (12). The vast number of protein-protein interactions from these data can be used to create methods for categorizing protein-protein interactions, assisting researchers in correlating sequence to function (13–15), and predicting protein-protein interactions (16). In addition, because these experiments can create variable results, cross-validating the results from different experiments and extracting overlapping information can increase the accuracy of interaction prediction (17). Each of the databases listed in **Table 1** is reviewed in more detail. The descriptions are organized to explain the general purpose and contents of the database, followed by an explanation of a database's search abilities and the type of results to expect.

2.1.1. Database of Interacting Proteins

Developed at University of California, Los Angeles, the Database of Interacting Proteins (DIP) is a relational database that provides experimentally determined protein-protein, protein-nucleotide, and protein-ligand interactions (18). The homepage is shown in **Fig. 1A**. Interactions from a wide variety of experimental methods, including high and low throughput and some quantified results, were obtained manually (19) and by automated methods (12). In addition to the primary sources, the DIP draws on data from a number of other databases: organism-specific databases, like the Yeast Protein Database (YPD) (20), EcoCyc (21), and FlyNet (22), as well as pathways databases, like Kyoto Encyclopedia of Genes and Genomes (KEGG) (23) and the Italian National Council of Students in Biotechnology (CNSB) (24). A curator first manually enters information from journals. These data are automatically tested to show that the proteins and citations exist. A second curator then rechecks the data. The DIP is regularly updated and has grown from 1089 proteins in 1999 to 7141 proteins covering 31,777 interactions in 2003 (**Fig. 1B** and **Table 1**, respectively). New additions to the dataset can be checked in the “News” link from the DIP’s homepage. This database is useful in identifying interacting partners with your protein of interest and visualizing the interactions within and between different pathways. The available maps also provide researchers with a confidence level for every interaction, as indicated by the widths of lines connecting interacting proteins.

Members of the academic community can gain free access to the DIP by following a registration process, linked directly on the homepage. The “Help” link on the sidebar provides definitions and guidance and is a useful resource for new users. In the search page shown in **Fig. 2A**, proteins can be found by protein name (node), sequence motif, or article. When performing a “node search” by protein name, it is useful to start with a very general term and then skim the list of results for the protein of interest. For example, if the protein of interest is mitogen-activated protein kinase (MAPK), a search would begin with the word “kinase” and the organism of interest. The search results are listed as in **Fig. 2B**. Once the protein of interest has been found in the node search, choosing the “links” field will produce a new window showing interactions. The interacting proteins can be browsed as seen in **Fig. 2C**. To view details and eventually interaction maps of a protein, the link under the DIP Node column on the search results page (**Fig. 2B**) can be selected. In the top right-hand corner of the resulting DIP Node page (**Fig. 2D**) you can click the “graph” link to view the interaction map (**Fig. 2E**). This is an interactive map, in which the various nodes connected to your protein can be clicked on to view their details. The map is also useful in that the different widths

B
A

Database of Interacting Proteins					
DATA SOURCE/LEVEL		INTERACTIONS			
Protein		Protein		Protein	
Source	Source	Source	Source	Source	Source
Number of proteins					
Scopdb	Scopdb	Scopdb	Scopdb	Scopdb	Scopdb
Number of articles	4724	1,5158	1,6705	1,6705	1,6705
Number of organisms	104	104	104	104	104
Number of interactions	18700	18700	18700	18700	18700
Number of distinct experiments describing an interaction	22971	22971	22971	22971	22971
Number of articles	2519	2519	2519	2519	2519
DATABASE STATISTICS					
ORGANISM		PROTEINS		INTERACTIONS	
Saccharomyces cerevisiae (baker's yeast)		1,5158		1,6705	
<i>Helicobacter pylori</i>		1425		1425	
<i> Homo sapiens</i> (Human)		988		1356	
<i> Escherichia coli</i>		516		969	
<i> Mus musculus</i> (house mouse)		249		321	
<i> Rattus norvegicus</i> (Norway rat)		105		150	
<i> Drosophila melanogaster</i> (fruit fly)		98		115	
<i> Bos taurus</i> (cow)		50		54	
Others (90)		265			

Fig. 1. Database of Interacting Proteins (DIP). (A) The DIP homepage provides links to current updates, registration, statistics, search page, related links, and a help page. (B) The statistics link displays basic figures for the current holdings.

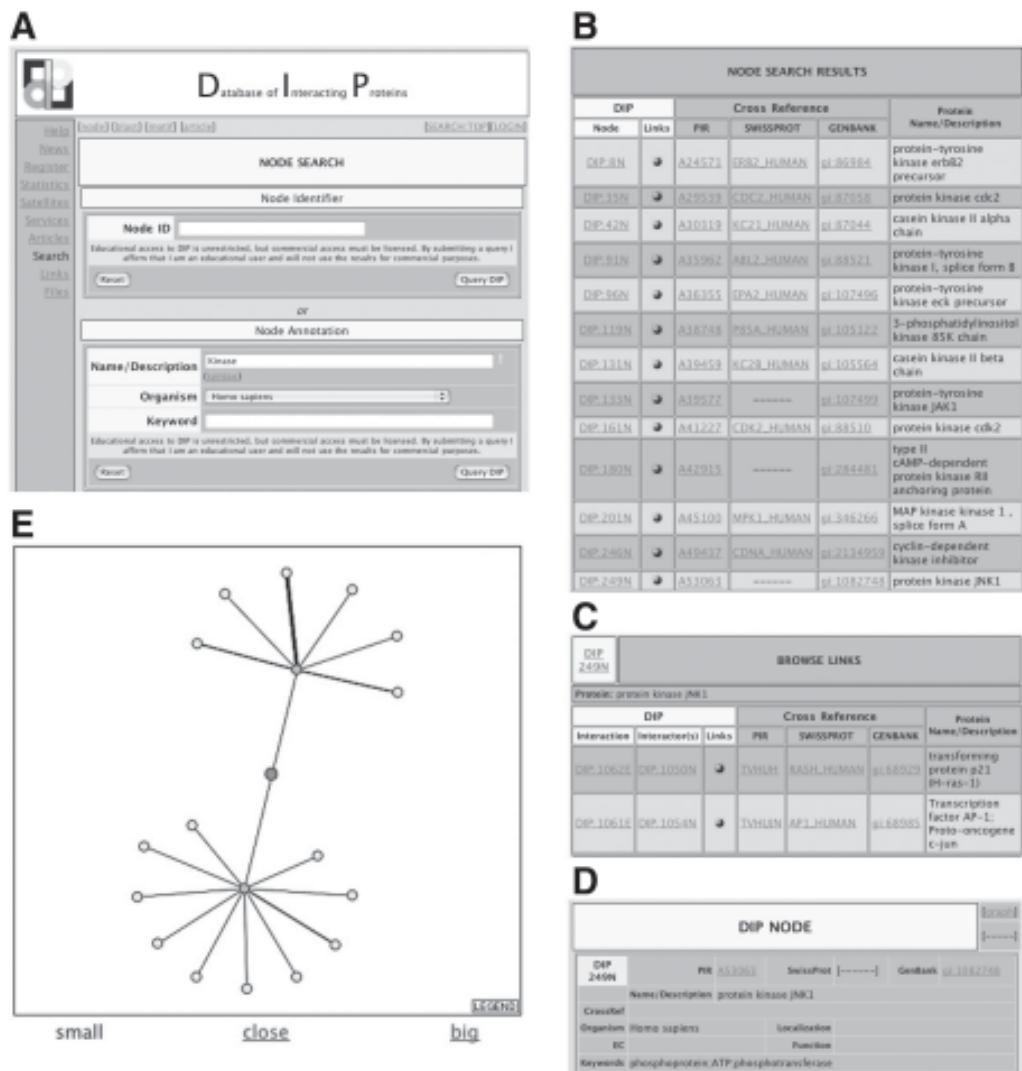


Fig. 2. Search and results of DIP. The search link on the DIP homepage takes you to a site that allows you to perform searches by protein under the node link, sequence, motif, or by article. (A) When performing a “node search” by protein name, it is useful to start with a very general term and then browse the list of results for the protein of interest. For example, if your proteins of interest are mitogen-activated protein kinases (MAPK), then you would search for kinase and the organism of interest, and browse the results, in which JNK1 would appear. (B) Once you have found your protein in the node search, interactions can be found by choosing the “links” field. The interacting proteins will be listed as seen in (C). To view details and interaction maps of a protein, in the results page you can click the DIP node for a particular protein, as seen in the top left-hand corner of (C) for “DIP249N.” In the top right-hand corner of the “DIP NODE” page (D) you can click the “graph” link to view the interaction map (E). This is an interactive map, in which you may click on various nodes connected to your protein to view details. The map is also useful in that the different widths connecting various nodes demonstrate the confidence level of that interaction.

connecting various nodes demonstrate the confidence level of that interaction being valid. The information in the DIP is available for download.

2.1.2. Biomolecular Interaction Network Database

The Biomolecular Interaction Network Database (BIND) offers tremendous amounts of protein interaction data to investigators in the biological sciences (25). This database was created at the University of Toronto. Currently, the BIND is undergoing a transition (see **Fig. 3A,B**) to a new look. With approx 41,776 (**Table 1**) and growing interactions, the BIND is a monolith in the arena of interaction databases. As stated in **Fig. 3B**, interactions are defined between two objects that can be a protein, DNA, RNA, or ligand (26). Interactions contained in BIND are a result of literature searches or investigator submissions, which are reviewed by curators before incorporation. BIND has both a help page and a FAQ list for new users. Visualization tools are also available to help visualize complex multiprotein interactions.

Querying BIND is an easy task, and an example is shown in **Fig. 4** for a basic search. The basic search uses text entries to query BIND (**Fig. 4A**). The results page from a search shows interactions between what was queried and what is in the databases in a binary set (**Fig. 4B**). Querying the database for interactions based on a protein of interest generates a results page with interactions listed in sets of two—the protein of interest with its potential partners. The binary interaction sets have six descriptive columns that provide information about the specifics of the protein: description, molecular function, cellular component, biological process, experiments, and links. Under the description column is a brief summary of the protein or subunit. Molecular function describes the protein's activity (for instance, a kinase with regulatory activity). The cellular component column identifies where within a cell the protein is found, and the biological process column provides information regarding a protein's role within a cell, such as involvement in the G/S cell-cycle transitions. Experiments provide the method used to determine the interaction. The links column presents paths to additional information on the protein of interest, like the National Center for Biotechnology Information (NCBI), Saccharomyces Genome Database (SGD), and SeqHound, as well as links to PubMed and “other BIND data.” The results of the BLAST search produce a list of proteins of similar sequence with links to their individual BIND entries. The preBIND result page lists the “potential interactors” in a column to the right and links to their original references.

A similar search page to the one shown in **Fig. 4A** has been created for the newer look of the BIND (**Fig. 5C**). Entries can range from a gene name to an open reading frame (ORF) to a biochemical pathway like glycolysis. Also, other methods are available to the user for finding interactions, such as a browsing function, an advanced search, a field-specific search tool, a Basic Local Alignment Search Tool (BLAST) search, and a preBIND search option (some of which are shown in **Fig. 5**). As shown in **Fig. 5A**, browsing can be done by the BIND record number or by the NCBI accession number. The advanced search is named the BIND Field Specific Search (**Fig. 5B**). With this tool, the user can specify whether to search interactions, complexes, or pathways, or any combination of the three (**Fig. 5B**). The number of results per page can be limited. On the BIND Field Specific Search page is an accession query tool. Accession query searches the BIND using any number of identification numbers: BIND record number, NCBI GI, NCBI publication, or NCBI taxonomy ID. The field makes use of



Blueprint

BluePrint is a new Cytoscape plugin for biomolecular interaction network analysis.



Biomolecular Interaction Network Database

SEARCH

Recent News

Prefers

BIND

BIND Development

Curations

Downloads

Feedback and Stats

Help

FAQ

Training

BioRxiv

Related Databases

Credits

Publications

Contact BIND

SEARCH

Recent News

Prefers

BIND

BIND Development

Curations

Downloads

Feedback and Stats

Help

FAQ

Training

BioRxiv

Related Databases

Credits

Publications

Contact BIND

LAUNCH SERVICE

- In a new browser window
- In the browser window

SEARCH

Recent News

Prefers

BIND

BIND Development

Curations

Downloads

Feedback and Stats

Help

FAQ

Training

BioRxiv

Related Databases

Credits

Publications

Contact BIND

STATISTICS

Recent News

Prefers

BIND

BIND Development

Curations

Downloads

Feedback and Stats

Help

FAQ

Training

BioRxiv

Related Databases

Credits

Publications

Contact BIND

B

Data Manager Menu

Version 2.0

About Help Browse

BIND Statistics

Search Browse

Administration

BIND stands for the Biomolecular Interaction Network Database.

Browse the BIND v3.1 ASN.1 Specification
The BIND specification has been published.

The Biomolecular Interaction Network Database (BIND) is a database designed to store full descriptions of interactions, molecular complexes and pathways. The BIND database is freely available to both academics and commercial researchers. Please click on the links at the bottom of this page to view our policies.

BIND contains interaction, molecular complex and pathway records.

An **Interaction** record is based on the interaction between two objects. An object can be a protein, DNA, RNA, ligand or molecular complex. Description of an interaction encompasses cellular location, experimental conditions used to observe the interaction, conserved sequence, molecular location of interaction, chemical action, kinetics, thermodynamics, and chemical state.

Molecular complexes are defined as collections of more than two interactions that form a complex, with extra descriptive information such as complex topology.

Pathways are defined as collections of two or more interactions that form a pathway, with extra descriptive information such as cell cycle stage.

If you use BIND, please cite:

Bader GD, Betel D, Hogue CW (2003) BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.* 31(1):248-50 PMID: 12519993

Mail: info@bind.ca to get more info.

Disclaimer **Privacy policy** **Data policy**

This file was last updated: Wednesday, June 18, 2003 20:28:47

Fig. 3. Biomolecular Interaction Network Database (BIND). (A) This is the BIND homepage; the simple search tool is located on the right side. Links and drop-down boxes for other BIND features are located on the left side of the page. (B) Original BIND search page.

A

Data Manager Menu
Version 2.0

[About](#) [Help](#)
[Search](#) [Browse](#)

BIND Text Query

Text Query:
Find in: Interactions Molecular Complexes Pathways
Number of records to show per page: 10

Field Specific Text Query Builder

Text Searching Help

Accession Query

Search by: BIND Accession ID
Integer ID:

Comments and suggestions to: <info@bind.ca>

[SLRI Bioinformatics Research Group Homepage](#)

B

Print Page

1 2 3 4 5 6 7 8 9 - ≥
Results 1 - 5 out of 41 in 9 total pages.

BIND Text Query

Text Query:
Find in: Interactions Molecular Complexes Pathways
Number of records to show per page: 5

Field Specific Text Query Builder

Results for CDC28

Interaction	378	Saccharomyces cerevisiae	Full BIND Record	Launch Viewer:	Select Below		
Molecule	CLB1	Gi(sub)2-specific B-type cyclin, Involved in mitotic induction	Molecular Function	Cellular Component	Biological Process	Experiment(s)	Links
			• cyclin-dependent protein kinase, intrinsic regulator activity	• cellular component unknown	• mitotic spindle assembly (sensu Saccharomyces) • regulation of CDK activity • G2/M transition of mitotic cell cycle • meiotic G2/M transition	Two Hybrid Test	NCBI SGD SeqHound Abstract [Pubmed] [Other BIND data]
CKS1		subunit of the Cdc28 protein kinase	• molecular function unknown	• cellular component unknown	• regulation of cell cycle		NCBI SGD SeqHound

Interaction	379	Saccharomyces cerevisiae	Full BIND Record	Launch Viewer:	Select Below		
Molecule	CLB2	Gi(sub)2-specific B-type cyclin, Involved in mitotic induction	Molecular Function	Cellular Component	Biological Process	Experiment(s)	Links
			• cyclin-dependent protein kinase, intrinsic regulator activity	• nucleus • cytoplasm	• regulation of CDK activity • G2/M transition of mitotic cell cycle	Two Hybrid Test	NCBI SGD SeqHound Abstract [Pubmed] [Other BIND data]
CKS1		subunit of the Cdc28 protein kinase	• molecular function unknown	• cellular component unknown	• regulation of cell cycle		NCBI SGD SeqHound

Fig. 4. Simple search of and results from BIND. (A) Simple search page for BIND. (B) Here is a BIND results page, generated from either a simple, advanced, or field-specific search.

A

Browse the BIND database

Browse by BIND record:	Interaction <input type="button" value="▼"/>
Browse by NCBI accession:	Select Below <input type="button" value="▼"/>
Number of records to show per page:	10 <input type="button" value="▼"/> <input type="button" value="Submit"/>

Comments and suggestions to: <info@bind.ca>

[SLRI Bioinformatics Research Group Homepage](#)

B

BIND Field Specific Search

Search for <input type="text" value="CLN3"/>	in record type <input type="button" value="Interaction"/> within <input type="button" value="All"/>	Interaction Description A Short Label B Short Label A Molecule Description	<input type="button" value="field(s)"/>
AND <input type="radio"/> OR <input type="radio"/> NOT <input type="radio"/>			

Search for <input type="text" value="CDC28"/>	in record type <input type="button" value="Interaction"/> within <input type="button" value="All"/>	Interaction Description A Short Label B Short Label A Molecule Description	<input type="button" value="field(s)"/>
AND <input type="radio"/> OR <input type="radio"/> NOT <input type="radio"/>			

Full Text Query

Comments and suggestions to: <info@bind.ca>

[SLRI Bioinformatics Research Group Homepage](#)

C

 **Blueprint** 

Bioinformatics.ca	Contact Us	Help	Search	Mount Sinai
Home	BIND	SebHound	Protein Folding	Downloads
About	Products	Services	Technical Support	Exhibitions
				News
				People
				Jobs

Blueprint Home

>**BIND**

>**BIND Search**

[Search Help](#)

BIND Search

BIND Text Query

Text Query:

Find in: Interactions Molecular Complexes Pathways

Number of records to show per page:

Accession Query

Search By:

[Policies](#) [Authors](#) [Feedback](#)

Boolean operators to refine a search of BIND. Multiple object searches can be constructed by defining their interaction type; interaction, complex, or pathway and selecting one of 30 different fields. The BLAST search option allows the user to search the BIND database using a particular sequence with either BLASTx or BLASTp. As its name suggests, preBIND implements a literature-mining algorithm based on a query for a particular interaction or protein and produces a list of articles meant for eventual human curation.

2.1.3. Molecular Interactions Database

The Molecular Interaction Database (MINT), developed by the University of Rome Tor Vergata (27), has recently moved from its original form (Fig. 6A) to adhering to the Proteomics Standard Initiative (28), which endeavors to create consistency in data format and exchange of proteomic data (Fig. 6B). Although it incorporates interactions with enzymatic modifications of one of the partners, the MINT focuses primarily on protein–protein interactions (Fig. 6C). Interactions are experimentally verified, encompassing both direct and indirect relationships, and are mined from scientific literature with the aid of a literature-mining program, the MINT assistant; the putative interactions are then established by expert curators. As of November 2003, the MINT contained 22 experimental procedures with a total of 15,068 interactions (Table 1) on two yeast partners and 2149 interactions with at least one mammalian partner (Fig. 6C).

MINT combines basic protein and gene information with binary interactions and compiles the data into a user-friendly database. Data queries are divided into three categories: accession number, protein or gene name, and keyword (Fig. 7A). As shown in Fig. 7B, the search result is a compilation of interacting proteins related to the search criteria as well as basic information for the protein, binary interactions, pathways, and complexes. Basic information consists of protein and gene names, protein description, amino acid length, accession numbers, posttranslational modifications, subcellular location and function, as well as links to other major Websites for sequence and domain information. Binary interactions offer several partner proteins that interact with the protein and explain the domain organization, functional relationships, and enzymatic modifications based on the experimental methods. One of the MINT's most innovative features is an interactive, Java-based viewer for the interactions (Fig. 7C), which can be accessed via the links below the interaction list (left side of Fig. 7B). The MINT viewer illustrates the relationship between interacting proteins using proteins as nodes and interactions as edges. Complex interaction maps can be formed. The original protein is displayed first, with a link to its information page, and its interacting partners bonded to it, which are linked to their corresponding binary interactions page (Fig. 7C). If the interacting partners contain interactions to other proteins, these can be built into

Fig. 5. (opposite page) Other Biomolecular Interaction Network Database (BIND) search options. (A) This is a browse option available to BIND users. Through this page, users can browse the BIND for interactions of interest. (B) This is a picture of the field-specific search option. As can be seen in the picture, Boolean operators allow the user to define a more thorough search. (C) This a picture of the advanced search option. The advanced search option allows the user to define whether to query interaction, molecular complex, or pathways. The user can also define the number of results displayed per page.

A



B



C



Number of interactions	
Number of interactions on two year partner:	12584
Number of interactions with at least one eponymous partner:	2148
Total interactions:	18068

Different experiments reporting the same interaction	
interactions verified 1 time	12887
interactions verified 2 time	1456
interactions verified 3 or more time	745

Relations to partners	
interactions related to 1 partner	14015
interactions related to 2 partner	827
interactions related to 3 or more partner	226

Number of interactions by detection method	
using method two hybrid pooling approach	4521
using method tag tag coimmunoprecipitation	3952
using method flag tag coimmunoprecipitation	2638
using method two hybrid array	1967
using method column coimmunoprecipitation	1513
using method two hybrid	1264
using method pull down	1028
using method prediction based on phage display consensus	296
using method experimental	288
using method peptide array	211
using method colocalization/localization technologies	194
using method filter binding	180
using method cross-linking studies	82
using method beta lactamase complementation	88
using method copurification	81
using method x-ray crystallography	36
using method phage display	33
using method enzyme-linked immunosorbent assay	28
using method ubiquitin reconstruction	14
using method fluorescent resonance energy transfer	13
using method cocondensation	3
using method beta galactosidase complementation	1

Number of interactions by organism	
interactions with at least one <i>Arabidopsis thaliana</i> (Mouse-ear cress) partner	25
interactions with at least one <i>Bacillus phage T7</i> partner	25
interactions with at least one <i>Bacillus</i> partner	35
interactions with at least one <i>Candida albicans</i> partner	69
interactions with at least one <i>Cards familiaris</i> (Drosophila) partner	20
interactions with at least one <i>Drosophila melanogaster</i> (Fruit fly) partner	60
interactions with at least one <i>Escherichia coli</i> partner	26
interactions with at least one <i>Escherichia coli</i> O197:H7 partner	29
interactions with at least one <i>Galleria gallin</i> (Silkworm) partner	33
interactions with at least one <i>Hom sapiens</i> (Human) partner	1728
interactions with at least one <i>Mac musculus</i> (Mouse) partner	422
interactions with at least one <i>Rattus norvegicus</i> (Rat) partner	245
interactions with at least one <i>Saccharomyces cerevisiae</i> (Baker's yeast) partner	12589
interactions with at least one <i>Vaccinia virus</i> (smallpox) partner	30
interactions with at least one <i>Xenopus laevis</i> (African clawed frog) partner	25

Fig. 6. Molecular Interactions Database (MINT). **(A)** The older version of MINT contains a brief description about the database, including information about the interactions housed, and how they are mined from the literature. Also included on the homepage are links for searches, statistics, submitting new information, and the database reference. **(B)** The test version of MINT was designed to incorporate their new query and results face for data searching. Also included in the changes are their method of data mining, the MINT viewer applet, and the ability to download the results as a flat file or XML file (coming soon!). **(C)** The Statistics link provides detailed information on the number of protein interactions and divides the list into categories: experiments reporting the same interaction, relations to PubMed, by detection method, and by organism.

the map by simply clicking on the circled “+” and then on the node itself (Fig. 7D). This can lead to the discovery of more interacting partners, resulting in a comprehensive interaction map containing the original protein as the base. In addition to its text representation of the data, this method to browse the interaction map is quite simple, but is also a very powerful and intuitive tool for finding relationships between proteins.

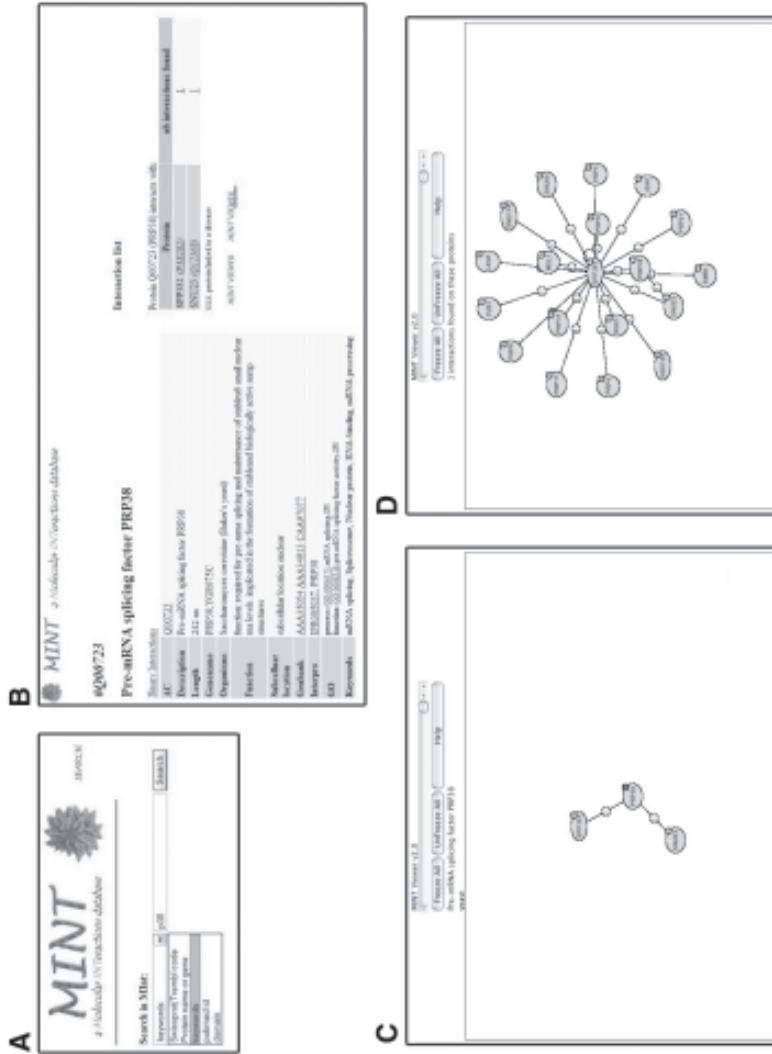


Fig. 7. Molecular Interactions Database (MINT) search and results. **(A)** The new test version includes a simplified search method—a drop-down list enabling the user to pick one of five query criteria for their searching. **(B)** The tabular format of the results displays the important features of a single protein in a user-friendly fashion. By simply clicking on the Binary Interaction link, a list of interacting proteins is displayed, as well as links for the second protein's information table. The two MINT viewer buttons allow a contact map to be built (see **C** and **D**) in the same field or in a new window. **(C)** The original contact map includes the original protein (PRP38 in this example) bonded to its interacting proteins. Attached to each bond is an encircled number pertaining to the number of interactions found in the database for the two proteins. Clicking on this circle brings up a table describing the interaction between the proteins. **(D)** Attached to an individual protein is a (+) sign, which does two things: it pulls up the protein information table as well as adding its interacting proteins to the contact map. After numerous hits, you can build a complex contact map like the example shown here.

2.1.4. Kinetic Data Bio-Molecular Interactions

The Kinetic Data of Bio-Molecular Interactions Database (KDBI), administered at the National University of Singapore, contains a plethora of kinetic data for protein–protein, protein–DNA, protein–RNA, protein–ligand, RNA–ligand, and DNA–ligand kinetic data (29). The KDBI homepage is shown in **Fig. 8A**. The database is comprised of 8273 entries and 1231 distinct bi-molecular binding events (**Table 1**), which involves 1380 proteins, 143 nucleic acids, and 1395 small molecules. Staff scientists manually procure KDBI entries from the literature. The interactions are verified twice to ensure validity. In addition to literature searches, KDBI welcomes independent investigator submissions. Help is available to users new to KDBI. Next to the search tools are help icons that provide a specific example on how to use that particular search tool. The kinetic data within KDBI are from 11 species and will continue to increase as more interaction entries are added.

KDBI has four main fields in which to search the database directly on its homepage (**Fig. 8A**): molecule 1, molecule 2, bioevent, and protein list. The molecule fields allow the user to search the database by entering a specific molecule name: an example is ATP. The bioevent field searches the database using a particular bioevent as a query. A bioevent is anything describing the biochemical process that the protein is involved in, such as ATP synthesis, glucose metabolism, and so on. The protein list indicates those entries that are in the contents of the KDBI. The fields can be used simultaneously for a more specific search. It is suggested that the user begin searches using one or two fields. The page resulting from a search is easy to follow because the hits are listed in numerical order (**Fig. 8B**). Each result shows the following: molecule(s) queried, the bioevent involving the molecule, and a link to kinetic data. The kinetic data link opens a separate page titled “Detailed Information.” This page contains the kinetic data available for the particular molecule and a reference to the paper from which these data were obtained (**Fig. 8C**).

2.1.5. Binding Database

The Binding Database (BindingDB) is maintained by the Center for Advanced Research in Biotechnology (**Fig. 9A**). Its purpose is to organize a thorough dataset of experimental measurements of protein–protein, protein–nucleotide, and protein–ligand interactions based on isothermal titration calorimetry and enzyme inhibition (30). The database is updated regularly and currently describes 438 interactions (**Table 1**). This database provides various thermodynamic characteristics of interactions as well as a detailed description of the experimental conditions and procedures used to obtain the data (31). Additionally, submissions of new data can be done through the Website, such that the BindingDB’s holdings are in part submitted and reviewed by members of the scientific community. The extensiveness of the quantitative information given in this database allows easy comparison of current to previous results for verification or to draw inferences about mechanistic behavior of related proteins. The data presented by the BindingDB have other potential applications (30,31), such as quantitative models of signaling pathways, identifying drug candidates, and prediction of side effects. The accumulation of easily searchable thermodynamic characteristics also allows researchers to create programs and training datasets for programs. These data can be used for creating parameters for structure-based drug design (30,31), predicting the ΔG and $\Delta\Delta G$ for protein stability (32–34), and the stability of molecular interactions.

Fig. 8. Kinetic Data Bio-Molecular Interactions (KDBI). (A) The KDBI homepage. The search options and the help tools are available on this page. (B) This is the result page for KDBI searches. The search page shows the molecule queried, what molecule(s) it interacts with, and events that occur between the molecules. There is a link to the kinetic data. (C) This page is accessed using the Kinetic Data link provided on the results page. It contains all the available kinetic data available for the molecule queried.

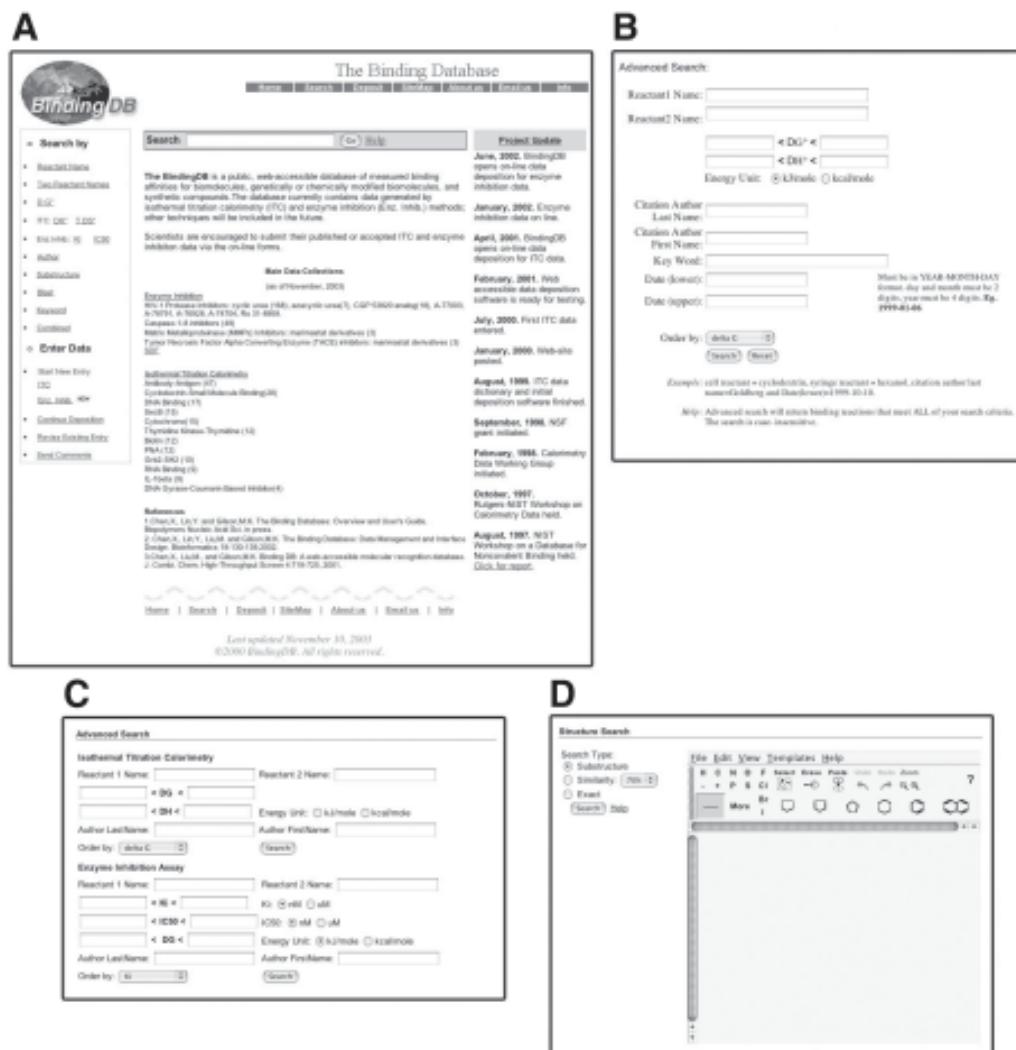


Fig. 9. The Binding Database (BindingDB). **(A)** The BindingDB homepage lists an extensive set of search options in the left-most column. The list of protein types in the center of the page allows a quick search for your protein of interest. **(B)** The old advanced search page for BindingDB. **(C)** An advanced search can be preformed by choosing the “search” button on the top-most tool bar or the “combined” button on the left-most column. **(D)** A unique feature of this database is the ability to search by substructure using a molecule builder program from the homepage “substructure” link.

The BindingDB is very easy to search because all protein types are listed on the homepage. Also, the database can be searched in a number of different ways, as listed on the sidebar and shown in [Fig. 9B–D](#) and [Fig. 10A](#). Keyword and sequence can also be used, as well as general terms such as a particular method or a particular thermodynamic parameter. Once the protein of interest is found in the database, an advanced or combined search form can be used to limit your search results. Two forms of this advanced search are shown in [Fig. 9B,C](#). After using the simple search function directly

A

Expert Search

Your Query is: IL-1β

Currently Listing 0 hits from Isothermal Titration Calorimetry match your query.

View results (data 1-10)

Sort by: ID Score

Search

Currently Listing 0 hits from Bioprocess Inhibition Assays match your query.

View results (data 1-10)

Sort by: ID Score

Search

B

Query result: (filtered by score >0)

7 matches data (unsorted by score)

Reaction	ΔG ²	ΔG ² Automatic	ΔH ²	ΔH ² Automatic	pH	Temp	Experim.
IL-1β + IL-1β	-47	19.0	-46.3	8.21	7	25	
IL-1β + IL-1β	-46.2	17.8	-45.8	8.09	7	25	
IL-1β + IL-1β	-46.2	17.8	-44.6	8.09	7	25	
IL-1β + IL-1β	-46.2	47.7	-49.2	7.67	7	37	
IL-1β + IL-1β	-45.7	35.8	-39.8	7.72	7	37	
IL-1β + IL-1β	-45.3	25.8	-46.8	7.59	7	37	
IL-1β + IL-1β	-45.3	31.6	-76.8	7.81	7	37	
IL-1β + IL-1β	-46.5	19.6	-44.2	7.20	7	25	
IL-1β + IL-1β	-46.5	17.5	-41.3	7.92	7	20	

Query result: (filtered by score >0)

7 matches data (unsorted by score)

Reaction	ΔG ²	ΔG ² Automatic	ΔH ²	ΔH ² Automatic	pH	Temp	Experim.
IL-1β + IL-1β	-47	19.0	-46.3	8.21	7	25	
IL-1β + IL-1β	-46.2	17.8	-44.6	8.09	7	25	
IL-1β + IL-1β	-46.2	47.7	-49.2	7.67	7	37	
IL-1β + IL-1β	-45.7	35.8	-39.8	7.72	7	37	
IL-1β + IL-1β	-45.3	25.8	-46.8	7.59	7	37	
IL-1β + IL-1β	-45.3	31.6	-76.8	7.81	7	37	
IL-1β + IL-1β	-46.5	19.6	-44.2	7.20	7	25	
IL-1β + IL-1β	-46.5	17.5	-41.3	7.92	7	20	

Your query is:
Key word is "IL-1β"

C

Expert Search

Reaction: Delta 35

Cell Reactant: IL-1β

Stringent Reactant: IL-1β

Entry Date: 12/05/00

D (G²) = -47 ± 1.26 (ΔH²)
Δ (ΔG²) = 8.21 ± 1.99
D (ΔH²) = -49.3 ± 1.26 (ΔΔH²)

Current or Previous: yes

Previous Method: 0.35
Δ ΔH²: -1.39 ± 0.49 (ΔΔH²)
Δ ΔG²: 0.66 ± 0.097 (ΔΔG²)

Comments: n/a

Source:
pH Buffer: 100 PBS
CellType: Chang, Basic, Rosent, Michel, Cogn, Yang, McCall, Alexander, Ulfhake, Ivan, Danay, Gotoh
Description: investigating protein-protein interactions: a model for antagonist design Biochemistry 39: 7042-7049 (2000) [200001]

More Info: IL-1β (ON data - System Info - Instrument Info)

hit-1/total: Source: n/a
Parity: n/a
Prep. Method: n/a
Synonym: IL-1β
Other Links: IL1B
Type: Cytokine
Topology: n/a
Mol. Mass: 17361 Dalton
Organism: Human
Description: n/a
Residue Count: 153
Source: avian sonic hedgehog and leucokinin homologs: a phylogenetic analysis Protein Sci 1995; 5: 103-110
Title: Leucokinin homologs: avian sonic hedgehog and leucokinin homologs: a phylogenetic analysis Protein Sci 1995; 5: 103-110
Parity: n/a
Prep. Method: Cell suspension was purified by bound to anti-p72^GM2 affinity column, and then a IL1α column
Synonym: scIL1β
Type: Single-Chain Antibody
Topology: Linear
Mol. Mass: 28590 Dalton
Organism: Escherichia coli
Description: single-chain antibody that blocked binding between IL-1^β and IL-1 receptor
Residue Count: 0
Synonym: n/a

From this search in this entry please send an Email

Fig. 10. The Binding database (BindingDB) search and results. (A) When a protein is searched for, such as interleukin (IL)-1 β from the basic homepage search box, another page will appear that will allow you to limit your search results by the experiment used and results. (B) When your limits have been defined, a results page shows up, containing a list of all the matches and the quantities results for each. Each hit can be further explored by choosing “more” on the far right-hand side. (C) The detailed results are displayed.

on the homepage, a page opens that allows the refinement of the initial search (**Fig. 10A**). This function is quite useful, because it is a hands-on method for teaching novice users to understand and eventually create their own complex searches. After refining a search, a list of hits that match your criteria will be displayed, where each hit is linked to a detailed description (**Fig. 10B**). An example of the end result is provided in **Fig. 10C**, which provides detailed information, such as ΔG , $\log K_b$, ΔH , ΔC_p , and ΔS . A detailed user's guide is available (35). ProTherm, AseDB, the Binding Interface Database (BID), and ProNIT are databases providing similar content to BindingDB. While each of these databases contains quantitative experimental results, they present data describing different types of interactions. Protherm has thermodynamic results from intramolecular mutations, whereas AseDB and BID give thermodynamic results for intermolecular interactions as a result of mutations. ProNIT provides the thermodynamic results from protein-nucleotide interactions and is further described in **Subheading 2.3.1.1**. The BID database is described further in **Subheading 2.2.2.3**.

2.2. Protein–Protein Interaction Databases

Table 2 lists the numerous databases that contain protein–protein interaction information. An honest attempt was made to find all the protein interaction databases, and most of the major databases have been found. This list has been divided up loosely into three sections based on the following criteria: “Organism-specific” are those databases providing protein interactions from one or a few organisms. “Structural/mutational” are those looking at structural and mutational aspects of protein interactions. The “general” category consists of all other databases. Because this chapter would become quite cumbersome if every database were covered, a few from each section are described in more detail based on the criteria listed in the introduction: Website clarity in navigation, search, and data presentation. Even so, all of the databases have their merits and each provides a service to the scientific community by collecting, organizing, and presenting the somewhat disparate information concerning protein interactions.

2.2.1. Organism-Specific Databases

Because of the wealth of proteomic data, many databases focus their efforts on organizing information from one or a few organisms. In many cases, the protein interaction data are presented along with a host of other information concerning that one organism, and the protein interaction information is a part of a whole description. This is certainly the case for two of three databases described below: the Comprehensive Yeast Genomic Database (CYGD) and the Human Protein Reference Database (HPRD). Other databases like the Human Unidentified Gene-Encoded Protein–Protein Interaction Database (HUGE-ppi) primarily focus on protein interactions from a single organism.

2.2.1.1. COMPREHENSIVE YEAST GENOMIC DATABASE

The CYGD, provided by the Munich Information Center for Protein Sequences (MIPS) is an organism-specific database cataloging all the bioinformatics data concerning *Saccharomyces cerevisiae* (36), including protein–protein interactions and protein complexes from high-throughput methods. The database has recently undergone a major revision in terms of look and organization on the homepage. This will probably be extended to all the parts in due time. The protein interaction portion of the database

can be accessed by the “Protein–Protein Interaction/Complex Viewer” link on the homepage. Alternately, the protein interaction page can be found through first clicking on the “Search/View” link on the sidebar and then on the “Protein–Protein Interaction/Complex Viewer” on the presented list. In this part of the CYGD, 9030 physical and 1540 genetic protein–protein interactions have been culled from literature containing results from yeast two-hybrid experiments, tandem-affinity purification (TAP), and mass spectrometry. Protein–protein interactions and complexes can be searched using the functions on the sidebar. Various search terms can be used, such as the ORF name (preferentially), gene name, part of description (such as the experiment type used), or PubMed reference number. Results for both the protein interaction and protein complex searches are lists of interacting partners with functional interaction descriptions and links to protein-specific descriptions (Fig. 11A,B, respectively).

2.2.1.2. HUMAN PROTEIN REFERENCE DATABASE

The HPRD was built as a cooperative effort between Johns Hopkins University and the Institute of Bioinformatics (37). This database “integrates information pertaining to domain architecture, post-translational modifications, interaction networks and disease association for each protein in the human proteome” (directly from the homepage). Presently, a very extensive 3187 proteins and 13,194 protein–protein interactions exist in the HPRD (Table 2). All of these are manually extracted from the scientific literature, and combine information from various experimental methods. The number of entries grows continually. The HPRD presents very clean Web pages, which facilitates site navigation (Fig. 12A). The database can be searched by keywords from the “Query” page or by sequence using the “BLAST” page. The “Browse” page contains a number of lists that organize the proteins accordingly. These lists can be simply accessed via tabs. The “Pathways” link leads to tabbed images of comprehensive interaction maps for nine signaling pathways. These maps arrange protein–protein interactions in a diagrammatic format, but at present the names are not linked to individual protein information. Protein description pages provide basic information (Fig. 12B), including molecular weight, protein sequence, and function, as well as providing in-depth information on the interacting proteins and their domains and regions. This wealth of data is elegantly broken up by the use of a number of tabs to access specific data. Another unique approach of the HPRD is that the database authors allow for and actively seek input from the biological community to help in accurately annotating the database. This can be done from a link on the sidebar to the comment link at the bottom of each protein annotation page (Fig. 12B).

2.2.1.3. THE HUMAN UNIDENTIFIED GENE-ENCODED PROTEIN–PROTEIN INTERACTION DATABASE

The HUGE-ppi database was built according to results obtained from the Human cDNA project from the Kazusa DNA Research Institute (38). HUGE-ppi “should be a very useful resource for detecting previously unidentified interactions because it complements conventional expression libraries, which seldom contain large cDNAs” (from the homepage). These large cDNA clones are the cytoplasmic domains of trans-membrane proteins and are used to identify proteins interacting with these cytoplasmic domains. The database presents yeast two-hybrid results from over 100 “prey” and 30 “bait” clones, reporting interacting clones and their domain organization (39). Linked from the homepage (Fig. 13A), the search function is limited to a table of all interac-

A

Protein-Protein-Interaction

Partner 1	Interaction	Partner 2	Description	Reference
YBR128c Apg14p	physical < >physical	YPL120w Vps30p	Two distinct Vps34 PtdIns 3-kinase complexes exist: one, containing Vps15p, Vps30p, and Apg14p, functions in autophagy and the other, containing Vps15p, Vps30p, and Vps35p, functions in CPY sorting. EMI	9712845 11157979
YBR160w Cdc28p	physical < >physical	YMR199w Cdc1p	interacts with CDC28 protein kinase to control events at START EMI	YAL038w
YBR160w Cdc28p	physical < >physical	YPL246c Cdc28p	interacts with Cdc28p protein kinase to control events at START EMI	YAL038w

PPI
 Complex
 Go
 Reset

B

Protein Complex

Bait	Prey	Description	Reference
YIL033c	Bcy1p, YAL164c, Tpk1p, YKL166c, Ypk1p, YPL203w	cAMP-dependent protein kinase	DOKU
-	-	-	-
YER123w	Yck3p, YHR135c	-	DOKU
-	Yck1p, YNL154c, Yck2p, YPL284m, Hnr25p	-	-
YGL019w	Cbb1p, YIL035c	-	DOKU
Cka1p, YOR039w, Ckl2p, YOR061w	Cka2p	Casein kinase II	DOKU
-	-	-	-
YGL253w	Hsk2p	Hexokinase 2	DOKU
-	-	-	-
YGR248w	Tal1p, YMR205c	Phosphofructokinase	DOKU
-	Phc1p	-	-
YAL038w	Cdc19p, YOR347c	Pyruvate kinase	DOKU
-	Pyk1p	-	-
YJL006c	Ctk2p, YKL139w	TFIIK (CTD kinase)	DOKU
Ctk1p, YML112w	Ctk2p	-	-

erase
 Select:
 ORF/Gene
 Description
 Reference
 Type of Interaction
 Physical
 Genetic
 PPI
 Complex
 Go
 Reset

Fig. 11. Comprehensive Yeast Genome Database (CYGD). The open reading frame (ORF) name (preferentially), Gene name, part of Description, such as the experiment type used or PubMed reference number, can be used as search terms. Results are listed of interacting partners with functional interaction descriptions in the form of (A) protein–protein interactions or (B) protein complexes.

A

B

Fig. 12. The Human Protein Reference Database (HPRD). **(A)** The HPRD contains the majority of its information on the homepage of the database. A brief description explains the type of interaction contained in the database, how the literature is mined, and a reference. Statistics on the number of proteins, protein interactions, domains, and PubMed links are displayed right on the homepage. Also offered are links for running a query, a browse, a blast, for pathways, and frequently asked questions for new users. **(B)** Set up in a wonderfully organized fashion, HPRD displays large amounts of data in an easily accessible manner. Tabs separate information based on basic information, sequence, interactions, expression, alternate names, diseases, posttranslational modifications and substrates, and external links. Information is arranged in tabular format.

tions. Selecting an interaction pair produces an overview of the interacting protein pairs. Choosing one interacting member brings up a page from the Gene/Protein Characteristics Table maintained by HUGE, the institute's broad-based database (**Fig. 13B**). This page provides an overview of the protein and genetic/genomic make up for the baits and preys as well as experimental results from Northern blots, reverse transcriptase (RT)-polymerase chain reaction (PCR), RT-enzyme-linked immunosorbent assay (ELISA), and radiation hybrid (RH) mapping. Further information about DNA sequence, protein sequence, expression, and mapping are on the top of the page as well as on the page itself. The top of the page also provides links to other databases holding information about the entry. Help pages are obtainable from the "Description" links (**Fig. 13B**) to assist the user in navigating the results.

2.2.2. Structural/Mutational Databases

Protein structure is often used to determine how two proteins interact with the aid of site-directed mutagenesis experiment. The primary source repository for structures of proteins is the Protein Data Bank (PDB) (**10**). While a structure file containing a pair of interacting proteins is most informative, coordinates of individual members can be used for speculation. Usually, the mutations are focused at "hot spot" areas (**40**), clusters of primarily charged residues that make the largest contribution to binding energy. An important consideration for mutagenesis studies is that the mutation does not alter the protein's structure. Otherwise, changes in measurements could be attributed to the structural change and not to the individual amino acid. Most often, these studies employ alanine scanning, where sets of individual residues are successively mutated into alanine and the effect is measured. Alanine is a nonpolar amino acid with a small side chain, and minimizes structural changes upon mutation, making it well suited for initial mutagenesis studies. Replacement of a single residue in an amino acid sequence with alanine can determine the importance of the residue to the protein–protein interaction or the activity of the protein. Other site-directed mutagenesis studies mutate a residue to one that is considered an opposite of the original residue, to further characterize a protein interface—i.e., replacement of a polar amino acid with a nonpolar amino acid or changing charge states. In these experiments, the mutations can alter binding by causing conformational changes. Such conformational changes need to be ruled out so that the experimental differences can be attributed to the mutation's effects on binding.

Fig. 13. (*opposite page*) Human Unidentified Gene-Encoded Large Protein Analyzed by Kazusa cDNA (HUGE-ppi). **(A)** HUGE-ppi has everything you need for understanding and searching the database contained nicely on its main page. A list of all interacting baits and their interacting preys from yeast two-hybrids are displayed in tabular format, unless you have a specific KIAA bait or clone in mind. Offered on this page is a description of the yeast two-hybrid screening results to help a novice user. **(B)** The Gene/Protein Characteristic Table contains impressive amounts of information on the protein and genetic/genomic make up for the baits and preys, as well as experimental results from Northern blots, reverse transcriptase (RT)-polymerase chain reaction, RT-enzyme-linked immunosorbent assay, and radiation hybrid mapping. The description pages listed to the right help the user sort and interpret the data received.

A

HUGE ppi

A Database of protein-protein interactions between large proteins
 (LARGE): Human Unidentified Gene-Encoded Large Proteins Analyzed by Keio's cDNA Project

Thank you for your interest in our protein-protein interaction database.

The HUGO gene protein database has been created to publicize the results of our Human cDNA project at the Kansas DNA Research Institute. Large proteins have multiple domains potentially capable of binding many kinds of proteins. It is conceivable, therefore, that a single protein could function in an intricate framework of extracellular protein complexes. To comprehensively study protein-protein interactions between large cDNA proteins, we have constructed a library composed of cDNA clones, categorized on prior functional classifications, and cloned into a vector that facilitates cDNA sequencing. Our representative library should be a very useful resource for detecting potentially redundant function(s) by one cDNA, or for determining cDNA expression patterns.

METHACRYLATE

更多資訊請上

- **Mitsuru Nakamura, Reiko Kikuchi, and Osamu Obata**
Protein-Protein Interaction Between Large Protein: Two-Hybrid Screening Using a Functionally Classified Library Composed of Long cDNAs
Genome Research (2007) 17: 1772–1784

ГИБДД АЛМЕРЫ

ACKNOWLEDGMENTS

- List of protein-protein interaction pairs
 - Description of the experimental result of two-hybrid screening
 - Direct access to a table of interest
Enter KIAA number
 Both Prox Dist
(e.g. KIAA0318)
 - Supplemental table
 - Bibliography

How to obtain KIAA clone(s)

B

Gene/Protein Characteristic Table for KIAA0319

Link to: [GTOP](#) | [SWISS-PROT](#) | [TrEMBL](#) | [OmmeCach](#)
[Prokaryotic DNA sequence](#) | [Protein sequence](#) | [Expression](#) | [Mammalian](#)

Accession No.: AB000317

Album Name:

HUGO Gene Name :

Close Name: Ig00378 [Delete]

卷之二

117

- Length: 6791 bp
 - Physical map

Physical map showing the scale from 0 to 748. A dark bar indicates the coding region, spanning approximately from position 100 to 600.

 - Restriction map
 - Prediction of protein coding region (GeneMark analysis) for :
closed DNA seq.
 - Warning for N-terminal truncation: ND
 - Warning for coding interruption: ND

Length of 5'UTR	3061 bp
Genome contig ID	gi 2982457 tr_2450657
PolyA signal sequence (AAA-AAA, -18)	TTTTTTTAAAATTTTAAATGGACATATCC
Flanking genome sequence (99914 - 99865)	ACATTCTACCTTAAATGATCTCTGGCTGTCTTAAATGACAA

Databases of structural and mutational data concerning protein interactions are built to report the findings of various experimental methods. They act as meta-servers to the PDB and to PubMed, respectively. Examples of these databases are described below, and their homepages are shown in **Fig. 14**. The Surface Properties of Interfaces—Protein–Protein Interfaces (SPIN-PP) database catalogs and analyzes structures with interacting proteins. The Alanine Scanning Energetics Database (ASEdb) and the BID both collect and present data from mutational studies of protein interactions.

2.2.2.1. SURFACE PROPERTIES OF INTERFACES—PROTEIN–PROTEIN INTERFACES

Hosted by Columbia University, SPIN-PP is a structural database for all PDB (10) entries involved in protein interfaces (41). The database was built to “assist researchers in inspecting various properties of protein–protein interfaces available in the Protein Data Bank” and “aims at using the computed properties for the purpose of gaining insights into the structural basis of protein–protein interactions, prediction of protein active sites and approaching the problem of protein function prediction based on structural analysis” (Nayal, personal communication). SPIN-PP classifies protein–protein interfaces as a pair of PDB chains of at least 20 residues in length and buries an accessible surface area of at least 200 Å₂ between them. SPIN-PP contains two datasets: (1) the unique dataset consisting of 855 entries all containing interface pairs with less than 80% sequence identity removed (Table 2) and (2) the full dataset including all 6460 interfaces (nonunique) made between two chains in the PDB.

A wide variety of search functions are shown directly on the homepage (Fig. 14A). Within the “Unique” or “Full Dataset,” the user can search SPIN-PP for a protein–protein interface or by author and source. Results are built by selectively searching the database for function, PDB ID, method, compound name, chain names, and by surface property values (size, hydrophobicity, curvature, electrostatic potential, and sequence variability). Documentation on searching tips is provided to aid the user in their search. The first part of the analysis reports the essentials of your search: PDB code, chains and number of residues involved in the interface, percent identity, and method. If multiple chains are involved, each interface is analyzed separately. A Surface Picture Gallery was created using the GRASS server (trantor.bioc.columbia.edu/cgi-bin/GRASS/surfserv_info.cgi), which depicts the molecular surface of the interface individually colored by curvature, electrostatic potential, sequence variability, or hydrophobicity and the transparent molecular surface with a backbone ribbon.

SPIN-PP depicts molecular surface properties using GRASP (42), showing the side views and open book views for distance and transparency. Open-book views offer information on hydrophobicity, curvature, electrical potential, and sequence variability relative to the average from a set of surfaces. Classification of buried accessible surface area by residue and residue type are separated by chain and then by critical residue before and after the complex is formed. Lastly, the amino acid sequences for both chains are presented, indicating the residues involved in the interface—those that form hydrogen bonds across the interface, possess a hydrophobic contact, or line an interface cavity.

2.2.2.2. THE ALANINE SCANNING ENERGETICS DATABASE

“In a protein–protein interface, a small subset of the buried amino acids typically contribute the majority of binding affinity as determined by the change in the free

Fig. 14. Protein interaction databases of structure and energy. **(A)** The massive search engine offered by Surface Properties of Interfaces—Protein–Protein Interfaces (SPIN-PP) allows for both broad and specific interaction searches using either the complete or the unique set of proteins. For a specific search, information pertaining to function, method, and structure can be included to narrow the results received. **(B)** The Alanine Scanning Energetics Database (ASEdb) contains links for protein interaction searches, listings of the protein systems housed within the database, reference lists used in data mining, and links to add new mutations. The database description offers informative explanations on what the database offers and how it is maintained. **(C)** The Binding Interface Database (BID) houses mutagenesis data for protein–protein interactions mined from current literature and offers easy searching tools for finding interactions of interest.

energy of binding ($\Delta\Delta G$) upon mutation of the residue to alanine”—termed the *hot spots* (43). The ASEdb was developed to house mutational information on the thermodynamics of side-chain interactions determined by alanine-scanning mutagenesis (Fig. 14B). The ASEdb was built as an outgrowth of previous work, with the goal of determining how hotspots of binding energy arise (44). After analyzing the data, they found that the free binding energy is poorly correlated with the change in buried surface area when the individual side chains are considered instead of the whole interface. They discovered that the degree of surface accessibility is related to the free binding energy, in that the more the surface is buried, the higher the binding energy. Currently, the ASEdb contains over 98 protein pairs (Table 2) and 2919 mutations, including protein–protein interactions as well as protein–DNA/RNA and protein–small molecule interactions. Researchers are also invited to submit entries.

ASEdb offers several methods for searching for an interaction in the database. After selecting protein, nucleic acid, or small molecules as a partner type, all protein interactions in the database that have a partner match will be listed. Entering the mutated protein or partner name helps to narrow the search, which will result in all interactions pertaining to those proteins. For an even more refined search, entering information on free energy change, monomer, complex, and change in accessible surface area (ASA) results in entries with specific mutational information. Results are displayed in tabular form, ordered first by mutated protein and then by residue number of the mutated amino acid. Each entry states the mutated protein name, protein partner name and type, PDB ID, mutated amino acid, and residue number, then the mutational information gained from the references. Mutational changes in binding energy are presented as free binding energy ($\Delta\Delta G$) and the accessible surface area, which includes the area for the protein alone (monomer) when it is combined with the protein partner (complex), showing solvent accessibility, and the change in the area.

2.2.2.3. THE BINDING INTERFACE DATABASE

The BID is a comprehensive mutational database that compiles data on site-directed mutagenesis studies of protein–protein interactions (45). The BID was developed to describe how the protein–protein pairs interacted, mainly focusing on hot-spot regions of binding using both direct and indirect experimental methods (see Fig. 14C). To organize vast amounts of protein interaction information, the BID structures the data in terms of protein systems. Using this approach, the database is continually augmented by manual mining of the scientific literature for protein interaction descriptions as well as wild-type and mutational binding energies. Currently, the BID contains 17 protein systems that encompass 308 interacting proteins (Table 2) and over 2891 hot spots. Data are represented using a tabular form, graphical contact maps, and descriptive functional profiles.

Users can look for data by browsing a list of all proteins in the database or by searching for a specific protein name. Searching through the BID for a protein–protein interaction is handled through the five search categories offered. Organism name, structure code (PDB ID), or interacting protein information can be specified as search criteria to narrow the results received. The protein systems are listed in tabular form, including information about the number of interacting pairs, which allows easy searching of all protein interactions within a system. Other search options include references, PubMed ID, or the protein’s NCBI gene ID. Results reveal an extensive display of protein–

protein interactions listed in tabular form by reference. The mutational information is displayed in two ways: graphical contact maps and tabular format. A contact map represents the two proteins on each axis, and the residue numbers of the mutations are shown in the graph based on the type of interaction—i.e., hydrogen bonds and/or salt bridges. The tabular format of the mutational information reveals the bulk of the research. Residue number and amino acid name are organized by reference and include the amino acid the residue was mutated to as well as the domain it is located in. Mutational information includes the type of interaction (hydrogen bond, hydrophobic contact, or salt bridge); the strength of the original interaction, based on the change in strength after the mutation (strong, intermediate, weak, or negative); binding energy of the interaction after the mutation (k_a , k_d , K_d , and percent binding); and the experimental method used in the article. Interacting amino acids are derived from structural data and listed separately, identifying only the type of interaction and experimental method—X-ray or nuclear magnetic resonance (NMR).

2.2.3. General Databases

These databases do not quite fall into any category. Many cover information from a number of organisms (GRID), hold data from alternative experimental methods (Doodle), or contain theoretical interactions (GKD). The GRID will be discussed in further detail.

2.2.3.1 GENERAL REPOSITORY FOR INTERACTION DATASETS

GRID, short for General Repository of Interacting Proteins, is an excellent protein interaction database, developed at Mount Sinai Hospital (46,47). The homepage is shown in [Fig. 15A](#). Currently, the GRID contains approx 21,236 unique protein interactions available to query ([Table 2](#)). Protein interactions found in the GRID are from *S. cerevisiae* and *Drosophila melanogaster*. Protein interactions from other organisms are a future possibility. GRID data are an accumulation of the current protein interactions found in the literature. The literature is searched for novel protein interactions and then screened by staff curators to examine the validity of the interactions. GRID management encourages submission of unique protein interactions from independent investigators. GRID has both HELP and FAQ pages to assist new users. GRID also provides a multiprotein interaction visualization tool, available for users to interpret complex protein interactions.

Searching the GRID is a simple task, with a basic search available to the user. Using the simple search, the user enters the gene name or ORF into the search box to query the database. Although no longer available, at one time an advanced search allowed the user to search for protein interactions using five different criteria: interacting gene, experimental system, interacting source, interaction ID, and publication ID. Boolean operators allow the user to search multiple proteins per advanced search. Results are displayed according to what type of search is used ([Fig. 15B](#)). Simple searches display results in a tabular format, while advanced searches give the user four different choices on how to display the results. The result options for the advanced search are a tab-delimited text file, online summary, downloadable form to view on the Osprey network visualization system, or immediately view the network via a browser window using Osprey (48,49).

A
B

Fig. 15. The General Repository for Interaction Datasets (GRID). **(A)** This is the GRID homepage. Available on this page are the search tools for both the yeast and fly datasets. Links to the Osprey visualization software, Flybase, and *Saccharomyces Genome Database* (SGD) are located at the bottom of the picture. **(B)** A typical results page for a GRID search.

2.3. Other Protein Interaction Databases

2.3.1. Databases of Nucleic Acid–Protein Interactions

Nucleic acid databases can be categorized into three types of databases (50): (1) protein interactions at the atomic level like the Protein–Nucleic Acid Recognition Database; (2) direct protein interactions with quantified results, like the ProNIT (51); and (3) transcriptional regulation databases like PathoDB (52) and RegulonDB (53), which are typified by indirect interactions. Below, the ProNIT database is covered in more detail.

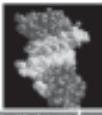
2.3.1.1. THERMODYNAMIC DATABASE FOR PROTEIN–NUCLEIC ACID INTERACTIONS

ProNIT contains thermodynamic data for wild-type interactions and mutants in a range of different experimental conditions (50). This database was released in 2000 and is updated at intervals. ProNIT currently contains 164 proteins (Table 3). To check for new updates, a bulletin board is available from the “What’s New” link on the homepage (Fig. 16A). The information stored in this database can be used to quantitatively describe gene regulation. This information can also be used for predicting nucleotide-binding proteins, binding sites, and possibly identifying residue-binding hot spots. ProNIT has been integrated into 3DInSight to relate structure, function, and properties of biomolecules, in which mutations, motifs, and binding sites can be mapped onto protein structures. Structures solved for various protein-nucleotide interactions can be linked to the recognition database to identify the interactions of specific atoms. It is best to view the stats link from the homepage before you begin searching this database. This will allow you to quickly skim through the proteins that are in the database to see whether your protein of interest is available. Once you find your protein, you can use the name listed to perform further searches. One particularly useful feature of this database is that the home link takes you to a page of various computational tools for analyzing proteins.

2.3.2. Protein–Ligand Databases

Protein–ligand databases are specific for protein interactions with small molecules. These databases fall into basically two groups: databases of enzyme ligands and databases on peptides binding to the major histocompatibility complex (MHC). Enzyme-ligand databases provide detailed information about reaction mechanisms, kinetics, metabolic pathways, and several other variables. There are many good databases available to researchers who are interested in protein-ligand interactions. Most importantly, these databases, each with its strengths, make information about protein–ligand interactions more accessible. They provide content to answer basic protein–ligand interaction questions. The Protein Ligand Database (PLD) or the Comprehensive Enzyme Information System (BRENDA) provides dissociation- or binding-constant information. The University of Minnesota’s Biodegradation and Biocatalyst Database (UM-BBD) contains proteins in their metabolic pathways. The MHC peptide databases simply catalog and organize information about these complexes that are fundamental to many immunological processes and pathways. As an example, the MHC Peptide Interaction Database (MPID) integrates many types of information into one site. Of course, which database will be informative is dependent on what the researcher is investigating. In the following section, four of the aforementioned databases will be profiled. An overview of what to expect and how they might help will be given.

A



Thermodynamic Database for Protein-Nucleic Acid Interactions

Home 3Dinsight ProTherm ProNIT Protein-DNA Recognition Biomolecules Gallery

Last Update: 24-July-2002

Go

Advanced search

Overview

What's New

Statistics

Tutorial

More about ProNIT

Acknowledgement

Members

Reference

Contact Us

Welcome to ProNIT Database

Overview of ProNIT

Protein-nucleic acid interactions play important role in the regulation of gene expression. Experimental interaction data provide valuable information for understanding the protein-nucleic acid interactions. However those data are scattered among literature. Thus, we have started developing Thermodynamic Database for Protein-Nucleic Acid Interactions (ProNIT) by collecting experimentally observed thermodynamic data of protein-nucleic acid binding. ProNIT contains information about sequence, structure, bibliographic information and several thermodynamic parameters such as dissociation constant, association constant, changes in Gibbs free energy, enthalpy and heat capacity, activity (K_m and $Kcat$) etc., along with experimental method and conditions. ProNIT is implemented in 3Dinsight, an integrated relational database and search tool for the structure, function and properties of biomolecules. It is also a part of Protein-Nucleic Acid Recognition Database, which consists of Protein-Nucleic Acid Complex Database, a collection of structural data of protein-nucleic acid complex, and Database of Base-Amino Acid Interactions, which enables users to search specific base-amino acid interactions. A WWW interface for ProNIT enables users to search data based on different conditions with various sorting and display options. Furthermore, ProNIT is cross-linked with other important archives like Protein Data Bank (PDB), Nucleic Acid Database (NDB), Enzyme code (EC), Protein Information Resource (PIR), ProTherm and PubMed literature database. For more detail about ProNIT, please see [here](#).

Please note that this database is under constant development. There will be changes without prior notice. We welcome your comments and suggestions to improve this database.

Home | 3Dinsight | ProTherm | ProNIT | Protein-DNA Recognition | Biomolecules Gallery

B

PLD

Search On: Keyword Binding Energy Ligand Tanimoto Score Protein Percentage Sequence Author

Protein Ligand Complexes Binding Energy CATH EC

Protein Ligand Databases

Home

Quick Start: Search PDB_CODE

Submit Query

Welcome to the Protein Ligand Database (v1.2)

The PLD is a resource containing biomolecular data, including binding energies, Tanimoto ligand similarity scores and protein sequence similarities of protein-ligand complexes. The PLD(v1.2) currently has data on 273 protein-ligand complexes.

PLD dataset (PDB Codes).

Latest update : 273 Complexes 22nd September 2003

The database can be searched using:

- PDB Code
- Protein and ligand name
- Protein function
- Ligand origin
- Protein-ligand binding energies
- Protein sequence similarity
- Ligand fingerprint Tanimoto similarity
- Author
- EC family
- CATH classification

In citing the PLD please refer to:

D. Puvanendrampillai, J.B.O. Mitchell. (2003) *Bioinformatics* 19, 1856-1857

Theory underlying the PLD

BEST VIEWED WITH NETSCAPE 7 / INTERNET EXPLORER 6 OR HIGHER

Home | Search | PLD Developed by D. Puvanendrampillai | Group Leader Dr John Mitchell Contact us | Site Map

Copyright 2003 University of Cambridge

Last Updated: 22nd September 2003

Powered by MySQL

Fig. 16. Databases of protein interactions with nonproteins. (A) Thermodynamic Database for Protein Nucleic Acid Interactions (ProNIT) homepage. (B) Protein Ligand Database (PLD) homepage.

2.3.2.1. THE PROTEIN LIGAND DATABASE

The PLD is a straightforward collection of proteins and their small-molecule ligands with some thermodynamic data and links to structural data (54). Besides the basic PDB search on the homepage, the site makes use of a number of simple pages to search for protein–ligand interactions. These searches can be done using various criteria, all of which are accessible via links from the top of the homepage (Fig. 16B). The criteria range from keyword search to binding energies of the protein–ligand complex to Class, Architecture, Topology, and Homology (CATH) classification (55). For protein interactions found within the database, a straightforward results page is displayed with experimental and theoretical binding energies, binding constant, and links for the protein structure from the PDB (10), as well as links of the protein–ligand complex from the PDBsum Website (56,57). The database, maintained at Cambridge University, is regularly updated.

2.3.2.2. THE MINNESOTA BIODEGRADATION AND BIOCATALYST DATABASE

UM-BBD is a protein–ligand database of “microbial biocatalytic reactions and biodegradation pathways” (58–61). Its homepage is shown in Fig. 17A. UM-BBD combines individual reactions to form larger metabolic pathways. The site contains outstanding graphic representations of these reactions and their mechanisms, within which the molecules and enzymes are linked directly to their respective pages. The “Guided Tour” link at the top of the homepage (Fig. 17A) provides a very descriptive overview of the database’s holdings with links to examples. Searching the database has been designed to be very straightforward. From the homepage (Fig. 17A), a pathway can be directly accessed from the scrolling text box, or compiled lists of holdings under various subjects are available. The “Search” link sends the user to a well-documented search and browse page (Fig. 17B). Above the search’s input field, a pull-down menu allows the user to limit the search to a topic. Browsing is underneath and consists of another set of compiled lists, many of which are different from those on the homepage. Figure 18A shows a pathway selected from the scroll box on the homepage. A text representation of the enzymatic pathway is shown, with links to pages for more detailed information. A graphic representation is also available (Fig. 18B), with all of the same links as on the text page. Even more polished graphics are available for some pathways. These graphics are much easier to navigate and digest than the text versions. Clicking on an enzyme produces a page with more information about that particular reaction in the pathway (Fig. 18C). A noninteractive graphic representation is available, as well as links to the original reference and related citations. At this point, the user can choose to grow out a pathway from this point. Clicking on individual reactant(s) or product(s) leads to the page shown in Fig. 18D. Basic information and an image of the molecule are shown. Reactions involving the substance are listed, allowing the user to enter into alternative pathways. Overall, the UM-BBD presents data in clear and unobtrusive format and provides information at various levels, from pathways to individual molecules. Interactive graphic and text pathways have been created, where all reactants, enzymes, and products are linked to their respective pages. The UM-BBD, maintained at the University of Minnesota, is regularly updated.

2.3.2.3. THE COMPREHENSIVE ENZYME INFORMATION SYSTEM

BRENDA, or Comprehensive Enzyme Information System, is just that: a regularly updated database and retrieval system containing exhaustive descriptions of enzymes,

4

UM-BBD Search and Browse Page

UM-BBD Search and Browse Page

BBB0 Main Menu | Search | About the UMBBD | User's Guide | User Email List | Contact Us | Related Tools | Help/Feedback | Search | Recent Searches | Favorites | Help |

Search the UM-BBD

From a list below, search for chemical structures using SMILES strings or drawing them using a drawing, or search for the following:

Search for **compounds** by full or partial name of substance (e.g., volume or methyl benzene), CAS registry number (e.g., 108-83-7), or formula (e.g., C7H8, H3C7, or even C4H10C1). Search for **enzymes** by full or partial name or abbreviation name (e.g., a kinase or kinase), or full or partial name (e.g., 3.1.1.5 or 3.4.1.1). Search for **biologics** by full or partial name of starting or ending organic functional group (e.g., aldehyde or carboxylic). Search for **interrogations** by full or partial name (e.g., 2.2.1. Peptidomimetics, Peptides or Peptides). Search for **reactions** by full or partial name (e.g., 3.5.1.5 or 3.4.1.1). Search for **biocatalysis** by full or partial name (e.g., 3.4.1.1.1, other or threefold).

Select a search type and enter what you want to search for. A **compound** search will retrieve all matching compounds and the reactions in which they are produced or consumed. An **enzyme** search will retrieve enzyme names and each enzyme page will list the reactions catalyzed by that enzyme. A **biocatalysis** search will retrieve biocatalysis names and each biocatalysis page will list the reactions catalyzed by that biocatalysis. A **reactions** search will retrieve all matching reactions. A **pathway** search will retrieve pathway pages and each pathway page will list the reactions, compounds and enzymes in that pathway.

You can use the Boolean operators AND, OR, or NOT in compound, enzyme, reaction, reaction and pathway name searches. Parenthesized Boolean expressions are not allowed, but can be logical. For example, aldehyde OR kinase AND (not) isopropyl in (aldehyde OR kinase) AND isopropyl. Searches are case-insensitive.

Find the compound reaction search controls 

List all UM-BBD

- Compounds
- Reactions
- Enzymes
- Reactions and modified enzymes
- Compounds and further degraded
- Biocatalysis and reactions
- Entries ordered by functional group
- Entries for meth., meth., and related chelation
- Reactions of琅琊 (琅琊) 1,3-dioxanes
- Reactions of thione dioxepane
- Organic Functional Groups in UM-BBD compounds
- Biochemical Principles Tables

BBB0 Main Menu | Search | About the UMBBD | User's Guide | User Email List | Contact Us | Related Tools | Help/Feedback | Search | Recent Searches | Favorites | Help |

Page: Authors: Ryan McNamee and Doug Bartholomew
Received: September 22, 2003 Contact Us
0. 24/24, University of Michigan -

Fig. 17. The University of Minnesota Biocatalysis/Biodegradation Database (UM-BBD). (A) The UM-BBD homepage contains a link to the search tool, a scroll menu for the UM-BBD pathways, and links to various UM-BBD features. (B) The UM-BBD search page provides a drop box in which the user can search for a molecule of interest under a specific category. Under the search tool are links to lists of the available database contents.

A

Nitrilotriacetate (an/aerobic) Pathway Map

[Compounds and Reactions](#) [BBD Main Menu](#)

This pathway is contributed by Guang Yan, University of Minnesota.

Nitrilotriacetate (NTA) is an aminotriacrylic acid which readily binds bivalent metal ions in a ratio of 1:1. As an important industrial chelating agent, NTA has been widely used for various sulfonamide processing and decontamination procedures, such as textile, paper and pulp processing and water treatment. The excessive use of NTA has led to considerable concern about the release of NTA to the environment. Nitrilotriacetate is relatively stable and inert in acidic or neutral aqueous media, so it would not pose a contamination of drinking water with metal-NTA complexes. Fortunately, NTA is relatively easy to degrade. Many microorganisms, either in aerobic environments or under anaerobic or severely oxygen-limited conditions, can use NTA as sole source of carbon, nitrogen and energy. NTA degradation is initiated by nitrilotriacetate monooxygenase (EC 1.14.12.1), which requires Fe^{2+} and O_2 as component protein domains A and B. Component A is an NADH:FMN oxidoreductase that reduces FMNH₂ to FMN to oxidize NTA, component B is an NADH:FMN oxidoreductase that provides FMNH₂ for NTA oxidation by using NADH to reduce FMN to FMNH₂. In the following step, DTA is further converted to glycine and glyoxylate by nitrilotriacetate dehydrogenase.

The following is a text-based nitrilotriacetate acid pathway map. Organisms which can initiate the pathway are given, but other organisms may also carry out later steps. Follow the links for more information on compounds or reactions. This map is also available in graphic (GML) format.

Reactions

Pathway

Compounds

[Page Author\(s\):](#) Guang Yan
[July 13, 2002](#) [Contact Us](#)
 © 2002, University of Minnesota.
 All rights reserved.

C

Nitrilotriacetate (an/aerobic) Graphical Pathway Map

[Compounds and Reactions](#) [Text Map](#) [BBD Main Menu](#)

Click on the boxed compound or enzyme names for further information.

[Page Author\(s\):](#) Guang Yan
[July 13, 2002](#) [Contact Us](#)
 © 2002, University of Minnesota.
 All rights reserved.

C

From Iminodiacetate to Glyoxylate and Glycine

Graphic of the reaction

Reference

Ueda T, Egli T. Biodegradation (1993) 2: 423-434.

Search [Molfile for Iminodiacetate dehydrogenase](#)
 12 citations on July 13, 2003.

Pathway

Compounds

[Page Author\(s\):](#) Guang Yan
[July 13, 2002](#) [Contact Us](#)
 This is the UM-BBD reaction, reactionID 6089.
 It was generated on November 17, 2003 12:52:45 PM CST.
 © 2002, University of Minnesota.

D

Iminodiacetate

[Page Author\(s\):](#) Guang Yan
[December 20, 2002](#) [Contact Us](#)
 This is the UM-BBD compound, compoundID 6089.
 It was generated on November 17, 2003 12:52:45 PM CST.

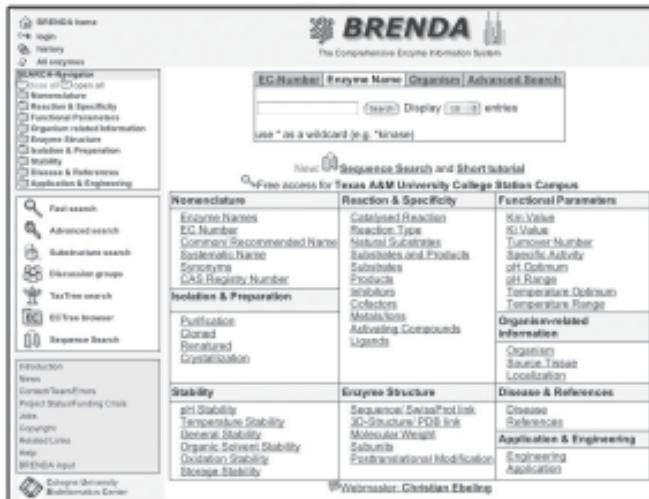
Fig. 18. The University of Minnesota Biocatalysis/Biodegradation Database (UM-BBD) search and results. (A) The results page from a UM-BBD search is shown. A hypertext-based representation of a typical enzyme-catalyzed substrate-to-product map is provided. The links produce further detail about that enzyme or molecule. (B) Accessible via the “graphic” link on the first page, this page depicts the same substrate-to-product map as the previous page; however, in this figure the structure of the substrates and products are in a graphic format. All images are linked as in the previous page. (C) In this picture, the individual substrate-to-product reaction, including the enzyme, is shown in a text-style format. (D) This picture shows the substrate of one of the reactions in the previous diagrams. The structure, molecular formula, and other names of the molecule are listed. The structure of both substrate and product are available in UM-BBD.

hosted by the University of Cologne (62). Protein–ligand data make up only a part of BRENDA. The major strength with BRENDA is its multitude of searching utilities and its detailed help/user guides. BRENDA’s homepage (**Fig. 19A**) is almost completely devoted to different means of searching the database. All 47 database fields are searchable. These simple searches are considered “Fast Searches,” as they take very little time to return results. Finding/accessing these searches can be done in more than one way. Not only are all of these searches placed prominently on the homepage, they are also organized in a “Search-Navigator,” with searches grouped into nine subjects within expandable folders in the upper left-hand corner (**Fig. 19A,B**). This arrangement allows a user to browse through subjects to find the correct search. In addition, underneath the “Search-Navigator,” several of the more useful searches have been developed into more intuitive/complex searching/browsing tools, including a substrate substructure search similar to that of BindingDB (**Fig. 9D**) and an advanced search function that combines 16 user-definable fields. Being more complicated, these searches are slower to return results. The most common searches have been placed directly under the homepage’s banner. Without BRENDA’s thorough user’s guide, available from the “Help” link, the sheer number of searches would be very disconcerting. This guide is very user friendly and informative, with pictures and text to lead users through any type of search (example page shown in **Fig. 19B**). Also, most of the search pages provide simple explanations and examples of how to search that topic (**Fig. 19C**). The diversity of searches and clearly written guides facilitate the access to the database. The result pages for BRENDA are quite thorough, displaying a complete listing of experimental information about an enzyme. A few topics are substrate name, organism, localization, and inhibitors (**Fig. 20**). While a long list of data is generally very cumbersome, BRENDA’s results pages are simple to navigate because of the topic menu in the right frame. Each subject jumps to that specific point in the list of data. The menu is dynamic in that its contents are determined by the information contained within that protein entry, such that not all topic menus are alike. Within the results page, further links provide images of molecules and information on literature sources.

2.3.2.4. THE MHC PEPTIDE INTERACTION DATABASE

The MPID incorporates structural and “sequence-structure-function information” on peptide interactions of MHC alleles (63). Released by the Bioinformatics Centre at the National University of Singapore, the database contains 90 entries from human, rat, or mouse sources (**Table 3**). These protein complexes are obtained from the PDB (10) and are manually verified. One of this database’s strengths is the help and description that are designed into the site. A straightforward search page is provided from the Query MPID link on the homepage (**Fig. 21A**). Help links from the homepage as well as the search page bring up a page that is very descriptive about how to search the site. The search results are a simple list of hits with information about each MHC-peptide pair (**Fig. 21B**). Navigation around this page can be done by scrolling or by using the floating sidebar, which moves the display directly to particular entries in the list (**Fig. 21B**). The structures are analyzed for atomic interactions of the peptide complex, including calculations for hydrogen bonds, gap volume, gap index, and interface, using such programs as SURFNET (64) and NACCESS (65). All of this information is shown as a long list on an entry page. Most of the row titles are links to descriptive pop-up windows. Interactions of the MHC with peptides are presented as instructive ligplots

A



The BRENDA homepage features a central search bar with fields for EC Number, Enzyme Name, Organism, and Advanced Search. Below the search bar is a "Sequence Search and Short Tutorial" section. The page is divided into several sections with "SEARCH-Navigator" boxes:

- Enzyme Names:** EC Number, Enzyme Name, Organism, Advanced Search.
- Reaction & Specificity:** Reaction Type, Natural Substrates, Substrates and Products.
- Functional Parameters:** K_M Value, K_I Value, Turnover Number, Specific Activity, pH Optimum, pH Range, Temperature Optimum, Temperature Range.
- Isolation & Preparation:** Isolation, Crystallization.
- Organism-related Information:** Organism, Source, Tissue, Localization.
- Stability:** pH Stability, Temperature Stability, General Stability, Organic Solvent Stability, Storage Stability.
- Enzyme Structure:** Sequence, SwissProt link, 3D-Structure, PDB link, Molecular Weight.
- Disease & References:** Disease References, Application & Engineering.

At the bottom, there is a "Webmaster: Christian Ebeling" link.

B



This help page for the "Sequence Search" section shows the SEARCH-Navigator on the left. The "SEARCH-Navigator" box for "Sequence Search" is expanded, showing sub-options like "Reaction & Specificity" and "Organism-related Information". Other sections in the navigator include "Enzyme Names", "Functional Parameters", "Isolation & Preparation", "Stability", "Enzyme Structure", and "Disease & References".

C



The "Sequence Search" page displays search examples and a search form. The examples include:

- Sequences which contains "keal"
- Sequences which begins with "keal"
- Sequences which ends with "keal"
- Sequences which contains "ke"
- Sequences which begins with "keal" in a not defined distance
- Sequences with the consensus PROSITE pattern PD003952: GDA[1]CDS9 family of nucleotide phosphatases [LIVM][D-E][D-E]-G-[F-Y]-[F-W][A,V][T/A][S-N-H]
- You have a DNA or RNA sequence and want to know if the translation sequence codes for an enzyme

The search form includes fields for "Amino acid Sequence", "No. of results", "Recommended name", "EC Number", "Entry name", "Organism", "Number of amino acids", and "Molecular weight [Da]".

Fig. 19. The Comprehensive Enzyme Information System (BRENDA). (A) The homepage of BRENDA, showing the many and varied search possibilities. Simple, direct searches or “Fast searches” are at the top of the page, just under the title. Below this are the individual searches on every possible database entry field. These are also organized in the upper-right box in a “SEARCH-Navigator” box, so that the searches can be easily browsed by context. Underneath this box is another one with more complex and common searches. The bottom box contains various information, including links for help and tutorials. (B) Example of a BRENDA help page, explaining how to use the SEARCH-Navigator box to browse and find the correct search. Like files organized in folders, the searches are grouped under expandable folders. (C) Example of the “Sequence Search” page. The simple example is very helpful at the top of the page, and shows a search with potential results.

[BRENDA home](#) [BACK](#) [History of your search](#)

BRENDA
The Comprehensive Enzyme Information System

Entry of calmodulin-lysine N-methyltransferase (EC-Number 2.1.1.60)

Are you interested in using the BRENDA Discussion process?
Mark a special word or phrase in this record: [Match](#) (at least 3 characters)

[PRINT](#)

Selected one or more organism in this record: [All organisms](#) [Escherichia coli](#) [Ovis aries](#) [Parsmeum tetraurelia](#) [Rattus norvegicus](#) [Search](#)

EC NUMBER	COMMENTARY
2.1.1.60	-

RECOMMENDED NAME	Gene/Proteome No.
calmodulin-lysine N-methyltransferase	EC.001.1803

SYSTEMATIC NAME
S-adenosyl-L-methionine:calmodulin-L-lysine N-methyltransferase

SYNONYM	ORGANISM	COMMENTARY	LITERATURE
calmodulin-lysine N-methyltransferase	-	-	-
calmodulin N-methyltransferase	-	-	-
lysine methyltransferase, calmodulin (lysine)	-	-	-
S-adenosylmethionine:calmodulin N-methyltransferase (lysine)	-	-	-

CAS REGISTRY NUMBER	COMMENTARY
TM681-29-2	-

REACTION	COMMENTARY
S-adenosyl-L-methionine + calmodulin-L-lysine = S-adenosyl-L-homocysteine + calmodulin-N-methyl-L-lysine	Ratios homogeneous: Bi Bi sequential mechanism \rightarrow

REACTION TYPE	ORGANISM	COMMENTARY	LITERATURE
methionyl group transfer	-	-	-

ORGANISM	COMMENTARY	LITERATURE
Escherichia coli	-	9, 9, 9
Ovis aries	-	8
Parsmeum tetraurelia	-	2
Rattus norvegicus	-	2, 9, 9, 9

SUBSTRATE	PRODUCT	ORGANISM	COMMENTARY	LITERATURE	COMMENTARY	LITERATURE
more	?	Rattus norvegicus	trypsin C has low activity at very high trypsin concentration \rightarrow	2	-	-
more	?	Escherichia coli	structural requirements of calmodulin \rightarrow	8	-	-
S-adenosyl-L-methionine + calmodulin-L-lysine	S-adenosyl-L-homocysteine + calmodulin-N-methyl-L-lysine	Rattus norvegicus	Calmodulin: trimethyllysine: calmodulin, major product; + dimethyllysine + monomethyllysine \rightarrow 3 mol of methyl per mol of calmodulin are incorporated into lysine: 115 of trimethyllysine, 114 of dimethyllysine, 113 of monomethyllysine. Dimethyllysine: trimethyllysine: calmodulin: major product; + trimethyllysine + formation of epsilon-N-mono-, epsilon-N-di- and epsilon-N-trimethyllysine. The labelled N-methyllysine lies in the 107-126 peptide \rightarrow enzyme methylates a specific lysine residue of endogenous calmodulin \rightarrow	1, 2, 3, 5, 6	1, 2, 3, 5, 6	
S-adenosyl-L-methionine + calmodulin-L-lysine	S-adenosyl-L-homocysteine + calmodulin-N-methyl-L-lysine	Ovis aries	Plasma calmodulin: trimethyllysine calmodulin, major product; + dimethyllysine + monomethyllysine \rightarrow	6	-	6
S-adenosyl-L-methionine + calmodulin-L-lysine	S-adenosyl-L-homocysteine + calmodulin-N-methyl-L-lysine	Parsmeum tetraurelia	-	2	trimethyllysine calmodulin, major product; + dimethyllysine + monomethyllysine \rightarrow	2

INHIBITOR	ORGANISM	COMMENTARY	LITERATURE	IMAGE
calmodulin	Rattus norvegicus	-	8	
calmodulin	Parsmeum tetraurelia	-	2	
EGTA	Rattus norvegicus	-	1	
esselin	Parsmeum tetraurelia	-	2	
more	Parsmeum tetraurelia	not inhibitory: sinatungol, tubendorf, calmodulin antagonist IV \rightarrow	2	-
γ-chloronorbornyl sulfonic acid	Rattus norvegicus	-	3	
S-adenosylhomocysteine	Rattus norvegicus	-	8	
S-adenosylhomocysteine	Parsmeum tetraurelia	-	2	
VU-3 calmodulin	Rattus norvegicus	-	8	-
VU-5 calmodulin	Ovis aries	site-specific mutant of calmodulin, Lys 10 Arg-115 \rightarrow	6	-
VU-6 calmodulin	Ovis aries	site-specific mutant of calmodulin, Lys 10 Ile-115 \rightarrow	6	-

Fig. 20. Example of a Comprehensive Enzyme Information System (BRENDA) results page. The page shown has been abridged, although BRENDA stores a wealth of information. All of these data can be navigated from the floating menu on the left of the page. Links in the tables guide users to further information like references and/or images of molecules (see the inhibitors table on the bottom).



Fig. 21. Major Histocompatibility Complex (MHC)-Peptide Interaction Database (MPID). **(A)** The very simplified homepage for the MPID offers straightforward information on the database, including statistics, searching, and structural alignment information. Also offered on the homepage is a help section to introduce the novice user to searching for MHC peptides. **(B)** At the top of the results page, the search parameters are displayed with the number of entries found that match the user's criteria. Following this are the actual MHC peptide results, including MHC type and allele, structural information, and links for alignments and ligplots (66). **(C)** The ligplot link from the results page illustrates the MHC peptide with ligand and nonligand bonds, hydrophobic contacts, and accessibility.

(66): plots which incorporate ligand and nonligand bonds, hydrogen bond locations and lengths, solvent accessibility, and hydrophobic contacts (Fig. 21C). Structural alignments of MHC proteins by class and then by peptide length are also available from the “Structural alignment” link on the homepage (Fig. 21A), but display requires the use of the Chime Plug-in software (available from MDL Information Systems, Inc., at www.mdlchime.com).

3. Discussion

All of the protein interaction databases listed in Tables 1–3 have required a great deal of work to compile and present on the Internet. In general, databases require a lot of effort to design and to populate with data. Adding a robust user interface adds another level of complexity. For all this work, it should be noted that many databases are not directly funded, having been created as the byproduct of another research project. The majority of protein interaction databases remain free and open to the public. While privatizing protein interaction databases is not a workable model, funding has become a concern, as exemplified by the BRENDA database. BRENDA is now free only to academic users and must charge others (for more detail see the “Project Status/Funding Crisis” link on the BRENDA homepage, Fig. 19A). Funding issues may explain why some of the databases have become static or have irregular updates. Even so, as evidenced by the recent improvements to the GRID, MINT, and BIND, many databases are under continual change and development. As the field of proteomics expands, new protein interaction databases will certainly be created, but also, certain databases will cease to exist or become integrated into another, like DIP or BIND. This expansion also portends that the amount of information within databases will certainly grow. For the general biological community, finding pertinent information in these even larger datasets will become even more challenging. To provide such services, database authors need to take the time and effort to design transparent user interfaces for the searching and browsing of their information. After reviewing the many protein interaction databases, certain common features were identifiable in the databases that were easy to navigate. Such transparency includes an informative description, user-friendly search tools, and clear results presentation. In many ways, this section is oriented more for database architects of protein interaction databases than for the end-user.

Before even a search for information is attempted, potential users of the database need to know whether the database contains data that will be useful for their research. A brief introduction and information page is necessary. This description should include the following: (1) the purpose and goals of the database; (2) the types (including species) of interactions stored; (3) how the data were obtained, when applicable; (4) statistics about the database’s current holdings; (5) last and frequency of updates; and (6) how the results can potentially be used.

This page should be written so that it is easy to find all the information necessary to convince a user that the database contains relevant data and is a reliable source for interactions they are seeking. The purpose and goals are especially important, as they provide a clear point of reference for the database as well as define keywords/phrases by which the database will be classified in the various search engines.

The search/browse tools act as the bridge between the user and the data. Both simple and advanced methods for finding data provide the necessary variety to satisfy all lev-

els of users. The simple search should be designed for the novice user, who doesn't want to spend too much time trying to figure out how to define their search. A good place to start is a basic one- or two-word search with limits, or browsing with a pre-defined list of protein names like the UM-BBD (**Fig. 17A**). Such basic searches help the novice user gain confidence with the database. Because of their simplicity, the search is fast, so that the user obtains immediate feedback: type in the word and get a result. Also, the results should not be too overwhelming, such that the user has to cull through them manually. Faced with a daunting search tool and no results or too many after the first couple of tries, the novice user is unlikely to persist as a result of frustration. In conjunction with these search pages, there should be examples of searches either on the same page as the search (like BindingDB [**Fig. 9B**] or BRENDA [**Fig. 19C**]) or on a separate page. As with all the help pages, these guides to searching should be clear and concise. Then, there is the general approach to searching. The UM-BBD breaks the search into multiple levels, going from complicated pathways to simple substrates. In addition, the graphic representations used by UM-BBD (**Fig. 18B**), DIP (**Fig. 2E**), and MINT (**Fig. 7C,D**) allow the user to traverse pathways and search more intuitively by making visual connections. Most of the other databases take a more direct route, but BRENDA (**Fig. 19**) and the BindingDB (**Fig. 9**) provide more than one way to get there. Even with the advanced methods, the search or browse tool should not be an obstacle to a user in locating relevant protein interaction data.

As the consequence of a search, the results should also be presented clearly and be easy to manage. This applies to the general and specific results. For results at the general and specific levels, the basic strategy with large amounts of information is to split it up into smaller, digestible portions. With general results, the user acquires a list of potential matches. When the number of hits is too large, the user should be able to refine the search or select a subset of the results. Another possibility is that the user can select one match and use it as a starting point to find other similar matches (much like the "Related Article" link in PubMed). Once the user has found specific hits, the amount of information at this level can be as overwhelming as the large list from the general search. While most of the protein interaction databases' individual entries take up less than a page, in some cases, the entry is many pages long, as in the HPRD (**Fig. 12**) and BRENDA (**Fig. 20**). The HPRD solves this wealth of data by breaking it up into sections using tabs. BRENDA uses a topic menu to navigate its considerable amount of enzyme data. Help or definition pages for the results are also necessary, especially where abbreviations and unique terms are used. These allow a user to quickly parse the database and understand the kind of information that they can retrieve from it. Overall, the goal is for the results to be understandable to the user, whether it be smaller amounts or an explanation of data.

Some last miscellaneous points: All links should lead to something, because dead links can frustrate end-users. While not always necessary, a flat text-file version is always useful to parse, especially when a researcher wants to download the data and model it. A step up from the flat text file is to use XML, as suggested by the Proteomics Standards Initiative (28,67). This movement is attempting to create a ubiquitous format across many databases, so that combining information from two different databases becomes easier. In general, databases with clean and simple Web front ends are easier to navigate through. As a step in that direction, changing the predominant text versions

of data into more graphic representations of the data would help in interpretation of data. For large amounts of data, this is especially true, since it is easier to absorb large amounts of data visually than in text format. It is appreciated that improving databases and adding graphics requires a large commitment from database designers. These are suggestions and do not fit the requirements of all the protein interaction databases.

In this chapter, protein interaction databases have been generally reviewed. As with the protein interaction databases, the goal of this work is to act as a general guide for protein interaction databases. A number of these databases have been evaluated in detail as to their contents and search methods. Protein interaction databases have been covered from comprehensive protein pathway databases to protein interaction structure to proteins with nucleic acids and ligands. The discussion of these databases has attempted to highlight the strengths found in the databases in searching and presenting the data to a researcher in as transparent manner as possible. This discussion should also be useful to those who have designed protein interaction databases and those who are going to. It is hoped that these descriptions will help the general biological community and individual researchers in particular to find the appropriate protein interaction database as well as learn to use them.

Acknowledgments

This work was supported by NSF Biological Informatics Starter Grant DBI-0202599, Welch Foundation grant A-1459, and Texas A&M University.

References

1. Zhu, H., Bilgin, M., and Snyder, M. *Proteomics*. (2003) *Annu. Rev. Biochem.* **72**, 783–812.
2. Ito, T. et al. (2002) Roles for the two-hybrid system in exploration of the yeast protein interactome. *Mol. Cell. Proteomics* **1**, 561–566.
3. Fodor, S. P. et al. (1993) Multiplexed biochemical assays with biological chips. *Nature* **364**, 555–556.
4. Schena, M. et al. (1996) Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc. Natl. Acad. Sci. USA* **93**, 10,614–10,619.
5. Fields, S., and Song, O. (1989) A novel genetic system to detect protein-protein interactions. *Nature* **340**, 245–246.
6. O'Farrell, P. Z., Goodman, H. M., and O'Farrell, P. H. (1977) High resolution two-dimensional electrophoresis of basic as well as acidic proteins. *Cell* **12**, 1133–1141.
7. Figeys, D., McBroom, L. D., and Moran, M. F. (2001) Mass spectrometry for the study of protein-protein interactions. *Methods* **24**, 230–239.
8. Yandell, M. D. and Majoros, W. H. (2002) Genomics and natural language processing. *Nat. Rev. Genet.* **3**, 601–610.
9. Hirschman, L., Park, J. C., Tsujii, J., Wong, L., and Wu, C. H. (2002) Accomplishments and challenges in literature data mining for biology. *Bioinformatics* **18**, 1553–1561.
10. Berman, H. M. et al. (2002) The Protein Data Bank. *Acta Crystallogr. D. Biol. Crystallogr.* **58**, 899–907.
11. Marino-Ramirez, L., Campbell, L., and Hu, J. C. (2003) Screening peptide/protein libraries fused to the lambda repressor DNA-binding domain in *E. coli* cells. *Methods Mol. Biol.* **205**, 235–250.
12. Marcotte, E. M., Xenarios, I., and Eisenberg, D. (2001) Mining literature for protein-protein interactions. *Bioinformatics* **17**, 359–363.

13. McDermott, J. and Samudrala, R. (2003) Bioverse: Functional, structural and contextual annotation of proteins and proteomes. *Nucleic Acids Res.* **31**, 3736–3737.
14. Nanao, M. H., Zhou, W., Pfaffinger, P. J., and Choe, S. (2003) Determining the basis of channel-tetramerization specificity by x-ray crystallography and a sequence-comparison algorithm: Family Values (FamVal). *Proc. Natl. Acad. Sci. USA* **100**, 8670–8675.
15. Puntervoll, P. et al. (2003) ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res.* **31**, 3625–3630.
16. Kim, W. K., Park, J., and Suh, J. K. (2002) Large scale statistical prediction of protein-protein interaction by potentially interacting domain (PID) pair. *Genome Inform. Ser. Workshop Genome Inform.* **13**, 42–50.
17. Deane, C. M., Salwinski, L., Xenarios, I., and Eisenberg, D. (2002) Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol. Cell. Proteomics* **1**, 349–356.
18. Xenarios, I. et al. (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* **30**, 303–305.
19. Xenarios, I. et al. (2000) DIP: the database of interacting proteins. *Nucleic Acids Res.* **28**, 289–291.
20. Hodges, P. E., Payne, W. E., and Garrels, J. I. (1998) The Yeast Protein Database (YPD): a curated proteome database for *Saccharomyces cerevisiae*. *Nucleic Acids Res.* **26**, 68–72.
21. Karp, P. D. et al. (2002) The EcoCyc Database. *Nucleic Acids Res.* **30**, 56–58.
22. Giot, L. et al. (2003) A protein interaction map of *Drosophila melanogaster*. *Science* **302(5651)**, 1727–1736.
23. Kanehisa, M. (2002) The KEGG database. *Novartis Found. Symp.* **247**, 91–101; discussion 101–103, 119–128, 244–252.
24. Takai-Igarashi, T., Nadaoka, Y., and Kaminuma, T. (1998) A database for cell signaling networks. *J. Comput. Biol.* **5**, 747–754.
25. Bader, G. D., Betel, D., and Hogue, C. W. (2003) BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.* **31**, 248–250.
26. Bader, G. D. and Hogue, C. W. (2000) BIND—a data specification for storing and describing biomolecular interactions, molecular complexes and pathways. *Bioinformatics* **16**, 465–477.
27. Zanzoni, A. et al. (2002) MINT: a Molecular INTeraction database. *FEBS Lett.* **513**, 135–140.
28. Orchard, S., Hermjakob, H., and Apweiler, R. (2003) The proteomics standards initiative. *Proteomics* **3**, 1374–1376.
29. Ji, Z. L. et al. (2003) KDBI: Kinetic Data of Bio-molecular Interactions database. *Nucleic Acids Res.* **31**, 255–257.
30. Chen, X., Lin, Y., Liu, M., and Gilson, M. K. (2002) The Binding Database: data management and interface design. *Bioinformatics* **18**, 130–139.
31. Chen, X., Liu, M., and Gilson, M. K. (2001) BindingDB: a Web-accessible molecular recognition database. *Comb. Chem. High Throughput Screen.* **4**, 719–725.
32. Guerois, R., Nielsen, J. E., and Serrano, L. (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.* **320**, 369–387.
33. Carter, C. W., Jr., LeFebvre, B. C., Cammer, S. A., Tropsha, A., and Edgell, M. H. (2001) Four-body potentials reveal protein-specific correlations to stability changes caused by hydrophobic core mutations. *J. Mol. Biol.* **311**, 625–638.
34. Kwasigroch, J. M., Gilis, D., Dehouck, Y., and Roonman, M. (2002) PoPMuSiC, rationally designing point mutations in protein structures. *Bioinformatics* **18**, 1701–1702.
35. Chen, X., Lin, Y., and Gilson, M. K. (2001) The binding database: overview and user's guide. *Biopolymers* **61**, 127–141.

36. Mewes, H. W. et al. (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* **30**, 31–34.
37. Peri, S. et al. (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res.* **13**, 2363–2371.
38. Kikuno, R., Nagase, T., Waki, M., and Ohara, O. (2002) HUGE: a database for human large proteins identified in the Kazusa cDNA sequencing project. *Nucleic Acids Res.* **30**, 166–168.
39. Kikuno, R. et al. (2000) HUGE: a database for human large proteins identified in the Kazusa cDNA sequencing project. *Nucleic Acids Res.* **28**, 331–2.
40. DeLano, W. L. (2002) Unraveling hot spots in binding interfaces: progress and challenges. *Curr. Opin. Struct. Biol.* **12**, 14–20.
41. Nayal, M., Hitz, B., and Honig, B. SPIN-PP: trantor.bioc.columbia.edu/cgi-bin/SPIN.
42. Nicholls, A., Sharp, K. A., and Honig, B. (1991) Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins* **11**, 281–296.
43. Thorn, K. S. and Bogan, A. A. (2001) ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics* **17**, 284–285.
44. Bogan, A. A. and Thorn, K. S. (1998) Anatomy of hot spots in protein interfaces. *J. Mol. Biol.* **280**, 1–9.
45. Fischer, T. B. et al. (2003) The binding interface database (BID): a compilation of amino acid hot spots in protein interfaces. *Bioinformatics* **19**, 1453–1454.
46. Breitkreutz, B. J., Stark, C., and Tyers, M. (2003) The GRID: the General Repository for Interaction Datasets. *Genome Biol.* **4**, R23.
47. Breitkreutz, B. J., Stark, C., and Tyers, M. (2002) The GRID: The General Repository for Interaction Datasets. *Genome Biol.* **3**, PREPRINT0013.
48. Breitkreutz, B. J., Stark, C., and Tyers, M. (2003) Osprey: a network visualization system. *Genome Biol.* **4**, R22.
49. Breitkreutz, B. J., Stark, C., and Tyers, M. (2002) Osprey: a network visualization system. *Genome Biol.* **3**, PREPRINT0012.
50. Sarai, A. et al. (2001) Thermodynamic databases for proteins and protein-nucleic acid interactions. *Biopolymers* **61**, 121–126.
51. Prabakaran, P. et al. (2001) Thermodynamic database for protein-nucleic acid interactions (ProNIT). *Bioinformatics* **17**, 1027–1034.
52. Heinemeyer, T. et al. (1999) Expanding the TRANSFAC database towards an expert system of regulatory molecular mechanisms. *Nucleic Acids Res.* **27**, 318–322.
53. Salgado, H. et al. (2001) RegulonDB (version 3.2): transcriptional regulation and operon organization in *Escherichia coli* K-12. *Nucleic Acids Res.* **29**, 72–74.
54. Puvanendrampillai, D., and Mitchell, J. B. (2003) L/D Protein Ligand Database (PLD): additional understanding of the nature and specificity of protein-ligand complexes. *Bioinformatics* **19**, 1856–1857.
55. Orengo, C. A. et al. (2002) The CATH protein family database: a resource for structural and functional annotation of genomes. *Proteomics* **2**, 11–21.
56. Laskowski, R. A. (2001) PDBsum: summaries and analyses of PDB structures. *Nucleic Acids Res.* **29**, 221–222.
57. Laskowski, R. A. et al. (1997) PDBsum: a Web-based database of summaries and analyses of all PDB structures. *Trends Biochem. Sci.* **22**, 488–490.
58. Ellis, L. B., Hershberger, C. D., Bryan, E. M., and Wackett, L. P. (2001) The University of Minnesota Biocatalysis/Biodegradation Database: emphasizing enzymes. *Nucleic Acids Res.* **29**, 340–343.

59. Ellis, L. B., Hershberger, C. D., and Wackett, L. P. (1999) The University of Minnesota Biocatalysis/Biodegradation Database: specialized metabolism for functional genomics. *Nucleic Acids Res.* **27**, 373–376.
60. Ellis, L. B., Hershberger, C. D., and Wackett, L. P. (2000) The University of Minnesota Biocatalysis/Biodegradation database: microorganisms, genomics and prediction. *Nucleic Acids Res.* **28**, 377–379.
61. Ellis, L. B., Hou, B. K., Kang, W., and Wackett, L. P. (2003) The University of Minnesota Biocatalysis/Biodegradation Database: post-genomic data mining. *Nucleic Acids Res.* **31**, 262–265.
62. Pharkya, P., Nikolaev, E. V., and Maranas, C. D. (2003) Review of the BRENDA Database. *Metab. Eng.* **5**, 71–73.
63. Govindarajan, K. R., Kangueane, P., Tan, T. W., and Ranganathan, S. (2003) MPID: MHC-Peptide Interaction Database for sequence-structure-function information on peptides binding to MHC molecules. *Bioinformatics* **19**, 309–310.
64. Laskowski, R. A. (1995) SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graph.* **13**, 323–30, 307–308.
65. Hubbard, S. J. and Thornton, J. M. (1992) Naccess, <http://wolf.bms.umist.ac.uk/naccess/>.
66. Wallace, A. C., Laskowski, R. A., and Thornton, J. M. (1995) LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. *Protein Eng.* **8**, 127–134.
67. Pandey, A. (2001) Common standards for genomics and proteomics. *Trends Genet.* **17**, 442.

Deriving Function From Structure

Approaches and Limitations

Annabel E. Todd

1. Introduction

The fold and biochemical activity of a protein are tightly coupled. Once a protein is characterized, it is usual to determine its structure in order to derive an atomic description of its molecular mechanism. The fold reveals interaction surfaces, ligand-binding pockets, and the precise juxtaposition of functional groups.

The derivation of function directly from structure has gained widespread attention in recent years. Genome sequencing has seen the advent of international structural genomics initiatives, which aim to derive a structural representative for all homologous protein families (1,2). A considerable fraction of the structures solved by these projects will have little, if any, functional annotation, and biologists are faced with a complete reversal of the traditional approach to protein characterization (3).

To maximize the information return from these structural genomics projects, accurate methods to derive function from structure are essential. Critical to the success of these methods is a complete understanding of the complex relationships between protein sequence, structure, and function.

2. What Is Function?

Function is a somewhat vague term in that a protein may be described in one of several distinct and different ways: biochemical or molecular activity, cellular function, and physiological role. For example, the biochemical activity of the human protein trypsin is as a serine protease, its cellular function is protein degradation, and its physiological role is to aid digestion. Although the native structure may hint at little more than biochemical function, this functional assignment can usefully guide experiments to determine function at all levels. In this chapter, the word *function* largely refers to biochemical function.

3. Challenges of Inferring Function From Structure

Bioinformatics sequence search tools are used routinely to identify homologs of uncharacterized proteins. Evolutionary relationships are exploited to extrapolate function, on the basis that sequence homologs commonly exhibit some functional similarity. Having structural data, however, is more advantageous than sequence data alone. Not only do the structures reveal regions of functional significance, they uncover evo-

Table 1
One Structure, Many Functions
and One Function, Many Structures Paradox

One structure, many functions: α/β hydrolase fold	One function, many structures: Glycosyl hydrolase
Triacylglycerol lipase	α/α toroid
Cholesterol esterase	Concanavalin A-like 2-layer β-sandwich
Dienelactone hydrolase	Double psi β-barrel
Haloalkane dehalogenase	6-bladed β-propeller
Serine carboxypeptidase	(βα) ₈ or TIM barrel
Non-heme chloroperoxidase	Cellulase-like β/α-barrel
Neurotactin (cell-cell adhesion)	Orthogonal α-bundle

lutionary relationships between hitherto apparently unrelated proteins, since protein structure is conserved even after all trace of sequence similarity disappears (4). Such distant relationships can provide functional insights that are impossible to glean at the sequence level. These two benefits that structural data bring are two principal motivations behind the structural genomics projects.

However, there is no straightforward relationship between structure and function. For example, homologs sharing the α/β hydrolase fold have a rich variety of enzyme and non-enzyme functions (5), while the glycosyl hydrolase enzyme activity is associated with at least seven different scaffolds (6) (see Table 1). This “one structure, many functions and one function, many structures” paradox presents a major challenge to biologists attempting to infer function from structure.

4. Methods of Functional Evolution

In the discussion of protein structure and function, it is useful to consider the possible routes to new functions.

4.1. Gene Duplication

Biological complexity implies structural diversity. However, the α/β hydrolase fold and other superfamilies provide evidence for multiple gene duplication events and the reuse of “old” genes, and therefore folds, for new functions. Even in the very small genome of *Mycoplasma genitalium*, 60% of genes arose via gene duplication, with one ancestral domain (discussed later) having more than 50 copies (7). In the more complex eukaryotic organisms, the proportion of domains produced by duplications increases significantly, to 88% for *Saccharomyces cerevisiae* and 95% for *Caenorhabditis elegans* (8). With two identical copies of the same gene, one of them retains its function, while the second explores other functional possibilities through incremental mutations.

4.2. Gene Fusion

Proteins comprise one or more compact substructures called “domains.” Domains are fundamental building blocks in that they may occur in isolation, or in combination with different partners via gene fusion, or both. Accordingly, each domain has its own evolutionary history. Additionally, domains that are not fused may function together in oligomeric assemblies.

The importance of domain accretion in evolution is illustrated by the high percentage (>30%) of polypeptide chains within the Protein Data Bank (PDB) (9) that comprise more than one domain. (This probably represents a lower limit, since for many proteins, the structure of only one domain of several has been determined.) This “mix and match” method is probably a fast route to new functions, and has made a substantial contribution to the greater complexity of vertebrate proteomes compared to those of simpler eukaryotes and prokaryotes (10).

4.3. One Gene, Two or More Functions

An increasing number of genes are known to be multifunctional. The acquisition of a new function by an existing gene product is referred to as *gene recruitment*. This evolutionary strategy is exemplified by the recruitment of enzymes as crystallins, structural proteins in the eye lens (11), where this new non-catalytic role has been acquired by modifications in gene expression. There are several methods by which proteins “moonlight” (12). In addition to differential expression, the use of alternative binding sites, and changes in cellular localization, oligomeric state, and substrate concentration can all lead to functional variations.

Other genes owe their multifunctionality to posttranslational modifications, alternate translation initiation, and alternative splicing (AS). The gene products do not moonlight; the functions are carried out by non-identical proteins, which are nevertheless derived from the same gene. At least a third, and possibly as many as 60%, of human genes undergo AS (13), and a recent study (14) showed that AS tends to alter the domain architecture and functional sites of proteins, illustrating its importance in increasing the functional diversity of proteomes.

The multifunctionality of genes can explain the surprisingly small number of genes (<30,000) in the human genome. The use of one gene for two or more functions clearly simplifies the genome, but complicates the process of genome annotation.

5. From Structure to Function

The section outlines the functional information that can be derived from protein structure coordinates and from the local or global structural similarities the protein may share with others of known fold.

5.1. Basic Structure

The PDB file, which lists the coordinates of all atoms in a protein structure, provides little functional information. It gives the protein name, when known. Some files contain one or more “SITE” records that provide a formatted list of residues of functional relevance, such as those involved in catalysis, ligand binding, or protein–protein interactions. From the coordinates themselves, however, we can derive the functional information summarized in **Fig. 1**.

The structure reveals the spatial organization of the polypeptide chain(s), in terms of helices, strands, and coils, and the packing arrangement and connectivities of these elements. From this, one can identify those residues that contribute to the hydrophobic core and those that contact solvent, as well as the shape and electrostatic properties of the protein surface. Long, loopy regions lacking regular secondary structure may be functionally important (15,16). The quaternary structure can reveal the protein’s biologically relevant oligomeric state. This is not always true for X-ray crystal structures,

and algorithms are available to distinguish biological from non-biological interfaces in the crystal (17).

Protein–ligand complexes are particularly useful since they reveal the precise location of the binding site, the conformation of the ligand in its bound state, functional residues and their relative disposition, and, for enzymes, a possible catalytic mechanism. Typically, ligands are chosen with some knowledge of the protein’s function. In structural genomics, the ligand is often unknown, but examples have been reported in which the structure fortuitously includes a ligand scavenged from the expression host (18). Such data can provide valuable clues to function.

5.2. Structural Class

The most gross level of structure, as defined by secondary structure content, can tell us little about the precise biochemical activity of a protein. Nevertheless, there are clear biases with respect to general function at the class level (6,19). There is a dominance of mixed α/β folds in enzymes compared with non-enzymes, while there are very few enzyme superfamilies in the mainly α class. This may be historical in origin, as well as having some physicochemical basis.

Several common biological ligands have a notable bias towards certain protein classes defined by the stereochemical requirements for ligand binding (19). For example:

- a. Heme (α): the preferred binding mode is for the heme to slot between two or more helices.
- b. DNA (α and α/β): DNA recognition is dominated by the helix motif binding in the major groove.
- c. Nucleotides (α/β): many of the nucleotide-binding folds are $\alpha/\beta/\alpha$ sandwiches.

However, in terms of general function, the fold itself, rather than class, can sometimes be a better guide. For example, the ubiquitous triphosphate isomerase (TIM) barrel and P-loop nucleotide triphosphate hydrolase folds of the α/β class are dedicated catalytic folds, and the few noncatalytic proteins that adopt these structures were probably derived from ancestral enzymes (20).

5.3. Global or Local Structural Similarity

Currently, the best way to infer function from structure is the knowledge-based approach, which involves comparison of the protein fold, or spatial motifs contained within it, to other structures in the PDB (21). The functional information that can be derived depends on the nature of the relationship of the structure and its particular features to those already known (22).

The complexity of protein structure and function relationships was discussed under **Subheading 3.** and is further highlighted by the eight pairs of proteins illustrated in **Fig. 2.** The protein pairs are labeled as one of two types: homologs or analogs. Homologs, by definition, share a common ancestry. Analogs are proteins that are similar in some way, yet exhibit no evidence of an evolutionary relationship. Structural analogs share the same fold, and functional analogs perform the same function.

5.3.1. Homologous Relationship

Homologous relationships, if undetectable at the sequence level, can be identified using one of the numerous protein structure comparison algorithms now available, and

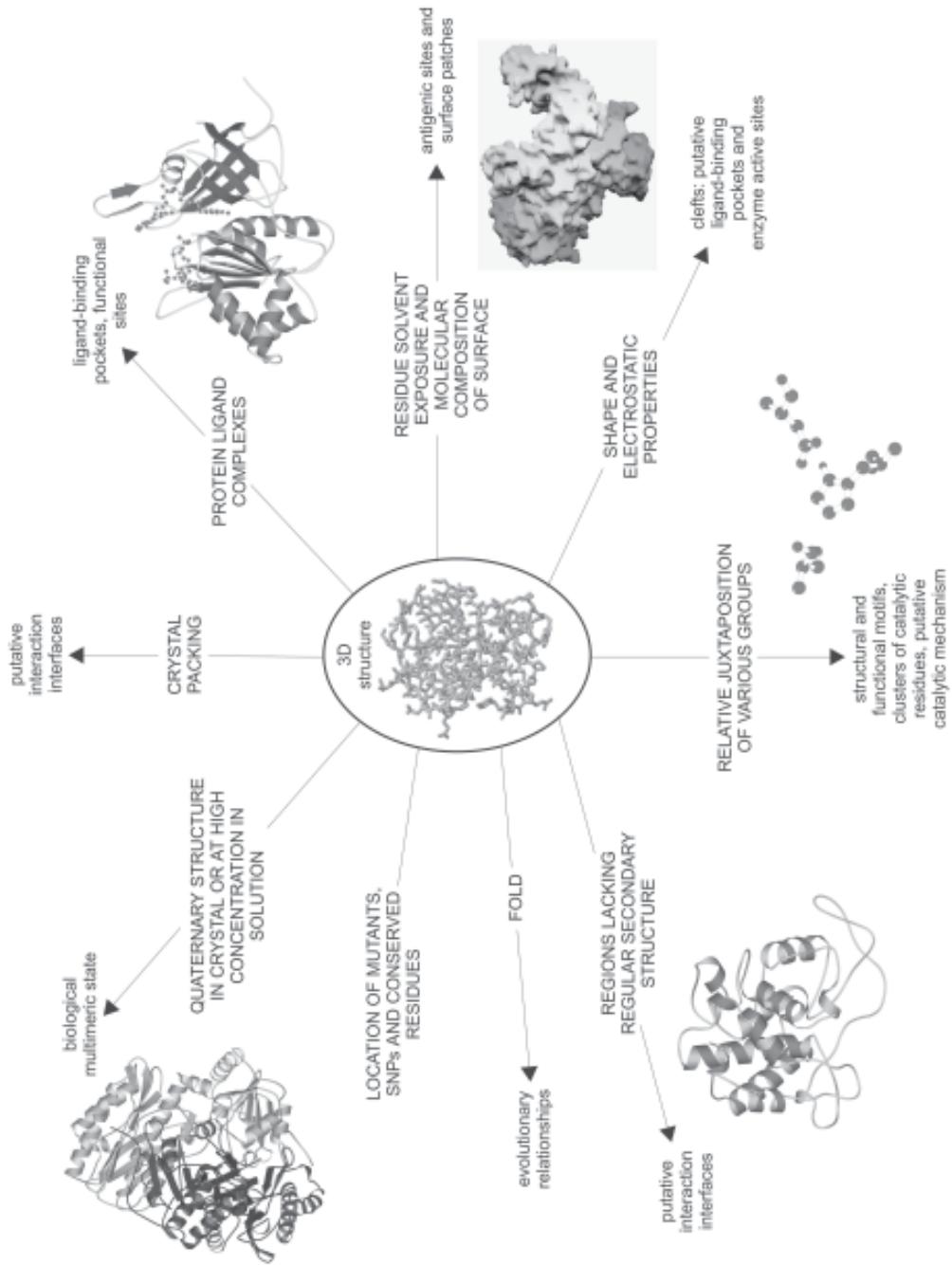


Fig. 1. From structure to function: A summary of information, relating to biochemical function, that can be derived from three-dimensional structure.

various classifications that rationalize known structures into domain superfamilies, such as Class, Architecture, Topology, and Homology (CATH) (23) and Structural Classification of Proteins (SCOP) (24), facilitate this process. The hemoglobins provide an excellent example of the benefit of identifying homologs in this way to infer function (see **Fig. 2C**). Sequence analysis fails to detect their common ancestry, yet their structures are very well conserved. Hypothetically, if the structure of an uncharacterized hemoglobin family member was solved, one could assign to it a putative oxygen-transport function.

However, considerable caution must be exercised in the transfer of function between relatives, as exemplified by the α/β hydrolases. The assumption that homologous proteins have related functions appears to be largely applicable to orthologs—equivalent genes in different species that evolved by speciation, but the function of paralogs—genes related by duplication within a genome—may be quite different in divergent families. Even at high sequence identity, changes in function occur (see **Fig. 2A,B**), multifunctional proteins representing a special case. Studies have shown, however, that

Fig. 2. (opposite page) Complexities of protein sequence, structure, and function relationships. **A–E** and **F–H** illustrate homologous and analogous relationships, respectively. Enzyme commission (EC) (33) numbers are included where they usefully illustrate similarities or differences in enzyme function. Enzyme reactions are assigned a four-digit EC number, where the first digit denotes the class of reaction (e.g., hydrolase, transferase). The meaning of subsequent levels depends upon the primary EC number, and these levels describe substrates, cofactors, and other reaction details. **(A)** Duck crystallins δ I and δ II function as structural proteins in the eye lens, and δ II also has argininosuccinate lyase activity; this activity is lacking in δ I despite the conservation of the catalytic residues (gray-shaded). **(B)** Lysozyme functions as an *O*-glycosyl hydrolase, but α -lactalbumin lacks this activity and instead regulates the substrate specificity of galactosyltransferase; the active site is disrupted in α -lactalbumin. **(C)** Despite their shared function, hemoglobins exhibit remarkable sequence diversity; the heme molecule bound to each structure is shown in ball-and-stick representation. **(D)** Given their different sizes and insignificant sequence similarity, the evolutionary origin of adenyllyl cyclase and the “palm” domain of DNA polymerase was a subject of controversy (108,109) until the identification of their co-located Mg^{2+} -binding sites (110); despite their different overall enzyme activities, they share some similarities in their catalytic mechanism and act on chemically similar substrates; the “fingers” domain of DNA polymerase interrupts the “palm” domain at the top, and it is represented by the long loops; Mg^{2+} -binding sites are shown in ball-and-stick representation. **(E)** Alanine racemase and eukaryotic ornithine decarboxylase have the same domain architecture, a pyridoxal 5'-phosphate-binding TIM barrel and a β -barrel domain, and their common ancestry is detectable from sequence; the topologies of their β -barrel domains differ and are related by a β -hairpin swap (29); a conserved active site Cys residue is shown in ball-and-stick representation. **(F)** Acylphosphatase and the DNA-binding domain of bovine papillomavirus-1 E2 both adopt the α/β -plait fold; there is no evidence for a common ancestry, and so they belong to different superfamilies. **(G)** Class B β -lactamases are metal-dependent, while the structurally distinct class A, C, and D enzymes have a Ser nucleophile, which is essential for catalysis and forms a covalent intermediate with the substrate; the two metal ions bound to class B β -lactamase are represented by the gray spheres, and the Ser nucleophile of the class A, C, and D fold is shown in ball-and-stick representation. **(H)** Subtilisin and chymotrypsin have different folds, yet they have the same Ser-His-Asp catalytic triad, and both function as serine endopeptidases via the same catalytic mechanism; the triad residues are shown in ball-and-stick representation. (**Fig. 2.** continued on page 808)

cryst. I BASE-GERLMOORGSTOPTIMELSTTSGLSVEVDQSTAYAKRERAGILERTTRKILISLSEERLSEELSHQIVVYTGQESQ-TANSGVFLEL100TAGLHTGR
cryst. II PAAASRANDEELVQVPPSTOPTIMELSTTSGLSVEVDQSTAYDQRLSVVQCSMAYAKRERAGILERTTRKILISLSEERLSEELSHQIVVYTGQESQ-TANSGVFLEL100TAGLHTGR

cryst. I: SGNQVTELEELFNSQGSLIISIETHLQLIKTIVKAKKIDIVLPLPNTLQEQADPIMKSFYLSQHANVLTQPGKGLGENVQHNLIPGQGQGALALSPFLTDSQMLKLEPAFSL
cryst. II: SGNQVTELEELFNSQGSLIISIETHLQLIKTIVKAKKIDIVLPLPNTLQEQADPIMKSFYLSQHANVLTQPGKGLGENVQHNLIPGQGQGALALSPFLTDSQMLKLEPAFSL

cryst. I: EMDA1280DPVQFPLKTYLTDLKLEMAGLIL1TETRQFPLTLQSAFPTTQGLMPVQD9P111LIGKAGKTFYKLA1LAVVQSLPFTYNGLQEDRANV1LDDVYTTATNL
cryst. II: EMDA1280DPVQFPLKTYLTDLKLEMAGLIL1TETRQFPLTLQSAFPTTQGLMPVQD9P111LIGKAGKTFYKLA1LAVVQSLPFTYNGLQEDRANV1LDDVYTTATNL

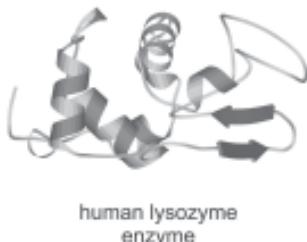
cryst. I QATUVIESTLQISQSGEGERALTFEMLATILALFLAVKQGFFRCQNTAERGEGTETGIAWNLTELEQIISPLPQGTVQ-AQVWVHEVTAEGTATGEGVFTQGIGQELM
cryst. II QATUVIESTLQISQSGEGERALTFEMLATILALFLAVKQGFFRCQNTAERGEGTETGIAWNLTELEQIISPLPQGTVQ-AQVWVHEVTAEGTATGEGVFTQGIGQELM

crys. I ~~1000000000~~
crys. II ~~1000000000~~

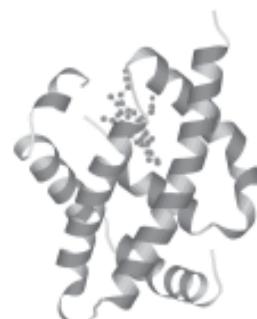
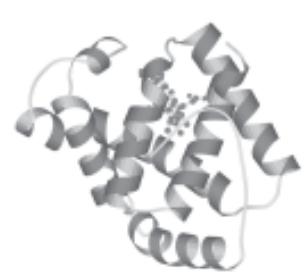
A

duck crystallin $\delta 1$ non-enzyme
duck crystallin $\delta 1$ /argininosuccinate lyase enzyme

HOMOLOGS
LOSS OF ENZYME ACTIVITY
94% seq ID
conserved active site



B
HOMOLOGS
ENZYME / NON-ENZYME
40% seq ID
disruption of active site

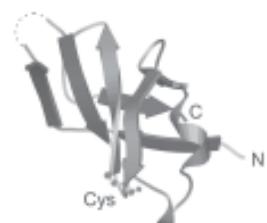


C
HOMOLOGS
IDENTICAL FUNCTIONS
8% seq. ID



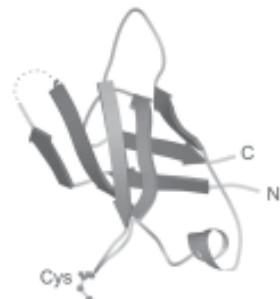
D
HOMOLOGS
DIFFERENT ENZYME ACTIVITIES
12% seq ID
co-located Mg-binding site
similarity in catalytic mechanism

'palm' domain of DNA polymerase
EC 2.7.7.7



alanine racemase

E
HOMOLOGS
DIFFERENT FOLDS
significant sequence similarity
conserved active site Cys in
'swapped' β -strands



eukaryotic ornithine decarboxylase



acylphosphatase

F
STRUCTURAL ANALOGS
SIMILAR FOLDS
DIFFERENT FUNCTIONS
no shared functional attributes

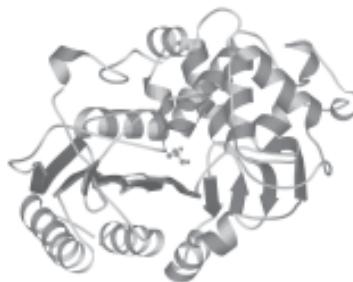


bovine papillomavirus-1 E2 transcription regulation protein, DNA-binding domain

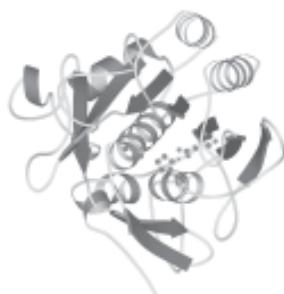


β -lactamase class B
EC 3.5.2.6
metal-dependent

G
FUNCTIONAL ANALOGS
DIFFERENT FOLDS
IDENTICAL ENZYME ACTIVITY
different active sites



β -lactamase classes A, C, D
EC 3.5.2.6
catalytic Ser nucleophile



subtilisin
EC 3.4.21.62

H
FUNCTIONAL ANALOGS
DIFFERENT FOLDS
SERINE ENDOPEPTIDASES
identical Ser-His-Asp catalytic triad



chymotrypsin
EC 3.4.21.1

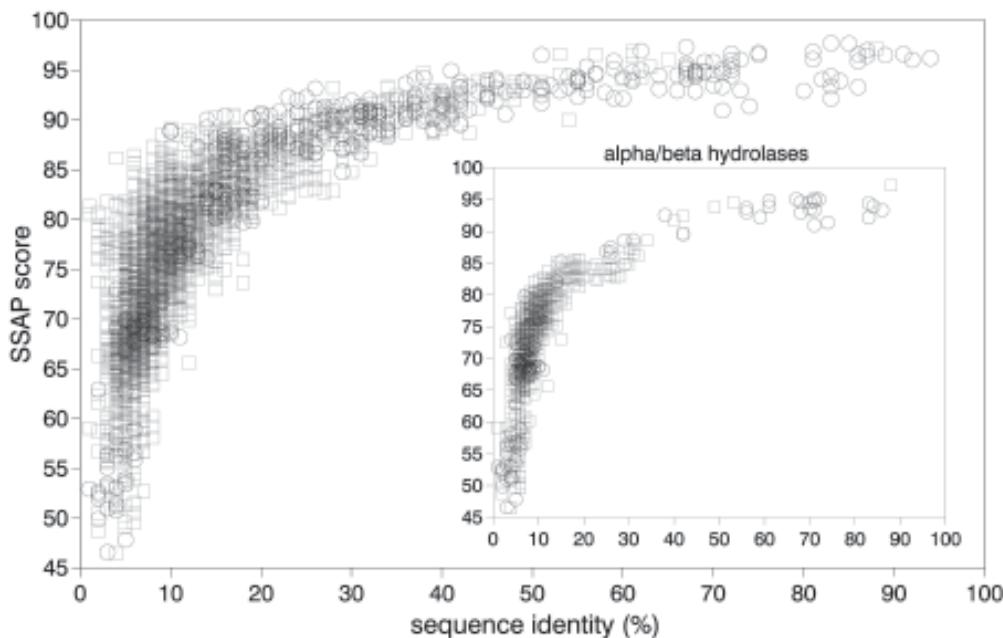


Fig. 3. Correlation between sequence, structural, and functional similarity in protein homologs (90). The graph plots pairwise sequence identity vs structural similarity score for all unique homologous domain pairs within 31 enzyme superfamilies (27), with points colored to distinguish pairs having identical (black circles) and different (gray squares) functions. The gray squares also include pairs in which one protein has a second activity not catalyzed by the other. 99% of functionally distinct pairs share less than 40% sequence identity. Structural similarity is measured by the SSAP algorithm (111), which returns a score of 0–100, a pair of identical proteins scoring 100. The inset illustrates the distribution for the α/β hydrolase superfamily only.

functional variation is significant only below 40% sequence identity (25–27), although more recent work suggests that functional conservation was overestimated in these analyses (28). Obviously the challenge is for biologists to differentiate sequence and structural variation that has specific functional consequences from variation that results from neutral drift. This requirement is highlighted further by Fig. 3, which plots the sequence identity and structural alignment score for homologous domain pairs, where pairs with the same (black circles) and different (gray squares) functions are distinguished.

A powerful method to infer shared ancestry and function is to assess the conservation of functional residues. In the inset in Fig. 3, all black circles in the bottom left corner correspond to pairs of triacylglycerol lipases of the α/β hydrolase superfamily. This lipase function is compatible with an assortment of structures; the folds of these homologs differ in β -sheet length (6–10 strands), topological connectivities close to the edges of the sheet, and in the number and sizes of insertions within the sheet. Nevertheless, they all have a Ser-His-Asp triad that is critical for catalysis. Fold change during evolution has become an increasingly recognized phenomenon (29) with the growth in the PDB (see Fig. 2E).

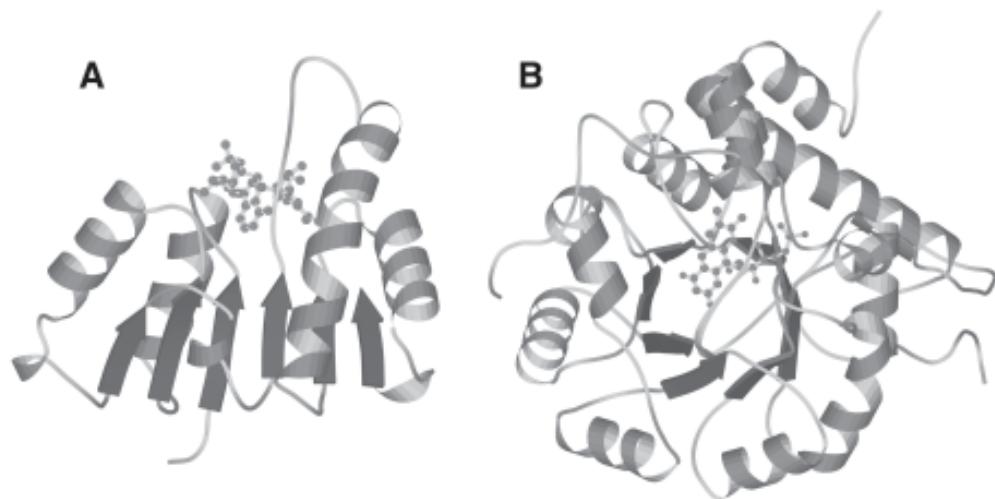


Fig. 4. The location of (A) Rossmann and (B) TIM barrel supersites (34). In the doubly wound α/β Rossmann fold, ligands bind at the crossover point between the two identical halves of the structure, where the β -strand order is reversed. In the TIM barrel, the enzyme active site is invariably formed by the carboxy ends, or the loops that follow, the β -strands in the central β -barrel. The existence of supersites may be associated with the general principles of protein structure. For example, the dipole moment of a helix could assist a catalytic residue located at one end of it.

5.3.2. Fold Similarity and Structural Analogs

Fold similarity does not necessarily imply a common ancestry. Evidence suggests that there is a limited number of folds in nature, perhaps as a result of physicochemical constraints on protein folding; there may be as few as 1000 (30). The vast majority of folds has one homologous superfamily associated with it, but a few structures, termed *superfolds* (31), such as the TIM barrel, appear to have recurred multiple times during evolution. Functions performed by these multifamily folds may be numerous and diverse (see Fig. 2F). For example, at least 60 enzyme functions are associated with TIM barrel proteins of known structure, and these activities belong to as many as five of the six enzyme classes (32), as defined by the Enzyme Commission (EC) (33). With the inclusion of sequence data, the number of functions is considerably higher. Therefore, it is useful to establish whether a common ancestry exists between an uncharacterized protein structure and its structural matches in the PDB.

On a cautionary note, divergence from a common ancestor, rather than the parallel evolution of a stable folding motif, may account for the existence of superfolds, and insertions, deletions, and substitutions have simply obscured their evolutionary history. This is just one benefit of classifying structures by fold similarity, as well as by superfamily, in CATH and SCOP; it highlights a possible ancestry between superfamilies of the same fold.

While functional prediction of an uncharacterized protein belonging to a highly populated fold may be fraught with difficulties, for a few superfolds there are preferred binding sites, termed *supersites* (34) (see Fig. 4). Therefore, in the absence of evolu-

tionary homologs, a fold similarity can suggest a putative site of functional relevance, and this information can usefully guide biochemical experiments. A statistical analysis (34) has confirmed and identified binding preferences in nine superfolds in the SCOP database, although the ligands to which each supersite binds can differ markedly in size and chemical properties.

5.3.3. Structural Motifs and Functional Analogs

It is sometimes possible to identify local spatial motifs of functional relevance, like the catalytic triad in the triacylglycerol lipases of the α/β hydrolase superfamily, and these may be used to infer molecular recognition and aspects of catalysis. Similar three-dimensional motifs may be identified in unrelated proteins having different folds, as well as in evolutionary relatives. For example, the lipase Ser-His-Asp triad appears to be a useful catalytic framework for other hydrolytic activities, because it exists in several scaffolds (35), including those of subtilisin and chymotrypsin. These two nonhomologous proteins both function as serine endoproteases via the same catalytic mechanism (see Fig. 2H). Local structural motifs are not restricted to catalytic sites, and other well-known examples include the DNA-binding helix-turn-helix motif and the calcium-binding EF hand. A wide variety of pattern-matching algorithms is available for the automatic detection of recurring motifs, as described in the next section.

Given that catalytic activity usually depends upon the precise spatial orientation of just a few amino acids, it is perhaps not surprising that the same active site can recur in different scaffolds. Moreover, only a subset of residues has the required functionality for most catalytic tasks, such as proton abstraction (36). Indeed, just six amino acids (H, C, D, E, R, K) account for 70% of all catalytic residues (see Fig. 5).

6. Methods to Identify Functional Sites

Structural genomics initiatives have fuelled the development of algorithms to identify functional sites in protein structures (37–41). The “resolution” of the data returned by these methods varies, from surface patches to specific atoms; nonetheless, all results are hypotheses and, ultimately, they must be tested by direct experimentation. A few methods are described in Box 1. Traditional docking methods predict protein–biomolecule interactions in a known functional site and are beyond the scope of this review, but it is worth drawing the reader’s attention to new docking applications that attempt to locate the site as well (see ref. 40 for a review).

The numerous approaches to the problem of automated site detection differ widely, but they can be loosely classified as follows: (a) pattern matching/similarity-based methods (42–54); (b) evolutionary/conservation-based (55–60); (c) physicochemical (61,62); and (d) geometric (63–66). Some newer methods adopt an integrated approach, incorporating various types of complementary data (67,68), so they fall into two or more of these broad categories. Both the first two types of approaches imply the exploitation of other sequence and structural data, but many structural genomics targets bear no global or local similarities with other known folds. Furthermore, some of these are orphans, having no (identifiable) sequence relatives in any genome. Accordingly, such methods are not applicable to these proteins, and good progress has been made in the development of *ab initio* techniques that identify functional sites (61,62) and general function (69) from hypothetical structures in isolation.

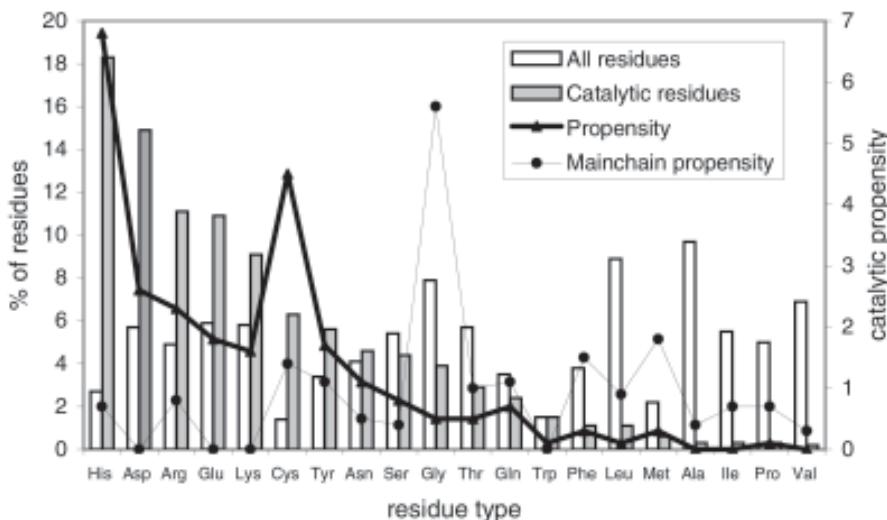


Fig. 5. Observed frequency distribution of catalytic residue types vs all residues (histogram bars), and catalytic propensity (line graphs). This figure was adapted from an analysis of 178 nonhomologous enzyme active sites (36). Catalytic propensity is defined as the percentage of catalytic residues constituted by a particular amino acid, divided by the percentage of all residues in the dataset constituted by the same amino acid. Just six amino acids (H, C, E, D, R, K) account for 70% of all catalytic groups. His and Cys have the highest catalytic propensities, possibly because their pK_{a} s are close to physiological pH and they can easily change protonation state. Eight percent of functional residues use main-chain atoms for catalysis, and for these interactions, for which small Gly accounts for a high proportion, amino acid identity is often irrelevant. The study also showed that catalytic residues have limited exposure to solvent despite their polarity, have reduced mobility although half of them reside in loops, and that they and the residues local in structure are highly conserved, much more so than those local in sequence.

6.1. Ab Initio: *Geometric and Physicochemical Approaches*

Methods to detect functional sites, notably *ab initio* methods, must reflect the nature of that site. Strategies that incorporate geometric criteria (67,68) exploit the observation that ligand-binding sites tend to occur in surface depressions and, for enzymes, the active site pocket usually corresponds to the largest cavity (70). A large cleft maximizes the number of complementary geometric and electrostatic interactions, and ensures protection of reactive intermediates from bulk solvent. Several different algorithmic approaches to locate surface depressions, and therefore putative binding sites, have been developed (63–66).

Another characteristic of active sites is that they tend to correspond to regions of instability in the structure. Active sites are “preorganized” to recognize the substrate (71) and they contain exposed hydrophobic patches, strained conformations, unfavorable electrostatic interactions, and unfulfilled hydrogen bond donors and acceptors (72). This concept of stability-function tradeoffs in enzymes has been exploited in an integrated method (67) that computes stability changes upon residue mutation. A purely *ab initio* physicochemical approach calculates the electrostatic energy of charged residues

Box 1**Methods to Detect Functional Sites***TESS, Jess, Catalytic Site Atlas*

Template Search and Superposition (TESS) (44) has been applied to enzyme active sites, but in theory it is applicable to any type of functional site. An active site template is derived manually and includes those atoms that are essential for catalysis. TESS identifies proteins in the Protein Data Bank (PDB) containing a similar cluster of atoms that overlap with the search template within a user-defined root mean square deviation (RMSD). More recently, the same group has developed an improved constraint-based search algorithm with rigorous statistics (Jess) (80), which uses a geometric partition tree to search efficiently for template matches. Catalytic Site Atlas, an inventory of literature-derived catalytic sites, is available at <http://www.ebi.ac.uk/thornton-srv/databases/CSA> (85), from which templates will be derived in the future to facilitate “reverse” searches, which involve the scanning of a newly determined structure against the template library.

FFF

Fuzzy Functional Forms (FFFs) are derived by the superposition of functionally significant residues in a few selected protein structures that have related functions (45). An FFF comprises the $C\alpha$ to $C\alpha$ distances of these residues and their tolerated variances, and may incorporate other information such as secondary structure. Using a search algorithm, residues that satisfy the constraints of the FFF may be identified in other experimentally determined structures, as well as in low-to-moderate-resolution models produced by fold recognition or *ab initio* structure prediction algorithms. Therefore, given an uncharacterized sequence, one can predict its fold, identify a putative active site that matches an FFF, and then predict its function. The method has been applied to the identification of disulphide oxidoreductases in the *Saccharomyces cerevisiae* genome (86). All previously known thioredoxins, glutaredoxins, and disulphide isomerases were correctly identified, and three novel predictions have been subsequently validated by experiment.

PINTS

Patterns In Nonhomologous Tertiary Structures (PINTS) (46,79) detects recurring side-chain patterns in protein structures without having any prior knowledge of the functional residues. It uses a depth-first search algorithm to find all spatially adjacent and similar clusters of residues common to two sets of coordinates, as measured by relative distances of $C\alpha$, $C\beta$, and functional atoms. To reduce search time, the method uses multiple sequence alignments to concentrate on conserved polar residues, i.e., those likely to be involved in function. A server at <http://pints.embl.de> (87) offers three search types: pairwise comparison, protein vs an automatically derived pattern database, and pattern vs protein database. Matches are ranked by statistical significance (79), and associated RMSDs are provided.

Evolutionary Trace (ET) and Related Methods

ET uses a phylogenetic tree derived from a multiple sequence alignment (MSA) and assumes that distinct branches approximate to functional classes of family members (56). ET iteratively partitions the family into classes at different branch points, and those residues identified as universally conserved or class-specific (invariant within a class but for which identity varies among classes) are mapped onto a representative

(continued)

Box 1 (Continued)**Methods to Detect Functional Sites**

structure to manually assess spatial proximity. So-called trace residues that form surface clusters are likely to define functional sites. The recent introduction of statistical measures has removed the need for human assessment, so it is now more applicable to automatic site prediction on a large scale (81,88). A server is available at <http://imgen.bcm.tmc.edu/molgen/labs/lichtarge/>. Recent variants use weighting schemes to better quantify residue replacements (58), employ better tree-building methods (ConSurf, <http://consurf.tau.ac.il> (59)), and explicitly consider branch lengths in calculating conservation (60).

In a novel variation of the ET method, 3-D cluster analysis (58), the reliance on phylogenetic trees is removed and structural data play an integral part in the analysis. For each residue in an MSA, the method computes a regional conservation score in which its spatial neighbors are considered too. It has been successfully applied to nucleic acid, small ligand, and protein–protein interfaces.

(61), and another computes theoretical titration curves of ionizable residues, since abnormal curves typically correspond to active site groups (62). Spatial proximity of putative functional residues gives a good indication of a relevant site.

Protein–protein interfaces are more difficult to characterize than small ligand-binding sites (73–75). Protein–protein interactions are structurally and functionally diverse, differing in composition, affinity, and lifetime (76). However, some generalizations can be made. Interfaces are relatively planar and accessible, and those in permanent assemblies are hydrophobic (73,74,77). A method that analyses surface patches on a protein in terms of six parameters (solvation potential, residue interface propensity, hydrophobicity, planarity, protrusion, and accessible surface area) has successfully identified the interaction site in various protein complexes (78). A recent analysis using a larger dataset found significant differences in amino acid composition and residue–residue contact preferences for six types of protein–protein interfaces, suggesting that prediction of interface type from sequence alone may be feasible (77).

6.2. Pattern Matching/Similarity-Based Methods

Pattern-matching methods identify spatial motifs common to two or more proteins. The ability to detect recurring patterns in different folds as well as in evolutionary relatives is a major advantage of this approach to functional site prediction. Traditional strategies capture structural features of a site in a three-dimensional template, which is then used to search for the same constellation of residues in other proteins (42–45,47). Given the size of the PDB, a simple global calculation that involves fitting of the template to all possible atom combinations within each structure is computationally prohibitive. A popular solution to improve search efficiency is to use reduced spatial information, with each residue represented by a subset of atoms that includes one or more of the following: C α atom, C β atom, a pseudo-atom (typically located at the residue's center of mass), functional atoms. Another advantage of simplified residue representations is that, to some extent, they accommodate the structural variations that exist even between different coordinate sets of the same or closely related proteins (42).

A major disadvantage of most pattern-matching strategies is that they require user-defined search templates (42–45,47), which usually necessitates referral to the literature to identify residues of functional significance, and the definition of template-specific root mean square deviation (RMSD) or distance tolerances. Mathematical data-mining techniques (49), a novel graph theoretic approach (50), and the exploitation of evolutionary information to restrict the search to conserved polar residues (46) overcome this limitation so that recurring spatial clusters can be identified automatically. Secondly, it is usually necessary to manually assess the search hits to discern biologically relevant similarities from random matches. To reduce background noise, and thereby improve specificity, some methods place constraints on sequence order (45,48), but this limits their usefulness in detecting clusters common to unrelated proteins. New rigorous statistical measures of significance (79,80) discriminate meaningful similarities from noise, thus reducing the need for human intervention.

Some new approaches to identifying functional site similarities focus on the physicochemical properties of known and putative functional cavities, rather than on the atomic positions of the residues that flank them (52–54). These approaches are particularly suitable for identifying cofactor- and other ligand-binding pockets, since binding residues, in general, are less strictly conserved than catalytic residues; nevertheless, consensus ligand interaction patterns do exist. Two related methods (53,54) conduct searches of a query cavity against a database of binding surfaces, and matching surface motifs are identified by clique detection algorithms.

6.3. Evolutionary Approaches

Exploiting the evolutionary conservation and spatial clustering of functional residues is a powerful approach for the identification of functional sites (38). A limitation of this approach is that it requires a sufficiently diverse set of sequences for meaningful conservation patterns, but the relentless growth in sequence data should increasingly widen its applicability.

The popular evolutionary trace (ET) method (56) and its variants (57,59,60,81), described in more detail in **Box 1**, use a phylogenetic tree derived from a multiple sequence alignment (MSA), and evolutionary conserved residues are then mapped onto a representative structure. Experimental confirmation of ET predictions (82–84) demonstrate the potential of this approach. A complementary MSA-dependent method conducts a vectorial analysis of sequence profiles to identify functional residues (55).

The relative performance of comparable methods is difficult to assess without independent analysis. The selection of criteria for the definition of a correct prediction is somewhat arbitrary (88), and authors choose different protein targets, as well as different datasets of functional residues, regarded as the “true positives,” to assess their predictions—e.g., interface residues in co-crystals (58), catalytic residues identified from the literature (68), SITE records in PDB files (57), and ACT_SITE residues in SWISS-PROT (89) files (67). What is apparent is that residue conservation, when available, improves performance, as illustrated by the testing of integrated methods both with and without these data (67,68). All methods have their strengths and limitations, and using them in combination is advised to obtain reliable consensus predictions. The ultimate test is mutagenesis studies to confirm or refute predictions.

7. Evolution of Protein Function From a Structural Perspective

Because evolution is heavily exploited in inferring function in genome sequencing and structural genomics projects, an understanding of the underlying mechanisms of evolving new functions through sequence and structural changes is vital. Todd et al. (27,90) conducted a detailed, collective analysis of 31 functionally diverse enzyme superfamilies in the PDB, assessing the conservation and variation of catalytic residues, substrate specificity, and reaction chemistry, as well as changes in domain composition and subunit assembly. Typically, functional analyses are done for a single family in isolation, but together, they provide a better insight into protein evolution by allowing the identification of preferred mechanisms of functional diversification. These 31 structural superfamilies, in combination with other modules, support over 200 protein functions. The principal findings of their work are summarized in **Fig. 6**. The analysis and descriptions in the text are restricted to the structural data, unless stated otherwise.

7.1. Substrate Specificity and Reaction Chemistry

The role that substrate specificity and catalytic activity play in enzyme recruitment has been a subject of discussion (91–95). Jensen (91) proposed that enzyme recruitment exploited the substrate ambiguity of ancestral proteins. Enzymes with improved selectivity have evolved through gene duplication and specialization of active site architectures. In more recent work (93–95), it has been proposed that chemistry, as opposed to substrate specificity, has dictated the choice of ancestral enzymes for the evolution of new activities, where chemistry refers to the strategy of transforming substrate into product and the nature of the intermediates involved.

Twenty-eight of the 31 domain superfamilies are directly involved in substrate binding, and in only one of these is the substrate absolutely conserved. Enzymes in six superfamilies bind to a common substrate type (e.g., DNA, sugars, phosphorylated proteins), but in three of these families, variations within these ligand types are extensive. In as many as 19 superfamilies, substrate specificity is completely diverse, in that the substrates bound vary in their size, chemical properties, and/or structural scaffolds (e.g., aromatic versus linear-chain hydrocarbons). If any similarity exists within the substrates, it is limited to a small chemical moiety (e.g., carbonyl group, peptide bond), typically the center of reactivity during catalysis.

These findings illustrate the plasticity of folds with respect to ligand binding. Commonly, substrates are bound by surface loops, which can evolve and adapt while the structural core is maintained. For example, in some TIM barrel superfamilies, substrate differences are accommodated by long loop excursions at the carboxyl termini of the central β -strands.

Far more common than the conservation of substrate binding is the conservation of reaction chemistry, in support of the “chemistry first strategy.” There are 28 superfamilies involved directly in catalysis and for which some details of catalysis are known, and in two, the reaction chemistry is conserved. In a further 21, the chemistry is “semi-conserved,” in that enzyme members utilize a common chemical strategy in the contexts of different overall transformations. Typically, it is the initial catalytic step that is conserved, while the reaction paths that follow vary, sometimes extensively. In the enolase superfamily, for example, the initial step invariably involves metal-assisted

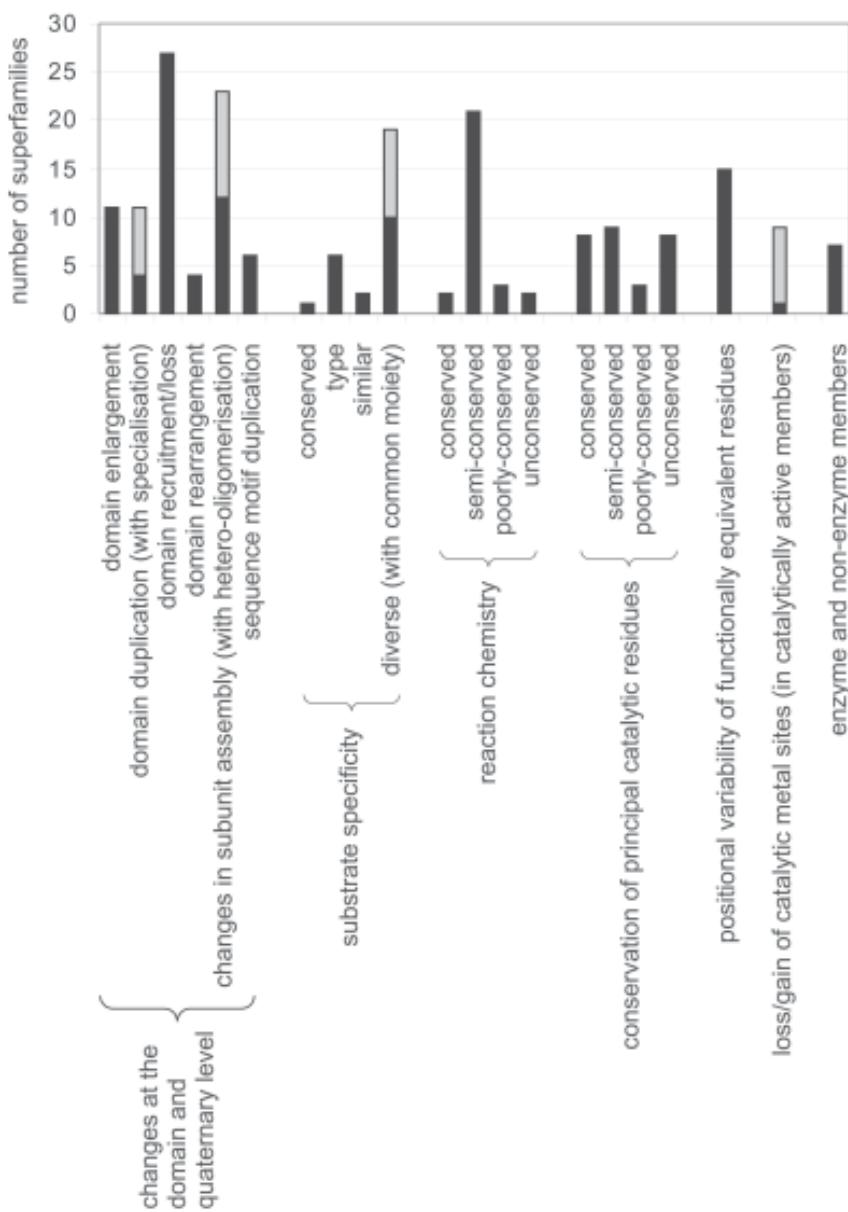


Fig. 6. Summary of the frequency with which particular types of changes occur, as derived from the structural data only, and the extent of conservation of functional attributes within the 31 superfamilies analyzed (27,90). Parts of histogram bars in gray correspond to the *x*-axis label in brackets (e.g., substrate specificity is diverse in 19 superfamilies, but in 9 of these superfamilies the substrates all share a common chemical moiety).

general base-catalyzed removal of a proton α to a carboxylate (96). Overall, activities include racemization, epimerization, cycloisomerization, and β -elimination of water.

In three superfamilies the chemistry is “poorly conserved,” in that just one small aspect of the reactions catalyzed is conserved, such as a common intermediate. The

crotonase-like superfamily includes structural and sequence members having hydratase, dehydratase, dehalogenase, isomerase, esterase, decarboxylase, and peptidase activities. Stabilization of an oxyanion by a conserved oxyanion hole, in which the intermediate is hydrogen-bonded to two backbone amide groups, is the only functional similarity conserved across all members, and in contrast to many of the “semi-conserved” superfamilies, the reaction paths to this intermediate vary widely (94) (see **Fig. 7**). They include peptide bond addition, proton abstraction, and nucleophilic aromatic addition, and involve different residues in the active site. In two superfamilies, reaction chemistry is completely dissimilar in at least one pair of enzymes, and in one of these, substrates are diverse too.

7.2. Catalytic Residues

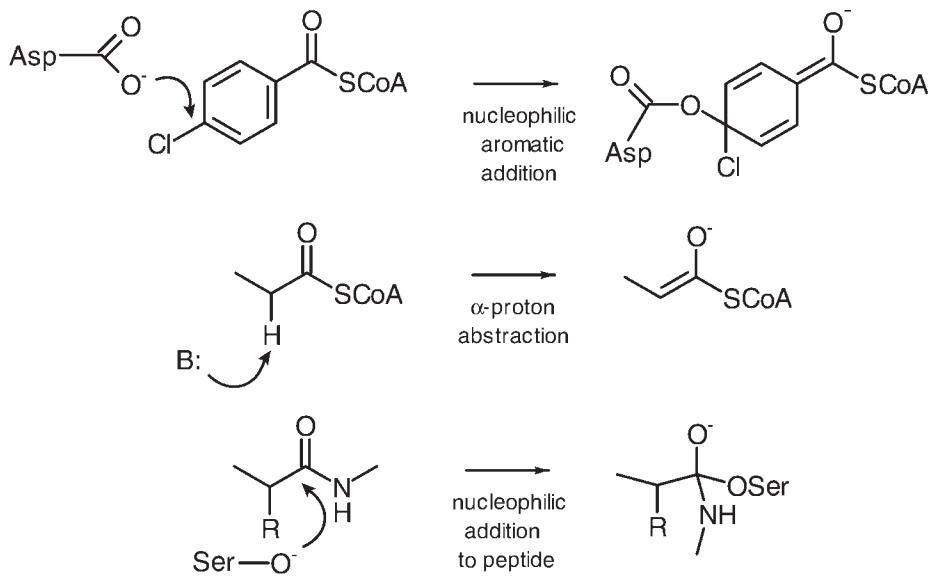
The expectation is that a conserved functional feature is associated with one or more key residues that are invariant across a family of proteins; owing to their important functional role, these residues are subject to evolutionary constraints, and their loss during evolution would be deleterious to function. However, in as many as one-half of the superfamilies analyzed, there is poor positional conservation of residues that play equivalent catalytic roles in related proteins (90,97). The reverse scenario, the use of the same active site framework to catalyze diverse activities, is also observed—e.g., Ser-His-Asp triad in the α/β hydrolase superfamily.

Two examples are illustrated in **Fig. 8**. The glutathione-S-transferases (GST) have been classified into several subfamilies on the basis of sequence and substrate specificity. Members of most classes have an essential Tyr residue involved in glutathione (GSH) activation. This is present in a few θ -class GSTs, but it is not part of the active site (98). Instead, members of this class have a Ser for the same job, located four positions downstream. Nevertheless, the relative dispositions of the stabilizing hydroxyl group and GSH thiolate anion are similar (see **Fig. 8A**).

For most superfamilies, the active site pockets of family members are co-located, despite the residue variability. The ferritins are an exception (see **Fig. 8B**). They have ferroxidase activity associated with a di-iron site, and in classical ferritins this site is located within a four-helix bundle core. *Listeria innocua* ferritin is unusual in that the site is situated at the interface between two subunits related by two-fold symmetry. Nevertheless, the two di-iron centers are chemically and structurally similar (99). In the pyridine nucleotide disulfide oxidoreductase superfamily, members use similar sulfur redox chemistries in one of four distinct active sites (97).

Several possible origins of this variability have been suggested (97). It may reflect evolutionary optimization of the catalytic efficiency of these enzymes. As new groups fortuitously evolve in the active site, they might take over the roles of “old” residues if they are better suited to the job. Alternatively, this variability may indicate that the enzyme functions in question have evolved completely independently. That is, nature has devised two or more distinct solutions to the same catalytic problem within homologous folds.

Other active site “anomalies” besides those discussed here are also observed. For example, in eight superfamilies, catalytically active proteins differ in their metal cofactor requirements, in terms of number of metal ions, and in four of these, some enzyme members lack a metal cofactor site altogether.



Sequence relatives:

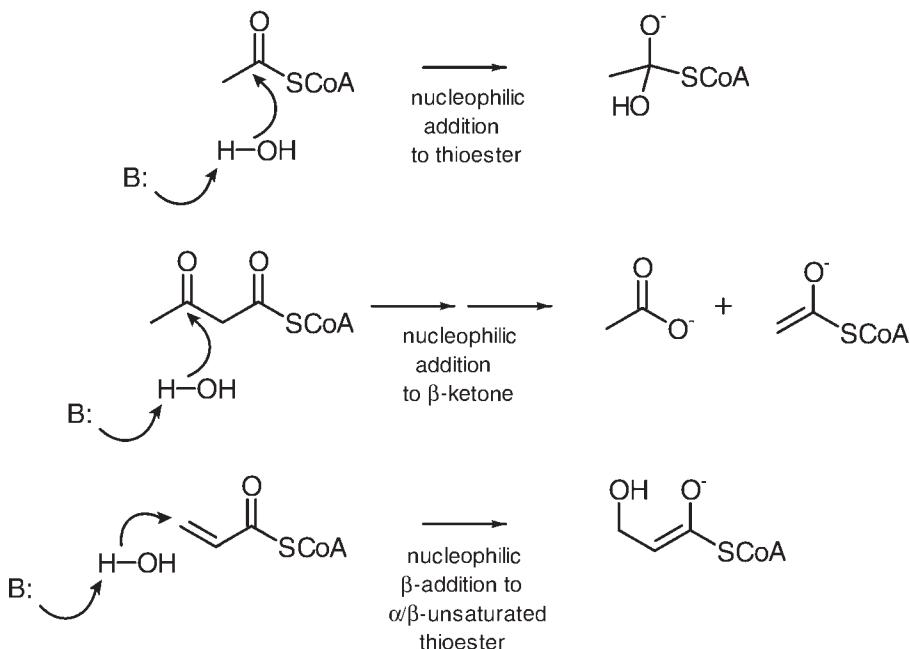


Fig. 7. Initial reaction steps catalyzed by members of the crotonase-like superfamily. These reaction steps are varied, but all members stabilize an oxyanion intermediate using a structurally conserved oxyanion hole (94). B denotes an enzyme base. Initial steps include nucleophilic aromatic addition (4-chlorobenzoyl-CoA dehalogenase), α -proton abstraction (enoyl-CoA hydratase, $\Delta 3,5\text{-}\Delta 2,4$ -dienoyl-CoA isomerase), and nucleophilic addition to a peptide bond (Clp protease). The inclusion of sequence relatives expands the repertoire of reaction steps catalyzed, and steps include nucleophilic addition to a thioester, nucleophilic addition to a β -ketone, and β -addition to an α/β -unsaturated thioester.

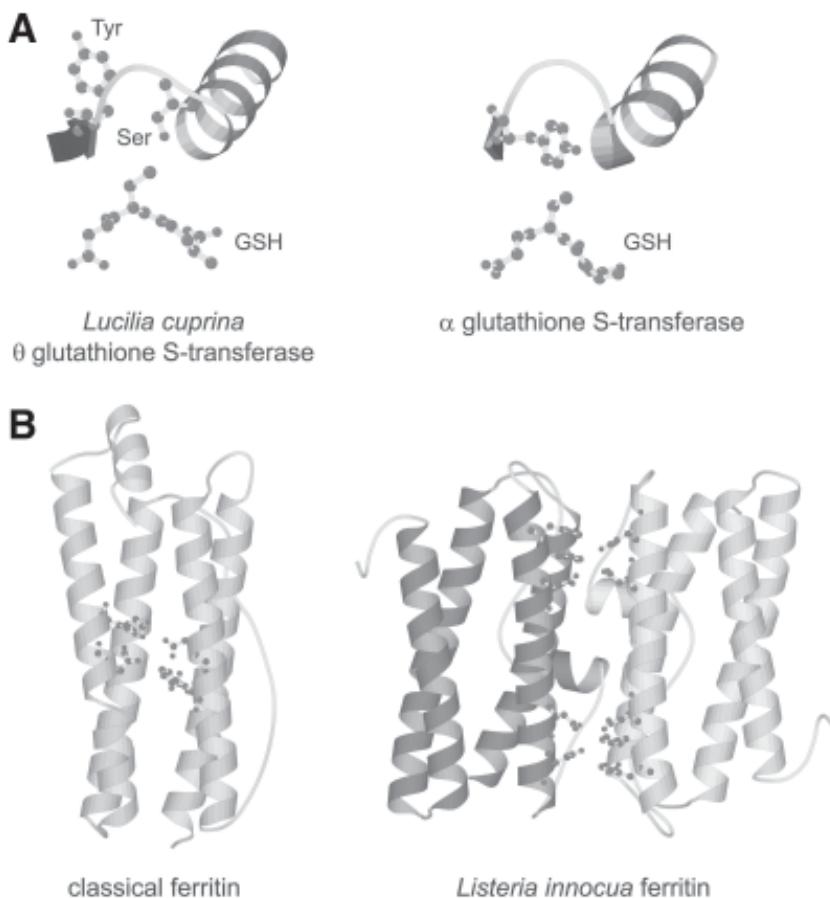


Fig. 8. Positional variability of functionally equivalent residues. (A) Glutathione S-transferases (GST) in the θ class use a Ser for glutathione (GSH) activation, but members of other subfamilies, such as the α class, use a nonequivalent Tyr residue. In *Lucilia cuprina* θ GST, and in a few other θ GSTs, the same Tyr is present, but it is not part of the active site (98). (B) In classical ferritins, the ferroxidase di-iron site is located in the core of a four-helix bundle, but in *Listeria innocua* ferritin it is situated at the interface between two subunits (shown in different shades of gray) (99). Residues involved in iron-binding are shown in ball-and-stick representation.

7.3. Domain Enlargement, Domain Organization, and Subunit Assembly

Holm and Sander (100) introduced the concept of the minimal structural and functional core of related proteins. Provided the key structural elements of the functional site remain intact, considerable deviations around this core are permissible. In some superfamilies, there are small differences in topological connectivities, and in a third, the size of the domain(s) common to all family members varies by at least twofold. Size differences may be attributed to the addition/loss of subdomains, changes to the core such as β -sheet extension, and/or variations in loop length. These deviations usually influence substrate preference and oligomeric contacts. In the evolution of organism complexity and functional specialization, the embellishment of simple folds by

residue and intron insertions is probably more common than the reverse scenario, domain reduction.

In most superfamilies the domain organization varies between members, illustrating the importance of domain accretion as a route to new functions (see **Fig. 9**). Modules attached to catalytic domains may play a role in oligomerization, cofactor dependency, regulation, subcellular targeting or, commonly, substrate specificity. In one-third of superfamilies, one or more members function as hetero-oligomers, interacting with different partners, while others in the same family function simply as monomers or homo-oligomers.

7.4. Summary

This section has briefly illustrated the extent and the mechanisms by which proteins evolve new functions. Some superfamilies have accommodated diverse functions on a common scaffold by a broad variety of routes, while others have achieved their diversity through incremental mutations in the active site alone, the domain and subunit content remaining constant. These diverse changes were observed in the limited structural data available. With the growth in the PDB, we are likely to observe even more extensive and unexpected variations in structure and function in these superfamilies. Indeed, with the inclusion of sequence data, the number of functions attributed to these families triples to over 600.

8. Structural Genomics

With the recent technological revolution in all aspects of structure determination, protein structures are now solved at an unprecedented rate, making viable the structural genomics projects that depend upon high-throughput structure determination on a genome-wide scale. Outlined in this section are several real-life examples of structure-driven functional analysis, which have been met with mixed degrees of success. Examples have been categorized according to the relationship, if any, the target protein shared with others of known fold when its structure was solved. Clearly structural information complements experimental data, and these together help to assign function (101).

8.1. Homologous Relationship

BioH is involved in biotin biosynthesis in *Escherichia coli*, but its biochemical function was unknown until its structure was determined (102). The structure revealed an evolutionary relationship with members of the α/β hydrolase superfamily. It was scanned against a database of three-dimensional active site templates using TESS (44), and this search identified the Ser-His-Asp catalytic triad, common to various related and unrelated hydrolases. BioH was subjected to several hydrolase assays, and it demonstrated carboxylesterase activity with a preference for short acyl-chain substrates (102). This example illustrates the importance of automated methods to help assign function when the experimentalists responsible for solving a particular hypothetical target are not familiar with the area of biology in question, a very common scenario in structural genomics.

Structure determination of MJ0577 from *Methanococcus jannaschii* illustrates the usefulness of bound ligands to infer function (18). The significant structural simili-

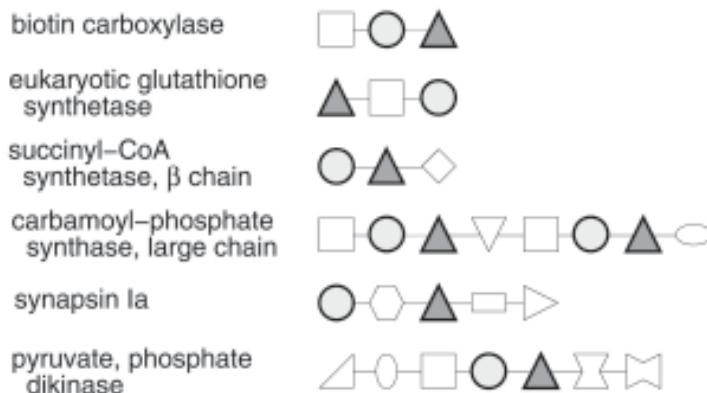


Fig. 9. Domain organization of members of the ATP-dependent carboxylate-amine/thiol ligase superfamily. Each shape corresponds to a structural domain, and domains of the same shape are homologous. The two domains that together define the ATP-binding fold of this superfamily are highlighted in gray. In eukaryotic glutathione synthetase, these domains occur in a different order, owing to a genetic permutation event. In carbamoyl-phosphate synthase, there are two copies of the ATP-binding fold, and each copy has a distinct role in the reaction catalyzed. This fold has been combined with a wide variety of other domains, including the ubiquitous Rossmann and TIM barrel folds, leading to the broad functional diversity of this superfamily.

ties MJ0577 shares with several proteins in the PDB provided no functional clues, while the fortuitous binding of ATP, scavenged from its *Escherichia coli* host during over-expression, suggested the protein hydrolyzes ATP. Biochemical experiments showed that MJ0577 can hydrolyze ATP in the presence of *Methanococcus jannaschii* cell extract, but not by itself, and so the protein probably functions as an ATP-mediated molecular switch (18). The ATP-binding pocket in MJ0577 contains some motifs commonly found in nucleotide-binding proteins, but they had not been detected from sequence owing to their different sequential arrangement (18).

YjgF from *Escherichia coli* belongs to an ubiquitous family of hypothetical proteins. Its structure revealed a distant relationship with chorismate mutase, with which it shares a similar fold and the same quaternary structure (103). Invariant residues in the YjgF family cluster in three cavities in the homotrimer, and these putative active sites are in the same relative locations as those in chorismate mutase. However, they are completely unrelated in amino acid composition, so these proteins are likely to have different functions. Interestingly, while YjgF and protein tyrosine phosphatases share no global structural similarity, their active sites have several common features. YjgF shows no detectable phosphatase activity, however. Although the structure-to-function approach failed in this instance, the YjgF structure has guided subsequent mutagenesis studies.

8.2. Fold Similarity

Structure determination of MJ0266 did not uncover any distant homologies (104). Nevertheless, the structure is an elaborated version of a known fold, with structural

comparisons revealing similarities to the anticodon-binding domains of two tRNA synthetases and to the nucleotide-binding domains of other proteins. Nucleotide-binding assays revealed the protein's ability to hydrolyze nucleotide triphosphates to mono-phosphates, and xanthine triphosphate was the best substrate. These results indicated strongly that MJ0266 is a pyrophosphate-releasing XTPase (104).

The example of MTH538 from *Methanobacterium thermoautotrophicum* illustrates the difficulty in assigning function to a protein with a common fold (105). Furthermore, since MTH538 lacks close sequence homologs, putative functional residues could not be identified. The structure revealed similarities to functionally diverse proteins in ten superfamilies in SCOP having the flavodoxin fold. It was not clear whether MTH538 has a function similar to one of its fold analogs, or one that is distinct. It was most similar to the flavodoxins and to the receiver domains of two-component response regulator systems, such as CheY. Tests on MTH538 for characteristic activities of these proteins, including flavin binding and phosphorylation, were negative. However, the compelling similarity to CheY may indicate a role in signal transduction as a phosphorylation-independent conformational switch (105).

8.3. Novel Fold

The structure of YrdC from *E. coli* revealed a novel fold (106). The size, curvature and positive electrostatic potential of a concave depression on the surface, and a cluster of conserved basic residues at the floor of the cavity, suggested a nucleic acid-binding function. Tests of binding affinities for single- and double-stranded RNA and DNA, and tRNAs, demonstrated a preference for dsRNA, suggesting a role in translation (106).

9. Outlook

As the structural genomics initiatives have gained pace, the prediction of function from structure has become an increasingly pressing issue. If structure-based functional annotation is to proceed in a high-throughput way to keep pace with the flood of structural data, automated approaches are crucial. Currently, functional site prediction is a very active research area. While the methods available at present only occasionally provide insights into the specific biochemical activity of a protein, they are invaluable in focusing mutagenesis studies, assays, protein engineering, and drug design to the relevant site. Encouragingly, it is expected that the systematic targeting of those families without structural data will expand the current inventory of known protein folds, or will reveal many new and exciting discoveries of remote evolutionary kinships. These structures, in time, will establish novel protein structure-function relationships, and thereby improve the scope and success of knowledge-based approaches to predict function from structure.

For many hypothetical proteins, their interacting partners, if any, are unknown, and this hampers the assignment of function (105). If the structural genomics projects are to achieve their full potential, an integrated approach to assign function, which includes high-throughput proteomics methods and “non-homology” bioinformatics approaches (107) to identify interacting partners, is essential. These data complement structure-derived aspects of biochemical activity, and taken together, these two types of data help to assign function at all levels.

References

1. Burley, S. K., Almo, S. C., Bonanno, J. B., et al. (1999) Structural genomics: beyond the human genome project. *Nature Genet.* **23**, 151–157.
2. Brenner, S. E. (2001) A tour of structural genomics. *Nature Rev. Genet.* **2**, 801–809.
3. Shapiro, L. and Harris, T. (2000) Finding function through structural genomics. *Curr. Opin. Biotech.* **11**, 31–35.
4. Chothia, C. and Lesk, A. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823–826.
5. Ollis, D. L., Cheah, E., Cygler, M., et al. (1992) The α/β hydrolase fold. *Protein Eng.* **5**, 197–211.
6. Hegyi, H. and Gerstein, M. (1999) The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J. Mol. Biol.* **288**, 147–164.
7. Teichmann, S. A., Park, J., and Chothia, C. (1998) Structural assignments to the proteins of *Mycoplasma genitalium* show that they have been formed by extensive gene duplications and domain rearrangements. *Proc. Natl. Acad. Sci. USA* **95**, 14,658–14,663.
8. Teichmann, S. A., Chothia, C., and Gerstein, M. (1999) Advances in structural genomics. *Curr. Opin. Struct. Biol.* **9**, 390–399.
9. Berman, H. M., Westbrook, J., Feng, Z., et al. (2000) The Protein Data Bank. *Nucl. Acids Res.* **28**, 235–242.
10. International Human Genome Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.
11. Wistow, G. J., Mulders, J. W. M., and Dejong, W. (1987) The enzyme lactate dehydrogenase as a structural protein in avian and crocodilian lenses. *Nature* **326**, 622–624.
12. Jeffery, C. (1999) Moonlighting proteins. *Trends Biochem. Sci.* **24**, 8–11.
13. Mondrek, B. and Lee, C. (2002) A genomic view of alternative splicing. *Nat. Genet.* **30**, 13–19.
14. Kriventseva, E. V., Koch, I., Apweiler, R., et al. (2003) Increase of functional diversity by alternative splicing. *Trends Genet.* **19**, 124–128.
15. Wright, P. E. and Dyson, J. (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.* **293**, 321–331.
16. Liu, J., Tan, H., and Rost, B. (2002) Loopy proteins appear conserved in evolution. *J. Mol. Biol.* **322**, 53–64.
17. Postigl, H., Henrick, K., and Thornton, J. M. (2000) Discriminating between homodimeric and monomeric proteins in the crystalline state. *Proteins: Struct. Func. Genet.* **41**, 47–57.
18. Zarembinski, T. I., Hung, L.-W., Mueller-Dieckmann, H.-J., et al. (1998) Structure-based assignment of the biochemical function of a hypothetical protein: a test case for structural genomics. *Proc. Natl. Acad. Sci. USA* **95**, 15189–15193.
19. Martin, A. C. R., Orengo, C. A., Hutchinson, E. G., et al. (1998) Protein folds and functions. *Structure* **6**, 875–884.
20. Anantharaman, V., Aravind, L., and Koonin, E. V. (2003) Emergence of diverse biochemical activities in evolutionary conserved structural scaffolds of proteins. *Curr. Opin. Chem. Biol.* **7**, 12–20.
21. Thornton, J. M., Todd, A. E., Milburn, D., Borkakoti, N., and Orengo, C. A. (2000) From structure to function: approaches and limitations. *Nat. Struct. Biol.* **7**, 991–994.
22. Moult, J. and Melamud, E. (2000) From fold to function. *Curr. Opin. Struct. Biol.* **10**, 384–389.
23. Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., and Thornton, J. M. (1997) CATH - a hierachic classification of protein domain structures. *Structure* **5**, 1093–1108.

24. Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995) SCOP - A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
25. Wilson, C. A., Krychman, J., and Gerstein, M. (2000) Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J. Mol. Biol.*, **297**, 233–249.
26. Devos, D. and Valencia, A. (2000) Practical limits of function prediction. *Proteins: Struct. Func. Genet.*, **41**, 98–107.
27. Todd, A. E., Orengo, C. A., and Thornton, J. M. (2001) Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.*, **307**, 1113–1143.
28. Rost, B. (2002) Function less conserved than anticipated. *J. Mol. Biol.*, **318**, 595–608.
29. Grishin, N. V. (2001) Fold change in evolution of protein structures. *J. Struct. Biol.*, **134**, 167–185.
30. Chothia, C. (1992) One thousand families for the molecular biologist. *Nature*, **357**, 543–544.
31. Orengo, C. A., Jones, D. T., and Thornton, J. M. (1994) Protein superfamilies and domain superfolds. *Nature*, **372**, 631–634.
32. Nagano, N., Orengo, C. A., and Thornton, J. M. (2002) One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *J. Mol. Biol.*, **321**, 741–765.
33. Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB) (1992) *Enzyme Nomenclature*, Academic Press, New York, NY.
34. Russell, R. B., Sasieni, P. D., and Sternberg, M. J. E. (1998) Supersites within superfolds. Binding site similarity in the absence of homology. *J. Mol. Biol.*, **282**, 903–918.
35. Wallace, A. C., Laskowski, R. A., and Thornton, J. M. (1996) Derivation of 3D coordinate templates for searching structural databases: application to the Ser-His-Asp catalytic triads of the serine proteinases and lipases. *Protein Sci.*, **5**, 1001–1013.
36. Bartlett, G. J., Porter, C. T., Borkakorti, N., and Thornton, J. M. (2002) Analysis of catalytic residues in enzyme active sites. *J. Mol. Biol.*, **324**, 105–121.
37. Orengo, C. A., Todd, A. E., and Thornton, J. M. (1999) From protein structure to function. *Curr. Opin. Struct. Biol.*, **9**, 374–382.
38. Lichtarge, O. and Sowa, M. E. (2002) Evolutionary predictions of binding surfaces and interactions. *Curr. Opin. Struct. Biol.*, **12**, 21–27.
39. Sottriffer, C. and Klebe, G. (2002) Identification and mapping of small-molecule binding sites in proteins: computational tools for structure-based drug design. *Il Farmaco*, **57**, 243–251.
40. Campbell, S. J., Gold, N. D., Jackson, R. M., and Westhead, D. R. (2003) Ligand binding: functional site location, similarity and docking. *Curr. Opin. Struct. Biol.*, **13**, 389–395.
41. Kinoshita, K. and Nakamura, H. (2003) Protein informatics towards function identification. *Curr. Opin. Struct. Biol.*, **13**, 396–400.
42. Artymiuk, P. J., Poirrette, A. R., Grindley, H. M., Rice, D. W., and Willett, P. (1994) A graph-theoretic approach to the identification of 3-dimensional patterns of amino-acid side-chains in protein structures. *J. Mol. Biol.*, **243**, 327–344.
43. Spriggs, R. V., Artymiuk, P. J., and Willett, P. (2003) Searching for patterns of amino acids in 3D protein structures. *J. Chem. Inf. Comp. Sci.*, **43**, 412–421.
44. Wallace, A. C., Borkakorti, N., and Thornton, J. M. (1997) TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases: application to enzyme active-sites. *Protein Sci.*, **6**, 2308–2323.
45. Fetrow, J. S. and Skolnick, J. (1998) Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T₁ ribonucleases. *J. Mol. Biol.*, **281**, 949–968.

46. Russell, R. B. (1998) Identification of protein three-dimensional side-chain patterns: new examples of convergent evolution. *J. Mol. Biol.* **279**, 1211–1227.
47. Kleywegt, G. (1999) Recognition of spatial motifs in protein structures. *J. Mol. Biol.* **285**, 1887–1897.
48. Jonassen, I., Eidhammer, I., and Taylor, W. R. (1999) Discovery of local packing motifs in protein structures. *Proteins* **34**, 206–219.
49. Oldfield, T. J. (2002) Data mining the protein data bank: residue interactions. *Proteins: Struct. Func. Genet.* **49**, 510–528.
50. Wangikar, P. P., Tendulkar, A. V., Ramya, S., Mali, D. N., and Sarawagi, S. (2003) Functional sites in protein families uncovered via an objective and automated graph theoretic approach. *J. Mol. Biol.* **326**, 955–978.
51. Hamelryck, T. (2003) Efficient identification of side-chain patterns using a multidimensional index tree. *Proteins: Struct. Func. Genet.* **51**, 96–108.
52. Zhao, S., Morris, G. M., Olson, A. J., and Goodsell, D. S. (2001) Recognition templates for predicting adenylate-binding sites in proteins. *J. Mol. Biol.* **314**, 1245–1255.
53. Schmitt, S., Kuhn, D., and Klebe, G. (2002) A new method to detect related function among proteins independent of sequence and fold homology. *J. Mol. Biol.* **323**, 387–406.
54. Kinoshita, K., Furui, J., and Nakamura, H. (2001) Identification of protein functions from a molecular surface database, eF-site. *J. Struct. Func. Genomics* **2**, 9–22.
55. Casari, G., Sander, C., and Valencia, A. (1995) A method to predict functional residues in proteins. *Nat. Struct. Biol.* **2**, 171–178.
56. Lichtarge, O., Bourne, H. R., and Cohen, F. E. (1996) An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257**, 342–358.
57. Aloy, P., Querol, E., Aviles, F. X., and Sternberg, M. J. E. (2001) Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *J. Mol. Biol.* **311**, 395–408.
58. Landgraf, R., Xenarios, I., and Eisenberg, D. (2001) Three-dimensional clustering analysis identifies interfaces and functional residue clusters in proteins. *J. Mol. Biol.* **307**, 1487–1502.
59. Armon, A., Graur, D., and Ben-Tal, N. (2001) ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J. Mol. Biol.* **307**, 447–463.
60. Pupko, T., Bell, R. E., Mayrose, I., Glaser, F., and Ben-Tal, N. (2002) Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics* **18 Suppl.**, S71–S77.
61. Elcock, A. H. (2001) Prediction of functionally important residues based solely on the computed energetics of protein structure. *J. Mol. Biol.* **312**, 885–896.
62. Ondrechen, M. J., Clifton, J. G., and Ringe, D. (2001) THEMATICS: a simple computational predictor of enzyme function from structure. *Proc. Natl. Acad. Sci. USA* **98**, 12,473–12,478.
63. Laskowski, R. A. (1995) SURFNET: a program for visualising molecular surfaces, cavities and intermolecular interactions. *J. Mol. Graph.* **13**, 323–330.
64. Peters, K. P., Fauck, J., and Frommel, C. (1996) The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. *J. Mol. Biol.* **256**, 201–213.
65. Hendlich, M., Rippmann, F., and Barnickel, G. (1997) LIGSITE: automatic and efficient detection of potential small-molecule binding sites in proteins. *J. Mol. Graph. Model.* **15**, 359–363.

66. Brady Jr, G. P. and Stouten, P. F. W. (2000) Fast prediction and visualization of protein binding pockets with PASS. *J. Comput. Aided Mol. Des.* **14**, 383–401.
67. Ota, M., Kinoshita, K., and Nishikawa, K. (2003) Prediction of catalytic residues in enzymes based on known tertiary structure, stability profile, and sequence conservation. *J. Mol. Biol.* **327**, 1053–1064.
68. Gutteridge, A., Bartlett, G. J., and Thornton, J. M. (2003) Using a neural network and spatial clustering to predict the location of active sites in enzymes. *J. Mol. Biol.* **330**, 719–734.
69. Dobson, P. D. and Doig, A. J. (2003) Distinguishing enzyme structures from non-enzymes without alignments. *J. Mol. Biol.* **330**, 771–783.
70. Laskowski, R. A., Luscombe, N. M., Swindells, M. B., and Thornton, J. M. (1996) Protein clefts in molecular recognition and function. *Protein Sci.* **5**, 2438–2452.
71. Warshel, A. (1978) Energetics of enzyme catalysis. *Proc. Natl. Acad. Sci. USA* **75**, 5250–5254.
72. Beadle, B. M. and Schoichet, B. K. (2002) Structural bases of stability-function tradeoffs in enzymes. *J. Mol. Biol.* **321**, 285–296.
73. Jones, S. and Thornton, J. M. (1997) Analysis of protein-protein interaction sites using surface patches. *J. Mol. Biol.* **272**, 121–132.
74. Lo Conte, L., Chothia, C., and Janin, J. (1999) The atomic structure of protein-protein recognition sites. *J. Mol. Biol.* **285**, 2177–2198.
75. Nooren, I. M. A. and Thornton, J. M. (2003) Structural characterization and functional significance of transient protein-protein interactions. *J. Mol. Biol.* **325**, 991–1018.
76. Nooren, I. M. A. and Thornton, J. M. (2003) Diversity of protein-protein interactions. *EMBO J.* **22**, 3486–3492.
77. Ofran, Y. and Rost, B. (2003) Analysing six types of protein-protein interfaces. *J. Mol. Biol.* **325**, 377–387.
78. Jones, S. and Thornton, J. M. (1997) Prediction of protein-protein interaction sites using patch analysis. *J. Mol. Biol.* **272**, 133–143.
79. Stark, A., Sunyaev, S., and Russell, R. B. (2003) A model for statistical significance of local similarities in structure. *J. Mol. Biol.* **326**, 1307–1316.
80. Barker, J. A. and Thornton, J. M. (2003) An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis. *Bioinformatics* **19**, 1644–1649.
81. Madabushi, S., Yao, H., Marsh, M., et al. (2002) Structural clusters of evolutionary trace residues are statistically significant and common in proteins. *J. Mol. Biol.* **316**, 139–154.
82. Sowa, M. E., He, W., Wensel, T. G., and Lichtarge, O. (2000) A regulator of G protein signaling interaction surface linked to effector specificity. *Proc. Natl. Acad. Sci. USA* **97**, 1483–1488.
83. Sowa, M. E., He, W., Slep, K. C., Kercher, M. A., Lichtarge, O., and Wensel, T. G. (2001) Prediction and confirmation of a site critical for effector regulation of RGS domain activity. *Nat. Struct. Biol.* **8**, 234–237.
84. Slep, K. C., Kercher, M. A., He, W., Cowan, C. W., Wensel, T. G., and Sigler, P. B. (2001) Structural determinants for regulation of phosphodiesterase by a G protein at 2.0 Å. *Nature* **409**, 1071–1077.
85. Porter, C. T., Bartlett, G. J., and Thornton, J. M. (2004) The Catalytic Site Atlas (CSA): a resource of catalytic sites and residues identified in enzymes using structural data. *Nucl. Acids Res.* **32**, D129–D133.
86. Fetrow, J. S., Siew, N., Di Gennaro, J. A., Martinez-Yamount, M., Dyson, H. J., and Skolnick, J. (2001) Genomic-scale comparison of sequence- and structure-based meth-

- ods of function prediction: does structure provide additional insight? *Protein Sci.* **10**, 1005–1014.
87. Stark, A. and Russell, R. B. (2003) Annotation in three dimensions. PINTS: patterns in non-homologous tertiary structures. *Nucl. Acids Res.* **31**, 3341–3344.
 88. Yao, H., Kristensen, D. M., Mihalek, I., et al. (2003) An accurate, sensitive, and scalable method to identify functional sites in protein structures. *J. Mol. Biol.* **326**, 255–261.
 89. Boeckmann, B., Bairoch, A., Apweiler, R., et al. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucl. Acids Res.* **31**, 365–370.
 90. Todd, A. E. (2001) *Evolution of function in protein superfamilies*. PhD thesis, University College London.
 91. Jensen, R. A. (1976) Enzyme recruitment in evolution of new function. *Annu. Rev. Microbiol.* **30**, 409–425.
 92. O'Brien, P. J. and Herschlag, D. (1999) Catalytic promiscuity and the evolution of new enzymatic activities. *Chem. Biol.* **6**, r91–r105.
 93. Petsko, G. A., Kenyon, G. L., Gerlt, J. A., Ringe, D., and Kozarich, J. W. (1993) On the origin of enzymatic species. *Trends Biochem. Sci.* **18**, 372–376.
 94. Babbitt, P. C. and Gerlt, J. A. (1997) Understanding enzyme superfamilies—chemistry as the fundamental determinant in the evolution of new catalytic activities. *J. Biol. Chem.* **272**, 30,591–30,594.
 95. Gerlt, J. A. and Babbitt, P. C. (1998) Mechanistically diverse enzyme superfamilies: the importance of chemistry in the evolution of catalysis. *Curr. Opin. Chem. Biol.* **2**, 607–612.
 96. Babbitt, P. C., Hasson, M. S., Wedekind, J. E., et al. (1996) The enolase superfamily: a general strategy for enzyme-catalysed abstraction of the α -protons of carboxylic acids. *Biochemistry* **35**, 16,489–16,501.
 97. Todd, A. E., Orengo, C. A., and Thornton, J. M. (2002) Plasticity of enzyme active sites. *Trends Biochem. Sci.* **27**, 419–426.
 98. Wilce, M. C. J., Board, P. G., Feil, S. C., and Parker, M. W. (1995) Crystal structure of a theta-class glutathione transferase. *EMBO J.* **14**, 2133–2143.
 99. Ilari, A., Stefanini, S., Chiancone, E., and Tsernoglou, D. (2000) The dodecameric ferritin from *L. innocua* contains a novel intersubunit iron-binding site. *Nat. Struct. Biol.* **7**, 38–43.
 100. Holm, L. and Sander, C. (1997) New structure—novel fold? *Structure* **5**, 165–171.
 101. Zhang, C. and Kim, S.-H. (2003) Overview of structural genomics: from structure to function. *Curr. Opin. Chem. Biol.* **7**, 28–32.
 102. Sanishvili, R., Yakunin, A. F., Laskowski, R. A., et al. (2003) Integrating structure, bioinformatics and enzymology to discover function. *J. Biol. Chem.* **278**, 26,039–26,045.
 103. Volz, K. (1999) A test case for structure-based functional assignment: the 1.2 Å crystal structure of the yjgF gene product from *Escherichia coli*. *Protein Sci.* **8**, 2428–2437.
 104. Hwang, K. Y., Chung, J. H., Kim, S.-H., Han, Y. S., and Cho, Y. (1999) Structure-based identification of a novel NTPase from *Methanococcus jannaschii*. *Nat. Struct. Biol.* **6**, 691–696.
 105. Cort, J. R., Yee, A., Edwards, A. M., Arrowsmith, C. H., and Kennedy, M. A. (2000) Structure-based functional classification of hypothetical protein MTH538 from *Methanobacterium thermoautotrophicum*. *J. Mol. Biol.* **302**, 189–203.
 106. Teplova, M., Tereshko, V., Sanishvili, R., et al. (2000) The structure of the yrdC gene product from *Escherichia coli* reveals a new fold and suggests a role in RNA binding. *Protein Sci.* **9**, 2557–2566.
 107. Marcotte, E. M. (2000) Computational genetics: finding protein function by non-homology methods. *Curr. Opin. Struct. Biol.* **10**, 359–365.
 108. Artymiuk, P. J., Poirrette, A. R., Rice, D. W., and Willett, P. (1997) A polymerase I palm in adenylyl cyclase? *Nature* **388**, 33–34.

109. Bryant, S. H., Janin, J., Liu, Y., Ruoho, A. E., Zhang, G., and Hurley, J. H. (1997) A polymerase I palm in adenylyl cyclase?—reply. *Nature* **388**, 34.
110. Tesmer, J. J. T., Sunahara, R. K., Gilman, A. G., and Sprang, S. R. (1997) Crystal structure of the catalytic domains of adenylyl cyclase in a complex with G_{Sα}-GTPγS. *Science* **278**, 1907–1916.
111. Taylor, W. R. and Orengo, C. A. (1989) Protein structure alignment. *J. Mol. Biol.* **208**, 1–22.

Comparative Protein Structure Modeling

M. S. Madhusudhan, Marc A. Marti-Renom, Narayanan Eswar, Bino John, Ursula Pieper, Rachel Karchin, Min-Yi Shen, and Andrej Sali

1. Introduction

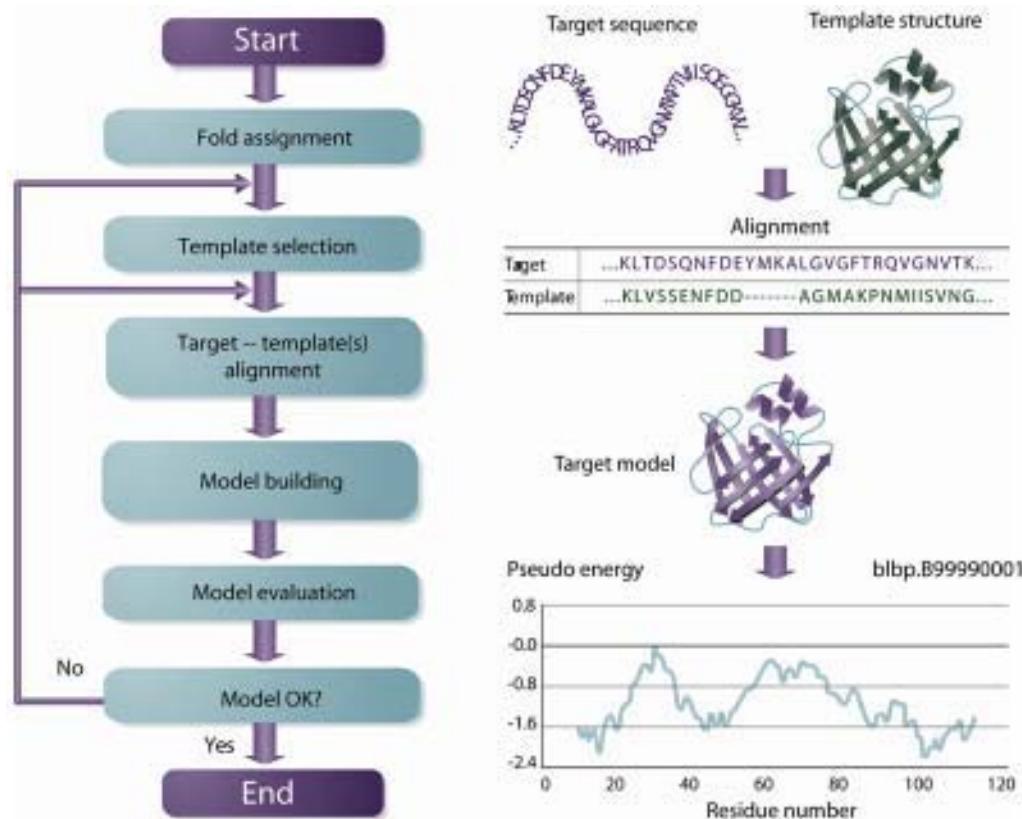
Three-dimensional protein structures are invaluable sources of information for the functional annotation of protein molecules. These structures are best determined by experimental methods such as X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy. However, the experimental methods cannot always be applied. In such cases, prediction of the protein structure by computational methods can frequently result in a useful model.

Protein structures can be modeled either *ab initio* from sequence alone or by comparative methods that rely on a database of known protein structures (1,2). *Ab initio* methods are largely based on the laws of physics, while comparative methods, including comparative (or homology) modeling and threading, are based primarily on statistical learning. Although there have been significant improvements in the *ab initio* (3) and threading methods (4), comparative modeling gives the most accurate results if a known protein structure that is sufficiently similar to the modeled sequence is available (1).

To predict protein structure by comparative modeling, two conditions have to be met (5,6). First, the sequence to be modeled (i.e., the target sequence) must have detectable similarity to another sequence of known structure (i.e., the template). Second, it must be possible to compute an accurate alignment between the target sequence and the template structure. The whole prediction process consists of fold assignment, target-template alignment, model building, and model evaluation (Fig. 1).

A simple predictor of the overall model accuracy is the degree of sequence similarity between the target and the template (Fig. 2). The higher is the sequence similarity to the template, the more accurate is the model. Although high-accuracy models are most informative, low-accuracy models may also provide coarse structural and functional annotation (Fig. 3) (1).

Comparative models can currently be built for domains in approx 57% of the approx 1.5 million protein sequences in the TrEMBL database (7). However, approximately two-thirds of the models are likely to contain significant errors because they are based on less than 30% sequence identity to the closest known protein structure. The primary sources of geometrical errors in the final models based on less than 30% sequence identity are the mistakes in the target-template alignment. Other errors include incor-



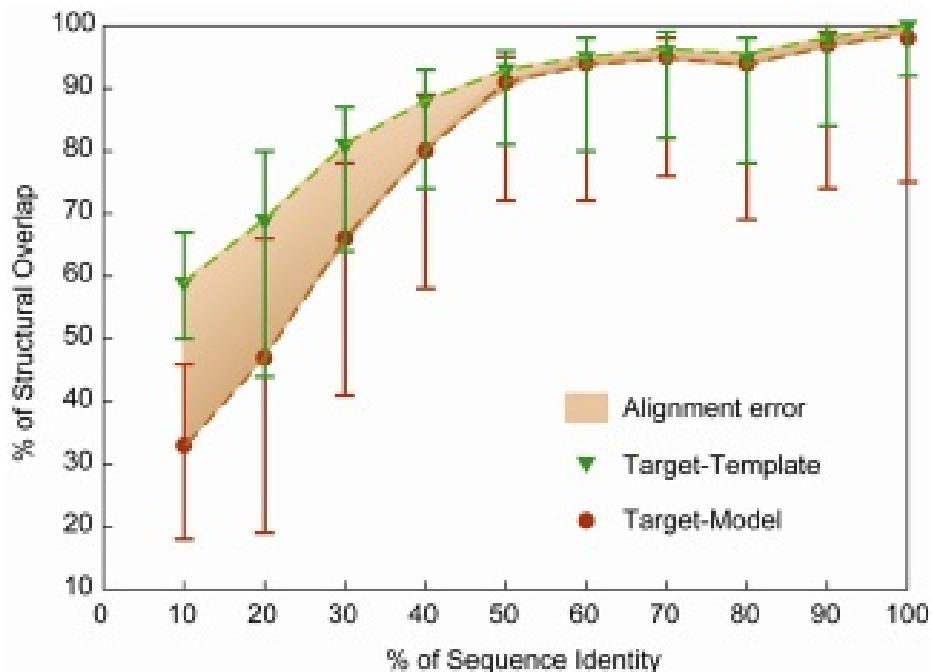


Fig. 2. Average accuracy of models calculated by ModPipe (218) with respect to the percentage sequence identity to the template. The average overlap of the experimentally determined protein structure with its calculated model (lower dashed line) and with the template on which the model was based (upper dashed line) are shown as a function of the target-template sequence identity. This sequence identity is calculated from the modeling alignment. The structure overlap is defined as the fraction of the equivalent C_{α} atoms after rigid superimposition of the two structures. Two C_{α} atoms are considered equivalent if they are within 3.5 Å of each other. The points in the curves correspond to the median values, and the error bars in the positive and negative directions correspond to the average positive and negative differences from the median, respectively. The shaded area between the two curves corresponds approximately to the model error that arises from the alignment error.

higher than 30% sequence identity (15,16). It was estimated that this cutoff requires a minimum of 16,000 targets to cover 90% of all protein domain families, including those of membrane proteins (16). These 16,000 structures will allow the modeling of a very much larger number of proteins. For example, New York Structural Genomics Research Consortium measured the impact of its structures by documenting the number and accuracy of the corresponding models for detectably related proteins in the non-redundant sequence database. For each new structure, on the average approx 100 protein sequences without any prior structural characterization could be modeled at least at the fold level (<http://www.nysgrc.org/>). This large leverage of structure determination by protein structure modeling illustrates and justifies the premise of structural genomics.

This chapter describes methods and computer programs used in all the steps of comparative modeling (Table 1). We conclude by reviewing several sample applications of the models.

APPLICATIONS

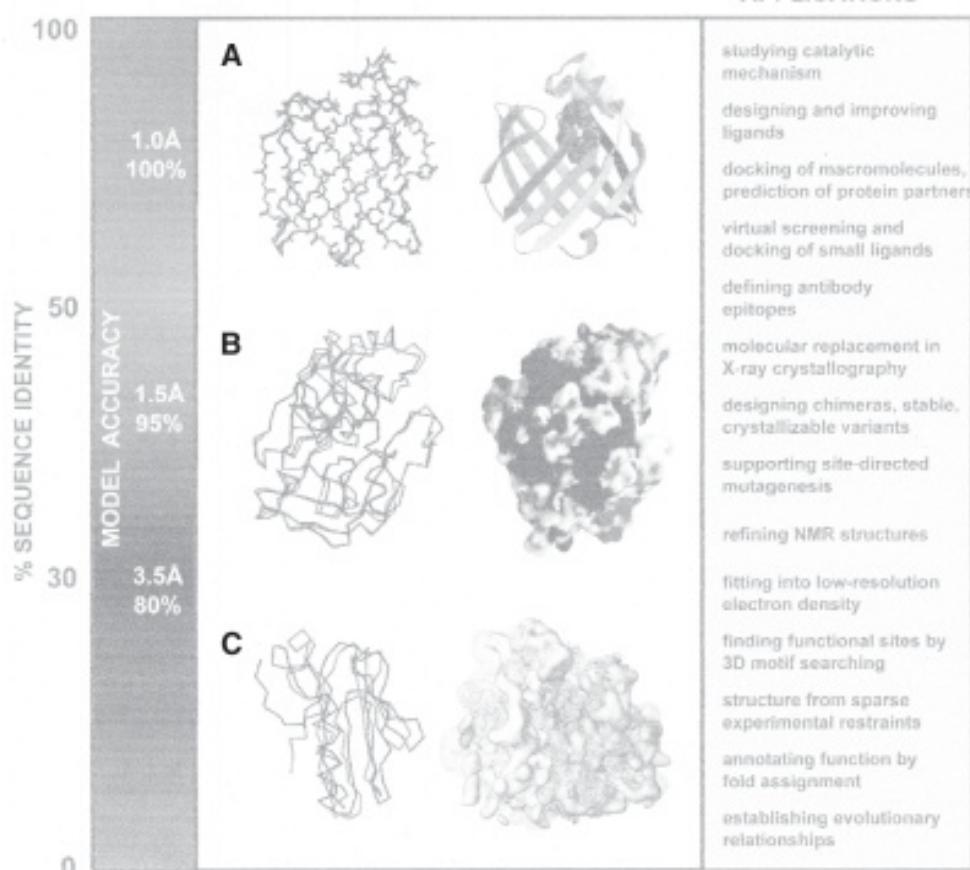


Fig. 3. Accuracy and applications of protein structure models. The vertical axis indicates the different ranges of applicability of comparative protein structure modeling, the corresponding accuracy of protein structure models, and their sample applications. (A) The docosahexaenoic fatty acid ligand was docked into a high accuracy comparative model of brain lipid-binding protein (right), modeled based on its 62% sequence identity to the crystallographic structure of adipocyte lipid-binding protein (PDB code *1adl*). A number of fatty acids were ranked for their affinity to brain lipid-binding protein consistently with site-directed mutagenesis and affinity chromatography experiments (194), even though the ligand specificity profile of this protein is different from that of the template structure (left). (B) A putative proteoglycan binding patch was identified on a medium accuracy comparative model of mouse mast cell protease 7 (right), modeled based on its 39% sequence identity to the crystallographic structure of bovine pancreatic trypsin (2*ptn*) that does not bind proteoglycans. The prediction was confirmed by site-directed mutagenesis and heparin-affinity chromatography experiments (193). Typical accuracy of a comparative model in this range of sequence similarity is indicated by a comparison of a trypsin model with the actual structure. (C) A molecular model of the whole yeast ribosome (right) was calculated by fitting atomic rRNA and protein models into the electron density of the 80S ribosomal particle, obtained by electron microscopy at 15 Å resolution (229). Most of the models for 40 out of the 75 ribosomal proteins were based on template structures that were approx 30% sequentially identical. Typical accuracy of a comparative model in this range of sequence similarity is indicated by a comparison of a model for a domain in L2 Protein from *Bacillus Stearothermophilus* with the actual structure (*1rl2*).

Table 1
Programs and Web Servers Useful in Comparative Protein Structure Modeling

Name	World-Wide Web address ^b	Reference ^c
<i>Databases</i>		
BALIBASE	http://www-igbmc.u-strasbg.fr/BioInfo/BALiBASE/	196
CATH	http://www.biochem.ucl.ac.uk/bsm/cath/	197
GENBANK	http://www.ncbi.nlm.nih.gov/Genbank/	198
GENECENSUS	http://bioinfo.mbb.yale.edu/genome/	199
MODBASE	http://www.salilab.org/modbase/	7
PDB	http://www.pdb.org	200
PRESAGE	http://presage.berkeley.edu	201
SCOP	http://scop.mrc-lmb.cam.ac.uk/scop/	202
SWISSPROT-TREMBL	http://www.expasy.org	203
<i>Template search</i>		
123D	http://123d.ncifcrf.gov/	204
3D PSSM	http://www.sbg.bio.ic.ac.uk/~3dpssm	77
BLAST	http://www.ncbi.nlm.nih.gov/BLAST/	22
DALI	http://www2.ebi.ac.uk/dali/	19
FASTA	http://www.ebi.ac.uk/fastat33/	23
MATCHMAKER	http://bioinformatics.burnham-inst.org	205
PREDICTPROTEIN	http://cubic.bioc.columbia.edu/predictprotein/	206
PROFIT	http://www.bioinfo.org.uk/software	207
THREADER	http://bioinf.cs.ucl.ac.uk/threader/threader.html	70
UCLA-DOE FOLD SERVER	http://fold.doe-mbi.ucla.edu	208
SUPERFAMILY	http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/	209
<i>Target-template alignment</i>		
BCM SERVERF	http://searchlauncher.bcm.tmc.edu	210
BLAST2	http://www.ncbi.nlm.nih.gov/blast/bl2seq/bl2.html	211
BLOCK MAKERF	http://blocks.fhcrc.org/	212
CLUSTALW	http://www2.ebi.ac.uk/clustalw/	62
MULTALIN	http://prodes.toulouse.inra.fr/multalin/	213
SEA	http://ffas.ljcrf.edu/sea/	214
FFAS03	http://ffas.ljcrf.edu/	26,64
SAM-T02	http://www.soe.ucsc.edu/research/compbio/	215
<i>HMM-apps/</i>		
FUGUE	http://www-cryst.bioc.cam.ac.uk/fugue	75
TCOFFEE	http://www.ch.embnet.org/software/TCoffee.html	216
COMPASS	ftp://iole.swmed.edu/pub/compass/	27
MUSCLE	http://www.drive5.com/muscle	217
SALIGN	http://www.salilab.org/modeller	218
USC SEQALN	http://www-hto.usc.edu/software/seqaln	219
<i>Modeling</i>		
COMPOSER	http://www.tripos.com/sciTech/inSilicoDisc/	87
CONGEN	http://www.congenomics.com/congen/congen_toc.html	94
ICM	http://www.molsoft.com/bioinfomatics/	^a 220
DISCOVERY STUDIO	http://www.accelrys.com/composer.html	^b

(continued)

Table 1 (Continued)**Programs and Web Servers Useful in Comparative Protein Structure Modeling**

Name	World-Wide Web address ^b	Reference ^c
MODELLER	http://www.salilab.org/modeller/	101
SYBYL	http://www.tripos.com	c
SCWRL	http://dunbrack.fccc.edu/SCWRL3.php	157
SNPWEB	http://salilab.org/snpweb	218
SWISS-MODEL	http://www.expasy.org/swissmod	221
WHAT IF	http://www.cmbi.kun.nl/whatif/	222
<i>Model evaluation</i>		
ANOLEA	http://protein.bio.puc.cl/cardex/servers/	188
AQUA	http://nmr.chem.uu.nl/users/jurgen/Aqua/server	184
BIOTECH	http://biotech.embl-heidelberg.de:8400	183
ERRAT	http://www.doe-mbi.ucla.edu/Services/ERRAT/	223
PROCHECK	http://www.biochem.ucl.ac.uk/~roman/procheck/procheck.html	178
PROSAIL	http://www.came.sbg.ac.at	181
PROVE	http://www.ucmb.ulb.ac.be/UCMB/PROVE	224
SQUID	http://www.ysbl.york.ac.uk/~oldfield/squid/	185
VERIFY3D	http://www.doe-mbi.ucla.edu/Services/Verify_3D/	74
WHATCHECK	http://www.cmbi.kun.nl/gv/whatcheck/	225
<i>Methods evaluation</i>		
CASP	http://predictioncenter.llnl.gov	226
CAFASP	http://bioinfo.pl/cafasp.html	170
EVA	http://cubic.bioc.columbia.edu/eva/	173
LIVEBENCH	http://bioinfo.pl/LiveBench/	171
CASA	http://capb.dbi.udel.edu/casa	227
AMAS	http://www.compbio.dundee.ac.uk/	228

Some of the sites are mirrored on additional computers.

^aMolSoft Inc., San Diego.

^bAccelrys Inc., San Diego.

^cTripos Inc., St. Louis.

The BIOTECH server uses PROCHECK and WHATCHECK for structure evaluation.

2. Steps in Comparative Modeling

Comparative modeling consists of four sequential steps: fold assignment, target-template alignment, model building, and model assessment (Fig. 1). If an assessment of the model is not positive, the model can be rebuilt by selecting different templates, refining the target-template alignment, or changing model-building parameters. The sections below deal with each one of the four steps in the modeling protocol.

2.1. Fold Assignment and Template Selection

The initial step in comparative modeling is to assign the likely fold of the target sequence. Template identification can be achieved using any one of the many programs that scan sequence and structure databases, such as Protein Data Bank (PDB) (17), structural classification of proteins (SCOP) (18), distance-matrix alignment

(DALI) (19), and Class, Architecture, Topology, and Homology (CATH) (20,21) (Table 1). Template search methods can be categorized into three different classes:

First, pairwise comparison methods, which include the popular programs Basic Local Alignment Search Tool (BLAST) (22) and FASTA (23), align the target sequence with all the sequences in the database of known structures. The performance and efficiency of this class of methods has been studied extensively (24). Second, sequence profile methods, such as position specific iterative (PSI)-BLAST (25) and HMMER (<http://hmmer.wustl.edu>), rely on profiles derived from multiple sequence alignments to increase the sensitivity and accuracy of the template search. The profile enhances the sensitivity of the search (26–29). Profiles are also utilized by the intermediate sequence search algorithms that establish a homology between two remotely related sequences through an intermediary sequence (30–36). Third, the so-called threading methods use a combination of sequence and structure considerations to detect similarities between sequences and structures (37–41). In these methods, the target sequence is threaded through a library of 3-D profiles or folds, and each threading is assessed based on a certain scoring function. Commonly used methods and servers in this category include Superfamily (42) and GenThreader (43). The threading methods are more effective in detecting homology at low sequence similarity than the methods relying on sequence information alone (44).

The three different classes of methods are best suited for identifying templates in different regimes of the sequence-identity spectrum. The pairwise sequence comparison methods are the least sensitive and are best used to detect close homologs. The profile-based methods are usually capable of recognizing homologs sharing only approx 25% sequence identity. Threading methods can sometimes recognize common folds even in the absence of any statistically significant sequence similarity. Because most of the fold assignment methods involve sequence alignment, some of them are discussed in more detail in the following section about sequence-structure alignment.

While a correct fold assignment can be used to build a useful model, an incorrect fold assignment renders the resulting model useless. Thus, when using a fold-recognition method, it is crucial to be aware of the accuracy of the method. In an assessment of different fold-recognition methods, the best method detected 75% of the closest structures correctly for a set of sequences related at the “family” level in the SCOP database (18). However, at the superfamily and fold levels, the accuracy dropped to 29 and 15%, respectively (44).

Once a list of all related protein structures is obtained, templates that are appropriate for the given modeling problem have to be selected. Usually, a higher overall sequence identity between the target and the template sequence yields a better template. Several other factors should also be considered in selecting templates.

Constructing a phylogenetic tree for the whole family can frequently help in selecting a template from the subfamily that is closest in structure to the target sequence. Databases of structure-based phylogenies, such as the database of Phylogeny and Alignment (PALI) (45), are useful in making a distinction between the sequence and structure similarity, which can be a key consideration for template identification.

Accuracy of the template structure is another important factor in template selection. The resolution and the R-factor of a crystallographic structure and the number of restraints per residue for an NMR structure are indicative of structure accuracy.

It is also crucial to compare the environment of the template to the required environment for the model. The term *environment* is used in a broad sense and includes all factors that determine protein structure, except its sequence (e.g., solvent, pH, ligands, and quaternary interactions). For example, if the objective of the model-building exercise is to dock ligands in the model, it is usually best to use a template that is itself bound to an identical or similar ligand. In general, prior biological information about the target sequence can be valuable in identifying an appropriate template (46,47).

Prioritization of the criteria for template selection depends on the purpose of the comparative model. For instance, if a protein–ligand model is to be constructed, the choice of the template that contains a similar ligand is probably more important than the resolution of the template. On the other hand, if a model is to be used to analyze the geometry of the active site of an enzyme, it is preferable to use a high-resolution template. It is not necessary to select only one template. In fact, the use of several templates approximately equidistant from the target sequence generally increases the model accuracy (48,49).

2.2. Target–Template Alignment

After identifying the template(s), the next crucial step in comparative modeling is to accurately align the target sequence to the template(s). Although most template-recognition methods produce a target–template alignment, there is frequently a need to use a specialized alignment method to realign the sequences because the template-identification step is often optimized to identify distant relationships, sometimes at the expense of alignment accuracy. The sequence-structure alignment is a vital step in the model-building process, and an erroneous alignment will almost certainly lead to the construction of an incorrect model.

An alignment between two sequences of residues is usually calculated by optimizing an alignment scoring function (50). The two common ingredients of the scoring function are a gap penalty function and a matrix of substitution scores for matching every residue in one sequence to every residue in the other sequence. The alignment score is usually a sum of the gap penalties, which depend linearly on the gap lengths, and the pairwise substitution scores, which depend on the matched residue types. The original and still widely used optimization method for sequence alignment is based on dynamic programming (51–53). Since its inception, the scoring function and its optimization by dynamic programming have been improved for alignment accuracy and speed, and applied to a variety of alignment problems.

In the next few paragraphs, we examine different methods to obtain substitution score matrices and gap penalties that optimize the accuracy of the output alignments. We examine the use of information from related multiple sequences and structures to enhance alignment accuracy and coverage, especially when target–template sequence identity decreases below 30%.

2.2.1. Using Multiple Sequence Information

The accuracy of a pairwise alignment method that uses dynamic programming greatly depends on the matrix of substitution scores and the gap penalties. Matrices with values for each of the possible residue type substitutions, such as Block Substitution Matrix (BLOSUM) (54) and point accepted mutation (PAM) (55), are useful only when sequence similarity is readily recognizable (e.g., above 30% sequence identity).

To increase the accuracy of the alignment between more divergent sequences, some methods construct the substitution scores by relying on substitution patterns revealed in a multiple sequence alignment (MSA) of many members of the corresponding protein family. A multiple sequence alignment is converted into a sequence profile that lists the likelihood of the 20 standard amino acid residue types at every position in a given MSA. Alignments based on sequence profiles rather than single sequences have been shown to be significantly more accurate (56–58) (Table 2). This improvement is reflected in the accuracy and extent of the resulting homology models.

Two popular profile alignment methods are PSI-BLAST (25) and SAM-T98 (59). Both methods take a single sequence as input and produce a sequence profile or a hidden Markov model (HMM) as output. PSI-BLAST relies on the BLAST algorithm (22) to collect homologs of a query sequence and to construct its profile by iteratively scanning a sequence database (25,32). SAM-T98 first uses BLAST to prefilter a large sequence database. It then constructs a multiple alignment and a HMM in parallel through several rounds of database searching and HMM building. The HMM is derived only from the sequences that score better than a specified threshold.

The latest generation of alignment methods extends sequence profile or MSA building to both sequences of interest, and aligns the two profiles or MSAs, rather than the individual sequences. These methods have been shown to be more sensitive than sequence-profile methods (26–28,60). The CLUSTALW program compares two MSAs by using a substitution matrix for all pairs of positions from the two alignments (61,62); each single value in this matrix is an average of residue-residue substitution scores over two matched alignment positions. The LAMA program aligns two MSAs by first transforming them into position-specific scoring matrices (PSSMs) and then comparing the two PSSMs to each other by the Pearson correlation coefficient (63). The FFAS program aligns two sequence profiles with each other using a dot product (26,64). A related approach, using mutual entropy, has been used by Yona and Levitt (65,66) to construct the ProtoMap database of protein sequence families (66–68). Most recently, the COMPASS program was developed to locally align two MSAs with assessment of statistical significance (27). The SALIGN command in the program MODELLER constructs a scoring matrix by comparing two profiles with mutual entropy and correlation coefficient measures (60). These methods compare two profiles by matching every position in one profile to each position in the other profile, followed by either local or global dynamic programming to calculate the optimal alignment.

2.2.2. Using Structural Information

Alignment accuracy can be significantly improved by incorporating information about protein structure. Threading and 3-D template-matching methods consider protein structure information for one of the sequences in a pairwise comparison (69–71). For a review of this class of methods, see (38–41,72). A combination of threading and sequence alignment scoring functions can also be used (43,73).

Another approach is to incorporate structural information into profile methods, by making substitution scores dependent on solvent exposure, secondary structure type, hydrogen bonding properties, and so on (74). Some methods in this category are FUGUE (75), 3D-PSSM (76,77), and SAM-T02 multitrack HMMs (78). These methods lie between traditional sequence-based algorithms and threading methods. The use of structural data is not restricted to the structure side of the aligned sequence-structure

Table 2
Different Programs for Aligning Two Protein Sequences, or a Protein Sequence and a Structure, Tested on a Benchmark of 200 Pairs of Related Known Structures*

Method	Type	Average alignment accuracy					
		1Å	2Å	3Å	4Å	5Å	Average
CE	structure/structure	18.81	49.09	68.02	78.77	84.54	59.85
BLAST	sequence/sequence	7.60	17.07	22.72	26.41	29.29	20.62
ALIGN	sequence/sequence	6.86	18.11	27.19	34.79	41.44	25.68
PSI-BLAST	sequence/profile	9.07	23.50	33.16	40.28	45.63	30.33
SAM	sequence/profile	7.76	21.60	31.40	38.72	45.26	28.95
LOBSTER	sequence/profile	8.81	23.32	33.82	41.51	48.17	31.13
SEA	profile/profile	9.02	24.15	34.90	43.27	50.43	32.36
CLUSTALW	profile/profile	7.41	19.31	28.02	35.36	41.87	26.40
COMPASS	profile/profile	10.37	26.06	36.08	42.35	46.65	32.30
SALIGN	profile/profile	9.63	27.05	39.81	49.64	57.55	36.74

*An alignment is assessed here by a degree of structure similarity that it implies. This criterion was calculated by first superposing the two compared structures according to the tested alignment, and then calculating the percentage of the C_α positions that were within the specified cutoff of 1, 2, 3, 4, or 5 Å; in addition, the average of these percentages at all cutoffs was also calculated. For comparison, the actual structure similarity calculated from the structure-based alignments produced by the CE program is also given in the first row.

pair. For example, SAM-T02 and HMAP (79) make use of the predicted local structure to enhance homolog detection and alignment accuracy.

To improve the alignment accuracy, gap penalties can be adjusted according to the local environment in which they occur (80). For example, the SALIGN command in MODELLER (81) scales the gap insertion penalty depending on the structural environment of the gap; the cost of opening a gap in a region of regular secondary structure is greater than opening a gap in a random coil region. The SALIGN command of MODELLER (82) can also use structure-dependent gap penalties in conjunction with a sequence profile, similar to FUGUE (75).

Even when algorithms are enriched with structural and multiple sequence information, it remains difficult to align distantly homologous proteins in the “twilight zone” of sequence identity below 30% sequence identity (83). In a comparative modeling setting, the MOULDER algorithm (84) uses an iterative approach to build better alignments between distant homologs. The method relies on a genetic algorithm to iteratively (1) build target-template alignments; (2) build structural models based on the alignments; (3) assess the models; and (4) select the alignments that produce the best models as seeds to generate further alignments (84). The method was shown to improve significantly the alignment accuracy of alignments that fell within the twilight zone.

2.3. Model Building

The target-template alignment, maps the sequence of the target on the template structure. This mapping is utilized in constructing the 3-D model of the target protein. There

are several methods of constructing the model, and some of these approaches are reviewed below. The various model-building procedures lead to the construction of models of similar accuracy when used optimally (85). In addition to the different schemes for building whole models, this review also examines techniques for constructing inserted loop segments of the target that have no corresponding template and for packing the side chains on a given backbone scaffold.

2.3.1. Modeling by Assembly of Rigid Bodies

The first and still widely used approach in comparative modeling is to assemble a model from a small number of rigid bodies obtained from the aligned protein structures (6,86). The approach is based on the natural dissection of the protein structure into conserved core regions, variable loops that connect them, and side chains that decorate the backbone. For example, the following semiautomated procedure is implemented in the computer program COMPOSER (87). First, the template structures are selected and superposed. Second, the “framework” is calculated by averaging the coordinates of the C_α atoms of structurally conserved regions in the template structures. Third, the main-chain atoms of each core region in the target model are obtained by superposing on the framework the core segment from the template whose sequence is closest to the target. Fourth, the loops are generated by scanning a database of all known protein structures to identify the structurally variable regions that fit the anchor core regions and have a compatible sequence (88). Fifth, the side chains are modeled based on their intrinsic conformational preferences and on the conformation of the equivalent side chains in the template structures (87). And finally, the stereochemistry of the model is improved either by a restrained energy minimization or a molecular dynamics refinement. The accuracy of a model can be somewhat increased when more than one template structure is used to construct the framework and when the templates are averaged into the framework using weights corresponding to their sequence similarities to the target sequence (48). Possible future improvements of modeling by rigid-body assembly include incorporation of rigid-body shifts, such as the relative shifts in the packing of α -helices and β -sheets (89).

2.3.2. Modeling by Segment Matching or Coordinate Reconstruction

The basis of modeling by coordinate reconstruction is the finding that most hexapeptide segments of protein structure can be clustered into only 100 structurally different classes (90,91). Thus, comparative models can be constructed by using a subset of atomic positions from template structures as “guiding” positions, and by identifying and assembling short, all-atom segments that fit these guiding positions. The guiding positions usually correspond to the C_α atoms of the segments that are conserved in the alignment between the template structure and the target sequence. The all-atom segments that fit the guiding positions can be obtained either by scanning all the known protein structures, including those that are not related to the sequence being modeled (92,93), or by a conformational search restrained by an energy function (94,95). For example, a general method for modeling by segment matching is guided by the positions of some atoms (usually C_α atoms) to find the matching segments in the representative database of all known protein structures (96). This method can construct both main-chain and side-chain atoms, and can also model unaligned regions (gaps). It is implemented in the program SegMod. Even some side-chain modeling methods (97)

and the class of loop construction methods based on finding suitable fragments in the database of known structures (98) can be seen as segment-matching or coordinate-reconstruction methods.

2.3.3. Modeling by Satisfaction of Spatial Restraints

The methods in this class begin by generating many constraints or restraints on the structure of the target sequence, using its alignment to related protein structures as a guide. The procedure is conceptually similar to that used in determination of protein structures from NMR-derived restraints. The restraints are generally obtained by assuming that the corresponding distances between aligned residues in the template and the target structures are similar. These homology-derived restraints are usually supplemented by stereochemical restraints on bond lengths, bond angles, dihedral angles, and nonbonded atom–atom contacts that are obtained from a molecular-mechanics force field. The model is then derived by minimizing the violations of all the restraints. This optimization can be achieved either by distance geometry or real-space optimization. For example, an elegant distance-geometry approach constructs all-atom models from lower and upper bounds on distances and dihedral angles (99).

We now describe our own approach to comparative modeling by satisfaction of special restraints in more detail (100–103). The approach was developed to use as many different types of data about the target sequence as possible. It is implemented in the computer program MODELLER (101). The comparative modeling procedure begins with an alignment of the target sequence with related known 3-D structures. The output, obtained without any user intervention, is a 3-D model for the target sequence containing all main-chain and side-chain nonhydrogen atoms.

In the first step of model building, distance and dihedral angle restraints on the target sequence are derived from its alignment with template 3-D structures. The form of these restraints was obtained from a statistical analysis of the relationships between similar protein structures. The analysis relied on a database of 105 family alignments that included 416 proteins of known 3-D structure (103). By scanning the database of alignments, tables quantifying various correlations were obtained, such as the correlations between two equivalent C_{α} - C_{α} distances, or between equivalent main-chain dihedral angles from two related proteins (101). These relationships are expressed as conditional probability density functions (PDFs) and can be used directly as spatial restraints. For example, probabilities for different values of the main-chain dihedral angles are calculated from the type of a residue considered, from main-chain conformation of an equivalent residue, and from sequence similarity between the two proteins. Another example is the PDF for a certain C_{α} - C_{α} distance given equivalent distances in two related protein structures. An important feature of the method is that the forms of spatial restraints were obtained empirically, from a database of protein structure alignments.

In the second step, the spatial restraints and the CHARMM22 force-field terms enforcing proper stereochemistry (104,105) are combined into an objective function. The general form of the objective function is similar to that in molecular dynamics programs, such as CHARMM22 (105). The objective function depends on the Cartesian coordinates of approx 10,000 atoms (3-D points) that form the modeled molecules. For a 10,000-atom system, there can be on the order of 200,000 restraints. The functional form of each term is simple; it includes a quadratic function, harmonic lower and

upper bounds, cosine, a weighted sum of a few Gaussian functions, Coulomb's law, Lennard-Jones potential, and cubic splines. The geometric features presently include a distance; an angle; a dihedral angle; a pair of dihedral angles between two, three, four atoms and eight atoms, respectively; the shortest distance in the set of distances; solvent accessibility in \AA^2 ; and atom density, expressed as the number of atoms around the central atom. Some restraints can be used to restrain pseudo-atoms such as the gravity center of several atoms.

Finally, the model is obtained by optimizing the objective function in Cartesian space. The optimization is carried out by the use of the variable target function method (106) employing methods of conjugate gradients and molecular dynamics with simulated annealing (107). Several slightly different models can be calculated by varying the initial structure, and the variability among these models can be used to estimate the lower bound on the errors in the corresponding regions of the fold.

Because the modeling by satisfaction of spatial restraints can use many different types of information about the target sequence, it is perhaps the most promising of all comparative modeling techniques. One of the strengths of modeling by satisfaction of spatial restraints is that constraints or restraints derived from a number of different sources can easily be added to the homology-derived restraints. For example, restraints could be provided by rules for secondary structure packing (108), analyses of hydrophobicity (109) and correlated mutations (110), empirical potentials of mean force (111), NMR experiments (112), cross-linking experiments, fluorescence spectroscopy, image reconstruction in electron microscopy, site-directed mutagenesis (113), intuition, and so on. In this way, a comparative model, especially in the difficult cases, could be improved by making it consistent with available experimental data and/or with more general knowledge about protein structure.

Accuracies of the various model-building methods are relatively similar when used optimally (85). Other factors such as template selection and alignment accuracy usually have a larger impact on the model accuracy, especially for models based on less than 40% sequence identity to the templates. However, it is important that a modeling method allow a degree of flexibility and automation to obtain better models more easily and rapidly. For example, a method should allow for an easy recalculation of a model when a change is made in the alignment; it should be straightforward to calculate models based on several templates; and the method should provide tools for incorporation of prior knowledge about the target (e.g., cross-linking restraints, predicted secondary structure) and allow *ab initio* modeling of insertions (e.g., loops), which can be crucial for annotation of function. Loop modeling is an especially important aspect of comparative modeling in the range from 30 to 50% sequence identity. In this range of overall similarity, loops among the homologs vary while the core regions are still relatively conserved and aligned accurately.

2.3.4. Loop Modeling

In comparative modeling, target sequences often have residues inserted relative to the template structures, or have regions that are structurally different from the corresponding regions in the templates. Thus, no structural information about these inserted segments can be extracted from the template structures. These regions frequently correspond to surface loops. Loops often play an important role in defining the functional specificity of a given protein framework, forming the active and binding sites. The

accuracy of loop modeling is a major factor determining the usefulness of comparative models in applications such as ligand docking. Loop modeling can be seen as a mini-protein folding problem, because the correct conformation of a given segment of a polypeptide chain has to be calculated mainly from the sequence of the segment itself. However, loops are generally too short to provide sufficient information about their local fold. Even identical decapeptides in different proteins do not always have the same conformation (114,115). Some additional restraints are provided by the core anchor regions that span the loop, and by the structure of the rest of a protein that cradles the loop. Although many loop-modeling methods have been described, it is still not possible to model correctly and confidently loops longer than approximately eight residues (102).

There are two main classes of loop-modeling methods: (1) the database search approaches that scan a database of all known protein structures to find segments fitting the anchor core regions (98,116); and (2) the conformational search approaches that rely on an optimization of a scoring function (117,118). There are also methods that combine these two approaches (119,120).

The database search approach to loop modeling is accurate and efficient when a database of specific loops is created to address the modeling of the same class of loops, such as β -hairpins (121), or loops on a specific fold, such as the hyper-variable regions in the immunoglobulin fold (116,122). There are attempts to classify loop conformations into more general categories, thus extending the applicability of the database search approach to more cases (123–125). However, the database methods are limited by the fact that the number of possible conformations increases exponentially with the length of a loop. As a result, only loops up to four to seven residues long have most of their conceivable conformations present in the database of known protein structures (126,127). Even according to the more optimistic estimate, approx 30% and 60% of all the possible eight- and nine-residue loop conformations, respectively, are missing from the database (126). This limitation is made even worse by the requirement for an overlap of at least one residue between the database fragment and the anchor core regions, which means that the modeling of a 5-residue insertion requires at least a 7-residue fragment from the database (92). Despite the rapid growth of the database of known structures, it does not seem possible to cover most of the conformations of a 9-residue segment in the foreseeable future. On the other hand, most of the insertions in a family of homologous proteins are shorter than 10–12 residues (102).

To overcome the limitations of the database search methods, conformational search methods were developed (117,128). There are many such methods, exploiting different protein representations, objective function terms, and optimization or enumeration algorithms. The search algorithms include the minimum perturbation method (129), molecular dynamics simulations (94,119), genetic algorithms (130), Monte Carlo and simulated annealing (131–133), multiple-copy simultaneous search (134), self-consistent field optimization (135), and an enumeration based on the graph theory (136). The accuracy of loop predictions can be further improved by clustering the sampled loop conformations and therefore partially accounting for the entropic contribution to the free energy (137). Another way to improve the accuracy of loop predictions is to consider the solvent effects. Improvements in implicit solvation models, such as the generalized Born solvation model (GB) (138) and surface-generalized Born with nonpolar

correction (SGB/NP) (139), motivated their use in loop modeling. The solvent contribution to the free energy can be added to the scoring function for optimization, or it can be used to rank the sampled loop conformations after they are generated with a scoring function that does not include the solvent terms (2,140–143).

The loop modeling module in MODELLER implements the optimization-based approach (2,102). The main reasons are the generality and conceptual simplicity of scoring function minimization, as well as the limitations on the database approach imposed by a relatively small number of known protein structures (126). Loop prediction by optimization is applicable to simultaneous modeling of several loops and loops interacting with ligands, which is not straightforward for the database search approaches. Loop optimization in MODELLER relies on conjugate gradients and molecular dynamics with simulated annealing. The pseudo-energy function is a sum of many terms, including some terms from the CHARMM22 molecular mechanics force field (104) and spatial restraints based on distributions of distances (111,144) and dihedral angles in known protein structures. The method was tested on a large number of loops of known structure, both in the native and near-native environments (102).

2.3.5. Side-Chain Modeling

Two simplifications are frequently applied in the modeling of side-chain conformations. First, amino acid replacements often leave the backbone conformation almost unchanged (145), allowing us to fix the backbone during the search for the best side-chain conformations. Second, most side chains in high-resolution crystallographic structures can be represented by a limited number of conformers that comply with stereochemical and energetic constraints (146). This observation motivated Ponder and Richards to develop the first library of side-chain rotamers for the 17 types of residues with dihedral angle degrees of freedom in their side chains, based on 10 high-resolution protein structures determined by X-ray crystallography (147). Subsequently, a number of additional libraries have been derived (148–152).

Rotamers on a fixed backbone are often used when all the side chains need to be modeled on a given backbone. This approach overcomes the combinatorial explosion associated with a full conformational search of all the side chains, and is applied by some comparative modeling (6) and protein design approaches (153). However, approx 15% of the side chains cannot be represented well by these libraries (154). In addition, it has been shown that the accuracy of side-chain modeling on a fixed backbone decreases rapidly when the backbone errors are larger than only 0.5 Å (155). Fortunately, these two approximations may be unnecessary in the modeling of a single-point mutation that in general does not trigger changes in many dihedral angles (152).

Earlier methods for side-chain modeling often put less emphasis on the energy or scoring function. The function was usually greatly simplified, and consisted of the empirical rotamer preferences and simple repulsion terms for non-bonded contacts (151). Nevertheless, these approaches have been justified by their performance. For example, a method based on a rotamer library compared favorably with that based on a molecular-mechanics force field (156), and more recently all the new and most efficient methods are also based on rotamer library (152,157). In contrast, a lot of attention has been paid to the optimization procedure. The various approaches include a Monte Carlo simulation (158), simulated annealing (159), a combination of Monte Carlo and simulated annealing (160), the dead-end elimination theorem (161,162), genetic algo-

rithms (148), neural network with simulated annealing (163), mean field optimization (164), and combinatorial searches (151,165,166). It was suggested that the modeling accuracy for up to 10-residue segments is currently limited by the accuracy of the scoring function, not by the thoroughness of the search algorithms (102). Several recent papers focused on the testing of more sophisticated potential functions for conformational search (166,167) and development of new scoring functions for side-chain modeling (168), report favorable performance compared to earlier studies.

3. Errors in Comparative Modeling

It is crucial for method developers and users alike to assess the accuracy of their methods. An attempt to address this problem has been made by the Critical Assessment of Techniques for Proteins Structure Prediction (CASP) (169) and the Critical Assessment of Fully Automated Structure Prediction (CAFASP) experiments (170). However, both CASP and CAFASP assess methods only over a limited number of target protein sequences (85,171). To overcome this limitation, two additional evaluation experiments have been described, LiveBench (171) and EVA (172,173). EVA is a large-scale and continuously running Web server that automatically assesses protein structure prediction servers in the categories of secondary structure prediction, residue-residue contact prediction, fold assignment, and comparative modeling. The aims of EVA are (1) to evaluate continuously and automatically blind predictions by prediction servers, based on identical and sufficiently large data sets; (2) to provide weekly updates of the method assessments on the Web; and (3) to enable developers, non-expert users, and reviewers to determine the performance of the tested prediction servers.

As the similarity between the target and the templates decreases, the errors in the model increase. Errors in comparative models can be divided into five categories as follows (49) (Fig. 4).

1. First, errors in side-chain packing. As the sequences diverge, the packing of the atoms in the protein core changes. Sometimes, even the conformation of identical side chains is not conserved, a pitfall for many comparative modeling methods. Side-chain errors are critical if they occur in regions that are involved in protein function, such as active sites and ligand-binding sites.
2. Second, distortions and shifts in correctly aligned regions. As a consequence of sequence divergence, the main-chain conformation changes, even if the overall fold remains the same. Therefore, it is possible that in some correctly aligned segments of a model, the template is locally different (<3 Å) from the target, resulting in errors in that region. The structural differences are sometimes not due to differences in sequence, but are a consequence of artifacts in structure determination, or structure determination in different environments (e.g., packing of subunits in a crystal). The simultaneous use of several templates can minimize this kind of an error (49,174).
3. Third, errors in regions without a template. Segments of the target sequence that have no equivalent region in the template structure (i.e., insertions or loops) are the most difficult regions to model. As mentioned in the section on loop modeling, this problem is akin to *ab initio* fold prediction. If the insertion is relatively short, less than nine residues long, some methods can correctly predict the conformation of the backbone (102,119,143,175). Conditions for successful prediction are the correct alignment and an accurately modeled environment surrounding the insertion.

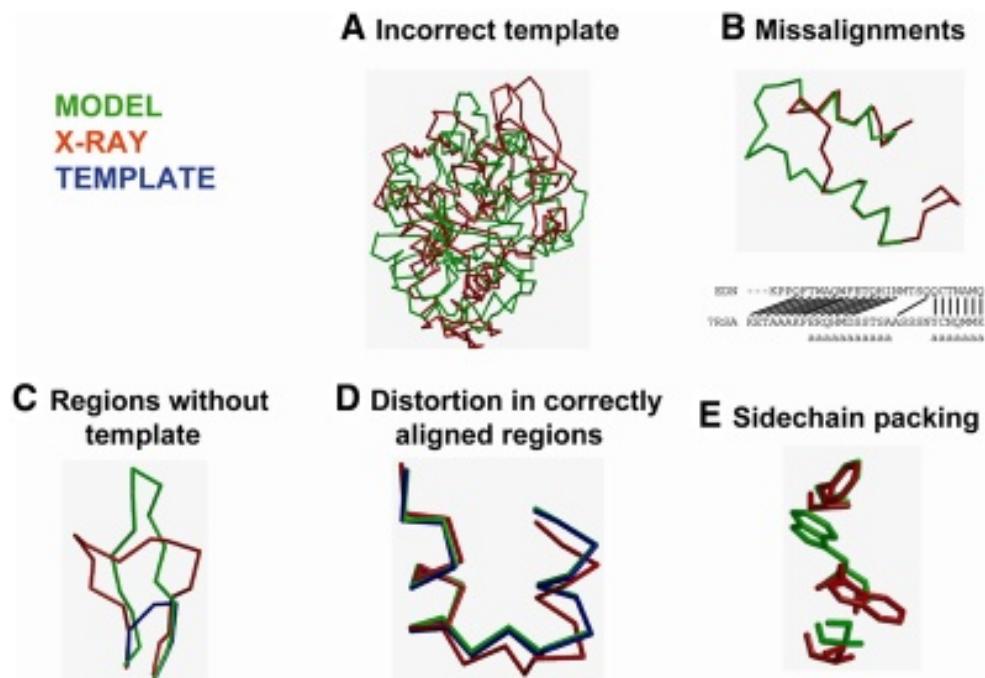


Fig. 4. Typical errors in comparative modeling.

- Fourth, errors resulting from misalignments. The largest source of errors in comparative modeling is misalignments, especially when the target–template sequence identity decreases below 30%. However, alignment errors can be minimized in two ways. First, it is usually possible to use a large number of sequences to construct a multiple alignment, even if most of these sequences do not have known structures. Multiple alignments are generally more reliable than pairwise alignments (176,177). A second way of improving the alignment is to iteratively modify those regions in the alignment that correspond to predicted errors in the model (49).
- Fifth, selection of incorrect templates. This error is a potential problem when distantly related proteins are used as templates (i.e., less than 25% sequence identity). Distinguishing between a model based on an incorrect template and a model based on an incorrect alignment with a correct template is difficult. In both cases, the evaluation methods will predict an unreliable model. The conservation of the key functional or structural residues in the target sequence increases the confidence in a given fold assignment.

4. Predicting Model Accuracy

The accuracy and extent of the predicted structure determines the information that can be extracted from it. Thus, estimating the accuracy of 3-D protein models in the absence of the known structures is essential for interpreting them. The model can be evaluated as a whole as well as in the individual regions. There are many model evaluation programs and servers (178,179) (Table 1).

The first step in model evaluation is to determine whether the model has the correct fold (180). A model will have the correct fold if the correct template is picked and if

that template is aligned at least approximately correctly with the target sequence. The confidence in the fold of a model is generally increased by a high sequence similarity with the closest template, an energy-based Z-score (180,181), or by conservation of the key functional or structural residues in the target sequence.

Once the fold of a model is accepted, a more detailed evaluation of the overall model accuracy can be obtained, based on the similarity between the target and template sequences (180). Sequence identity above 30% is a relatively good predictor of the expected accuracy, because the deviation from the least-squares curve relating sequence identity to the accuracy is relatively small. The reasons are the well-known relationship between structure and sequence similarities of two proteins (145), the “geometrical” nature of modeling (which forces the model to be as close to the template as possible) (101), and the inability of any current modeling procedure to recover from an incorrect alignment (49). The dispersion of the model-target structure overlap increases with the decrease in sequence identity. If the target-template sequence identity falls below 30%, the sequence identity becomes unreliable as a predictor of the model accuracy. Models that deviate significantly from the average accuracy are frequent. It is in such cases that model evaluation methods are particularly useful.

In addition to the target-template sequence similarity, the environment can strongly influence the accuracy of a model. For instance, some calcium-binding proteins undergo large conformational changes when bound to calcium. If a calcium-free template is used to model the calcium-bound state of the target, it is likely that the model will be incorrect irrespective of the target-template similarity or accuracy of the template structure (182). This observation also applies to the experimental determination of protein structure; a structure must be determined in the functionally meaningful environment.

A basic requirement for a model is to have good stereochemistry. Some useful programs for evaluating stereochemistry are PROCHECK (183), PROCHECK-NMR (184), AQUA (184), SQUID (185), and WHATCHECK (186). The features of a model that are checked by these programs include bond lengths, bond angles, peptide-bond and side-chain ring planarities, chirality, main-chain and side-chain torsion angles, and clashes between non-bonded pairs of atoms.

There are also methods for testing 3-D models that implicitly take into account many spatial features compiled from high-resolution protein structures. These methods are based on 3-D profiles and statistical potentials of mean force (74,111,144). Programs implementing this approach include VERIFY3D (74), PROSAII (181), HARMONY (187), ANOLEA (188), and DFIRE (189). These programs evaluate the environment of each residue in a model with respect to the expected environment as found in the high-resolution X-ray structures. There is a concern about the theoretical validity of the energy profiles for detecting regional errors in models (102). It is likely that the contributions of the individual residues to the overall free energy of folding vary widely, even when normalized by the number of atoms or interactions made. If this expectation is correct, the correlation between the prediction errors and energy peaks is greatly weakened, resulting in the loss of predictive power of the energy profile. Despite these concerns, error profiles have been useful in some applications (190).

5. Applications of Comparative Modeling

Comparative models have been used in a myriad of applications (1,191). The applicability of a model depends on its accuracy (Fig. 3). We now list typical applications of comparative models.

Models that are built using as templates protein structures with which they share less than approx 25% in sequence identity are usually used for fold assignment. Such models often have less than 50% of their C_{α} positions within 3.5 Å of the actual structure. Nevertheless, fold assignment is frequently sufficient to assign coarse protein function (20,192). At this level of target–template similarity, model evaluation can be used as a discriminator between correct and incorrect fold assignment (49,144,180).

Models built on approx 35% sequence identity to the templates, on the average cover about 85% of the residues to within 3.5 Å of their correct positions. Since the active and binding sites of proteins are frequently more conserved than the rest of the fold, they tend to be modeled more accurately than the rest of the fold (180). In general, medium-resolution models frequently allow a refinement of the functional prediction based on sequence alone, because ligand binding is most directly determined by the structure of the binding site rather than its sequence. It may be possible to correctly predict important features of the target protein that do not occur in the template structure. For example, the location of a binding site can be predicted from clusters of charged residues (193), and the size of a ligand may be predicted from the volume of the binding-site cleft (194). Medium-resolution models can also be used to construct site-directed mutants with altered binding capacity, which in turn could test hypotheses about the sequence-structure-function relationships. Other problems that can be addressed with medium-resolution comparative models include designing proteins that have compact structures—without long tails, loops, and exposed hydrophobic residues—for better crystallization; or designing proteins with added disulfide bonds for extra stability.

The high end of the accuracy spectrum corresponds to models based on 50% sequence identity or more. The average accuracy of these models approaches that of low-resolution X-ray structures (3 Å resolution) or medium-resolution NMR structures (10 distance restraints per residue) (49). The alignments on which these models are based generally contain almost no errors. In addition to the already listed applications, high-accuracy models can be used for docking of small ligands (130) or whole proteins onto the given protein (195). For an overall view of the scope of applicability of computational models, *see refs. 1,191*.

Acknowledgments

Our research has been supported by Sandler Family Supporting Foundation, NIH/NIGMS R01 GM 54762, NIH/NIGMS P50 GM62529, NIH/NCI R33 CA84699, Sun Academic Equipment Grant EDUD-7824-020257-US, an IBM SUR grant, and an Intel computer hardware gift.

References

1. Baker, D. and Sali, A. (2001) Protein structure prediction and structural genomics. *Science* **294**, 93–96.
2. Fiser, A., Feig, M., Brooks, C. L. III, and Sali, A. (2002) Evolution and physics in comparative protein structure modeling. *Acc. Chem. Res.* **35**, 413–421.
3. Bradley, P., Chivian, D., Meiler, J., et al. (2003) Rosetta predictions in CASP5: successes, failures, and prospects for complete automation. *Proteins* **53 Suppl. 6**, 457–468.
4. Kinch, L. N., Wrabl, J. O., Krishna, S. S., et al. (2003) CASP5 assessment of fold recognition target predictions. *Proteins* **53 Suppl. 6**, 395–409.
5. Marti-Renom, M. A., Stuart, A., Fiser, A., et al. (2000) Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 291–325.
6. Blundell, T. L., Sibanda, B. L., Sternberg, M. J., and Thornton, J. M. (1987) Knowledge-based prediction of protein structures and the design of novel molecules. *Nature* **326**, 347–352.
7. Pieper, U., Eswar, N., Braberg, H., et al. (2004) MODBASE, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Res.* **32**, D217–D222.
8. Chance, M. R., Bresnick, A. R., Burley, S. K., et al. (2002) Structural Genomics: A pipeline for providing structures for the biologist. *Protein Sci.* **11**, 723–738.
9. Burley, S. K., Almo, S. C., Bonanno, J. B., et al. (1999) Structural genomics: beyond the human genome project. *Nat. Genet.* **23**, 151–157.
10. Sanchez, R., Pieper, U., Mirkovic, N., et al. (2000) MODBASE, a database of annotated comparative protein structure models. *Nucleic Acids Res.* **28**, 250–253.
11. Sali, A. and Kuriyan, J. (1999) Challenges at the frontiers of structural biology. *Trends Cell Biol.* **9**, M20–M24.
12. Montelione, G. T. and Anderson, S. (1999) Structural genomics: keystone for a Human Proteome Project. *Nat. Struct. Biol.* **6**, 11–12.
13. Sali, A. (1998) 100,000 protein structures for the biologist. *Nat. Struct. Biol.* **5**, 1029–1032.
14. Gerstein, M., Edwards, A., Arrowsmith, C. H., and Montelione, G. T. (2003) Structural genomics: current progress. *Science* **299**, 1663.
15. Brenner, S. E. (2000) Target selection for structural genomics. *Nat. Struct. Biol.* **7 Suppl.**, 967–969.
16. Vitkup, D., Melamud, E., Moult, J., and Sander, C. (2001) Completeness in structural genomics. *Nat. Struct. Biol.* **8**, 559–566.
17. Westbrook, J., Feng, Z., Jain, S., et al. (2002) The Protein Data Bank: unifying the archive. *Nucleic Acids Res.* **30**, 245–248.
18. Lo Conte, L., Brenner, S. E., Hubbard, T. J., Chothia, C., and Murzin, A. G. (2002) SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.* **30**, 264–267.
19. Holm, L. and Sander, C. (1999) Protein folds and families: sequence and structure alignments. *Nucleic Acids Res.* **27**, 244–247.
20. Orengo, C. A., Bray, J. E., Buchan, D. W., et al. (2002) The CATH protein family database: a resource for structural and functional annotation of genomes. *Proteomics* **2**, 11–21.
21. Pearl, F. M., Lee, D., Bray, J. E., et al. (2002) The CATH extended protein-family database: Providing structural annotations for genome sequences. *Protein Sci.* **11**, 233–244.
22. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.
23. Pearson, W. R. (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol.* **183**, 63–98.

24. Brenner, S. E., Chothia, C., and Hubbard, T. J. (1998) Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl. Acad. Sci. USA* **95**, 6073–6078.
25. Altschul, S. F., Madden, T. L., Schaffer, A. A., et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402.
26. Rychlewski, L., Jaroszewski, L., Li, W., and Godzik, A. (2000) Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci.* **9**, 232–241.
27. Sadreyev, R. and Grishin, N. (2003) COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. *J. Mol. Biol.* **326**, 317–336.
28. Panchenko, A. R. (2003) Finding weak similarities between proteins by sequence profile comparison. *Nucleic Acids Res.* **31**, 683–689.
29. Wallner, B., Fang, H., Ohlson, T., Frey-Skott, J., and Elofsson, A. (2004) Using evolutionary information for the query and target improves fold recognition. *Proteins* **54**, 342–50.
30. Teichmann, S. A., Chothia, C., Church, G. M., and Park, J. (2000) Fast assignment of protein structures to sequences using the intermediate sequence library PDB-ISL. *Bioinformatics* **16**, 117–124.
31. Li, W., Pio, F., Pawlowski, K., and Godzik, A. (2000) Saturated BLAST: an automated multiple intermediate sequence search used to detect distant homology. *Bioinformatics* **16**, 1105–1110.
32. Park, J., Karplus, K., Barrett, C., et al. (1998) Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.* **284**, 1201–1210.
33. Gerstein, M. (1998) Measurement of the effectiveness of transitive sequence comparison, through a third 'intermediate' sequence. *Bioinformatics* **14**, 707–714.
34. Pipenbacher, P., Schliep, A., Schneckener, S., et al. (2002) ProClust: improved clustering of protein sequences with an extended graph-based approach. *Bioinformatics* **18 Suppl. 2**, S182–S191.
35. Salamov, A. A., Suwa, M., Orengo, C. A., and Swindells, M. B. (1999) Combining sensitive database searches with multiple intermediates to detect distant homologues. *Protein Eng.* **12**, 95–100.
36. John, B. and Sali, A. (2004) Detection of homologous proteins by an intermediate sequence search. *Protein Sci.* **13**, 54–62.
37. Jones, D. T. (1997) Successful ab initio prediction of the tertiary structure of NK-lysin using multiple sequences and recognized supersecondary structural motifs. *Proteins Suppl.* **1**, 185–191.
38. Smith, T. F., Lo Conte, L., Bienkowska, J., et al. (1997) Current limitations to protein threading approaches. *J. Comput. Biol.* **4**, 217–225.
39. Torda, A. E. (1997) Perspectives in protein-fold recognition. *Curr. Opin. Struct. Biol.* **7**, 200–205.
40. Levitt, M. (1997) Competitive assessment of protein fold recognition and alignment accuracy. *Proteins Suppl.* **1**, 92–104.
41. David, R., Korenberg, M. J., and Hunter, I. W. (2000) 3D-1D threading methods for protein fold recognition. *Pharmacogenomics* **1**, 445–455.
42. Gough, J., Karplus, K., Hughey, R., and Chothia, C. (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J. Mol. Biol.* **313**, 903–919.
43. Jones, D. T. (1999) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.* **287**, 797–815.

44. Lindahl, E. and Elofsson, A. (2000) Identification of related proteins on family, superfamily and fold level. *J. Mol. Biol.* **295**, 613–625.
45. Balaji, S., Sujatha, S., Kumar, S. S., and Srinivasan, N. (2001) PALI-a database of Phylogeny and ALignment of homologous protein structures. *Nucleic Acids Res.* **29**, 61–65.
46. Navaratnam, N., Fujino, T., Bayliss, J., et al. (1998) *Escherichia coli* cytidine deaminase provides a molecular model for ApoB RNA editing and a mechanism for RNA substrate recognition. *J. Mol. Biol.* **275**, 695–714.
47. Reva, B., Finkelstein, A., and Topiol, S. (2002) Threading with chemostructural restrictions method for predicting fold and functionally significant residues: application to dipeptidylpeptidase IV (DPP-IV). *Proteins* **47**, 180–193.
48. Srinivasan, S., March, C. J., and Sudarsanam, S. (1993) An automated method for modeling proteins on known templates using distance geometry. *Protein Sci.* **2**, 277–289.
49. Sanchez, R. and Sali, A. (1997) Evaluation of comparative protein structure modeling by MODELLER-3. *Proteins Suppl.* **1**, 50–58.
50. Barton, G. J. (1996) Protein sequence alignment and database scanning. In: (Sternberg, M. J. E., ed.) *Protein Structure Prediction: A Practical Approach*, IRL Press at Oxford University Press, Oxford, UK.
51. Needleman, S. B. and Wunsch, C. D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453.
52. Sellers, P. H. (1974) Theory and computation of evolutionary distances. *Siam Journal on Applied Mathematics* **26**, 787–793.
53. Smith, T. F. and Waterman, M. S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197.
54. Henikoff, S. and Henikoff, J. G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* **89**, 10,915–10,919.
55. Dayhoff, M., Schwartz, R., and BC, O. (1978) A model of evolutionary change in proteins, 345–352, National Biomedical Research Foundation, Washington, DC.
56. Gribskov, M., McLachlan, A. D., and Eisenberg, D. (1987) Profile analysis: detection of distantly related proteins. *Proc. Natl. Acad. Sci. USA* **84**, 4355–4358.
57. Gribskov, M. (1994) Profile analysis. *Methods Mol. Biol.* **25**, 247–266.
58. Gribskov, M., Luthy, R., and Eisenberg, D. (1990) Profile analysis. *Methods Enzymol.* **183**, 146–159.
59. Karplus, K., Barrett, C., and Hughey, R. (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics* **14**, 846–856.
60. Marti-Renom, M. A., Madhusudhan, M. S., and Sali, A. (2004) Alignment of protein sequences by their profiles. *Protein Sci.* **13**(4), 1071–1087.
61. Higgins, D. G. and Sharp, P. M. (1988) CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene* **73**, 237–244.
62. Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680.
63. Pietrokovski, S. (1996) Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Res.* **24**, 3836–3845.
64. Jaroszewski, L., Rychlewski, L., and Godzik, A. (2000) Improving the quality of twilight-zone alignments. *Protein Sci.* **9**, 1487–1496.
65. Yona, G., Linial, N., and Linial, M. (2000) ProtoMap: automatic classification of protein sequences and hierarchy of protein families. *Nucleic Acids Res.* **28**, 49–55.
66. Yona, G. and Levitt, M. (2002) Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. *J. Mol. Biol.* **315**, 1257–1275.

67. Yona, G., Linial, N., and Linial, M. (1999) ProtoMap: automatic classification of protein sequences, a hierarchy of protein families, and local maps of the protein space. *Proteins* **37**, 360–378.
68. Yona, G. and Levitt, M. (2000) Towards a complete map of the protein space based on a unified sequence and structure analysis of all known proteins. *ISMB* **8**, 395–406.
69. Bowie, J. U., Luthy, R., and Eisenberg, D. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **253**, 164–170.
70. Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992) A new approach to protein fold recognition. *Nature* **358**, 86–89.
71. Godzik, A. and Skolnick, J. (1992) Sequence-structure matching in globular proteins: application to supersecondary and tertiary structure determination. *Proc. Natl. Acad. Sci. USA* **89**, 12,098–12,102.
72. Jones, D. T. (1997) Progress in protein structure prediction. *Curr. Opin. Struct. Biol.* **7**, 377–387.
73. Teodorescu, O., Galor, T., Pillardy, J., and Elber, R. (2004) Enriching the sequence substitution matrix by structural information. *Proteins* **54**, 41–8.
74. Luthy, R., Bowie, J. U., and Eisenberg, D. (1992) Assessment of protein models with three-dimensional profiles. *Nature* **356**, 83–85.
75. Shi, J., Blundell, T. L., and Mizuguchi, K. (2001) FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J. Mol. Biol.* **310**, 243–257.
76. Bates, P. A., Kelley, L. A., MacCallum, R. M., and Sternberg, M. J. (2001) Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. *Proteins Suppl.* **5**, 39–46.
77. Kelley, L. A., MacCallum, R. M., and Sternberg, M. J. (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.* **299**, 499–520.
78. Karchin, R., Cline, M., Mandel-Gutfreund, Y., and Karplus, K. (2003) Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry. *Proteins* **51**, 504–514.
79. Tang, K. S., Fersht, A. R., and Itzhaki, L. S. (2003) Sequential unfolding of ankyrin repeats in tumor suppressor p16. *Structure (Camb.)* **11**, 67–73.
80. Zhu, Z. Y., Sali, A., and Blundell, T. L. (1992) A variable gap penalty function and feature weights for protein 3-D structure comparisons. *Protein Eng.* **5**, 43–51.
81. Madhusudhan, M. S., Marti-Renom, M. A., Sanchez, R., and Sali, A. (2004) Variable gap penalty function for protein sequence—structure alignment. in preparation.
82. Madhusudhan, M. S., Marti-Renom, M. A., Eswar, N., and Sali, A. (2004) SALIGN: a comprehensive sequence/structure alignment algorithm. in preparation.
83. Venclovas, C. (2003) Comparative modeling in CASP5: progress is evident, but alignment errors remain a significant hindrance. *Proteins* **53 Suppl. 6**, 380–388.
84. John, B. and Sali, A. (2003) Comparative protein structure modeling by iterative alignment, model building and model assessment. *Nucleic Acids Res.* **31**, 3982–3992.
85. Marti-Renom, M. A., Madhusudhan, M. S., Fiser, A., Rost, B., and Sali, A. (2002) Reliability of assessment of protein structure prediction methods. *Structure* **10**, 435–440.
86. Browne, W. J., North, A. C. T., Phillips, D. C., et al. (1969) A possible three-dimensional structure of bovine lactalbumin based on that of hen's egg-white lysosyme. *J. Mol. Biol.* **42**, 65–86.
87. Sutcliffe, M. J., Haneef, I., Carney, D., and Blundell, T. L. (1987) Knowledge based modelling of homologous proteins, Part I: Three-dimensional frameworks derived from the simultaneous superposition of multiple structures. *Protein Eng.* **1**, 377–384.

88. Topham, C. M., McLeod, A., Eisenmenger, F., et al. (1993) Fragment ranking in modeling of protein structure. Conformationally constrained environmental amino acid substitution tables. *J. Mol. Biol.* **229**, 194–220.
89. Nagarajaram, H. A., Reddy, B. V., and Blundell, T. L. (1999) Analysis and prediction of inter-strand packing distances between beta-sheets of globular proteins. *Protein Eng.* **12**, 1055–1062.
90. Unger, R., Harel, D., Wherland, S., and Sussman, J. L. (1989) A 3D building blocks approach to analyzing and predicting structure of proteins. *Proteins* **5**, 355–373.
91. Bystroff, C. and Baker, D. (1998) Prediction of local structure in proteins using a library of sequence-structure motifs. *J. Mol. Biol.* **281**, 565–577.
92. Claessens, M., Van Cutsem, E., Lasters, I., and Wodak, S. (1989) Modelling the polypeptide backbone with 'spare parts' from known protein structures. *Protein Eng.* **2**, 335–345.
93. Holm, L. and Sander, C. (1991) Database algorithm for generating protein backbone and side-chain co-ordinates from a C alpha trace application to model building and detection of co-ordinate errors. *J. Mol. Biol.* **218**, 183–194.
94. Brucolieri, R. E. and Karplus, M. (1990) Conformational sampling using high-temperature molecular dynamics. *Biopolymers* **29**, 1847–1862.
95. van Gelder, C. W., Leusen, F. J., Leunissen, J. A., and Noordik, J. H. (1994) A molecular dynamics approach for the generation of complete protein structures from limited coordinate data. *Proteins* **18**, 174–185.
96. Levitt, M. (1992) Accurate modeling of protein conformation by automatic segment matching. *J. Mol. Biol.* **226**, 507–533.
97. Chinea, G., Padron, G., Hooft, R. W., Sander, C., and Vriend, G. (1995) The use of position-specific rotamers in model building by homology. *Proteins* **23**, 415–421.
98. Jones, T. A. and Thirup, S. (1986) Using known substructures in protein model building and crystallography. *EMBO J.* **5**, 819–822.
99. Havel, T. F. and Snow, M. E. (1991) A new method for building protein conformations from sequence alignments with homologues of known structure. *J. Mol. Biol.* **217**, 1–7.
100. Sali, A., Overington, J. P., Johnson, M. S., and Blundell, T. L. (1990) From comparisons of protein sequences and structures to protein modelling and design. *Trends Biochem. Sci.* **15**, 235–240.
101. Sali, A. and Blundell, T. L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815.
102. Fiser, A., Do, R. K., and Sali, A. (2000) Modeling of loops in protein structures. *Protein Sci.* **9**, 1753–1773.
103. Sali, A. and Overington, J. P. (1994) Derivation of rules for comparative protein modeling from a database of protein structure alignments. *Protein Sci.* **3**, 1582–1596.
104. MacKerell, A. D., Jr., Bashford, D., Bellott, M., et al. (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **102**, 3586–3616.
105. Brooks, B. R., Brucolieri, R. E., Olafson, B. D., et al. (1983) CHARMM: A program for macromolecular energy minimization and dynamics calculations. *J. Comp. Chem.* **4**, 187–217.
106. Braun, W. and Go, N. (1985) Calculation of protein conformations by proton-proton distance constraints. A new efficient algorithm. *J. Mol. Biol.* **186**, 611–626.
107. Clore, G. M., Brunger, A. T., Karplus, M., and Gronenborn, A. M. (1986) Application of molecular dynamics with interproton distance restraints to three-dimensional protein structure determination. A model study of crambin. *J. Mol. Biol.* **191**, 523–551.
108. Cohen, F. E. and Kuntz, I. D. (1989) Tertiary structure prediction. In: (Fasman, G. D., ed.) *Prediction of Protein Structure and the Principles of Protein Conformations*, Plenum, New York, NY: 647–705.

109. Aszodi, A. and Taylor, W. R. (1994) Secondary structure formation in model polypeptide chains. *Protein Eng.* **7**, 633–644.
110. Taylor, W. R. and Hatrick, K. (1994) Compensating changes in protein multiple sequence alignments. *Protein Eng.* **7**, 341–348.
111. Sippl, M. J. (1990) Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.* **213**, 859–883.
112. Sutcliffe, M. J., Dobson, C. M., and Oswald, R. E. (1992) Solution structure of neuronal bungarotoxin determined by two-dimensional NMR spectroscopy: calculation of tertiary structure using systematic homologous model building, dynamical simulated annealing, and restrained molecular dynamics. *Biochemistry* **31**, 2962–2970.
113. Boissel, J. P., Lee, W. R., Presnell, S. R., Cohen, F. E., and Bunn, H. F. (1993) Erythropoietin structure-function relationships. Mutant proteins that test a model of tertiary structure. *J. Biol. Chem.* **268**, 15,983–15,993.
114. Kabsch, W. and Sander, C. (1984) On the use of sequence homologies to predict protein structure: identical pentapeptides can have completely different conformations. *Proc. Natl. Acad. Sci. USA* **81**, 1075–1078.
115. Mezei, M. (1998) Chameleon sequences in the PDB. *Protein Eng.* **11**, 411–414.
116. Chothia, C. and Lesk, A. M. (1987) Canonical structures for the hypervariable regions of immunoglobulins. *J. Mol. Biol.* **196**, 901–917.
117. Brucolieri, R. E. and Karplus, M. (1987) Prediction of the folding of short polypeptide segments by uniform conformational sampling. *Biopolymers* **26**, 137–168.
118. Shenkin, P. S., Yarmush, D. L., Fine, R. M., Wang, H. J., and Levinthal, C. (1987) Predicting antibody hypervariable loop conformation. I. Ensembles of random conformations for ringlike structures. *Biopolymers* **26**, 2053–2085.
119. van Vlijmen, H. W. and Karplus, M. (1997) PDB-based protein loop prediction: parameters for selection and methods for optimization. *J. Mol. Biol.* **267**, 975–1001.
120. Deane, C. M. and Blundell, T. L. (2001) CODA: a combined algorithm for predicting the structurally variable regions of protein models. *Protein Sci.* **10**, 599–612.
121. Sibanda, B. L., Blundell, T. L., and Thornton, J. M. (1989) Conformation of beta-hairpins in protein structures. A systematic classification with applications to modelling by homology, electron density fitting and protein engineering. *J. Mol. Biol.* **206**, 759–777.
122. Chothia, C., Lesk, A. M., Tramontano, A., et al. (1989) Conformations of immunoglobulin hypervariable regions. *Nature* **342**, 877–883.
123. Rufino, S. D., Donate, L. E., Canard, L. H., and Blundell, T. L. (1997) Predicting the conformational class of short and medium size loops connecting regular secondary structures: application to comparative modelling. *J. Mol. Biol.* **267**, 352–367.
124. Oliva, B., Bates, P. A., Querol, E., Aviles, F. X., and Sternberg, M. J. (1997) An automated classification of the structure of protein loops. *J. Mol. Biol.* **266**, 814–830.
125. Ring, C. S., Kneller, D. G., Langridge, R., and Cohen, F. E. (1992) Taxonomy and conformational analysis of loops in proteins. *J. Mol. Biol.* **224**, 685–699.
126. Fidelis, K., Stern, P. S., Bacon, D., and Moult, J. (1994) Comparison of systematic search and database methods for constructing segments of protein structure. *Protein Eng.* **7**, 953–960.
127. Lessel, U. and Schomburg, D. (1994) Similarities between protein 3-D structures. *Protein Eng.* **7**, 1175–1187.
128. Moult, J. and James, M. N. (1986) An algorithm for determining the conformation of polypeptide segments in proteins by systematic search. *Proteins* **1**, 146–163.
129. Fine, R. M., Wang, H., Shenkin, P. S., Yarmush, D. L., and Levinthal, C. (1986) Predicting antibody hypervariable loop conformations. II: Minimization and molecular dynam-

- ics studies of MCPC603 from many randomly generated loop conformations. *Proteins* **1**, 342–362.
130. Ring, C. S., Sun, E., McKerrow, J. H., et al. (1993) Structure-based inhibitor design by using protein models for the development of antiparasitic agents. *Proc. Natl. Acad. Sci. USA* **90**, 3583–3587.
131. Abagyan, R. and Totrov, M. (1994) Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. *J. Mol. Biol.* **235**, 983–1002.
132. Collura, V., Higo, J., and Garnier, J. (1993) Modeling of protein loops by simulated annealing. *Protein Sci.* **2**, 1502–1510.
133. Higo, J., Collura, V., and Garnier, J. (1992) Development of an extended simulated annealing method: application to the modeling of complementary determining regions of immunoglobulins. *Biopolymers* **32**, 33–43.
134. Zheng, Q., Rosenfeld, R., Vajda, S., and DeLisi, C. (1993) Determining protein loop conformation using scaling-relaxation techniques. *Protein Sci.* **2**, 1242–1248.
135. Koehl, P. and Delarue, M. (1995) A self consistent mean field approach to simultaneous gap closure and side-chain positioning in homology modelling. *Nat. Struct. Biol.* **2**, 163–170.
136. Samudrala, R. and Moult, J. (1998) A graph-theoretic algorithm for comparative modeling of protein structure. *J. Mol. Biol.* **279**, 287–302.
137. Xiang, Z., Soto, C. S., and Honig, B. (2002) Evaluating conformational free energies: the colony energy and its application to the problem of loop prediction. *Proc. Natl. Acad. Sci. USA* **99**, 7432–7437.
138. Still, W. C., Tempczyk, A., Hawley, R. C., and Hendrickson, T. (1990) Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* **112**, 6127–6129.
139. Ghosh, A., Rapp, C.S., and Friesner, R.A. (1998) Generalized born model based on a surface integral formulation. *J. Phys. Chem. B* **102**, 10,983–10,990.
140. de Bakker, P. I., DePristo, M. A., Burke, D. F., and Blundell, T. L. (2003) *Ab initio* construction of polypeptide fragments: Accuracy of loop decoy discrimination by an all-atom statistical potential and the AMBER force field with the Generalized Born solvation model. *Proteins* **51**, 21–40.
141. DePristo, M. A., de Bakker, P. I., Lovell, S. C., and Blundell, T. L. (2003) *Ab initio* construction of polypeptide fragments: efficient generation of accurate, representative ensembles. *Proteins* **51**, 41–55.
142. Felts, A. K., Gallicchio, E., Wallqvist, A., and Levy, R. M. (2002) Distinguishing native conformations of proteins from decoys with an effective free energy estimator based on the OPLS all-atom force field and the Surface Generalized Born solvent model. *Proteins* **48**, 404–422.
143. Jacobson, M., Pincus, D., Rapp, C. S., et al. (2004) A hierarchical approach to all-atom loop prediction. *Proteins* **55**, 351–367.
144. Melo, F., Sanchez, R., and Sali, A. (2002) Statistical potentials for fold assessment. *Protein Sci.* **11**, 430–448.
145. Chothia, C. and Lesk, A. M. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823–826.
146. Janin, J. and Chothia, C. (1978) Role of hydrophobicity in the binding of coenzymes. Appendix. Translational and rotational contribution to the free energy of dissociation. *Biochemistry* **17**, 2943–2948.
147. Ponder, J. W. and Richards, F. M. (1987) Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.* **193**, 775–791.

148. Tuffery, P., Etchebest, C., Hazout, S., and Lavery, R. (1991) A new approach to the rapid determination of protein side chain conformations. *J. Biomol. Struct. Dyn.* **8**, 1267–1289.
149. Mendes, J., Baptista, A. M., Carrondo, M. A., and Soares, C. M. (1999) Improved modeling of side-chains in proteins with rotamer-based methods: a flexible rotamer model. *Proteins* **37**, 530–543.
150. Dunbrack, R. L., Jr. and Cohen, F. E. (1997) Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci.* **6**, 1661–1681.
151. Dunbrack, R. L. and Karplus, M. (1993) Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *J. Mol. Biol.* **230**, 543–574.
152. Xiang, Z. and Honig, B. (2001) Extending the accuracy limits of prediction for side-chain conformations. *J. Mol. Biol.* **311**, 421–430.
153. Desjarlais, J. R. and Handel, T. M. (1999) Side-chain and backbone flexibility in protein core design. *J. Mol. Biol.* **290**, 305–318.
154. De Filippis, V., Sander, C., and Vriend, G. (1994) Predicting local structural changes that result from point mutations. *Protein Eng.* **7**, 1203–1208.
155. Chung, S. Y. and Subbiah, S. (1996) How similar must a template protein be for homology modeling by side-chain packing methods? *Pac. Symp. Biocomput.* 126–141.
156. Cregut, D., Liautard, J. P., and Chiche, L. (1994) Homology modelling of annexin I: implicit solvation improves side-chain prediction and combination of evaluation criteria allows recognition of different types of conformational error. *Protein Eng.* **7**, 1333–1344.
157. Canutescu, A. A., Shelenkov, A. A., and Dunbrack, R. L., Jr. (2003) A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci.* **12**, 2001–2014.
158. Eisenmenger, F., Argos, P., and Abagyan, R. (1993) A method to configure protein side-chains from the main-chain trace in homology modelling. *J. Mol. Biol.* **231**, 849–860.
159. Lee, G. M., Varma, A., and Palsson, B. O. (1991) Application of population balance model to the loss of hybridoma antibody productivity. *Biotechnol. Prog.* **7**, 72–75.
160. Holm, L. and Sander, C. (1992) Fast and simple Monte Carlo algorithm for side chain optimization in proteins: application to model building by homology. *Proteins* **14**, 213–223.
161. Lasters, I. and Desmet, J. (1993) The fuzzy-end elimination theorem: correctly implementing the side chain placement algorithm based on the dead-end elimination theorem. *Protein Eng.* **6**, 717–722.
162. Looger, L. L. and Hellinga, H. W. (2001) Generalized dead-end elimination algorithms make large-scale protein side-chain structure prediction tractable: implications for protein design and structural genomics. *J. Mol. Biol.* **307**, 429–445.
163. Hwang, J. K. and Liao, W. F. (1995) Side-chain prediction by neural networks and simulated annealing optimization. *Protein Eng.* **8**, 363–370.
164. Koehl, P. and Delarue, M. (1994) Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. *J. Mol. Biol.* **239**, 249–275.
165. Bower, M. J., Cohen, F. E., and Dunbrack, R. L., Jr. (1997) Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. *J. Mol. Biol.* **267**, 1268–1282.
166. Petrella, R. J., Lazaridis, T., and Karplus, M. (1998) Protein sidechain conformer prediction: a test of the energy function. *Fold Des.* **3**, 353–377.
167. Jacobson, M. P., Kaminski, G. A., Friesner, R. A., and Rapp, C. S. (2002) Force field validation using protein side chain prediction. *J. Phys. Chem. B* **106**, 11,673–11,680.
168. Liang, S. and Grishin, N. V. (2002) Side-chain modeling with an optimized scoring function. *Protein Sci.* **11**, 322–331.
169. Zemla, A., Venclovas, Moult, J., and Fidelis, K. (2001) Processing and evaluation of predictions in CASP4. *Proteins* **45 Suppl. 5**, 13–21.

170. Fischer, D., Elofsson, A., Rychlewski, L., et al. (2001) CAFASP2: The second critical assessment of fully automated structure prediction methods. *Proteins* **45 Suppl. 5**, 171–183.
171. Bujnicki, J. M., Elofsson, A., Fischer, D., and Rychlewski, L. (2001) LiveBench-1: continuous benchmarking of protein structure prediction servers. *Protein Sci.* **10**, 352–361.
172. Eyrich, V. A., Marti-Renom, M. A., Przybylski, D., et al. (2001) EVA: continuous automatic evaluation of protein structure prediction servers. *Bioinformatics* **17**, 1242–1243.
173. Koh, I. Y., Eyrich, V. A., Marti-Renom, M. A., et al. (2003) EVA: Evaluation of protein structure prediction servers. *Nucleic Acids Res.* **31**, 3311–3315.
174. Srinivasan, N. and Blundell, T. L. (1993) An evaluation of the performance of an automated procedure for comparative modelling of protein tertiary structure. *Protein Eng.* **6**, 501–512.
175. Coutsias, E. A., Seok, C., Jacobson, M. P., and Dill, K. A. (2004) A kinematic view of loop closure. *J. Comput. Chem.* **25**, 510–528.
176. Barton, G. J. and Sternberg, M. J. (1987) A strategy for the rapid multiple alignment of protein sequences. Confidence levels from tertiary structure comparisons. *J. Mol. Biol.* **198**, 327–337.
177. Taylor, W. R., Flores, T. P., and Orengo, C. A. (1994) Multiple protein structure alignment. *Protein Sci.* **3**, 1858–1870.
178. Laskowski, R. A., MacArthur, A. G., and Thornton, J. M. (1998) PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **26**, 283–291.
179. Wilson, C., Gregoret, L. M., and Agard, D. A. (1993) Modeling side-chain conformation for homologous proteins using an energy-based rotamer search. *J. Mol. Biol.* **229**, 996–1006.
180. Sanchez, R. and Sali, A. (1998) Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. *Proc. Natl. Acad. Sci. USA* **95**, 13,597–13,602.
181. Sippl, M. J. (1993) Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. *J. Comput. Aided Mol. Des.* **7**, 473–501.
182. Pawlowski, K., Bierzynski, A., and Godzik, A. (1996) Structural diversity in a family of homologous proteins. *J. Mol. Biol.* **258**, 349–366.
183. Laskowski, R. A., MacArthur, M. W., and Thornton, J. M. (1998) Validation of protein models derived from experiment. *Curr. Opin. Struct. Biol.* **8**, 631–639.
184. Laskowski, R. A., Rullmann, J. A., MacArthur, M. W., Kaptein, R., and Thornton, J. M. (1996) AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J. Biomol. NMR* **8**, 477–486.
185. Oldfield, T. J. (1992) SQUID: a program for the analysis and display of data from crystallography and molecular dynamics. *J. Mol. Graph.* **10**, 247–252.
186. Hooft, R. W., Vriend, G., Sander, C., and Abola, E. E. (1996) Errors in protein structures. *Nature* **381**, 272.
187. Topham, C. M., Srinivasan, N., Thorpe, C. J., Overington, J. P., and Kalsheker, N. A. (1994) Comparative modelling of major house dust mite allergen Der p I: structure validation using an extended environmental amino acid propensity table. *Protein Eng.* **7**, 869–894.
188. Melo, F. and Feytmans, E. (1998) Assessing protein structures with a non-local atomic interaction energy. *J. Mol. Biol.* **277**, 1141–1152.
189. Zhou, H. and Zhou, Y. (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci.* **11**, 2714–2726.
190. Miwa, J. M., Ibanez-Tallón, I., Crabtree, G. W., et al. (1999) lynx1, an endogenous toxin-like modulator of nicotinic acetylcholine receptors in the mammalian CNS. *Neuron* **23**, 105–114.

191. Johnson, M. S., Srinivasan, N., Sowdhamini, R., and Blundell, T. L. (1994) Knowledge-based protein modelling. *CRC Crit. Rev. Biochem. Mol. Biol.* **29**, 1–68.
192. Orengo, C. A., Michie, A. D., Jones, S., et al. (1997) CATH—a hierachic classification of protein domain structures. *Structure* **5**, 1093–1108.
193. Matsumoto, R., Sali, A., Ghildyal, N., Karplus, M., and Stevens, R. L. (1995) Packaging of proteases and proteoglycans in the granules of mast cells and other hematopoietic cells. A cluster of histidines on mouse mast cell protease 7 regulates its binding to heparin serglycin proteoglycans. *J. Biol. Chem.* **270**, 19,524–19,531.
194. Xu, L. Z., Sanchez, R., Sali, A., and Heintz, N. (1996) Ligand specificity of brain lipid-binding protein. *J. Biol. Chem.* **271**, 24,711–24,719.
195. Vakser, I. A. (1995) Protein docking for low-resolution structures. *Protein Eng.* **8**, 371–377.
196. Thompson, J. D., Plewniak, F., and Poch, O. (1999) BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics* **15**, 87–88.
197. Pearl, F. M., Bennett, C. F., Bray, J. E., et al. (2003) The CATH database: an extended protein family resource for structural and functional genomics. *Nucleic Acids Res.* **31**, 452–455.
198. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Wheeler, D. L. (2004) GenBank: update. *Nucleic Acids Res.* **32 Database issue**, D23–D26.
199. Lin, J., Qian, J., Greenbaum, D., et al. (2002) GeneCensus: genome comparisons in terms of metabolic pathway activity and protein family sharing. *Nucleic Acids Res.* **30**, 4574–4582.
200. Bourne, P. E., Addess, K. J., Bluhm, W. F., et al. (2004) The distribution and query systems of the RCSB Protein Data Bank. *Nucleic Acids Res.* **32 Database issue**, D223–D225.
201. Brenner, S. E., Barken, D., and Levitt, M. (1999) The PRESAGE database for structural genomics. *Nucleic Acids Res.* **27**, 251–253.
202. Andreeva, A., Howorth, D., Brenner, S. E., et al. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.* **32 Database issue**, D226–D229.
203. Boeckmann, B., Bairoch, A., Apweiler, R., et al. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 365–370.
204. Alexandrov, N. N., Nussinov, R., and Zimmer, R. M. (1996) Fast protein fold recognition via sequence to structure alignment and contact capacity potentials. *Pac. Symp. Biocomput.* 53–72.
205. Godzik, A., Kolinski, A., and Skolnick, J. (1992) Topology fingerprint approach to the inverse protein folding problem. *J. Mol. Biol.* **227**, 227–238.
206. Rost, B. and Liu, J. (2003) The PredictProtein server. *Nucleic Acids Res.* **31**, 3300–3304.
207. Flockner, H., Braxenthaler, M., Lackner, P., et al. (1995) Progress in fold recognition. *Proteins* **23**, 376–386.
208. Mallick, P., Weiss, R., and Eisenberg, D. (2002) The directional atomic solvation energy: an atom-based potential for the assignment of protein sequences to known folds. *Proc. Natl. Acad. Sci. USA* **99**, 16,041–16,046.
209. Gough, J. and Chothia, C. (2002) SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucleic Acids Res.* **30**, 268–272.
210. Worley, K. C., Culpepper, P., Wiese, B. A., and Smith, R. F. (1998) BEAUTY-X: enhanced BLAST searches for DNA queries. *Bioinformatics* **14**, 890–891.
211. Tatusova, T. A. and Madden, T. L. (1999) BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol. Lett.* **174**, 247–250.
212. Henikoff, J. G., Pietrovskii, S., McCallum, C. M., and Henikoff, S. (2000) Blocks-based methods for detecting protein homology. *Electrophoresis* **21**, 1700–1706.

213. Corpet, F. (1988) Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res.* **16**, 10,881–10,890.
214. Ye, Y., Jaroszewski, L., Li, W., and Godzik, A. (2003) A segment alignment approach to protein comparison. *Bioinformatics* **19**, 742.
215. Karplus, K., Karchin, R., Draper, J., et al. (2003) Combining local-structure, fold-recognition, and new fold methods for protein structure prediction. *Proteins* **53 Suppl. 6**, 491–496.
216. Notredame, C., Higgins, D. G., and Heringa, J. (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302**, 205–217.
217. Edgar, R. C. and Sjolander, K. (2004) COACH: profile-profile alignment of protein families using hidden Markov models. *Bioinformatics* **20(8)**, 1309–1318.
218. Eswar, N., John, B., Mirkovic, N., et al. (2003) Tools for comparative protein structure modeling and analysis. *Nucleic Acids Res.* **31**, 3375–3380.
219. Pearson, W. R. (1995) Comparison of methods for searching protein sequence databases. *Protein Sci.* **4**, 1145–1160.
220. Abagyan, R., Frishman, D., and Argos, P. (1994) Recognition of distantly related proteins through energy calculations. *Proteins* **19**, 132–140.
221. Schwede, T., Kopp, J., Guex, N., and Peitsch, M. C. (2003) SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Res.* **31**, 3381–3385.
222. Vriend, G. (1990) WHAT IF: a molecular modeling and drug design program. *J. Mol. Graph.* **8**, 56.
223. Colovos, C. and Yeates, T. O. (1993) Verification of protein structures: patterns of nonbonded atomic interactions. *Protein Sci.* **2**, 1511–1519.
224. Pontius, J., Richelle, J., and Wodak, S. J. (1996) Deviations from standard atomic volumes as a quality measure for protein crystal structures. *J. Mol. Biol.* **264**, 121–136.
225. Hooft, R. W., Sander, C., and Vriend, G. (1996) Verification of protein structure: side-chain planarity. *J. Appl. Crystallog.* 714–716.
226. Moult, J., Fidelis, K., Zemla, A., and Hubbard, T. (2001) Critical assessment of methods of protein structure prediction (CASP): Round IV. *Proteins* **45 Suppl. 5**, 2–7.
227. Kahsay, R. Y., Wang, G., Dongre, N., Gao, G., and Dunbrack, R. L., Jr. (2002) CASA: a server for the critical assessment of protein sequence alignment accuracy. *Bioinformatics* **18**, 496–497.
228. Livingstone, C. D. and Barton, G. J. (1993) Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Comput. Appl. Biosci.* **9**, 745–756.
229. Beckmann, R., Spahn, C. M., Eswar, N., et al. (2001) Architecture of the protein-conducting channel associated with the translating 80S ribosome. *Cell* **107**, 361–372.

Classification of Protein Sequences and Structures

S. Rackovsky

1. Introduction

With the advent of genome-scale sequencing and high-throughput structure determination, workers in bioinformatics and proteomics are faced with a very practical problem—the need to organize a vast and growing body of data, so that information of interest and important correlations are readily accessible.

Progress on several problems of fundamental biological importance depends on the development of appropriate classification methodology. The first is the delineation of functional relationships between sequences. The completion of any new genome yields a flood of sequences of proteins whose structure and function are unknown. The most reliable method of assigning a function to a new protein is to demonstrate appropriate relationships to known molecules. A second problem is the prediction of protein structure. Although significant progress has been made in *ab initio* methods for protein structure prediction, the most reliable methods remain those in which a model can be constructed based on a demonstrated homology to a protein of known structure. A third problem is the elucidation of the factors that determine fold choice in proteins (1). While this problem is very fundamental, its solution has important practical implications for genome analysis (2).

One can classify either structures (3–6) or sequences (7–11). The two classifications do not give the same results. Sequences that have no demonstrable homology are observed to occur in the same fold. In fact, a general problem in the classification of proteins is that the structure one observes for the protein universe depends on the features that one uses to classify the molecules. It is therefore useful to ask how the structure of the space of proteins depends on classification criteria.

In this chapter we will concisely survey some topics of current importance in the protein classification problem. First a broad outline will be given of the methods that have been developed to study this problem. We will then highlight recent work on some underlying problems.

2. General Methodology

The first requirement for the classification of a set of objects is a function that defines a degree of difference between the objects. Once such a function has been defined, it is possible to organize the differences in a distance matrix, each element of which gives the degree of dissimilarity between two members of the set.

From: *The Proteomics Protocols Handbook*
Edited by: J. M. Walker © Humana Press Inc., Totowa, NJ

Once a distance matrix is available, it is useful to think of the group of objects as occupying a set of points in a space. This space will, in general, be of high dimensionality, and the specified set of distances may not obey the constraints that characterize a Euclidean space. The study of such systems is the province of multivariate analysis, which provides several approaches that can be used to delineate the structure of this type of space. Among those that have proven useful in various protein contexts are clustering (both hierarchical and partitioning methods), principal component analysis (5), and graph-theoretical methods (3,12).

This analysis reveals the underlying organization of the set of proteins under consideration. The organization must be rationalized by relating it to some set of known properties of the proteins. Properties that may be reflected in the organization of a given space include biochemical function, fold family, and evolutionary (or at least homology) relationship.

The structure revealed by the foregoing approach can be thought of as a network of nearest-neighbor relationships between the proteins in the database. The physical significance of the network depends on the meaning of the distances between molecules. Furthermore, the structure of a network encodes important information about the process that gave rise to it. It is therefore important to develop precise descriptions of networks. The statistical mechanics of networks has been studied extensively in recent work (13), and several general classes of network have been delineated, with strikingly different mathematical properties, carrying correspondingly diverse physical implications. A surprising result of this work is the demonstration of the widespread occurrence of scale-free networks in diverse systems. Scale-free networks are characterized by a power-law distribution of the number of links experienced by network nodes (13). This indicates the presence in these systems of a small number of nodes with many links and a large number with only a few, a fact that has significant implications both for the mechanism of network generation (which is specific to a particular system) and for the resistance of the network to disruption (a general property of scale-free systems). These properties of scale-free networks have been reviewed in detail (13).

3. Protein Sequences

Sequence alignment is central to bioinformatics. It forms the basis for homology searching, in which the structure and/or function of a protein whose sequence alone is known is surmised by detecting similarity with other molecules of known structure and function. The objective of alignment is to blur the distinction between sequence and structure, by making it possible to relate three-dimensional structures using one-dimensional information.

The alignment of sequence pairs can be used to produce quantitative descriptors of the similarity between them, and can therefore be used to produce distance matrices and to delineate the organization of sequence space. It is clear that the structure of the space (or equivalently of the network of protein sequences) depends entirely on the detailed characteristics of the individual modules that together comprise the alignment process.

Sequence alignment is a modular task (14–16). The modules are:

- *An equivalence matrix.* An objective function is necessary to determine the quality of an alignment. This function is calculated as the sum of similarities between residues that are

declared to be in some sense equivalent. The investigator must therefore choose criteria for assigning a degree of similarity to amino acids in different sequences. These criteria are determined by those amino acid characteristics that are considered to be important in the context of the specific problem. Historically, the view was taken that one measured the probability that a given amino acid would be replaced by another in an evolutionary/mutation process. This led to the development of the PAM250 matrices (17). The use of a large number of blocks of aligned sequence segments to count replacements gave the Block Substitution Matrices (BLOSUM) (18). Other criteria are possible. Those based on physical rather than evolutionary properties are of particular interest, since they don't presuppose specific models or degrees of evolutionary distance (19). These criteria include similarity with respect to a specified physical property, or similarity in some structural sense (20–24). It should be noted that alignment can be performed using either similarity or distance matrices. The advantage of the former is that they make possible local alignment (25), which is the preferred approach to database searching.

- *A set of gap parameters.* It is often found that better correspondence between two sequences is obtained if account is taken of the presence of insertions and deletions in one or both. Such indels must be accounted for in calculating the alignment objective function, and this is usually done by means of a gap penalty function (14,26). A common form of the penalty is the affine penalty function, in which a fixed penalty is counted for the initiation of a gap, and a length-dependent propagation penalty is added to account for the size of the gap.
- *An alignment algorithm.* A method is required by which alignments between sequences are generated, a corresponding figure of merit calculated from the objective function, and the best alignment (and some of the runners-up (27–31)) selected. There are two general types of alignment—global, in which entire sequences are aligned, and local, in which a region of one sequence is matched with a region of another. The classical algorithm is some variant of dynamic programming (32). More recent methods include FASTA (33), the Basic Local Alignment Search Tool (BLAST) (34), hidden Markov models (35,36), and variants of these (37).

Simultaneous consideration of multiple sequences introduces a set of additional problems into the alignment algorithm, which are rapidly exacerbated as the number of sequences grows. This problem has been addressed in several ways. Methods of calculating limited subsets of the dynamic programming matrix have been developed (38), and genetic algorithms have been investigated (39).

- *A criterion for success.* A broadly accepted standard for successful sequence alignment is given by structure alignment (40–42). If a sequence alignment method, applied to two proteins with known structure, produces an alignment similar to that resulting from an independent, structure-based identification of equivalent residues, the sequence alignment method is considered able to identify biophysically (or evolutionarily) relevant correspondences in the two sequences.

An alternative, alignment-free approach to defining a distance between sequences is based on counting N -residue fragments (43,44). In this approach, a fingerprint for each sequence is provided by the distribution of frequencies of N -mers, and a distance function is constructed that measures the similarity of two frequency distributions. This approach has certain advantages over the alignment approach:

- Because normalized distributions are compared, it is straightforward to define a distance between sequences of different molecular weight.

- No gap parameters need be defined, since the comparison method automatically includes the effects of insertions and deletions. On the other hand, this approach does not identify residues that may be functionally or structurally equivalent, since no alignment is produced. It has been shown (44) that the distances produced by this approach are equivalent to those arising from alignment-based methods.

There have been several classifications of large databases of protein sequences (7–11,45). These have been directed almost entirely toward the goal of structure and function elucidation, and little if any attention has been paid to the network organization of the space. This is an important distinction because, as we remarked above, and will note again in connection with structure classification, the overall structure of a protein network is believed to carry important information about evolutionary processes.

In this section we wish to devote particular attention to some recent studies of an important but neglected aspect of sequence classification—the problem of information loss in sequence comparison. This point directly concerns the first module in the sequence alignment algorithm, and is equally relevant to alignment-free distance methods. We note that the construction of an equivalence matrix between amino acids is closely related to the use of reduced amino acid alphabets in protein studies. Adoption of an equivalence matrix is an implicit declaration that two amino acids, hitherto considered to be informatically distinct, are to some degree interchangeable. This step leads to the loss of information, and information loss must inevitably distort the structure of the sequence space/network. It is therefore of extreme interest to ask whether reduced alphabets (or amino acid equivalency matrices) can be constructed that are optimized to retain maximal information.

The first point that must be decided is what kind of information one wishes to retain. Since we are considering information content in sequences, the natural choice is to maximize the retention of structural information. In recent publications, we have developed methods for constructing reduced alphabets that encode the maximum possible amount of local structural information. In our initial studies (46) on the structural information content of sequence representations, we used information theoretical and statistical methods, and protein sequence and structure data, to demonstrate the following points:

- It is possible to quantitatively calculate the amount of structural information made available by knowledge of local sequence alone. This number depends on the representations used for both sequence and structure.
- A contracted amino acid alphabet of any specified size can be constructed in a manner that retains maximal structural information. The loss of information resulting from optimized alphabet contraction was calculated, and it was shown that, in practical applications, this loss is offset by statistical improvements resulting from the greatly decreased number of distinct sequence fragments.
- The optimal mapping of the 20 amino acids onto the reduced alphabet depends on the structure representation used. Structurally optimized alphabets as a function of size were produced for both the Database of Secondary Structure in Proteins (DSSP) and the Generalized Bond Matrix (GBM) (α -carbon backbone) representations. Examination of details of the clustering optimization process reveals that the former representation is able to detect only low-resolution properties of the amino acids, related to secondary structure

preference and hydrophobicity. The GBM representation, on the other hand, gives reduced alphabets that reflect subtle conformational nuances of the amino acids.

In subsequent work (47,48), we have extended this approach to consider the constraints imposed on the optimization of representations by the finite size of the databases from which we derive information. A serious problem in this regard is the presence of rare sequence fragments, for which it is not possible to construct statistically meaningful structure distributions. This problem was addressed by representing the structure distribution associated with a given sequence fragment as the superposition of two distributions—one specific to the sequence in question, and the other a background distribution with lower sequence specificity. The relative weights of these two components in the final structure distribution depend on the number of rare sequence fragments in the data set. When there are few rare fragments, the actual observed distribution is heavily weighted. When there are many, the background distribution is weighted more heavily, reflecting the lack of sequence-specific information.

A Monte Carlo procedure was developed that makes it possible to simultaneously optimize distribution weights and amino acid clustering for a given alphabet size. The information-theoretical and statistical machinery developed in the course of these studies give optimized reduced alphabets, and associated structural distributions, which have the following characteristics:

1. They compensate for the scarcity of structural data.
2. They use multiresidue (context specific) information.
3. They contain the maximum amount of local structure information that the underlying data set allows.

It was demonstrated that the maximum structural information is encoded in sequence fragments six residues long. This length scale represents the optimal compromise between additional sequence information, intrinsic in longer fragments, and statistical deterioration due to the finite size of the protein database. The distribution of the amount of structural information encoded in local sequences was analyzed, and it was shown that there is at least a 35% variance in structural entropy among different sequence fragments.

The result of these investigations is a set of contracted amino acid alphabets optimized to encode the maximum possible amount of structural information available in several commonly used structural representations. By construction, the structure of a sequence space based on these alphabets should represent structural relationships between sequences as accurately as possible. In work by Kuznetsov, Solis, and Rackovsky currently in progress, these alphabets are being incorporated into both alignment-based and alignment-free sequence distance functions, and the effect of this optimization on sequence classification is being explored.

4. Protein Structures

Structure comparison, like sequence comparison, is one of the cornerstones of protein bioinformatics. The need for appropriate tools is clear in both the experimental and theoretical domains. It is of obvious interest to compare a newly determined structure to those that have already been solved, in order to correctly trace evolutionary and functional relationships between molecules. The efficacy of a structure prediction

algorithm can be evaluated only by comparing the predicted and actual structures of test molecules.

An elegant and perceptive review of conformational comparison methods has been given by Brown et al. (49). The earliest approaches to structure comparison were based on optimal superposition. One structure is translated and rotated relative to another, fixed structure until a chosen figure of merit is optimized. In most cases, the figure of merit is the root mean square deviation (RMSD) between corresponding atoms of the two structures. This approach continues to be widely used in various incarnations. There are, however, a number of alternative metrics that can be used, and these have been compared recently by Wallin et al. (50).

There are several important concerns that should be noted here. The first is a general problem in the comparison of structures: The result that one obtains for a structure comparison depends critically on the method that is adopted, because structure comparison is a length-scale-dependent problem. The meaning of this point can be made clear by a simple example. Consider a protein consisting of two domains, and imagine generating an alternative conformation of the molecule by rotating one of the domains relative to the other around a single connecting bond. An attempt to compare the resulting conformation to the original conformation of the protein by optimized superposition will give a poor result, because it is no longer possible to bring the two domains into superposition simultaneously. Superposition algorithms are designed to operate on a length scale approximating the size of the molecules being compared. Imagine, however, comparing the two conformations using an algorithm that compares *local* conformations along the two chains. The result will be a chain plot that indicates identity of the two conformations everywhere except at the single bond around which rotation occurred. A *qualitatively* different answer is generated by a method that considers the problem at a different length scale. This is a feature of the comparison problem that investigators must keep in mind.

The second point that must be made is specific to the superposition method. While the method gives meaningful results when the molecules being compared are reasonably similar, it is somewhat difficult to know how to interpret results for the comparison of proteins that differ significantly in molecular weight and/or structure. What are the corresponding atoms in two unrelated structures? What is the meaning of a RMSD for such molecule pairs? It is in fact a question, as Godzik has pointed out (51), whether there is a unique answer to the optimal superposition problem.

In order to address these problems, other methods of comparison have been developed. An early alternative was based on the application of differential geometric (DG) methods to the description of protein conformation (52–59). It was first demonstrated that a DG-based representation of chain structure can be defined, and that it is possible to base a distance function on that representation, which describes local differences in chain folding. The output of the algorithm is a chain plot in which conformational differences at corresponding sites are revealed. This approach makes possible the detailed comparison of chain fragments of equal length, and can be used to compare chains of different lengths using a moving-window method. It was then demonstrated that the distribution of differential-geometric parameters can be used to define a length-independent fingerprint for any chain of known structure, and that these fingerprints can be used to compare the structures of proteins of different molecular weights. The

method was used to carry out an all-against-all comparison of a small group of structures, giving a sparse description of the structure of structure space. This was the earliest quantitative comparison of protein structures, and the earliest attempt to quantitatively delineate the characteristics of structure space, known to this author.

A limiting characteristic of the DG approach is the fact that it operates on a single, defined length scale within the molecule—the 4- α -carbon scale. All parameters are calculated from the coordinates of successive fragments of that size, and all information is therefore limited to structure at that scale. Attempts to create a representation in which the length scale is a definable parameter led to the development of the GBM representation (3). This representation of backbone structure is far more flexible than the differential geometric representation, in that fragments of any length, defined using any chemical or virtual bonds of interest, can be used as a basis for structure description and comparison. It shares with the DG representation several advantages over superposition methods. Both allow the definition of normalized (molecular-weight independent) structure fingerprints, making it possible to compare chains of arbitrarily different molecular weight. Both representations share with the alignment-free sequence comparison methods discussed above the characteristic that the presence of insertions and deletions is accounted for automatically, without the necessity for defining gap initiation and propagation penalties. At the same time, sequence-ordering information present in superposition algorithms is lost, or at least obscured, in the distribution-based methods. (There is, however, some evidence that, if the sequence fragments considered are sufficiently long, the correlations necessary to reconstruct the sequence from the fragment distribution are, in fact, preserved [Pevzner, P., personal communication].) The GBM representation was used (3,60) to carry out a detailed classification of a database of structures that well represented the known structure universe at the time. This involved the use of techniques from graph theory to study clustering in the space. One hundred and twenty-three structures were classified, and it was shown that structure space can be represented as a non-uniform continuum of structures, grading from all-helical structures at one edge of the space to sheet/barrel structures at the other. Details of the distribution of structures were investigated, as were the effect of the length scale and resolution of the chosen representation on the structure of the space.

A method for comparing structures based on intramolecular distance matrices was developed by Yee and Dill (4), and used to reanalyze structure space. Although this method is very different from the GBM approach, the anatomy of structure space that it revealed is substantially similar to that discussed in our own work (3). Holm and Sander (61,62) have also used distance matrices to compare structures.

An excellent summary and review of earlier studies of protein classification, together with a discussion of the coarse-grained statistical properties of protein space, has been given by Brenner et al. (63).

Nussinov, Wolfson, and collaborators have developed an approach to structure comparison based on tools of pattern recognition. The method is based on a hashing algorithm first applied to computer vision studies, and is able to carry out sequence-order-independent comparisons. This method has been used (64) to construct a non-redundant dataset of structures and investigate characteristics of the resulting space. In later work, the same group has used hashing methods to carry out multiple alignments (65,66) and detect common structural motifs (65).

Recently, Hou et al. (5) have revisited the structure of fold space using a method based on the distance-matrix alignment (DALI) comparison algorithm (61), which is distance-matrix based. Using a factor analysis of the resulting protein–protein distances, they constructed a picture in which the high-dimensional fold space was contracted to its three most significant dimensions. The folds cluster into disjoint regions corresponding to the classical low-resolution definition of fold types— α , β , α/β , and $\alpha + \beta$. The authors report that domain size is an important determinant of the structure of the space.

The statistical significance of a given comparison is an important point to address. Levitt and Gerstein (67) have given a general framework for the statistical validation of both sequence and structure comparisons.

A number of workers have investigated the network properties of structure space, which carries important implications for fold evolution. One approach to this question is to analyze the distribution of domain family sizes, and this has been studied by Qian et al. (68), Kuznetsov (69), and Karev et al. (70). The network of domain relationships has also been constructed directly by Dokholyan et al. (12) using DALI-generated distances. All evidence suggests that the distribution of contact numbers follows the power law characteristic of scale-free networks. Several dynamic simulations of the evolutionary process have been developed (12,70,71) that give this type of scale-free behavior in an model proteome.

Perhaps the central problem in protein science is the fact that sequence and structure classifications *do not* give the same picture of protein space. Various manifestations of this fact have been known for many years. It is widely recognized, for example, that some of the more common architectures are adopted by large groups of protein sequences, many of which exhibit no detectable mutual sequence similarity. An understanding of the mechanism that determines architecture choice will lead directly to a solution of the classical folding problem—the prediction of structure from sequence. This would therefore seem to be a problem worth study.

In recent work we have addressed this question (1). Our approach is based on the following observations:

- The set of proteins folding to a specified architecture frequently includes molecules that are not only unrelated by homology, but also differ widely in molecular weight.
- The choice of architecture must be determined by physical properties of the amino acids in the sequence.

The second of these points suggests that the architecture signal is expressed in some pattern of physical characteristics. The first suggests that the signal must scale with sequence length.

In order to investigate this problem, we therefore need to express protein sequences in terms of amino acid properties. There are many property sets available, and an arbitrary choice of properties leads to a twofold problem: the set chosen can be simultaneously incomplete and correlated. This problem has been solved by Scheraga and collaborators (72,73), who carried out a factor analysis of all available sets of amino acid attributes. They showed that the entire attribute data set can be described by property factors. Four major factors correspond essentially to individual amino acid properties, and the remaining six are superpositions of a limited number of properties. The 10 factors together carry 86% of the variance for the entire dataset. In mathematical terms,

this result means that the physics of the amino acids can be embodied in a set of 20 10-vectors, each of which gives the weights of the 10 factors for a particular amino acid. It follows that an N -residue protein sequence can be described by a set of 10 N -number strings, each of which traces the value of a particular property factor along the chain.

The next step is to construct a database of proteins suitable for the problem. Our approach is to assemble sets of proteins that fold to a common architecture but exhibit low sequence homology. We chose two architectures that are sufficiently populated that statistically meaningful samples of this type can be constructed—the trios phosphate isomerase (TIM) barrel and immunoglobulin folds. For each of these folds, an ensemble of sequences was chosen with pairwise similarities well below the homology limit.

Having written the amino acid sequence in a property-related numerical form, we wish to extract scalable signals that can be associated with protein architecture. We therefore carried out a Fourier analysis of the property strings for each protein in a specified architecture group. Note that the Fourier transform of a property string, for any wave number, is a function of the entire string. A consequence of this fact is that chain length is not a relevant variable in Fourier space, and the Fourier power spectra of chains of different sequence length can be directly compared. The Fourier analysis was followed by signal averaging over all proteins in the architecture group, which enabled us to distinguish Fourier components that are common to all members of the group from those that are characteristic of specific sequences.

It is important to ask whether the common Fourier components detected are statistically significant. In order to address this concern, randomized protein sequence groups were generated, by independently permuting the sequences of each of the proteins in the original architecture group. The entire Fourier analysis/signal averaging process was repeated on the permuted sequence groups. This was iterated 10,000 times for each architecture group, and each Fourier component arising from the actual sequence was compared to the average Fourier coefficient and standard deviation arising from the ensemble of random sequence groups. Only those Fourier components of the actual sequences that exceeded the average by two standard deviations were regarded as significant.

It was found (*I*, and Rackovsky, work in progress) that Fourier components that satisfy this requirement do indeed exist. A particularly dramatic result is observed in the TIM barrel group, in which a composite power spectrum signal was found at $k = 21$, which is 18σ above the average. A set of signals in the range $5\text{--}6\sigma$ was also found in the immunoglobulin group. A particularly fascinating insight into the mechanism of architecture selection emerges when we ask in which physical properties these signals are expressed. It is found that, while essentially all the proteins in a given architecture group exhibit statistically significant signals at the values of k identified by the signal-averaging procedure, these signals are expressed in *different* properties in the various proteins of the group. This suggests that an architecture can be generated by a well-defined set of periodicities, but that these periodicities can be expressed in a wide variety of physical properties. This constitutes a degeneracy in the architecture code.

The existence and characteristics of the architectural signals provide an understanding of certain fundamental observations about protein architecture and folding. The fact that proteins with no apparent sequence homology fold to common architectures (**74**) is an immediate consequence of the degeneracy of architecture signals, which

guarantees that there are many, dissimilar ways to produce a given architecture. It has also been noted (75–79) that proteins with similar architecture but no mutual homology fold with similar rates. A connection between this observation and the properties of architecture signals is readily made. The wavelength associated with a sinusoidal signal of wave number k in a sequence of length N is N/k . The sequence is composed of a set of k segments of this length, each of which contains a region in which the associated physical property of the amino acids is strongly expressed, flanked by regions in which it is weakly expressed. Note that, while the relevant physical property and the length of the segments differ in the various proteins of an architectural group, the *number* of segments is the same in all the sequences. These observations are consistent with a scenario (80,81) in which folding is governed by a number of early nucleation events, distributed over the entire sequence, each of which takes place in a relatively short, localized region—the segments delimited by the architectural signal. In each protein of a particular architecture group, the nature of these events is determined by the properties in which the folding signal is expressed. This suggestion is supported by the experimental demonstration (82) that proteins of similar architecture can fold by different mechanisms. It is possible (1), but not mandatory, that the segments defined by the architectural signals might be correlated with structures visible in the native fold. This possibility, and other implications of the present results, will be explored in forthcoming work.

These results suggest an alternate view of the classification problem. The reason that sequence and structure classification reveal different, parallel universes is that sequence classification, as currently practiced, is based on “incorrect” parameters. If one takes the reasonable (and widely held) view (40–42) that structure classification is the more fundamental process, it becomes clear that we should be searching for those sequence-related variables that give the closest correspondence possible between the two protein spaces. The approach we have just outlined, in looking for physical signals that encode architecture in sequence, represent a step in this direction, and away from a search for codes based on residue identity.

Acknowledgments

Our work in these areas was supported by Grant LM06789 from the National Library of Medicine of the National Institutes of Health. The author would like to acknowledge the contributions of Drs. Igor Kuznetsov and Armando Solis to the work summarized herein.

References

1. Rackovsky, S. (1998) “Hidden” sequence periodicities and protein architecture. *Proc. Nat. Acad. Sci. USA* **95**, 8580–8584.
2. Gerstein, M. (1997) A structural census of genomes: comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure. *J. Mol. Biol.* **274**, 562–576.
3. Rackovsky, S. (1990) Quantitative organization of the known protein X-ray structures. I. Methods and short length-scale results. *Proteins: Struct. Funct. Genet.* **7**, 378–402.
4. Yee, D. P. and Dill, K. A. (1993) Families and the structural relatedness among globular proteins. *Prot. Sci.* **2**, 884–899.
5. Hou, J., Sims, G. E., Shang, C., and Kim, S.-H. (2003) A global representation of the protein fold space. *Proc. Nat. Acad. Sci. USA* **100**, 2386–2390.

6. Holm, L. and Sander, C. (1997) Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acid Res.* **25**, 231–234.
7. Gonnet, G. H., Cohen, M. A., and Benner, S. A. (1992) Exhaustive matching of the entire protein sequence database. *Science* **256**, 1443–1445.
8. Linial, M., Linial, N., Tishby, N., and Yona, G. (1997) Global self-organization of all known protein sequences reveals inherent biological signatures. *J. Mol. Biol.* **268**, 539–556.
9. Gracy, J. and Argos, P. (1998) Automated protein sequence database classification. I. Integration of compositional similarity search, local similarity search, and multiple sequence alignment. *Bioinformatics* **14**, 164–173.
10. Wang, H.-C., Dopazo, J., De La Fraga, L. G., Zhu, Y.-P., and Carazo, J. M. (1998) Self-organizing tree-growing network for the classification of protein sequences. *Prot. Sci.* **7**, 2613–2622.
11. Yona, G., Linial, N., and Linial, M. (1999) ProtoMap: automatic classification of protein sequences, a hierarchy of protein families, and local maps of the protein space. *Proteins: Struct. Funct. Genet.* **37**, 360–378.
12. Dokholyan, N. V., Shakhnovich, B., and Shakhnovich, E. I. (2002) Expanding protein universe and its origin from the biological big bang. *Proc. Nat. Acad. Sci. USA* **99**, 14,132–14,136.
13. Albert, R. and Barabási, A.-L. (2002) Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74**, 47–97.
14. Myers, E. W. (1995) Seeing conserved signals: using algorithms to detect similarities between biosequences. In: (Lander, E.S., and Waterman, M.S., eds.) *Calculating the Secrets of Life* National Academy, Washington, DC.
15. Barton, G. J. (1998) Protein sequence alignment techniques. *Acta Cryst.* **D54**, 1139–1146.
16. Smith, T. F. (1999) The art of matchmaking: sequence alignment methods and their structural implications. *Structure* **7**, R7–R12.
17. Dayhoff, M. O. and Eck, R. V. (1996) *Atlas of Protein Sequence and Structure*. Volume 2, NBRF Press, Silver Spring, MD.
18. Henikoff, S. and Henikoff, J. (1992) Amino acid substitution matrices from protein blocks. *Proc. Nat. Acad. Sci. USA* **89**, 10915–10919.
19. Altschul, S. F. (1993) A protein alignment scoring system sensitive at all evolutionary distances. *J. Mol. Evol.* **36**, 290–300.
20. Naor, D., Fischer, D., Jernigan, R. L., Wolfson, H. J., and Nussinov, R. (1996) Amino acid pair interchanges at spatially conserved locations. *J. Mol. Biol.* **256**, 924–938.
21. Russell, R. B., Saqi, M. A. S., Sayle, R. A., Bates, P. A., and Sternberg, M. J. E. (1997) Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation. *J. Mol. Biol.* **269**, 423–439.
22. Johnson, M. S. and Overington, J. P. (1993) A structural basis for sequence comparison: an evaluation of scoring methodologies. *J. Mol. Biol.* **233**, 716–738.
23. Prlic, A., Domingues, F. S., and Sippl, M. J. (2000) Structure-derived substitution matrices for alignment of distantly related sequences. *Protein Eng.* **13**, 545–550.
24. Blake, J. D. and Cohen, F. E. (2001) Pairwise sequence alignment below the twilight zone. *J. Mol. Biol.* **307**, 721–735.
25. Smith, T. F. and Waterman, M. S. (1981) Identification of common molecular subsequences. *J. Mol. Evol.* **147**, 195–197.
26. Altschul, S. F. (1998) Generalized affine gap costs for protein sequence alignment. *Proteins: Struct. Funct. Genet.* **32**, 88–96.
27. Argos, P., Vingron, M., and Vogt, G. (1991) Protein sequence comparisons: methods and significance. *Protein Eng.* **4**, 375–383.
28. Saqi, M. and Sternberg, M. (1991) A simple method to generate non-trivial alternate alignments of protein sequences. *J. Mol. Biol.* **219**, 727–732.

29. Zuker, M. (1991) Suboptimal sequence alignment in molecular biology: alignment with error analysis. *J. Mol. Biol.* **221**, 403–420.
30. Agarwal, P. and States, D. (1996) A Bayesian evolutionary distance for parametrically aligned sequences. *J. Comput. Biol.* **3**, 1–17.
31. Vingron, M. (1996) Near-optimal sequence alignment. *Curr. Opin. Struct. Biol.* **6**, 346–352.
32. Horowitz, E. and Sahni, S. (1978) *Fundamentals of Computer Algorithms*, Computer Science Press, New York, NY: 198–247.
33. Pearson, W., Lipman, D. (1988) Improved tools for biological sequence comparison. *Proc. Nat. Acad. Sci. USA* **85**, 2444–2448.
34. Altschul, S., Gish, W., Miller, W., Myers, E., and Lipman, D. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.
35. Krogh, A., Brown, M., Mian, J., Sjölander, K., and Haussler, D. (1994) Hidden Markov models in computational biology: applications to protein modeling. *J. Mol. Biol.* **235**, 1501–1531.
36. Eddy, S. (1996) Hidden Markov models. *Curr. Opin. Struct. Biol.* **6**, 361–365.
37. Bucher, P. and Hoffman, K. (1996) A sequence similarity algorithm based on a probabilistic interpretation of an alignment scoring system. In: (States, D., Gaasterland, T., Hunter, L., and Smith, R., eds.) *ISMB-4 AAAI*, Menlo Park, CA.
38. Lipman, D. J., Altschul, S. F., and Kececioglu, J. (1989) A tool for multiple sequence alignment. *Proc. Nat. Acad. Sci. USA* **86**, 4412–4415.
39. Notredame, C. and Higgins, D. G. (1996) SAGA: sequence alignment by genetic algorithm. *Nucleic Acids Res.* **24**, 1515–1524.
40. Brenner, S. E., Chothia, C., and Hubbard, T. J. P. (1998) Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Nat. Acad. Sci. USA* **95**, 6073–6078.
41. Sauder, J. M., Arthur, J. W., and Dunbrack, Jr., R. L. (2000). Large-scale comparison of protein sequence alignment algorithms with structure alignments. *Proteins: Struct. Funct. Genet.* **40**, 6–22.
42. Panchenko, A. R. and Bryant, S. H. (2002) A comparison of position-specific score matrices based on sequence and structure alignments. *Protein Sci.* **11**, 361–370.
43. Blaisdell, B. E. (1986) A measure of the similarity of sets of sequences not requiring sequence alignment. *Proc. Nat. Acad. Sci. USA* **83**, 5155–5159.
44. Blaisdell, B. E. (1991) Average values of a dissimilarity measure not requiring sequence alignment are twice the averages of conventional mismatch counts requiring sequence alignment for a variety of computer-generated model systems. *J. Mol. Evol.* **32**, 521–528.
45. Yona, G. and Levitt, M. (2000) A unified sequence-structure classification of protein sequences: combining sequence and structure in a map of the protein space. In *Proceedings of the Fourth Annual Conference on Computational Molecular Biology*, Tokyo, pp. 308–317.
46. Solis, A. D. and Rackovsky, S. (2000) Optimized representations and maximal information in proteins. *Proteins: Struct. Funct. Genet.* **38**, 149–164.
47. Solis, A. D. and Rackovsky, S. (2002) Optimally informative backbone structural propensities in proteins. *Proteins: Struct. Funct. Genet.* **48**, 463–486.
48. Solis, A. D. (2002) *Structural Information From Local Sequence of Proteins and DNA*, Thesis, Mt. Sinai School of Medicine of New York University, pp. 148–191.
49. Brown, N. P., Orengo, C. P., and Taylor, W. R. (1996) A protein structure comparison methodology. *Computers Chem.* **20**, 359–380.
50. Wallin, S., Farwer, J., and Bastolla, U. (2003) Testing similarity measures with continuous and discrete protein models. *Proteins: Struct. Funct. Genet.* **50**, 144–157.

51. Godzik, A. (1996) The structural alignment between two proteins: is there a unique answer? *Protein Sci.* **5**, 1325–1338.
52. Rackovsky, S. and Scheraga, H. A. (1978) Differential geometry and polymer conformations. I. On the comparison of polymer conformations. *Macromolecules* **11**, 1168–1174.
53. Rackovsky, S. and Scheraga, H. A. (1980) Differential geometry and polymer conformations. II. Mathematical considerations and a conformational distance function. *Macromolecules* **13**, 1440–1453.
54. Rackovsky, S. and Scheraga, H. A. (1980) Intermolecular anti-parallel beta sheet: comparison of predicted and observed conformations of gramicidin S. *Proc. Nat. Acad. Sci. USA* **77**, 6965–6967.
55. Rackovsky, S. and Scheraga, H. A. (1981) Differential geometry and polymer conformations. III. Nearest-neighbor correlations and medium-range structure. *Macromolecules* **14**, 1259–1269.
56. Rackovsky, S. and Scheraga, H. A. (1982) Differential geometry and polymer conformations. IV. Conformational and nucleation properties of individual amino acids. *Macromolecules* **15**, 1340–1346.
57. Rackovsky, S. and Scheraga, H. A. (1984) Differential geometry and protein folding. *Acc. Chem. Res.* **17**, 209–214.
58. Rackovsky, S. and Goldstein, D. A. (1987) Differential geometry and protein conformation. V. Medium-range conformational influence of the individual amino acids. *Biopolymers* **26**, 1163–1187.
59. Rackovsky, S. and Goldstein, D. A. (1988) Protein comparison and classification: a differential geometric approach. *Proc. Natl. Acad. Sci. USA* **85**: 777–781.
60. Rackovsky, S. (1990) Quantitative classification of the known protein x-ray structures. *Polymer Preprints* **31**, 205.
61. Holm, L. and Sander, C. (1993) Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **223**, 123–138.
62. Holm, L. and Sander, C. (1996) Mapping the protein universe. *Science* **273**, 595–602.
63. Brenner, S. E., Chothia, C., and Hubbard, T. J. P. (1997) Population statistics of protein structures: lessons from structural classifications. *Curr. Opin. Struct. Biol.* **7**, 369–376.
64. Fischer, D., Tsai, C.-J., Nussinov, R., and Wolfson, H. (1995) A 3D sequence-independent representation of the protein data bank. *Protein Eng.* **8**, 981–997.
65. Leibowitz, N., Fligelman, Z., Nussinov, R., and Wolfson, H. (2001) Automated multiple structure alignment and detection of a common motif. *Proteins: Struct. Funct. Genet.* **43**, 235–245.
66. Dror, O., Benyamin, H., Nussinov, R., and Wolfson, H. (2003) MASS: multiple structure alignment by secondary structures. *Bioinformatics* **19 Suppl. 1**, i95–i104.
67. Levitt, M. and Gerstein, M. (1998) A unified statistical framework for sequence comparison and structure comparison. *Proc. Nat. Acad. Sci. USA* **95**, 5913–5920.
68. Qian, J., Luscombe, N. M., and Gerstein, M. (2001). Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. *J. Mol. Biol.* **313**, 673–681.
69. Kuznetsov, V.A. (2002) Statistics of the numbers of transcripts and protein sequences encoded in the genome. In: (Zhang, W., and Shmulevich, I., eds.) *Computational and Statistical Approaches to Genomics* Kluwer, Boston, MA: 125–171.
70. Karev, G. P., Wolf, Y. I., Rzhetsky, A. Y., Berezovskaya, F. S., and Koonin, E. V. (2003) The structure of the protein universe and genome evolution. In: Galperin, M. Y., and Koonin, E. V., eds.) *Computational Genomics From Sequence to Function* Horizon, Amsterdam.
71. Yanai, I., Camacho, C., and DeLisi, C. (2000) Predictions of gene family distributions in microbial genomes: evolution by gene duplication and modification. *Phys. Rev. Lett.* **85**, 2641–2644.

72. Kidera, A., Konishi, Y., Oka, M., Ooi, T., and Scheraga, H. A. (1985) Statistical analysis of the physical properties of the 20 naturally occurring amino acids. *J. Prot. Chem.* **4**, 23–55.
73. Kidera, A., Konishi, Y., Ooi, T., and Scheraga, H. A. (1985) Relation between sequence similarity and structural similarity in proteins. Role of important properties of amino acids. *J. Prot. Chem.* **4**, 265–297.
74. Yang, A.-S. and Honig, B. (2000) An integrated approach to the analysis and modeling of protein sequences and structures. II. On the relationship between sequence and structural similarity for proteins that are not obviously related in sequence. *J. Mol. Biol.* **301**, 679–689.
75. Alm, E. and Baker, D. (1999) Matching theory and experiment in protein folding. *Curr. Opin. Struct. Biol.* **9**, 189–196.
76. Shea, J. E., Onuchic, J. N., and Brooks, C. L. III. (1999) Exploring the origins of topological frustration: design of a minimally frustrated model of fragment B of protein A. *Proc. Nat. Acad. Sci. USA* **96**, 12,512–12,517.
77. Onuchic, J. N., Nymeyer, H., Garcia, A. E., Chaine, J., and Soccia, N. D. (2000) The energy landscape theory of protein folding: insights in folding mechanism and scenarios. *Adv. Protein Chem.* **53**, 87–152.
78. Micheletti, C., Banavar, J. R., Maritan, A., and Seno, F. (1999) Protein structures and optimal folding from a geometrical variational principle. *Phys. Rev. Lett.* **82**, 3372–3375.
79. Abkevich, V., Gutin, A., and Shakhnovich, E. (1994) Specific nucleus as the transition state for protein folding: evidence from the lattice model. *Biochemistry* **33**, 10,026–10,036.
80. Baldwin, R. L. (2001) Folding concensus? *Nat. Struct. Biol.* **8**, 92–94.
81. Fersht, A. R. (2000) Transition-state structure as a unifying basis in protein-folding mechanisms: contact order, chain topology, stability, and the extended nucleus mechanism. *Proc. Nat. Acad. Sci. USA* **97**, 1525–1529.
82. Burns, L. L., Dalessio, P. M., and Ropson, I. J. (1998) Folding mechanism of three structurally similar b-sheet proteins. *Proteins: Struct. Funct. Genet.* **33**, 107–188.

How to Use Protein 1-D Structure Predicted by PROFphd

Burkhard Rost

1. Introduction

The abbreviations used in this chapter are as follows:

- 1-D structure: one-dimensional structure, i.e., any structural feature that describes single residues, such as protein sequence or string of secondary structure and solvent accessibility assignments per residue.
- 3-D structure: three-dimensional coordinates of protein structure.
- EVA: server automatically evaluating structure prediction methods (1–3).
- META-PP: Internet service allowing access to a variety of bioinformatics tools through a single interface (4).
- PDB: Protein Data Bank of experimentally determined 3-D structures of proteins (5).
- PHDhtm: profile-based neural network prediction of transmembrane helices (6–8).
- PHDpsi: divergent profile (PSI-BLAST) based neural network prediction (9).
- PP (PredictProtein): Internet server for protein sequence analysis and protein structure prediction (7,10,11).
- PROFphd: advanced profile-based neural network prediction of secondary structure (PROFsec) and solvent accessibility (PROFacc) (11).
- SWISS-PROT: data base of protein sequences (12).
- Notations used:
 - Secondary structure: H = helix; E = strand; L = other.
 - Solvent accessibility: e = exposed ($\geq 16\%$ relative accessible surface); b = buried ($< 16\%$).
 - Transmembrane helices: T = transmembrane; N = globular.

No general prediction of 3-D structure from sequence, yet. The hypothesis that the 3-D structure of a protein (“the fold”) is fully determined by its sequence has been verified for many proteins (13). Although proteins such as chaperones often play an important role in folding (14,15), it is still generally assumed that the native structure is at a free-energy minimum (16,17). In principle, we should then be able to predict 3-D structure from physico-chemical principles (18,19). In practice, two obstacles have so far prevented accurate structure predictions for proteins—namely, the inaccuracy in experimentally determining the basic parameters such as the dielectric constant at a given position in the structure, and the limited computing resources (20,21). Hence, the only successful structure prediction tools are knowledge-based, using a combination of statistical theory and empirical rules. Although the field of protein structure prediction has advanced significantly over the last decade, we still cannot generally predict structure from sequence. Although the best current methods now get some of the aspects of

the fold sometimes somehow right (22–24), it remains unclear whether or not today's level of accuracy suffices for experimental biologists to benefit from such predictions. In fact, the only field that has undoubtedly profited from what is now referred to as "novel fold predictions" is comparative modeling (25,26), i.e., the prediction of 3-D structures for proteins with significant levels of sequence similarity to experimentally known structures.

Structure predictions in 1-D increasingly accurate and important. An extreme simplification of the prediction problem is to project 3-D structure onto 1-D strings of structural assignments. Examples are the assignments of a secondary structure state or a value for solvent-accessible surface to each residue; such strings of per-residue assignments are essentially 1-D. Methods predicting 1-D structure have been improved significantly over the last decade (27–31). The major key to this breakthrough was enabled by the increasing number of experimental high-resolution structures and by the wealth of information about protein evolution contained in ever growing databases. Evolutionary information is powerful because patterns of amino acid substitutions within protein families are highly specific for the 3-D structure of that family, as well as for aspects of function. Advance in computational methods, in particular in the field of artificial intelligence, created the means needed to explore these biological data. Improved 1-D predictions have been embedded into automatic meta-methods. Examples are target selection in structural genomics, methods predicting higher-dimensional aspects of structure, aspects of function, and more sensitive database search methods. Possibly more importantly, biologists have benefited from 1-D predictions to guide their experimental design, as amply demonstrated in thousands of publications (to give only a few recent examples from a variety of journals: [32–69]). One example of applications is the use of 1-D predictions to guide structure determination, in particular for chain tracing (an early example of this was the amazing GroES structure from the group of the late Paul Sigler [70]) and nuclear magnetic resonance (NMR) assignments. I find the GroES example particularly intriguing, since the most essential ingredient for making PROFphd successful is the availability of a great variety of experimental high-resolution structures. In other words, the example illustrates a cycle: experimental structures allow developing methods that contribute—albeit incrementally—to determining more high-resolution structures that in turn improve the methods. In this contribution, I also point to the increasing examples of another cycle confined to bioinformatics: 1-D predictions are used as input for methods predicting other aspects of structure, the results of which can then be used to improve the basic 1-D prediction to begin another cycle. An example is improving alignments of helical transmembrane proteins (71). Computational biologists and bioinformaticians increasingly need to master the whole spectrum of advanced tools in order to develop state-of-the-art methods.

Here, I describe PROFphd, a system combining three prediction methods that use evolutionary information as input to neural network systems to predict secondary structure (PROFsec), residue solvent accessibility (PROFacc), and transmembrane helices (PHDhtm). PROFphd constitutes a substantial improvement over the PHD methods that were state-of-the-art in the last century (1,2,7,30,72–76). Additionally, I illustrate some possibilities and limitations in practical applications of these methods. All methods are available through an automatic prediction server (PredictProtein [77]) that has

handled over a million requests from over 10,000 researchers. The programs are also available for expert computational biologists to run on their own machines (see **Subheading 3.3.** for details).

2. Materials

Accurate and diverse multiple alignments yield best predictions. The first step in PROFphd predictions is to generate multiple sequence alignments; these alignments are post-processed, translated into profiles, and then fed into the system of neural networks (**Fig. 1**). By default, PROFphd as run by the PredictProtein server (77) now uses PSI-BLAST alignments (78). Expert refinements of automatic alignments improve predictions (**Fig. 1**, and **Notes 1–3**).

Complex system of interconnected modules for multiple levels of computation. The PROFphd system processes the input information on multiple levels (**Fig. 2**). For all three methods (PROFsec, PROFacc, PHDhtm), the first level is a feed-forward neural network with three layers of units (input, hidden, and output). Input to this first-level sequence-to-structure network consists of two contributions: local information taken from a window of 13–21 sequence-consecutive residues, and global information reflecting average characteristics of the entire protein (**Fig. 2**). Output of the first-level network is the 1-D structural state of the residue at the center of the input window. For PROFsec and PHDhtm, the next level is a structure-to-structure network (discussed later). The output from PROFsec and PROFacc is combined in a next level with the original sequence information. All three methods then average over independently trained networks (jury decision). The final levels for PROFsec and PROFacc are simple filters; for PHDhtm, it is a dynamic programming-based filter (discussed later).

Number of output units is problem specific. Three output units code for secondary structure: one for helix (H; corresponding to states H, G, and I in DSSP [79]), one for strand (E; E and B in the Database of Secondary Structure in Proteins [DSSP]), and one for all other (L; S, T, and “ ” in DSSP). Two output units capture transmembrane predictions: one for residues in transmembrane helices, the other for non-membrane-bound residues. For solvent accessibility, the output coding is not as straightforward. First, the Connolly surface for residue solvent accessibility (80) is normalized to relative accessibility (observed accessibility divided by maximal accessibility of a given residue type [74]) to enable a comparison between residues of different sizes.

Second, the relative accessibility is projected onto ten states such that the nth state describes the following bin for percentage relative accessibility REL10(n):

$$(n-1)^2 = \text{REL10}(n) < n^2 \quad (1)$$

For example, the fourth output unit encodes relative solvent accessibility levels between 9% and 16%. Relative accessibility is filtered through a square-root function because it is more important to distinguish between two residues exposed to 0 or 16% (one is partially exposed, the other clearly buried) than it is to distinguish between two residues of 81 and 100% accessibility (both are undoubtedly exposed). The predicted relative solvent accessibility is re-converted into a prediction for the square Ångstrøm accessibility.

Better segment prediction by structure-to-structure networks. The output coding for the second-level network is identical to the one for the first (**Fig. 2**). The dominant

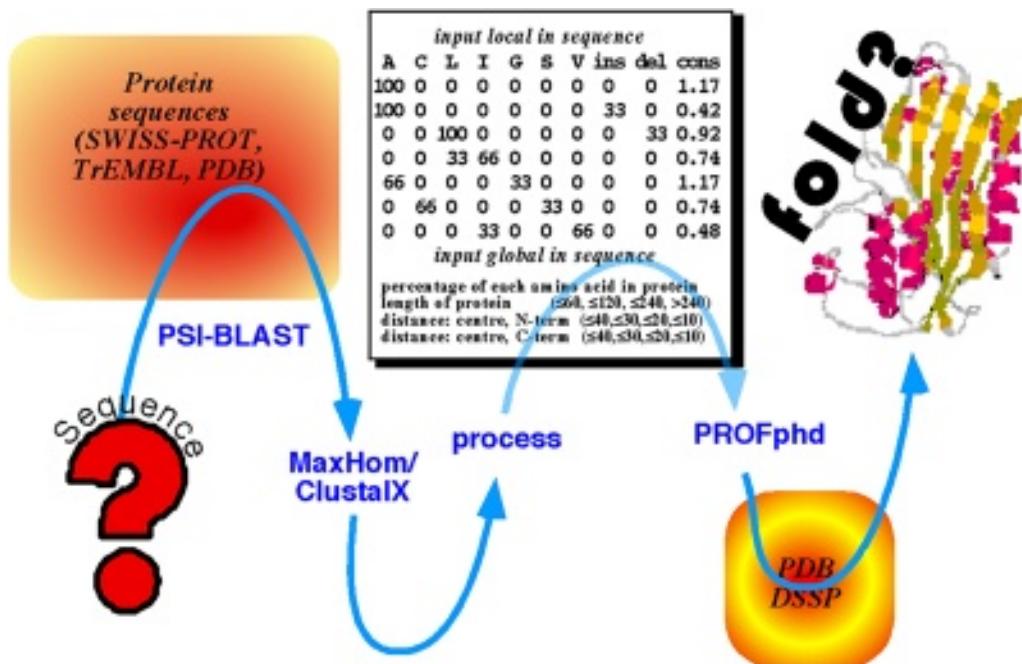


Fig. 1. (See Introduction for abbreviations.) From sequence to prediction. First, we maintain a nonidentical merger of the more reliable protein sequence databases locally (BIG = PDB [5] + SWISS-PROT [12] + TrEMBL [12]). We filter this database at a 98% identity cut-off (172) and mark low-complexity residues by the SEG method (173). Then we search by three PSI-BLAST (78) iterations in this subset of BIG_98_SEG (E values <10-10 for including sequences into the iterated profiles, and <10-3 for the alignment reported). Finally, we jump-start a single step of PSI-BLAST with the resulting profile against the full, unmarked BIG. Alignments can be improved through the dynamic programming alignment methods ClustalIX (174) or Maxhom (175). The final alignment is processed (removal of extreme redundancy, conversion to profiles) and fed into the PROFphd methods. A subsequent comparison of the predicted 1-D strings of accessibility and secondary structure can be used to unravel distant fold similarities (117,118,120,176–183).

input contribution to the second-level structure-to-structure network is the output of the first-level sequence-to-structure network. The reason for introducing a second level is the following. Networks are trained by changing the connections between the units such that the error is reduced for each of the examples successively presented to the network during training. The examples are chosen at random. Therefore, the examples chosen at time step t and at time step $t + 1$ are usually not adjacent in sequence. This implies that the network cannot learn that, e. g., helices contain at least three residues. The second-level structure-to-structure network correlates sequence-consecutive residues with the effect that predicted secondary structure segments have length distributions similar to the ones observed. For PHDhtm, the structure-to-structure network overshoots: membrane helices connected by frequently occurring short loops (81) are merged into overly long single membrane segments; this problem is partially addressed by a final step of dynamic programming (discussed later).

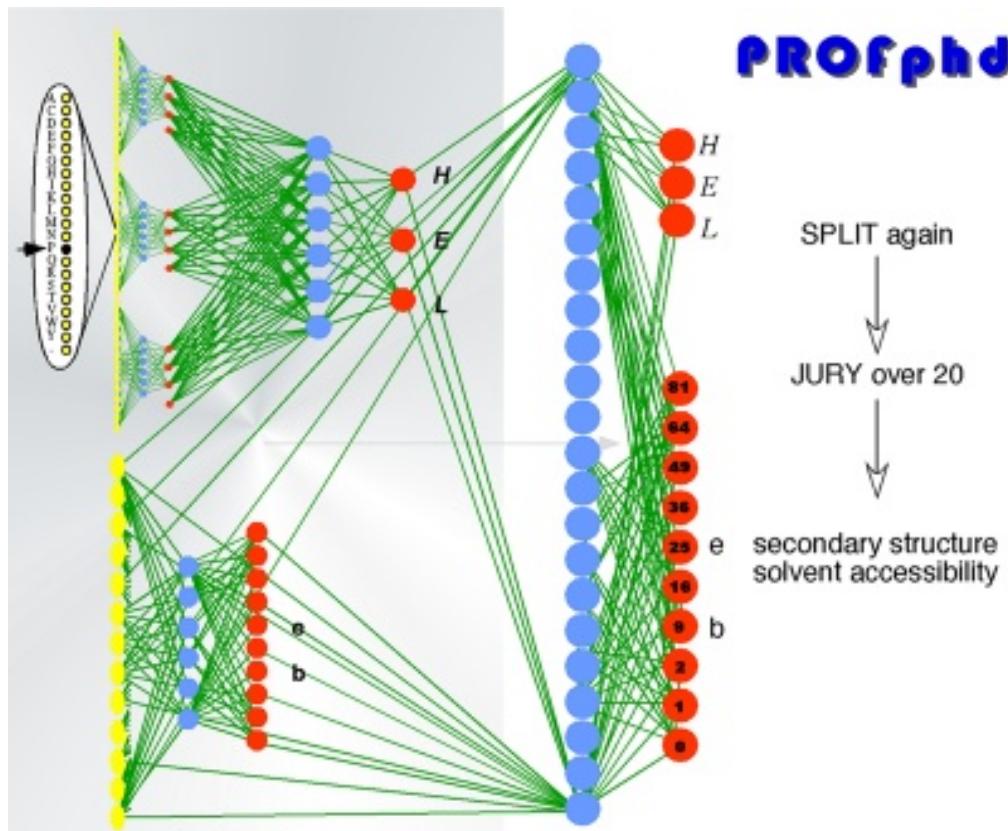


Fig. 2. Simplified sketch of PROFphd system. PROFphd arrives at the final prediction through many levels of computation, some of which are simplified here. First, local and global information (Fig. 1) for 13–21 sequence consecutive residues is represented by >1000 input units (13*21 input units sketched on the top left). Goal is to predict the secondary structure or accessibility for the residue at the center of the respective window (Proline marked by arrow on left). The second-level networks use the output of the first level (+ a spacer, i. e., 3*4 input units for the example shown) to identify correlations between adjacent predictions. The networks on the third level see the same sequence information as the first-level networks plus the previously predicted values for secondary structure and accessibility. The final prediction is obtained through averaging over a variety of such third-level networks.

Balanced predictions by balanced training. For the prediction of secondary structure and transmembrane helices, the distribution of the examples is rather uneven: about 32% of the residues are observed in helix, 21% in strand, and 47% in loop; about 18% of the residues in integral transmembrane proteins are located in transmembrane helices. Choosing the training examples proportional to the occurrence in the data set (unbalanced training), results in a prediction accuracy that mirrors this distribution, e. g., strands are predicted inferior to helix or loop. Traditionally, this has been attributed to beta-strand formation being determined by hydrogen bonds that are not local in sequence; in contrast, the standard helical $i \rightarrow i + 4$ pairing is “visible” in local sequence fragments for at least nine (four on either side of the central residue) consecutive resi-

dues. Although this explanation appears sound, balanced training is a simple way around the database bias: at each time step, one example is chosen from each class, i.e., one window with the central residue in a helix, one with the central residue in a strand, and one representing the loop class. Balanced training yields predictions well balanced between the output states. In particular, the basic balanced secondary structure network predicts strands as accurately as helices, implying that the local sequence preference for strand formation suffices. The consequence of this finding is that although hydrogen bonds stabilizing helices are more local in sequence than those stabilizing strands, the local sequence preferences suffice to predict both equally well. In other words, the explanation generally accepted for the poor performance of some prediction methods on strands was simply not valid. The final PROFsec system, however, re-introduces the bias somehow by a variety of averaging improving overall performance. Effectively, this averaging finds a better compromise between over- and under-prediction than the individual networks.

Detailed improvement by combining accessibility and secondary structure. Solvent-exposed helices have different physical environments than buried helices; this is reflected in different amino acid propensities. PROFacc and PROFsec account for this by introducing yet another combination-to-structure network that uses the same input features as the first-level networks and adds in the second-level output from secondary structure specialists and the first-level output from accessibility specialists. One hybrid network is trained to predict secondary structure (PROFsec), another to predict accessibility (PROFacc).

Final prediction by PROFsec: filtering out most obvious errors. The finally predicted secondary structure state (H, E, or L) for each residue is chosen according to the output unit that—after averaging—has the highest numerical value (state L chosen for a tie). This procedure may still result in predictions that expert users will immediately recognize as “unlikely correct,” such as single-strand residues. Most of these are not corrected automatically. Only the following drastic, unrealistic predictions are flipped: HEH → HHH; EHE → EEE, and LHL → LLL.

Final prediction by PROFacc: averaging over accessibility bins. For accessibility prediction (PROFacc), the final prediction requires the following three steps: (1) compute average over different networks for each of the ten output units; (2) average over neighboring output units ($n - 1, n, n + 1$); (3) determine the number of the unit n_{\max} with the maximal numerical value after the previous step. The predicted relative solvent accessibility for residue i $\text{PREL}(i)$ is reported as n_{\max}^2 , although it actually is more accurately described by:

$$\dagger (n_{\max} - 1)^2 = \text{PREL}(i) < n_{\max}^2$$

This filter performs an average over neighboring output units (n , Eq. 1; i.e., not over adjacent residues). The Connolly surface accessibility is obtained by simply multiplying $\text{PREL}(i)$ with the maximal accessibility of the amino acid at position i .

Final prediction by PHDhtm: dynamic programming significantly improves predictions. Only the filter used for predicting transmembrane helices (PHDhtm) is crucial for performance. In this final step, a dynamic programming algorithm finds the best local path through the “energy landscape” provided by the raw network output. The optimal path is constrained by allowing predicted membrane helices to span between

18 and 25 residues and to be separated by at least four predicted nonmembrane residues (82,8). We recently realized that these constraints are not fully appropriate (83,81); however, we still have not completed our improvements for the next version of PROFhtm. Once PHDhtm has completed the prediction of membrane helices, it simply counts all positively charged residues in even and odd numbered (counting from N- to C-terminal) nonmembrane regions and predicts the orientation of the helices (OUT if N-terminal translocated) according to the lower frequency of positively charged residues in extra-cytoplasmic regions (positive-inside rule [84]).

3. Methods

3.1 Using PROFphd to Predict 1-D Structure

Examples: prion and PYD domain. Prions are in many ways unusual proteins; not the least of the puzzles may be that we still do not entirely know their native function and structure (85,86). However, prion must be rather important, since it is related to many diseases, e. g., Creutzfeld–Jacob disease, Gerstmann–Straussler–Scheinker syndrome, fatal familial insomnia, kuru, Alpers syndrome, and scrapie, bovine spongiform encephalitis (BSE; “mad cow”). These diseases all appear to involve an aggregation through strands (87,88); in particular, what is labeled as the third helix, according to experimental structures of prion fragments, appears to be prone to switching conformation to β -strand (89–95). Interestingly, this region is “incorrectly” predicted as strand with high confidence (Fig. 3). In fact, this prediction mistake was used early on to help unravel the structure of prions (96). The NALP1 pyrin domain (PYD; 1b10 in PDB [97], Fig. 3) is another example of a protein the interest for which arose through sequence analysis that linked it to the DAPIN family of apoptosis-related domains (98). The particular aspect that differentiates PYD from other death domains is the absence of a helix around residues 90–100 (Fig. 3). Originally predicted, this absence has been confirmed experimentally (97).

How to estimate performance. Many developers fall victim to overoptimism by overestimating the performance of their programs. One of the important features of the old PHD programs for globular proteins was that the original estimates published still hold, more than a decade later (30). Ultimately, the only way of avoiding this problem is by testing methods on proteins for which the experimental information was not available when the method was developed. The structure prediction community makes a considerable effort in realizing such tests through the bi-annual CASP (Critical Assessment of Structure Prediction) meetings initialized by John Moult (CARB, Bethesda) (99–102). One severe problem with this concept is that the assessments usually have to be based on too small data sets (2,75). A similar concept is, however, realized by the EVA server, which continuously and objectively evaluates automatic prediction methods (1,3). Since it went online, EVA could for instance base the evaluation of secondary structure prediction methods on over 1500 new experimental structures, none of which were available to develop the method tested.

Accuracy of PROFphd. The EVA test demonstrated that PROFsec on average predicts over $76 \pm 10\%$ of all residues correctly in one of the three states (helix, strand, other) (Notes 4, 6, and 11–14). For a similar set of 1331 sequence-unique protein chains published after development, PROFacc predicted $76 \pm 10\%$ of the residues correctly as

1b10: RECOMBINANT SYRIAN HAMSTER PRION (fragment)

1pn5: NALP1 PYRIN DOMAIN (PYD)

Fig. 3. Examples of PROFsec and PROFacc predictions. “AA” gives the residues by their one-letter amino acid code; “SEC” and “ACC” mark the blocks of observed (1b10 and 1pn5) assignments and predictions (PROF) for secondary structure and solvent accessibility, respectively; “RI” mark residues by stars that are predicted at reliability levels above average; secondary structure is described by H (helix), E (strand), and L (other); accessibility by e (exposed: $\geq 25\%$), i (intermediate: 9–25%) and b (buried: <9%). The top gives the predictions for the globular regions of a prion fragment (PDB identifier 1b10 [90], numbering according to the convention in the field). This prion prediction is an example for a protein predicted well below average. However, the PROFsec error of overpredicting strand residues might in fact correctly reflect the *in vivo* aggregation of prion, in particular in the region of the third helix. The bottom shows the NALP1 pyrin domain (PYD, PDB identifier 1pn5 [97]) that has recently been discovered to constitute a link between apoptotic and interferon response [98]. Performance is above average for this molecule. Note that the examples show only some of the information provided by PROFphd—in particular, the alignment information and more details about the reliability indices, and the precise output states are omitted.

either buried (relative accessibility <16%) or exposed (relative accessibility \geq 16%); $77 \pm 10\%$ of the residues predicted on the surface were also observed, and $79 \pm 10\%$ of the observed surface residues were predicted (**Notes 1** and **4**). We previously overestimated the performance in predicting membrane helices (81,83). The main reason for this was that the high-resolution membrane structures determined over the last 8 y revealed details that were not anticipated by low-resolution data from proteolysis, fusion proteins, and antibody-binding sites. The problem is that there are still so few high-resolution structures that methods need to be developed considering the more amply available low-resolution data. From today's high-resolution data, we estimate the following performance: for approx $80 \pm 10\%$ of all proteins, all membrane helices are correctly predicted, i. e., the number of helices is correct and each predicted helix overlaps at least three residues with the corresponding experimentally observed trans-membrane segment). For approx $66 \pm 10\%$ of the proteins, all helices and the topology are correctly predicted and approx $80 \pm 5\%$ of all membrane residues are correctly predicted. For reasons that remain obscure to us, proteins with ≥ 6 membrane helices are predicted less accurately than those with fewer (83). This may also be related to the observation that predictions are less accurate for eukaryotic than for prokaryotic proteins (83,103). Whereas hydrophobicity index-based membrane predictions incorrectly

identify membrane helices in 30–100% of all globular proteins (83,104), PHDhtm using “good” PSI-BLAST alignments incorrectly identifies membrane helices in less than 2% of all globular proteins (Notes 17 and 18). However, one-fourth of all signal peptides are confused for membrane helices (most use hydrophobicity-based methods).

PDB sequences are not fully representative of the protein universe. The estimates for globular proteins are all based on protein sequences taken from the PDB (5) (Note 6). It is not clear to what extent proteins of known structure are fully representative of the universe of all proteins. First, structural biologists often choose fragments from longer native proteins. Second, structure determination favors soluble proteins. Third, proteins with known structure belong—on average—to larger sequence-structure families than proteins from entirely sequenced organisms (9,105). PROFacc is clearly less accurate in predicting the accessibility for residues at the interfaces between two domains from the same protein. (In contrast, PROFacc correctly captures many external protein–protein interfaces.) We have no idea whether the constraint in the type of protein that is suitable for structure determination influences prediction accuracy. Clearly, advances in structure determination over the last decade have yielded structures of globular proteins that differ in many ways from those we knew a decade ago. Nevertheless, the old PHD methods predicted these proteins accurately.

The third bias toward larger families severely impacts prediction accuracy. Unfortunately, we cannot easily correlate something like the information content of an alignment with prediction accuracy.

Estimates for prediction accuracy are averages over wide distributions. Statements such as “secondary structure is about 90% conserved within sequence-structure families” (106,107) or “solvent accessibility is about 85% conserved within sequence families” (74) refer to averages over distributions. In fact, for most features that we ever looked at, such distributions tend to be rather wide—provided we used sufficiently representative, varied proteins. The same holds true for the estimated performance of prediction methods (Fig. 4A). This implies that your particular protein may also fall into the extreme ends of the distribution, i.e., may be one of the 5% least or most accurately predicted proteins. Developers are tempted to fool themselves into believing they could have anticipated when things go wrong, once they know the experimental structure; and indeed, experts carefully using prediction methods often succeed in improving over machines by eye-balling what might be wrong (108). However, we have so far failed to automatically reproduce such expert-based case-by-case improvements (Note 8). This may or may not be a principal limitation of tools generalized to “all proteins”—e.g., while some folds stabilized by disulphide bridges are predicted less accurately, others are not. Often, unusually low performance originates from “poor” alignments, i.e., those that either have many errors and/or lack sufficiently diverse family members (Notes 1 and 3). Nevertheless, one important advantage of PROFsec and PROFacc over the previous PHD methods is that very wrong predictions are so exceptional that the 3% worst outliers still yield averages comparable to single-sequence-based methods from the past when tested over the 95% nonoutliers (76,109) (note: those ancient methods are still implemented in GCG).

Reliability index enables marking more accurately predicted residues. One of the most important aspects of PROFphd predictions is the estimate of the prediction reliability for each residue. For instance, the final networks used for PROFsec have three

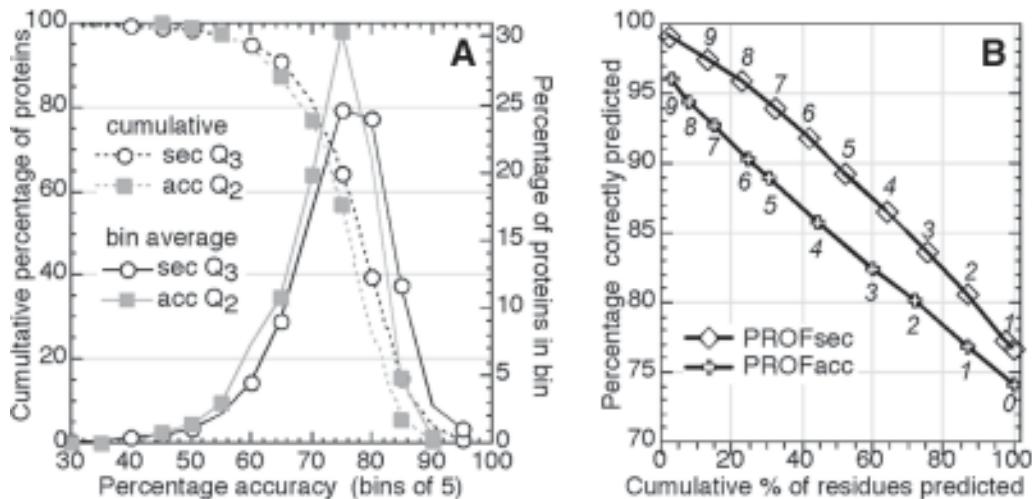


Fig. 4. Performance of PROFphd. Prediction accuracy varies between proteins; such a variation is also observed when computing how well secondary structure and accessibility are conserved in sequence-structure families. (A) Distribution of accuracy of PROFsec (percentage of residues correctly predicted as either helix, strand, or other; solid black line with open circles) and PROFacc (percentage of residues correctly predicted as buried or exposed, solid gray line with filled squares). The dotted lines show the cumulative values, e.g., 80% of all residues are predicted at levels around 70% accuracy. (B) Residues predicted at higher reliability are predicted more accurately. The x-axis gives the percentage of residues that are predicted above a given reliability index (actual values shown by numbers in italics), the y-axis shows the percentage of these that are correctly predicted. For instance, about half of all residues are predicted at indices >5 by PROFphd, and over 88% of these are predicted correctly (note: 88% happens to mark the average secondary structure similarity between proteins with similar folds [106,107]). The correlation between reliability and accuracy is slightly less impressive for PROFacc, reaching values around 88% for only 30% of the residues. Note that the reliability index averaged over the entire protein provides a good guess as to which side of the distribution (A) a particular prediction is likely to fall on: no protein predicted at an average reliability ≥ 6 has less than 76% accuracy, and only 3 out of 201 are below 70% accuracy, for an average index ≥ 5 .

output units (HEL, **Fig. 2**). The actual prediction is assigned to the unit with highest value, e.g., residue A with values $H = 0.9$, $E = 0.0$, $L = 0.1$, and residue B with $H = 0.5$, $E = 0.1$, $L = 0.4$ both result in the same prediction of helix, although A is predicted less unambiguously. This difference is reflected by the reliability index normalized to integers from 0 (ambiguous) to 9 (strong). Residues with higher reliability index are predicted with higher accuracy (**Fig. 4B**). In practice, the reliability index offers an excellent tool to focus on some key regions predicted at high levels of expected accuracy (**Notes 3, 8, and 10**).

Overall content of secondary structure accurately predicted by PROFsec. Proteins can be sorted roughly into four structural classes based on secondary structure content (all- α [helix $\geq 45\%$, strand $< 5\%$], all- β [strand $\geq 45\%$, helix $< 5\%$], α/β [helix $\geq 30\%$, strand $\geq 20\%$], and all others [110]). One experimental means of measuring secondary

structure content is circular dichroism spectroscopy (111,112). A simple alternative is to use the predictions of PROFsec to compile the overall prediction of secondary structure content. Based on the predicted content, over 80% of the proteins are correctly sorted into one of the four structural classes. The correlation between observed and predicted content is >0.92 for helix and >0.80 for strand. These values are comparable to results from circular dichroism spectroscopy (helix: 0.84, strand: 0.37–0.41 [112]). Of course, this does not imply that PROFsec can replace experiments. However, the high level of accuracy may recommend using PROFsec prediction as a complement.

Accurate and diverse multiple alignments yield best predictions. The more diverse and the more accurate the alignment used for the prediction, the better the performance (9,30); best performance results from inputting profiles generated by structure-based sequence alignments (113,114). While such alignments are obviously not available, this extreme example illustrates how much can be gained from carefully refining and extending database searches. Obviously, there is a trade-off between accuracy and divergence: the less sequence-similar a set of proteins, the higher the alignment errors (Notes 1, 3, and 5).

Prediction of porin-like membrane barrels. While PROFsec—trained on globular proteins—fails almost completely in predicting membrane helices (often predicted as strands), it does capture membrane strands more accurately (although still at levels slightly below the performance for globular proteins). We have explored this feature in developing PROFtmb, a method specialized to predicting membrane beta-strands. PROFtmb combines sequence information and PROFphd predictions in a hidden Markov model (115). Eighty to 90% of all β -strand residues predicted by PROFtmb are correct. Unfortunately, PROFtmb is less accurate in whole-proteome searches: our recent analysis yielded “only” about 100 unknown membrane-barrel proteins at acceptable levels of false positives.

3.2. Using PROFphd to Predict Other Aspects of Proteins

Widening the scope of database searches by threading. Most proteins with similar structure have levels of 5–15% pairwise sequence identity (midnight zone of database searches) (116); already at levels around 29% pairwise sequence identity (over 100 residues), 50% of all hits identified by database searches are proteins with different structure. Thus, intruding into the midnight zone becomes a struggle against noise. The most successful approaches are often referred to as “fold recognition” methods. Typically, such methods benefit from details in the structures of proteins of experimentally known structures. The most successful current techniques are all based on the idea of matching 1-D predictions to known structures introduced a while ago (117–120). Fold-recognition methods can be a powerful means of unraveling structural similarities between proteins of very different sequences; some methods even succeed in unravelling similarities without knowing the structure of any of the two proteins compared (D. Przybylski and B. Rost, manuscript in preparation). However, use such methods with extreme caution: most resulting models are likely to be mostly wrong (Notes 2, 3, and 19)!

1-D structure can improve comparative modeling alignments. The more different the sequences of two proteins, the more difficult they are to align. Alignment methods make considerable mistakes already at levels around 60% pairwise sequence identity over two entirely aligned homologous enzymes (26). Adding the 1-D predictions into

the comparison may help to considerably improve the alignment (121). Nevertheless, very few currently available modeling packages automatically explore this advantage.

Identify structural domains and predict globularity. Over 60–80% of all proteins appear to consist of more than one structural domain (defined as a globular subunit of a protein that could fold independently) (122). Secondary structure predictions can be explored to identify domain boundaries through comparisons with known structures (123). We have recently developed a method that predicts domain boundaries more directly from alignments and PROFphd predictions (Liu and Rost, manuscript in preparation; **Note 2**). Most globular, soluble protein domains are likely to adopt ellipsoid shapes. This implies a simple relation between the number of residues on the surface and the total length of a given domain. In fact, most single-chain proteins in the PDB fit a simple model assuming that proteins and amino acids are spheres, i. e., the number of surface residues is proportional to the number of residues by the power of two-thirds (124,125). Applying such a formula to PROFacc predictions, we can estimate how much a given protein deviates from typical globular domains. This simplistic prediction of protein globularity does not suffice to predict domains (domain fragments may be more compact than the full domain); however, it does capture some information about protein globularity (125) (**Note 15**).

Aiding annotation transfer. Most computational methods that extend the experimental data about protein function explore the fact that very closely related proteins have similar function. However, annotation transfer of function is a much more complex task than the transfer of structural annotations (126–129). For example, automatic methods that transfer the full enzyme classification are wrong 20% of the time even at PSI-BLAST expect (E) values <10–100 (130)! One prominent problem in transferring information originates from the domain organization of proteins (**Note 2**). Experts can correct many of these mistakes by critically revisiting the data employed for the transfer; of particular importance in this context are predictions of accessibility and secondary structure (131). Currently, we work on a method that inverts what developers of fold recognition usually do—instead of using functional information to improve fold recognition, we utilize our fold recognition method to improve the reliability of annotation transfer.

Identifying intrinsically unstructured proteins. Structural biologists are encountering an increasing number of proteins that appear intrinsically unstructured (132). We applied PROFphd to predict long exposed stretches with no regular secondary structure (NORS regions) in entirely sequenced proteomes (133). Three results were remarkable: (1) eukaryotes are over five times more abundant in such proteins than prokaryotes and archae (every fourth human protein had NORS regions, and 20,000 proteins had NORS regions longer than 150 residues); (2) the NORS regions were evolutionarily conserved; and (3) they often appeared to be involved in signaling processes. The NORSp program is available through our Web site (134).

Identifying regions undergoing local structural re-arrangements. Amyloid fibers often result from single sequence mutations that cause disease-related protein aggregation (135); often such changes involve regions that switch from helix or nonregular to strand (86,136–138). Highsmith and colleagues explored PROFsec to predict regions that more generally undergo conformational switches (139,140). The basic assumption of their method, ASP, is that local switches yield dips in the reliability of PROFsec,

and more specifically lead to predictions that are ambivalent between two output states (**Note 11**). This basic assumption is confirmed by a high reliability of identifying known conformational switches (140). One problem is that ASP cannot reliably distinguish proteins with and without such regions, i.e., is reasonably applied only if you do know that your protein contains such a region and wonder where this region could be.

Prediction of subcellular localization. Computational methods that predict protein subcellular localization are becoming increasingly important (141–144). While local sequence motifs often appear to determine the trafficking through the cell, even for the best-described example of nuclear localization motifs, the experimentally known signals are not even likely to cover half of all nuclear proteins (145). We recently improved significantly over previous methods that predict sub-cellular localization in the absence of known sequence motifs, by combining information from multiple alignments with PROFphd predictions (146). One particular result was that methods trained on full-length protein sequences as taken from SWISS-PROT (12) failed miserably when applied to the corresponding fragments taken from the PDB (5), and vice versa. This was also true for methods that did not use signal or transit peptides that are typically missing in PDB. One particular aspect of our method is that we can distinguish between proteins that contain signal peptides and are secreted, and those that will be retained in other compartments after this signal is cleaved. Overall, our method correctly sorts about 60% of the proteins into one of the four compartments—nucleus, cytoplasm, mitochondria, and extra-cellular space (we are currently addressing more fine-grained classifications); performance is slightly higher when using experimental information about structure. Note that in particular in the context of high-throughput structural genomics projects, we often have no information about subcellular localization.

Prediction of residues involved in protein–protein interactions. Interfaces between internally and externally bound residues differ significantly (147–150). We explored this difference to predict residues that are involved in external protein–protein interactions (as opposed to residues binding internally or “externally” between different sub-units of the same protein or at oligomerization sites) (151). Recently, we have significantly improved this prediction method by also using PROFphd predictions (Ofran and Rost, manuscript in preparation; **Notes 15** and **16**). We now correctly identify one-fifth of the residues involved in protein–protein interfaces at levels above 65% accuracy (i.e., 65% of the residues that we predict are also observed in high-resolution structures of complexes).

Prediction of functional classes. So far, very few groups have developed automatic methods that predict protein function in the absence of homology to experimental information (142,144). The most impressive exceptions to this statement are methods developed by the Brunak group in Denmark. These methods achieve low-resolution predictions of classes of cellular function (152) (e.g., “structural protein” vs “related to metabolism”), classes of enzymatic types (152) (e.g., enzyme vs nonenzyme, or transferase vs hydrolase), GeneOntology classes (153), and proteins involved in cell-cycle processes (154). All these methods use 1-D predictions as one of their input features.

3.3. Availability

PredictProtein Web server. PROFphd predictions (and the underlying alignments) are available upon request by the automatic prediction service PredictProtein (PP) (77).

For detailed information, send the word *help* as subject to PredictProtein@columbia.edu or access the World-Wide Web (WWW) site directly at <http://www.predictprotein.org>. Because PP handles requests on a first-come, first-served basis and occasionally handles over 1000 requests per day, returning the results may take a day. If you have no answer after two days, something has gone wrong (typical reasons: corrupted e-mail connection of sender or hardware/power problems at Columbia). In such a case, simply resubmit the request. Should the answer not appear after another two days, send a note to: Predict-Help@columbia.edu. Other comments that help us improve the server are welcome! If you want to run large data sets (>50 proteins), please split the jobs into chunks of 5–10, and wait for the answer before sending the next chunk. For other programs, databases, and computational biology resources from our group: <http://www.roslab.org>.

Downloadable program. The PROFphd methods are available for computational biologists who prefer to run the programs on their local machines. Users from academia may simply download and install the tools; colleagues from the industry are requested to confer with the author. Currently, the PROFphd suite is available for LINUX, SGI, and MAC OSX.

4. Notes

1. *Better alignments give better predictions.* The information content of the alignment is difficult to measure. Two important parameters are: (1) number of aligned sequences: the more sequences in the alignment, the better. The exact number of sequences needed for a “good prediction” depends on the variation and on characteristics of the particular protein family. As a rule of thumb: one is clearly not sufficient, more than five sequences can be enough. (2) Variation of aligned sequences: the aligned sequences should have a considerable variation with respect to the guide sequence (your protein). Ideally, the alignment should contain sequences at levels of 80%, 60%, 50%, 40%, and approx 30% pairwise sequence identity (with respect to the predicted protein). In general, more diverged sequences (30–40%) contribute more to the information content than do very similar ones (>80%). (Note: the levels of sequence identity are summarized in the alignment header of the output returned by PredictProtein [77].) Note, however, that the more distant relatives are difficult to align (actually below levels approx 40–60% pairwise sequence identity, some alignment errors are guaranteed). Furthermore, the three components of PROFphd behave rather differently with respect to the ratio accuracy/divergence: whereas secondary structure predictions tend to tolerate alignment errors and gain more from including diverged sequences than they lose from also including a few unrelated proteins, predictions of solvent accessibility and membrane helices are much more sensitive to alignment errors. In particular, membrane predictions become much less accurate when using standard PSI-BLAST alignments. One reason is that alignment methods—more precisely, the residue substitution matrices used to align proteins—are optimized for globular, soluble proteins. Although Ng and the Henikoffs have developed a substitution matrix tailored to helical membrane proteins (71), there is currently no method available that refines membrane protein alignments based on reliable membrane helix predictions (in fact, the only attempts that have been made explore hydrophobicity indices and should be handled with great caution).
2. *Chop very long proteins into structural domain-like fragments!* On the one hand, most proteins concatenate more than one structural domain, and most structural domains are about 100 residues long (122); eukaryotes stand out in that at least 20% of the eukaryotic

proteins have more than three structural domains (155). On the other hand, most modern alignment methods are more accurate when applied to domains than when applied to full-length proteins (the old pairwise BLAST [156] was an exception in this respect). Given these two observations, it is obviously beneficial to chop proteins into structural domain-like fragments before beginning the database searches. (PredictProtein [77] will incorporate methods that propose domain boundaries in the near future.)

3. *Playing with the alignment to monitor stable predictions.* Assume you have a protein that you suspect belongs to a certain family—e. g., to the family of death domains—although it is not sufficiently sequence similar to that family (such as PYD to the CARD domains). Assume further that your protein—say PYD—has one helix less than the known family members (as is the case for PYD, [Fig. 3](#)), or more generally that it differs locally in secondary structure segments. You may wonder how much you can trust the prediction and how much you should overrule it by your intuition. One way to investigate is by using the same prediction method for all family members—even if their structure were known. It could be that prediction methods have a “blind spot” when it comes to that helix. An example for such a “blind spot” is the prion protein ([Fig. 3](#)): although it is widely assumed that the third helix in [Fig. 3](#) actually is responsible for the disease-state aggregation of the molecule through a strand (85,157), all experimental structures reveal helices in that region (89,90,93,158–160). In other words, the predictions consistently fail to predict the observed helix. In fact, such consistent mistakes appear to be the reason why fold recognition based on secondary structure can be more successful when comparing predicted with predicted 1-D strings than when comparing predicted and observed secondary structure strings (Przybylski and Rost, manuscript in preparation). How can you gain confidence? One reasonable strategy is “playing with the alignment”: leave some members out and re-run the prediction method, investigate whether or not the closer relatives are more consistent (**Note 6**), monitor what happens upon gradually including more distant relatives. Often this procedure provides an excellent means of quickly spotting where predictions are more stable. (Note: for those who install PROFphd on their local machine, there is a little script that allows you to filter the input alignments by various criteria.)
4. *Seventy to 80% correct implies 20–30% incorrect.* The most accurate methods for predicting secondary structure reach sustained levels of approx 70–80% accuracy. When interpreting predictions for a particular protein, it is often extremely instructive to mark the 20–30% of the residues that you suspect to be falsely predicted.
5. *PROFphd does not directly predict “consensus structure.”* The PROFphd methods can only predict what they have been trained on. In particular, they have not encountered insertions in the guide sequence (the first sequence in your alignment is considered the guide sequence by the postprocessing step, [Fig. 1](#)). This is simply because insertions have no structure. Thus, although PROFphd uses the information from a sequence-structure family, it does not predict a consensus structure. Strictly speaking, the predictions are valid for the guide sequence only. The best way of generating something that resembles a “consensus structure”—ignoring that this concept may be intrinsically unphysical and hence misleading—is the following: Assume you have N proteins in a family. First use each of the N as guide sequences for one alignment, disallowing insertions in the guide sequence. Next, apply PROFphd to each of the resulting N alignments. Finally, align the N proteins by a method that conceptualizes consensus alignments, such as ClustalX (161), and pull in all the N PROFphd predictions according to that alignment (rather than using it directly as input to PROFphd).
6. *PROFphd predictions valid for the types of proteins encountered in training.* Prediction methods are usually derived from current experimental knowledge. Consequently, they may fail on classes of proteins that have not been included in the subsets. For example,

methods for predicting helices in globular proteins fail when applied to transmembrane helices. In general, results should be taken with caution for proteins with unusual features, such as P-, N-, or Q-rich regions, unusually many disulphide bridges, or for domain interfaces. The example of predicting unstructured regions (NORS), however, illustrated that it is not clear *a priori* what constitutes “unusual regions” (133). Novel folds are typically not unusual: the old PHD programs predicted those folds that were added since 1995 as accurately as estimated.

7. *Problems and advantages in using single sequences instead of alignments.* PROFphd predictions are more accurate when based on alignments than when based on single sequences: PROFsec reaches a three-state (helix, strand, other) per-residue accuracy of approx 69% without alignments; PROFacc reaches a two-state (buried/exposed) per-residue accuracy of approx 72% with single sequences. However, single-sequence-based predictions may help you to study trends in detail. For instance, you may want to study the effect of residue mutations using single sequences as input (otherwise the trend might be hidden totally in the alignment profile). Another possible application is to study differences between closely related family members.
8. *Aiding protein design.* The PROF networks are trained on naturally evolved proteins. However, the predictions have proven to be useful in some cases to investigate the influence of single mutations. In particular, PROFphd correctly predicted two different local conformations of the same 11-residue peptide in two different locations of the designed chameleon protein (162) and also correctly identified the secondary structure of the *Janus* protein that Regan et al. designed to have 50% pairwise sequence identity to two structurally different native proteins (163) (PROFsec also correctly identified the native protein with more similar structure and less similar sequence). Nevertheless, in particular when testing short polypeptides, the following should be taken into account: the PROFsec network input consists of 21 adjacent residues; thus, shorter sequences may be dominated by the ends (which are treated as solvent).
9. *Reliability indices extremely informative about predictions.* The PROFphd reliability indices correlate with prediction accuracy. Thus, they provide an excellent way to immediately focus on more reliably predicted residues/regions. Such indices answer the question: how reliably is the tryptophan at position 307 predicted in a surface loop? However, the reliability indices tend to be unusually high for poor alignments; an example for this is the unusual abundance of high values (stars in [Fig. 3](#)) for the prediction of the PYD domain. The previous PHD methods had a particular problem in this respect with single sequence-based predictions. PROFphd solved this problem by simple rescaling—i.e., in terms of the precision of the reliability indices, single sequences do not constitute “poor alignments.” The nontrivial problem that remains to be solved is to measure the information contained in multiple alignments in a meaningful way and to use this measure for rescaling the reliability indices. One particular way to use the reliability index is to guess whether or not your protein appears like many others to PROFphd, or whether it appears to be an outlier: average reliability indices above average (reported in the output) indicate more accurately predicted proteins (right-hand side of [Fig. 4A](#)), and vice versa.
10. *Averaging over many prediction methods may help.* PROFsec and PHDhtm are not the only state-of-the-art methods in their fields. Many users consider simply focusing on the consensus of all prediction methods. This is a reasonable idea provided it is used with caution. Firstly, when including methods that are not state-of-the-art, the consensus may actually be much less accurate than the single best method (164). Second, averaging over the best secondary structure prediction methods, on average, improves accuracy (164,165); however, a straightforward implementation of the many-method average reflects the reliability of predictions less accurately than PROFsec alone (164). In contrast, the average

over many membrane helix predictions clearly appears to improve accuracy and reliability of the predictions (103). Currently, there are no convincing results available that analyze the advantages of combining different methods predicting accessibility.

11. *Helix and strand are often confused.* One prominent application of secondary structure prediction methods is constraining the search space of methods that attempt predicting 3-D structures (23, 25, 166–168). Obviously, such methods will suffer significantly from initial 1-D predictions that confuse helices with strands. While one important improvement of PROFsec over PHDsec is the significant reduction of such bad predictions, still—on average—approx 2% of the residues are confused between helix and strands. The good news is that many of these bad predictions may actually not be that bad after all, in the sense that they may indicate regions undergoing local structural changes upon environmental changes (Subheading 3., and Fig. 3).
12. *Ends of regular secondary structure segments predicted less accurately.* On average, cores of helices and strands are predicted more accurately than their ends—often referred to as caps. This appears to contradict the finding that there are often strong sequence signals in particular for helix caps (169). Although overall approx 66% of the observed helix caps and 72% of the observed strand caps are predicted within a distance of two residues, only 21% of the helix and 42% of the observed strand caps are predicted right on spot. Part of this problem originates from underpredicting regular secondary structure segments (Fig. 5, and Note 13). We have recently addressed the task of distinguishing between secondary structure assignments that are unambiguous and those that may be more flexible (DSSPcont: an assignment of continuous secondary structure based on DSSP [107]). Although less flexible ends are predicted slightly more accurately than all other ends, this effect is rather small.
13. *Underprediction of strands and helices.* Although most helices and strands are correctly predicted, for most proteins some regular secondary structure elements are missed. In fact, for only approx 20% of the proteins is the number of strands and helices predicted correctly; for another 10%, one to three observed helices or strands are missing (Fig. 5). The majority of mis-predicted segments are either 3¹⁰ helices or strands shorter than three residues.
14. *Internal helices not predicted less accurately.* Traditional secondary-structure prediction methods fail at predicting internal helices (170). On average, this is not the case for PROFsec predictions; in fact, if anything core helices are predicted slightly more accurately.
15. *Prediction of accessibility may be wrong at hydrophobic external interfaces.* When estimating the performance of PROFacc, we consider residues predicted as buried and observed at the interface between two protein chains as errors. Recently, we noted that predicted accessibility is very useful to predict residues involved in protein–protein interactions (Ofran and Rost, manuscript in preparation). In contrast, PROFacc predictions tend to be rather inaccurate for residues involved in the interactions between different subunits of the same protein.
16. *PROFacc predictions useful to provide upper limits for internal contacts.* The predicted solvent accessibility (PROFacc) can be translated into a prediction of the number of water atoms around a given residue. Consequently, PROFacc can be used to derive upper and lower limits for the number of inter-residue contacts of a certain residue. Our preliminary results indicate that such an estimate can be explored to improve predictions of inter-residue contacts (Punta and Rost, manuscript in preparation).
17. *Signal peptides often confused with membrane helices.* Even when using carefully built alignments, PHDhtm still incorrectly predicts membrane helices for approx 20% of all tested signal peptides (83). Note furthermore that PredictProtein by default uses a rather

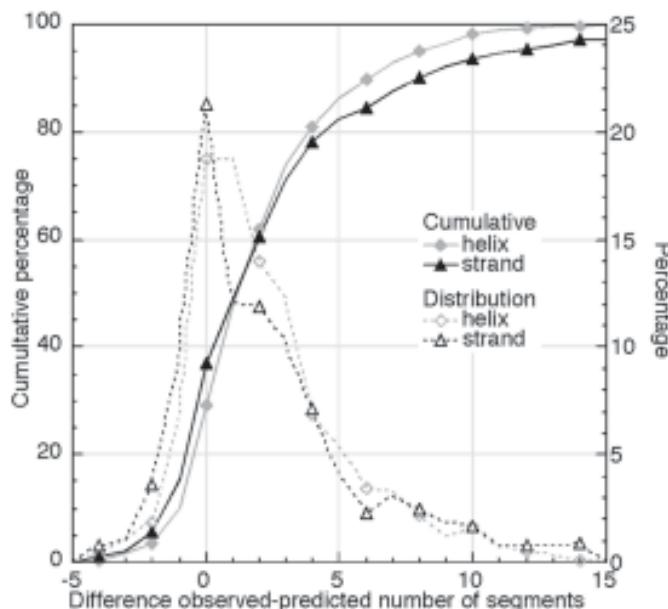


Fig. 5. PROFsec tends to underestimate the number of helices and strands. For how many proteins does PROFsec predict N helices (including 3^{10} helices, and helices shorter than five residues) and M strands (including strands shorter than two residues)—too few or too many? Shown are the differences observed—predicted per protein. Dotted lines give the raw distributions, solid lines the cumulative values. For example, for about 80% of all proteins, the numbers less than five of the observed helices or strands are missed. The vast majority of incorrect predictions imply missing strands or helices (most values >0). (Note: for the examples shown in [Fig. 3](#), the numbers observed-predicted are 1,1 for helix and -2 for strand.)

conservative threshold to minimize the error rate in predicting membrane helices in globular, soluble proteins. Consequently, no membrane helix might be detected even if your protein contains membrane helices. However, you can require a more permissive threshold that is likely to pick up most membrane helices.

18. *Short loops between membrane helices may not be predicted correctly.* Recent high-resolution structures have revealed that many very hydrophobic and short loops connect membrane helices; such loops constitute a particular problem for current membrane prediction methods (81). Another problem is posed by extremely long helices that cross the membrane and extend—without any discernible break—into globular domains (81).
19. *1-D structure predictions may or may not suffice to infer folds.* Assume that you find the following secondary structure prediction for your protein: helix-strand-strand-helix-strand-strand (H-E-E-H-E-E), and that you also find a protein of known structure with the same motif (H-E-E-H-E-E). Can you conclude that the two proteins have the same fold? Yes and no; your guess may be correct, but there are various ways to realize the given motif by completely different structures. For example, the secondary structure motif “H-E-E-H-E-E” is contained in at least 20 structurally unrelated proteins. Thus, you need additional constraints to decide which one is most similar to your protein. A particular example for how a careful application of 1-D predictions may contribute to guessing a fold and using this guess to explain experimental data, was the prediction of beta-propeller folds for the integrin α -subunits (171).

20. *Rely more on your intuition than on computer readouts.* The most important guideline for users of computational tools is: use your common sense and intuition! Although programs can help you, they are never as good as your expert intuition. On average, the best 1-D prediction methods now may be as good as the best experts, when not using these tools. However, experts with good tools will easily outperform machines.

Acknowledgments

Thanks to Jinfeng Liu and Megan Restuccia (Columbia) for computer assistance; to Yanay Ofran and Henry Bigelow (Columbia) for helpful comments on this manuscript; and to our entire group at Columbia for a wonderful working atmosphere. Thanks also to those who contributed crucially to the initial PHD methods—first of all, to Chris Sander (Sloan Kettering), then to Reinhard Schneider (LION Biosciences/Xapien) and Gerrit Vriend (Nijmegen). Our work is supported by grants P50-GM62413, RO1-GM63029-01, R01-GM64633-01, R01-LM07329-01 from the National Institute of Health (NIH), and grant DBI-0131168 from the National Science Foundation (NSF). Last, but not least, thanks to all those who deposit their experimental data in public databases, and to those who maintain these databases.

References

1. Eyrich, V., Martí-Renom, M. A., Przybylski, D., et al. (2001) EVA: continuous automatic evaluation of protein structure prediction servers. *Bioinformatics* **17**, 1242–1243.
2. Eyrich, V. A., Koh, I. Y. Y., Przybylski, D., et al. (2003) CAFASP3 in the spotlight of EVA. *Proteins* **53 Suppl 6**, 548–560.
3. Koh, I. Y. Y., Eyrich, V. A., Martí-Renom, M. A., et al. (2003) EVA: evaluation of protein structure prediction servers. *Nucl. Acids Res.* **31**, 3311–3315.
4. Eyrich, V. A. and Rost, B. (2003) META-PP: single interface to crucial prediction servers. *Nucl. Acids Res.* **31**, 3308–3310.
5. Berman, H. M., Westbrook, J., Feng, Z., et al. (2000) The Protein Data Bank. *Nucl. Acids Res.* **28**, 235–242.
6. Rost, B., Casadio, R., Fariselli, P., and Sander, C. (1995) Prediction of helical transmembrane segments at 95% accuracy. *Prot. Sci.* **4**, 521–533.
7. Rost, B. (1996) PHD: predicting one-dimensional protein structure by profile based neural networks. *Meth. Enzymol.* **266**, 525–539.
8. Rost, B., Casadio, R., and Fariselli, P. (1996) Topology prediction for helical transmembrane proteins at 86% accuracy. *Prot. Sci.* **5**, 1704–1718.
9. Przybylski, D. and Rost, B. (2002) Alignments grow, secondary structure prediction improves. *Proteins* **46**, 195–205.
10. Rost, B., Sander, C., and Schneider, R. (1994) PHD—an automatic server for protein secondary structure prediction. *CABIOS* **10**, 53–60.
11. Rost, B. (2000) PredictProtein—internet prediction service. Columbia University, New York.
12. Bairoch, A. and Apweiler, R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucl. Acids Res.* **28**, 45–48.
13. Anfinsen, C. B. (1973) Principles that govern the folding of protein chains. *Science* **181**, 223–230.
14. Gottesman, M. E. and Hendrickson, W. A. (2000) Protein folding and unfolding by *Escherichia coli* chaperones and chaperonins. *Curr. Opin. Microbiol.* **3**, 197–202.
15. Frydman, J. (2001) Folding of newly translated proteins in vivo: the role of molecular chaperones. *Annu. Rev. Biochem.* **70**, 603–647.

16. Dobson, C. M. and Karplus, M. (1999) The fundamentals of protein folding: bringing together theory and experiment. *Curr. Opin. Str. Biol.* **9**, 92–101.
17. Wales, D. J. and Scheraga, H. A. (1999) Global optimization of clusters, crystals, and biomolecules. *Science* **285**, 1368–1372.
18. Levitt, M. and Warshel, A. (1975) Computer simulation of protein folding. *Nature* **253**, 694–698.
19. Hagler, A. T. and Honig, B. (1978) On the formation of protein tertiary structure on a computer. *Proc. Natl. Acad. Sci. USA* **75**, 554–558.
20. van Gunsteren, W. F. (1993) Molecular dynamics studies of proteins. *Curr. Opin. Str. Biol.* **3**, 167–174.
21. Hansson, T., Oostenbrink, C., and van Gunsteren, W. (2002) Molecular dynamics simulations. *Curr. Opin. Str. Biol.* **12**, 190–196.
22. Koretke, K. K., Russell, R. B., and Lupas, A. N. (2001) Fold recognition from sequence comparisons. *Proteins* **45**, 68–75.
23. Bystroff, C. and Shao, Y. (2002) Fully automated ab initio protein structure prediction using I-SITES, HMMSTR and ROSETTA. *Bioinformatics* **18**, S54–S61.
24. Srinivasan, R. and Rose, G. D. (2002) *Ab initio* prediction of protein structure using LINUS. *Proteins* **47**, 489–495.
25. Baker, D. and Sali, A. (2001) Protein structure prediction and structural genomics. *Science* **294**, 93–96.
26. Tramontano, A., Leplae, R., and Morea, V. (2001) Analysis and assessment of comparative modeling predictions in CASP4. *Proteins Suppl.* **5**, 22–38.
27. Heringa, J. (2000) Computational methods for protein secondary structure prediction using multiple sequence alignments. *Curr. Protein Pept. Sci.* **1**, 273–301.
28. Jones, D. T. (2000) Protein structure prediction in the postgenomic era. *Curr. Opin. Str. Biol.* **10**, 371–379.
29. Bonneau, R. and Baker, D. (2001) *Ab initio* protein structure prediction: progress and prospects. *Annu. Rev. Biophys. Biomol. Struct.* **30**, 173–189.
30. Rost, B. (2001) Protein secondary structure prediction continues to rise. *J. Struct. Biol.* **134**, 204–218.
31. Chen, C. P. and Rost, B. (2002) State-of-the-art in membrane prediction. *Appl. Bioinf.* **1**, 21–35.
32. Ackerman, C. J., Harnett, M. M., Harnett, W., Kelly, S. M., Svergun, D. I., and Byron, O. (2003) 19 angstrom solution structure of the filarial nematode immunomodulatory protein, ES-62. *Biophys. J.* **84**, 489–500.
33. Alexandre, G. and Zhulin, I. B. (2003) Different evolutionary constraints on chernotaxis proteins CheW and CheY revealed by heterologous expression studies and protein sequence analysis. *J. Bacteriol.* **185**, 544–552.
34. Aravind, L. and Anantharaman, V. (2003) HutC/FarR-like bacterial transcription factors of the GntR family contain a small molecule-binding domain of the chorismate lyase fold. *FEMS Microbiol. Lett.* **222**, 17–23.
35. Balsera, M., Arellano, J. B., Gutierrez, J. R., Heredia, P., Revuelta, J. L., and De las Rivas, J. (2003) Structural analysis of the PsbQ protein of photosystem II by Fourier transform infrared and circular dichroic spectroscopy and by bioinformatic methods. *Biochem.* **42**, 1000–1007.
36. Bienstock, R. J., Skorvaga, M., Mandavilli, B. S., and Van Houten, B. (2003) Structural and functional characterization of the human DNA repair helicase XPD by comparative molecular modeling and site-directed mutagenesis of the bacterial repair protein UvrB. *J. Biol. Chem.* **278**, 5309–5316.
37. Bon, S., Ayon, A., Leroy, J., and Massoulie, J. (2003) Trimerization domain of the collagen tail of acetylcholinesterase. *Neurochem. Res.* **28**, 523–535.

38. Bonifati, V., Rizzu, P., van Baren, et al. (2003) Mutations in the DJ-1 gene associated with autosomal recessive early-onset parkinsonism. *Science* **299**, 256–259.
39. Cachot, J., Bultelle, F., Drouot, L., et al. (2003) Molecular cloning of flounder Xp18, a newly identified highly conserved protein mainly expressed in the ovary. *Gene* **307**, 13–21.
40. Campbell, J. D., Biggin, P. C., Baaden, M., and Sansom, M. S. P. (2003) Extending the structure of an ABC transporter to atomic resolution: Modeling and simulation studies of MsbA. *Biochem.* **42**, 3666–3673.
41. Carbone, M. A. and Robinson, B. H. (2003) Expression and characterization of a human pyruvate carboxylase variant by retroviral gene transfer. *Biochem. J.* **370**, 275–282.
42. Cavalcanti, A. R. O., Ferreira, R., Gu, Z. L., and Li, W. H. (2003) Patterns of gene duplication in *Saccharomyces cerevisiae* and *Caenorhabditis elegans*. *J. Mol. Evol.* **56**, 28–37.
43. Chereau, D., Kodandapani, L., Tomaselli, K. J., Spada, A. P., and Wu, J. C. (2003) Structural and functional analysis of caspase active sites. *Biochem.* **42**, 4151–4160.
44. Coffman, B. L., Kearney, W. R., Goldsmith, S., Knosp, B. M., and Tephly, T. R. (2003) Opioids bind to the amino acids 84 to 118 of UDP-glucuronosyltransferase UGT2B7. *Molec. Pharmacol.* **63**, 283–288.
45. Cordes, F. S., Komoriya, K., Larquet, E., et al. (2003) Helical structure of the needle of the type III secretion system of *Shigella flexneri*. *J. Biol. Chem.* **278**, 17,103–17,107.
46. da Fonseca, P. C. A., Morris, S. A., Nerou, E. P., Taylor, C. W., and Morris, E. P. (2003) Domain organization of the type 1 inositol 1,4,5-trisphosphate receptor as revealed by single-particle analysis. *Proc. Natl. Acad. Sci. USA* **100**, 3936–3941.
47. Desai, P., Akpa, J. C., and Person, S. (2003) Residues of VP26 of herpes simplex virus type 1 that are required for its interaction with capsids. *J. Virol.* **77**, 391–404.
48. Genevrois, S., Steeghs, L., Roholl, P., Letesson, J. J., and van der Ley, P. (2003) The Omp85 protein of *Neisseria meningitidis* is required for lipid export to the outer membrane. *EMBO J.* **22**, 1780–1789.
49. Grailles, M., Brey, P. T., and Roth, C. W. (2003) The *Drosophila melanogaster* multidrug-resistance protein 1 (MRP1) homolog has a novel gene structure containing two variable internal exons. *Gene* **307**, 41–50.
50. Huang, Y. P. J., Swapna, G. V. T., Rajan, P. K., et al. (2003) Solution NMR structure of ribosome-binding factor A (RbfA), a cold-shock adaptation protein from *Escherichia coli*. *J. Mol. Biol.* **327**, 521–536.
51. Huiskonen, J. T., Laakkonen, L., Toropainen, M., Sarvas, M., Bamford, D. H., and Bamford, J. K. H. (2003) Probing the ability of the coat and vertex protein of the membrane-containing bacteriophage PRD1 to display a meningococcal epitope. *Virology* **310**, 267–279.
52. Jin, W. Z., Kambara, O., Sasakawa, H., Tamura, A., and Takada, S. (2003) *De novo* design of foldable proteins with smooth folding funnel: automated negative design and experimental verification. *Structure* **11**, 581–590.
53. Juo, Z. S., Kassavetis, G. A., Wang, J. M., Geiduschek, E. P., and Sigler, P. B. (2003) Crystal structure of a transcription factor IIIB core interface ternary complex. *Nature* **422**, 534–539.
54. Kamada, K., Roeder, R. G., and Burley, S. K. (2003) Molecular mechanism of recruitment of TFIIF-associating RNA polymerase C-terminal domain phosphatase (FCP1) by transcription factor IIF. *Proc. Natl. Acad. Sci. USA* **100**, 2296–2299.
55. Kao, M. C., Di Bernardo, S., Matsuno-Yagi, A., and Yagi, T. (2003) Characterization and topology of the membrane domain Nqo10 subunit of the proton-translocating NADH-quinone oxidoreductase of *Paracoccus denitrificans*. *Biochem.* **42**, 4534–4543.
56. Kloetzel, J. A., Baroin-Tourancheau, A., Miceli, C., et al. (2003) Cytoskeletal proteins with N-terminal signal peptides: plateins in the ciliate *Euplotes* define a new family of articulins. *J. Cell Sci.* **116**, 1291–1303.

57. Mahdi, A. A., Briggs, G. S., Sharples, G. J., Wen, Q., and Lloyd, R. G. (2003) A model for dsDNA translocation revealed by a structural motif common to RecG and Mfd proteins. *EMBO J.* **22**, 724–734.
58. Maraver, A., Ona, A., Abaitua, F., et al. (2003) The oligomerization domain of VP3, the scaffolding protein of infectious bursal disease virus, plays a critical role in capsid assembly. *J. Virol.* **77**, 6438–6449.
59. Nam, Y., Weng, A. P., Aster, J. C., and Blacklow, S. C. (2003) Structural requirements for assembly of the CSL center dot Intracellular Notch1 center dot Mastermind-like 1 transcriptional activation complex. *J. Biol. Chem.* **278**, 21,232–21,239.
60. Orlova, E. V., Gowen, B., Droege, A., et al. (2003) Structure of a viral DNA gatekeeper at 10 angstrom resolution by cryo-electron microscopy. *EMBO J.* **22**, 1255–1262.
61. Payne, J. A., Rivera, C., Voipio, J., and Kaila, K. (2003) Cation-chloride co-transporters in neuronal communication, development and trauma. *Trends in Neurosciences* **26**, 199–206.
62. Pfannenschmid, F., Wimmer, V. C., Rios, R. M., et al. (2003) *Chlamydomonas* DIP13 and human NA14: a new class of proteins associated with microtubule structures is involved in cell division. *J. Cell Sci.* **116**, 1449–1462.
63. Rahaman, A., Srinivasan, N., Shamala, N., and Shaila, M. S. (2003) The fusion core complex of the Peste des petits ruminants virus is a six-helix bundle assembly. *Biochem.* **42**, 922–931.
64. Sheu, J. J. C., Cheng, T., Chen, H. Y., Lim, C., and Chang, T. W. (2003) Comparative effects of human Ig alpha and Ig beta in inducing autoreactive antibodies against B cells in mice. *J. Immunol.* **170**, 1158–1166.
65. van de Vosse, E., Lichtenauer-Kaligis, E. G. R., van Dissel, J. T., and Ottenhoff, T. H. M. (2003) Genetic variations in the interleukin-12/interleukin-23 receptor (beta 1) chain, and implications for IL-12 and IL-23 receptor structure and function. *Immunogenetics* **54**, 817–829.
66. van Swieten, J. C., Brusse, E., de Graaf, B. M., et al. (2003) A mutation in the fibroblast growth factor 14 gene is associated with autosomal dominant cerebral ataxia. *Am. J. Hum. Genet.* **72**, 191–199.
67. Zemojtel, T., Scheele, J. S., Martasek, P., Masters, B. S. S., Sharma, V. S., and Magde, D. (2003) Role of the interdomain linker probed by kinetics of CO ligation to an endothelial nitric oxide synthase mutant lacking the calmodulin binding peptide (residues 503–517 in bovine) *Biochem.* **42**, 6500–6506.
68. Zhang, Y., Corver, J., Chipman, P. R., et al. (2003) Structures of immature flavivirus particles. *EMBO J.* **22**, 2604–2613.
69. Zhulin, I. B., Nikolskaya, A. N., and Galperin, M. Y. (2003) Common extracellular sensory domains in transmembrane receptors for diverse signal transduction pathways in Bacteria and Archaea. *J. Bacteriol.* **185**, 285–294.
70. Braig, K., Otwinowski, Z., Hegde, R., et al. (1994) The crystal structure of the GroES co-chaperonin at 2.8 Å. *Nature* **371**, 578–586.
71. Ng, P., Henikoff, J., and Henikoff, S. (2000) PHAT: a transmembrane-specific substitution matrix. *Bioinformatics* **16**, 760–766.
72. Rost, B. and Sander, C. (1993) Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* **232**, 584–599.
73. Rost, B. and Sander, C. (1994) Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins* **19**, 55–72.
74. Rost, B. and Sander, C. (1994) Conservation and prediction of solvent accessibility in protein families. *Proteins* **20**, 216–226.

75. Rost, B. and Eyrich, V. (2001) EVA: large-scale analysis of secondary structure prediction. *Proteins* **45 Suppl 5**, S192–S199.
76. Rost, B. (2003) Prediction in 1D: secondary structure, membrane helices, and accessibility. *Methods Biochem. Anal.* **44**, 559–587.
77. Rost, B. and Liu, J. (2003) The PredictProtein server. *Nucl. Acids Res.* **31**, 3300–3304.
78. Altschul, S., Madden, T., Shaffer, A., et al. (1997) Gapped Blast and PSI-Blast: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402.
79. Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen bonded and geometrical features. *Biopolymers* **22**, 2577–2637.
80. Connolly, M. L. (1983) Solvent-accessible surfaces of proteins and nucleic acids. *Science* **221**, 709–713.
81. Chen, C. P. and Rost, B. (2002) Long membrane helices and short loops predicted less accurately. *Prot. Sci.* 2766–2773.
82. Rost, B., Casadio, R., and Fariselli, P. (1996) *Fourth International Conference on Intelligent Systems for Molecular Biology, St. Louis, MO*.
83. Chen, C. P., Kernytsky, A., and Rost, B. (2002) Transmembrane helix predictions revisited. *Prot. Sci.* **11**, 2774–2791.
84. von Heijne, G. (1994) Membrane proteins: from sequence to structure. *Annu. Rev. Biophys. Biomol. Struct.* **23**, 167–192.
85. Prusiner, S. B. (1998) Prions. *Proc. Natl. Acad. Sci. USA* **95**, 13,363–13,383.
86. Prusiner, S. B., Scott, M. R., DeArmond, S. J., and Cohen, F. E. (1998) Prion protein biology. *Cell* **93**, 337–348.
87. Harrison, P. M., Bamborough, P., Daggett, V., Prusiner, S., and Cohen, F. E. (1997) The prion folding problem. *Curr. Opin. Str. Biol.* **7**, 53–59.
88. Cohen, F. E. and Prusiner, S. B. (1998) Pathologic conformations of prion proteins. *Annu. Rev. Biochem.* **67**, 793–819.
89. Donne, D. G., Viles, J. H., Groth, D., et al. (1997) Structure of the recombinant fulllength hamster prion protein PrP(29–231): the N terminus is highly flexible. *Proc. Natl. Acad. Sci. USA* **94**, 13,452–13,457.
90. James, T. L., Liu, H., Ulyanov, N. B., et al. (1997) Solution structure of a 142-residue recombinant prion protein corresponding to the infectious fragment of the scrapie isoform. *Proc. Natl. Acad. Sci. USA* **94**, 10,086–10,091.
91. Kallberg, Y., Gustafsson, M., Persson, B., Thyberg, J., and Johansson, J. (2001) Prediction of amyloid fibril-forming proteins. *J. Biol. Chem.* **276**, 12,945–12,950.
92. Wuthrich, K. and Riek, R. (2001) Three-dimensional structures of prion proteins. *Adv. Protein Chem.* **57**, 55–82.
93. Nicholson, E. M., Mo, H., Prusiner, S. B., Cohen, F. E., and Marqusee, S. (2002) Differences between the prion protein and its homolog Doppel: a partially structured state with implications for scrapie formation. *J. Mol. Biol.* **316**, 807–815.
94. Wille, H., Michelitsch, M. D., Guenbaut, V., et al. (2002) Structural studies of the scrapie prion protein by electron crystallography. *Proc. Natl. Acad. Sci. USA* **99**, 3563–3568.
95. Qin, K., Coomaraswamy, J., Mastrangelo, P., et al. (2003) The PrP-like protein Doppel binds copper. *J. Biol. Chem.* **278**, 8888–8896.
96. Gasset, M., Baldwin, M. A., Lloyd, D. H., et al. (1992) Predicted alpha-helical regions of the prion protein when synthesized as peptides form amyloid. *Proc. Natl. Acad. Sci. USA* **89**, 10,950–10,944.
97. Hiller, S., Kohl, A., Fiorito, F., et al. (2003) NMR structure of the apoptosis- and inflammation-related NALP1 pyrin domain. *Structure* **11**, 1199–1205.
98. Staub, E., Dahl, E., and Rosenthal, A. (2001) The DAPIN family: a novel domain links apoptotic and interferon response proteins. *TIBS* **26**, 83–85.

99. Moult, J., Pedersen, J. T., Judson, R., and Fidelis, K. (1995) A large-scale experiment to assess protein structure prediction methods. *Proteins* **23**, ii–iv.
100. Moult, J., Hubbard, T., Bryant, S. H., Fidelis, K., and Pedersen, J. T. (1997) Critical assessment of methods of protein structure prediction (CASP): round II. *Proteins Suppl.* **1**, 2–6.
101. Moult, J., Hubbard, T., Bryant, S. H., Fidelis, K., and Pedersen, J. T. (1999) Critical assessment of methods of protein structure prediction (CASP): round III. *Proteins Suppl.* **3**, 2–6.
102. Moult, J., Fidelis, K., Zemla, A., and Hubbard, T. (2001) Critical assessment of methods of protein structure prediction (CASP): round IV. *Proteins Suppl.* **5**, 2–7.
103. Melen, K., Krogh, A., and von Heijne, G. (2003) Reliability measures for membrane protein topology prediction algorithms. *J. Mol. Biol.* **327**, 735–744.
104. Möller, S., Croning, D. R., and Apweiler, R. (2001) Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics* **17**, 646–653.
105. Liu, J. and Rost, B. (2001) Comparing function and structure between entire proteomes. *Prot. Sci.* **10**, 1970–1979.
106. Rost, B., Sander, C., and Schneider, R. (1994) Redefining the goals of protein secondary structure prediction. *J. Mol. Biol.* **235**, 13–26.
107. Andersen, C. A. F., Palmer, A. G., Brunak, S., and Rost, B. (2002) Continuum secondary structure captures protein flexibility. *Structure* **10**, 175–184.
108. Rost, B. (1997) Better 1D predictions by experts with machines. *Proteins Suppl.* **1**, 192–197.
109. Rost, B. (2003) Rising accuracy of protein secondary structure prediction. In: *Protein structure determination, analysis, and modeling for drug discovery*. (Chasman, D., ed.) Dekker, New York: 207–249.
110. Levitt, M. and Chothia, C. (1976) Structural patterns in globular proteins. *Nature* **261**, 552–558.
111. Johnson, C. W. J. (1990) Protein secondary structure and circular dichroism: a practical guide. *Proteins* **7**, 205–214.
112. Perczel, A., Park, K., and Fasman, G. D. (1992) Deconvolution of the circular dichroism spectra of proteins: the circular dichroism spectra of the antiparallel β -sheet in proteins. *Proteins* **13**, 57–69.
113. Levin, J. M., Pascarella, S., Argos, P., and Garnier, J. (1993) Quantification of secondary structure prediction improvement using multiple alignment. *Prot. Engin.* **6**, 849–854.
114. Al-Lazikani, B., Sheinerman, F. B., and Honig, B. (2001) Combining multiple structure and sequence alignments to improve sequence detection and alignment: application to the SH2 domains of Janus kinases. *Proc. Natl. Acad. Sci. USA* **98**, 14,796–14,801.
115. Bigelow, H., Petrey, D., Liu, J., Przybylski, D., and Rost, B. (2003) PROFtmb: prediction of transmembrane beta-barrels for entire proteomes. *Nucl. Acids Res.* **32**, 2566–2577.
116. Rost, B. (1995) Protein structures sustain evolutionary drift. *Folding Design* **2**, S519–S24.
117. Rost, B. (1995) TOPITS: Threading one-dimensional predictions into three-dimensional structures. In: Rawlings, C., Clark, D., Altman, R., Hunter, L., Lengauer, T., and Wodak, S. (eds.), *Third International Conference on Intelligent Systems for Molecular Biology*, Menlo Park, CA: AAAI, Cambridge, England. pp. 314–320.
118. Fischer, D. and Eisenberg, D. (1996) Fold recognition using sequence-derived properties. *Prot. Sci.* **5**, 947–955.
119. Russell, R. B., Copley, R. R., and Barton, G. J. (1996) Protein fold recognition by mapping predicted secondary structures. *J. Mol. Biol.* **259**, 349–365.
120. Rost, B., Schneider, R., and Sander, C. (1997) Protein fold recognition by prediction-based threading. *J. Mol. Biol.* **270**, 471–480.
121. Jennings, A. J., Edge, C. M., and Sternberg, M. J. (2001) An approach to improving multiple alignments of protein sequences using predicted secondary structure. *Prot. Engin.* **14**, 227–231.

122. Liu, J. and Rost, B. (2004) CHOP proteins into structural domain-like fragments. *Proteins* **55**, 678–688.
123. Marsden, R. L., McGuffin, L. J., and Jones, D. T. (2002) Rapid protein domain assignment from amino acid sequence using predicted secondary structure. *Prot. Sci.* **11**, 2814–2824.
124. Janin, J. (1976) Surface area of globular proteins. *J. Mol. Biol.* **105**, 13–14.
125. CUBIC, Columbia University, Dept. of Biochemistry & Mol. Biophysics. (1999) Short yeast ORFs: expressed protein or not? Rost, B. CUBIC-99-02. 1999.
126. Devos, D., and Valencia, A. (2001) Intrinsic errors in genome annotation. *TIGS* **17**, 429–431.
127. Koonin, E. V., Wolf, Y. I., and Karev, G. P. (2002) The structure of the protein universe and genome evolution. *Nature* **420**, 218–223.
128. Anantharaman, V., Aravind, L., and Koonin, E. V. (2003) Emergence of diverse biochemical activities in evolutionarily conserved structural scaffolds of proteins. *Curr. Opin. Chem. Biol.* **7**, 12–20.
129. Iliopoulos, I., Tsoka, S., Andrade, M. A., et al. (2003) Evaluation of annotation strategies using an entire genome sequence. *Bioinformatics* **19**, 717–726.
130. Rost, B. (2002) Enzyme function less conserved than anticipated. *J. Mol. Biol.* **318**, 595–608.
131. Whisstock, J. C. and Lesk, A. M. (2003) Prediction of protein function from protein sequence and structure. *Q. Rev. Biophys.* **36**, 307–340.
132. Wright, P. E. and Dyson, H. J. (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.* **293**, 321–331.
133. Liu, J., Tan, H., and Rost, B. (2002) Loopy proteins appear conserved in evolution. *J. Mol. Biol.* **322**, 53–64.
134. Liu, J. and Rost, B. (2003) NORSp: predictions of long regions without regular secondary structure. *Nucl. Acids Res.* **31**, 3833–3835.
135. Perutz, M. F. (1997) Amyloid fibrils. Mutations make enzyme polymerize. *Nature* **385**, 773–774.
136. Dobson, C. M. (1999) Protein misfolding, evolution and disease. *TIBS* **24**, 329–332.
137. Whisstock, J. C., Pike, R. N., Jin, L., et al. (2000) Conformational changes in serpins: II. The mechanism of activation of antithrombin by heparindagger. *J. Mol. Biol.* **301**, 1287–1305.
138. Whisstock, J. C., Skinner, R., Carrell, R. W., and Lesk, A. M. (2000) Conformational changes in serpins: I. The native and cleaved conformations of alpha(1)-antitrypsin. *J. Mol. Biol.* **296**, 685–699.
139. Kirshenbaum, K., Young, M., and Highsmith, S. (1999) Predicting allosteric switches in myosins. *Prot. Sci.* **8**, 1806–1815.
140. Young, M., Kirshenbaum, K., Dill, K. A., and Highsmith, S. (1999) Predicting conformational switches in proteins. *Prot. Sci.* **8**, 1752–1764.
141. Emanuelsson, O. and von Heijne, G. (2001) Prediction of organellar targeting signals. *Biochim. Biophys. Acta* **1541**, 114–119.
142. Nakai, K. (2001) Prediction of in vivo fates of proteins in the era of genomics and proteomics. *J. Struct. Biol.* **134**, 103–116.
143. Valencia, A. and Pazos, F. (2002) Computational methods for the prediction of protein interactions. *Curr. Opin. Str. Biol.* **12**, 368–373.
144. Rost, B., Liu, J., Nair, R., Wrzeszczynski, K. O., and Ofran, Y. (2003) Automatic prediction of protein function. *Cell Mol. Life Sci.* **60**, 2637–2650.
145. Cokol, M., Nair, R., and Rost, B. (2000) Finding nuclear localisation signals. *EMBO Rep.* **1**, 411–415.
146. Nair, R. and Rost, B. (2003) Better prediction of sub-cellular localization by combining evolutionary and structural information. *Proteins* **53**, 917–930.

147. Jones, S. and Thornton, J. M. (1997) Analysis of protein-protein interaction sites using surface patches. *J. Mol. Biol.* **272**, 121–132.
148. Lo Conte, L., Chothia, C., and Janin, J. (1999) The atomic structure of protein-protein recognition sites. *J. Mol. Biol.* **285**, 2177–2198.
149. Sheinerman, F. B. and Honig, B. (2002) On the role of electrostatic interactions in the design of protein-protein interfaces. *J. Mol. Biol.* **318**, 161–177.
150. Ofran, Y. and Rost, B. (2003) Analysing six types of protein-protein interfaces. *J. Mol. Biol.* **325**, 377–387.
151. Ofran, Y. and Rost, B. (2003) Predict protein-protein interaction sites from local sequence information. *FEBS Lett.* **544**, 236–239.
152. Jensen, L. J., Gupta, R., Blom, N., et al. (2002) Prediction of human protein function from posttranslational modifications and localization features. *J. Mol. Biol.* **319**, 1257–1265.
153. Jensen, L. J., Gupta, R., Staerfeldt, H. H., and Brunak, S. (2003) Prediction of human protein function according to Gene Ontology categories. *Bioinformatics* **19**, 635–642.
154. de Lichtenberg, U., Jensen, T. S., Jensen, L. J., and Brunak, S. (2003) Protein feature based identification of cell cycle regulated proteins in yeast. *J. Mol. Biol.* **329**, 663–674.
155. Liu, J., Hegyi, H., Acton, T. B., Montelione, G. T., and Rost, B. (2003) Automatic target selection for structural genomics on eukaryotes. *Proteins* **56**, 188–200.
156. Altschul, S. F. and Gish, W. (1996) Local alignment statistics. *Meth. Enzymol.* **266**, 460–480.
157. Leclerc, E., Peretz, D., Ball, H., et al. (2001) Immobilized prion protein undergoes spontaneous rearrangement to a conformation having features in common with the infectious form. *EMBO J.* **20**, 1547–1554.
158. Baldwin, M. A., James, T. L., Cohen, F. E., and Prusiner, S. B. (1998) The three-dimensional structure of prion protein: implications for prion disease. *Biochem. Soc. Trans.* **26**, 481–486.
159. Viles, J. H., Donne, D., Kroon, G., et al. (2001) Local structural plasticity of the prion protein. Analysis of NMR relaxation dynamics. *Biochem.* **40**, 2743–2753.
160. Kuwata, K., Li, H., Yamada, H., et al. (2002) Locally disordered conformer of the hamster prion protein: a crucial intermediate to PrPSc? *Biochem.* **41**, 12,277–12,283.
161. Chenna, R., Sugawara, H., Koike, T., et al. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucl. Acids Res.* **31**, 3497–3500.
162. Minor, D. L. J. and Kim, P. S. (1996) Context-dependent secondary structure formation of a designed protein sequence. *Nature* **380**, 730–734.
163. Dalal, S., Balasubramanian, S., and Regan, L. (1997) Protein alchemy: changing β -sheet into α -helix. *Nat. Struct. Biol.* **4**, 548–552.
164. Rost, B., Baldi, P., Barton, G., et al. (2001) Simple jury predicts protein secondary structure best. Columbia University. CUBIC_2001_10. 2001-10-01.
165. McGuffin, L. J. and Jones, D. T. (2003) Benchmarking secondary structure prediction for fold recognition. *Proteins* **52**, 166–175.
166. Eyrich, V. A., Standley, D. M., and Friesner, R. A. (1999) Prediction of protein tertiary structure to low resolution: performance for a large and structurally diverse test set. *J. Mol. Biol.* **288**, 725–742.
167. Ortiz, A. R., Kolinski, A., Rotkiewicz, P., Ilkowsk, B., and Skolnick, J. (1999) *Ab initio* folding of proteins using restraints derived from evolutionary information. *Proteins Suppl* **3**, 177–185.
168. Standley, D. M., Eyrich, V. A., An, Y., Pincus, D. L., Gunn, J. R., and Friesner, R. A. (2001) Protein structure prediction using a combination of sequence-based alignment, constrained energy minimization, and structural alignment. *Proteins Suppl* **5**, 133–139.
169. Aurora, R. and Rose, G. D. (1998) Helix capping. *Prot. Sci.* **7**, 21–38.

170. Benner, S. A., Cannarozzi, G., Gerloff, D., Turcotte, M., and Chelvanayagam, G. (1997) Bona fide predictions of protein secondary structure using transparent analyses of multiple sequence alignments. *Chem. Rev.* **97**, 2725–2844.
171. Springer, T. A. (1997) Folding of the N-terminal, ligand-binding region of integrin α subunits into a b-propeller domain. *Proc. Natl. Acad. Sci. USA* **94**, 65–72.
172. Li, W., Jaroszewski, L., and Godzik, A. (2001) Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* **17**, 282–283.
173. Wootton, J. C. and Federhen, S. (1996) Analysis of compositionally biased regions in sequence databases. *Meth. Enzymol.* **266**, 554–571.
174. Jeanmougin, F., Thompson, J. D., Gouy, M., Higgins, D. G., and Gibson, T. J. (1998) Multiple sequence alignment with Clustal X. *TIBS* **23**, 403–405.
175. Sander, C. and Schneider, R. (1991) Database of homology-derived structures and the structural meaning of sequence alignment. *Proteins* **9**, 56–68.
176. Jones, D. T., Tress, M., Bryson, K., and Hadley, C. (1999) Successful recognition of protein folds using threading methods biased by sequence similarity and predicted secondary structure. *Proteins* **37**, 104–111.
177. Thiele, R., Zimmer, R., and Lengauer, T. (1999) Protein threading by recursive dynamic programming. *J. Mol. Biol.* **290**, 757–779.
178. Xu, Y., Xu, D., Crawford, O. H., et al. (1999) Protein threading by PROSPECT: a prediction experiment in CASP3. *Prot. Engin.* **12**, 899–907.
179. Lindahl, E. and Elofsson, A. (2000) Identification of related proteins on family, superfamily and fold level. *J. Mol. Biol.* **295**, 613–625.
180. Xu, Y. and Xu, D. (2000) Protein threading using PROSPECT: design and evaluation. *Proteins* **40**, 343–354.
181. Bates, P. A., Kelley, L. A., MacCallum, R. M., and Sternberg, M. J. (2001) Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. *Proteins Suppl.* **5**, 39–46.
182. Deane, C. M., Kaas, Q., and Blundell, T. L. (2001) SCORE: predicting the core of protein models. *Bioinformatics* **17**, 541–550.
183. Karchin, R., Cline, M., Mandel-Gutfreund, Y., and Karplus, K. (2003) Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry. *Proteins* **51**, 504–514.

Classification of Protein Folds

Robert B. Russell

1. Introduction

The classification of three-dimensional (3D) structures now plays a central role in understanding the principles of protein structure, function, and evolution. Classification of new structures can provide functional details through comparison to others, which is of growing importance as X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy can now produce structures in advance of biochemical characterization (e.g., [ref. 1](#)). More generally, structure classifications themselves provide an excellent source of data for analyses of all kinds.

This chapter presents a strategy for classifying protein structures. Steps in the classification procedure — domains, structural class, folds, superfamilies — are discussed in turn by reference to examples and relevant literature. Methods are discussed for discerning when structural similarities are most likely to indicate an evolutionary and/or functional similarity when sequence similarity is absent. Finally, a review of the most widely used Internet-based classifications is given.

2. Methods

2.1. Secondary Structure

Protein folds are nearly always described in terms of the type and arrangement of secondary-structures (i.e., α -helices and β -strands), thus secondary-structure definition is a good first step in classification. A detailed review of methods for assigning secondary-structure is beyond the scope of this chapter. The reader is directed to references (2–4) and those therein.

2.2. Domain Assignment

Domains conveniently divide protein structures into discrete subunits, which are frequently classified separately. Protein domains are usually defined by one or more of the following criteria (*see* [ref. 5](#) and references therein):

1. Spatially separate regions of protein chains.
2. Sequence and/or structural resemblance to an entire chain from another protein.
3. A specific function associated with a region of the protein structure.
4. A substructure in another protein that meets one or more of requirements 1–3.
5. Repeating substructures within a single chain meeting one or more of requirements 1–3.

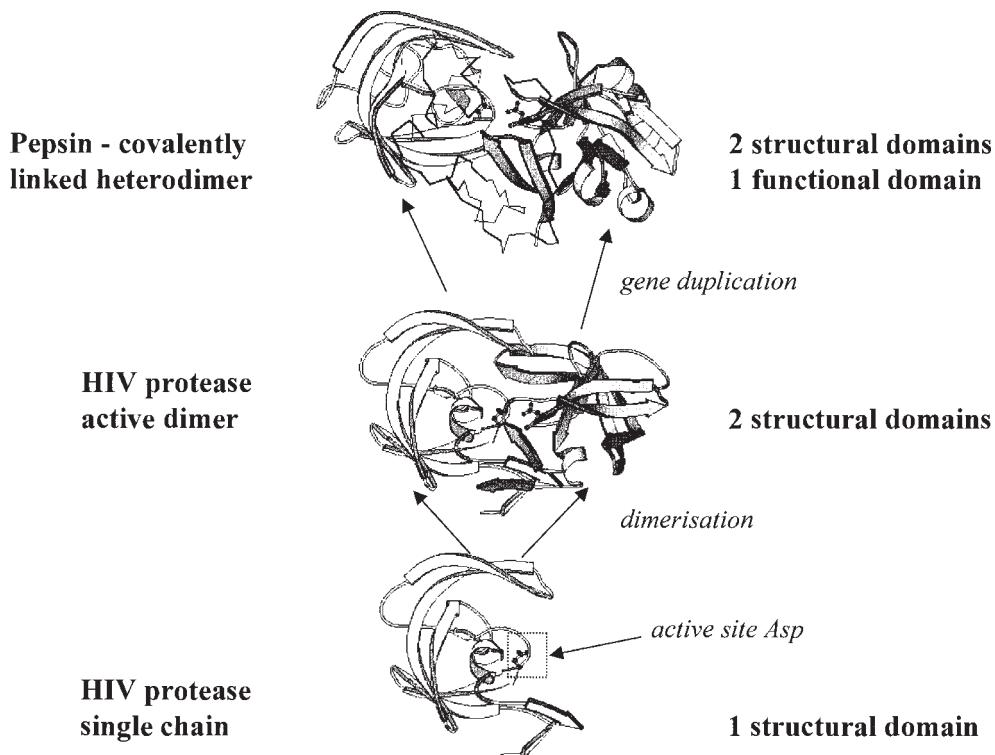


Fig. 1. Molscript (68) figure showing ambiguities in domain assignment for the aspartyl proteases. The bottom of the figure shows a single subunit from the HIV protease (PDB (69) code 1hiv-a) dimer shown in the middle of the figure. This dimer is equivalent to the covalently linked heterodimer found in the eukaryotic aspartyl protease (pepsin, PDB code 4pep) shown at the top of the figure. The single subunit corresponds to a structural domain, the homo/heterodimer corresponds to the functional domain.

Definitions 2 and 3 do not necessarily agree, as structural units may not be associated with a specific function. Some of the best examples of this structural/functional domain disagreement can be seen in the trypsin-like serine proteases and the pepsin-like aspartyl (acid) proteases (see Fig. 1). In both examples the functional domain (i.e., the catalytic domain) consists of two similar structural domains (presumed to be related by gene duplication; e.g., ref. 6). For the aspartyl proteases (see Fig. 1) there is further ambiguity as the retroviral (e.g., HIV) proteases consist of only a single copy of the structural domain that is functionally active as a homodimer, rather than the covalently linked heterodimer found in eukaryotes (e.g., pepsin).

For analysis of the principles of protein structure, use of structural domains is preferable, as these probably fold independently (e.g., each “lobe” of the proteases), and internal pseudo-symmetry (i.e., duplicated domains) can add to the understanding of the fold. It can be difficult to assign structural domains given only sequence data, and functional domains are frequently known in the absence of 3D structure data. It is thus best to consider functional domains during fold recognition/threading studies, where a protein of unknown structure (and often a functional domain) is compared to a database of known structures.

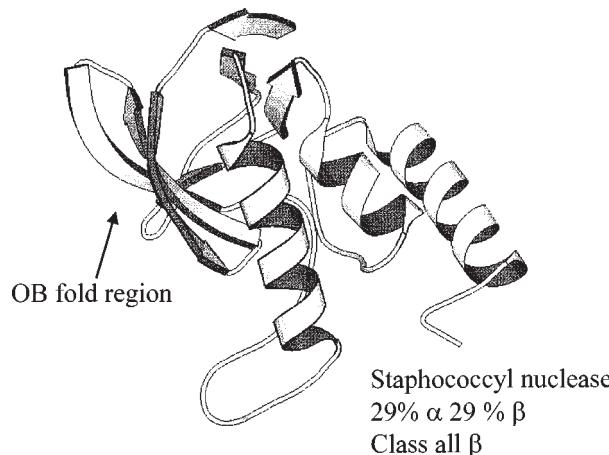


Fig. 2. Molscript (68) figure showing and example of difficulties in fold class assignment. Staphococcal nuclease (PDB code 1snc). The protein contains an equal proportion of β -strands and α -helices, but is generally classified as all β because of the β -barrel domain forming the core of the OB fold.

Automated methods have been developed for structural domain assignment, which look for spatially separated compact units (7–10) or hydrophobic cores (11). Methods can disagree even for relatively simple cases. A good strategy, adopted by the authors of Class Architecture Topology Homology (CATH) (12,13), is to combine the results of several algorithms with visual inspection, as often at least one of the methods will be correct. The property of recurrence is also very useful in defining domains. If a fragment of a larger protein is observed in isolation, or in a different domain context, then this adds confidence to the assignment of the segment as an independent folding unit (5,14).

Domains need not comprise single continuous segments of the polypeptide chain. Domain shuffling during protein evolution means that domains can be inserted into one another (15), making multisegmented (i.e., discontinuous) domains (see example 2 in Section 3.).

2.3. Assignment of Structural Class

After assignment of secondary-structures and domains, structural class can be assigned to domains. Structural classes divide proteins according to secondary-structure element content and organization. Globular proteins were first grouped into four classes (16): all α (or α/α), all β (or β/β), α/β , and $\alpha + \beta$. However, a fifth class, small or irregular, is now generally used to group those proteins with few secondary-structures (often containing multiple disulphides or metals).

2.3.1. One Secondary Structure Type: All α or All β

Class assignment is usually straightforward for domains with predominantly α -helices or β -sheets. Small elements of secondary-structure, such as 3^{10} helices, or small β hairpins are usually ignored in the assignment. Class is somewhat subjective, and may be based on the structure of the protein core rather than the abundance of α or β residues. Consider, e.g., staphococcal nuclease (see Fig. 2), which contains an equal proportion of residues in α helices and β strands, but is generally classed as all β (12,17).

because the core of the fold comprises an oligonucleotide/oligosaccharide (OB) binding-fold β barrel. Structural similarity may thus affect the assignment of class.

2.3.2. Both Secondary Structure Types: α/β or $\alpha+\beta$

Protein domains that contain a mixture of α helices and β sheets are more difficult to classify. Historically (16), α/β proteins are those containing both α helices and β sheets and where there is an intimate association of helices and strands. In contrast, $\alpha+\beta$ proteins define those consisting of segregated regions of helix and sheet. More recently, and perhaps most exemplified by the Structural Classification of Proteins (SCOP) database, α/β proteins tend to refer to those structures containing many $\beta\alpha\beta$ units, which consist of two adjacent (e.g., hydrogen bonded) β strands connected by a single α helix in a right-handed connection, whereas $\alpha+\beta$ proteins are those not falling easily into this definition. The authors of the CATH database (12) have done away with the distinction between α/β and $\alpha+\beta$, arguing that it is an architectural distinction, rather than an inherent difference in secondary structure content.

2.3.3. Other Classes

Proteins with few secondary structures form a category of their own. Frequently these proteins are small, with the tertiary structure dominated by multiple disulphide bonds, or one or more metal-binding sites. Other fold classes are used to contain peptides, or multidomain proteins for which no logical single class or domain divisions can be assigned.

Holm and Sander (18) positioned all structures in the protein database in a high-dimensional, abstract fold space. When multivariate scaling was used to project these positions onto a two-dimensional (2D) density plot, five “attractors” (peaks) were found to cover approx 40% of known folds. These attractors were found to correspond to architectural features: parallel β , β -meander, α -helical, β -zigzag, and $\alpha\beta$ -meander. Their, analysis thus provides an alternative means to define the “class” of a protein, although some of the attractors match the traditional classes closely.

2.4. Assignment of Fold

The number, type, connectivity, and arrangement of secondary-structures define the fold of a protein. Frequently, fold similarities are recognized by eye-following structure determination, although there are many papers published following a structure determination reporting a structure/function similarity not noted by the experimentalists (for examples *see refs.* 19,20; for reviews *see refs.* 21–23). Fold searching should thus be done with care, and similarities should be considered in a wider context that includes functional similarity.

It is best to compare each domain of a new structure to a database of those already known. Even in instances when the fold is known, such searches can reveal relationships that might be missed. For example, a protein may be easily seen to adopt an immunoglobulin (Ig)-like β -sandwich structure, but a structure with a similar function may be buried in the large group of Ig-like folds.

There are several means of searching protein structure databases with a probe structure (*see refs.* 21,22 for reviews). Programs such as SSAP (24,25), SARF (26), and STAMP (27) are available from the respective authors (*see Appendix* at the end of the chapter). It is also possible to run DALI (the engine of the Families of Structurally

Similar Proteins [FSSP] database [28,29] via the World Wide Web (see Appendix), and methods similar to the structure comparison technique of Artymiuk et al. (30) are encoded in QUANTA (31), and VAST (32) at the National Center for Biology Information (NCBI) (although VAST comparisons are only currently available for protein structures already in the database). Structure comparison is also possible within the O crystallographic package via the program *Déjà vu* (33). Different methods can give different results, particularly if structural similarity is slight. It is, therefore, prudent to run several algorithms and arrive at a consensus.

A phenomenon to consider during fold assignment is circular permutation, which relates domains that are similar in structure and/or sequence, yet whose N- and C-terminal portions have been exchanged. Permutations are real events in nature (see ref. 34 and references therein; for a recent example, see refs. 35,36). Although some structure comparison methods permit matches involving differences in connectivity, few, if any, are able to detect permutations directly.

2.5. Assignment of Superfamily

It is probably impossible to state definitively whether all proteins adopting a similar fold are descended from a similar common ancestor (i.e., related through divergence). For many proteins with similar folds, sequence, structure, or functional arguments suggest divergence from common ancestor; for others, no such conclusion can be drawn. Hence, some classifications distinguish between similarities that are due to divergent evolution, and those that may not be. It is clear that many homologous proteins have simply diverged beyond the point where sequence similarity can be detected. The term *superfamily* is often used in structure classification to refer to groups of proteins that appear to be homologous, even in the absence of significant sequence similarity.

Proteins with the same fold that are not thought to share a common ancestor are often referred to as *analogs* (to distinguish them from homologs), and are thought by many to be the result of a convergence to a stable structure. Although there is little hard evidence, there are some arguments that favor such a convergence. The number of proteins sampled during evolutionary time is vast, despite an estimated low number of possible folds (37–40), which may be due to restrictions on protein architecture. In addition, recent studies on sequence identity, calculated from structure-based sequence alignments (41,42), show a bimodal distribution. Although the results are very preliminary, the biomodality may suggest two origins for similarities between protein folds: analogy and homology.

How can analogy and homology be distinguished? A survey of recent literature (e.g., refs. 38,41–45) shows that one or more of the following features are often used to deduce a common ancestor (i.e., assign a common superfamily) given a pair of similar 3D structures:

1. Above a certain level of structural similarity, even if sequence similarity is insignificant, one can be largely confident of a homologous relationship (32,38,45).
2. The conservation of unusual structural features, sometimes outside the common core secondary structure elements. These features include functionally important turn conformations (46), left-handed $\beta\alpha\beta$ units (47), or others (e.g., ref. 43).
3. Low — but significant — sequence identity as calculated after structure superimposition (i.e., the identity from the structure-based alignment [41]). See ref. 43 for guide (illustrated by example) on how to calculate an associated statistical significance. It has been

- suggested (41) that sequence identities from a structure-based alignment of >12% are more likely to indicate a remote homology. Note also that structure similarities may confirm marginally significant sequence similarities seen prior to 3D structure determination.
4. The presence of key active site residues, even in the absence of global sequence similarity. This is most often applicable to enzymes, for examples, *see refs. 20,48–50*.
 5. Sequence similarity bridges, or *transitivity*. Even though two sequences may not be significantly similar to one another, inspection of homologs found in sequence searches with each sequence may reveal a “link” or “sequence bridge” linking the two sequences via significant sequence similarities (e.g., *refs. 49–51*). In other words, if domain A is significantly similar to domain B, and domain C is significantly similar to domain B, then domains A and C can generally be deemed homologous.

2.6. Superfolds, Superfamilies, and Supersites

Certain protein folds are populated by many different superfamilies, suggesting that the fold has arisen many times by convergent evolution. Such folds have been termed *superfolds* (38). For most of these folds, the core structure is highly symmetrical. Symmetry may imply an easier folding pathway and make convergence more likely than for less symmetrical folds, which often comprise only a single superfamily (e.g., aspartyl proteases). Examples of superfolds include β/α -triace phosphate isomerase (TIM)-barrels, Rossmann-like α/β -folds, ferredoxin-like folds, β -propellers, four-helical up-and-down bundles, Ig-like- β -sandwiches, β -jelly rolls, OB binding-folds, and SH3-like folds. All are adopted by groups of seemingly nonhomologous protein, which perform different functions (17,38).

For some of these superfolds, including the β/α -(TIM)-barrels, Rossmann-like α/β -folds, ferredoxin-like folds, β -propellers, four-helical up-and-down bundles, proteins from different superfamilies show a tendency to bind ligands in a common location (even when the nature of the bound atoms is different). These locations have been termed supersites, as they occur, by definition, within superfolds (52). Rather than being due to a divergence, supersites are thought to be a property of the protein fold, such as the alignment of nonhydrogen-bonded main-chain atoms, or the “helix dipole” (53), that dictates the best location for binding non-protein atoms, regardless of evolutionary origin. For some superfolds, it is thus possible to make predictions as to binding-site locations even in the absence of evidence of a common ancestor.

2.7. Predicting Function From Structural Similarity

Proposals have been made to determine large numbers of protein structures with the explicit aim of assigning function (54). After analysis of all structures within the SCOP database, Russell et al. (52) estimated the fraction of new structures (ignoring those that are obviously sequence similar to a known structure) that will currently show binding site or functional similarity via structure comparison. This estimate (illustrated by a pie chart in **Fig. 3**) was based on the distribution of homologous and analogous similarities within SCOP, and the fraction of superfolds containing supersites (i.e., how often do analogs have a common binding site?). Currently just under half of new structures will have accurate binding-site information predicted through structure comparison, which highlights the danger of interpreting every structural similarity as an indication of a common function.

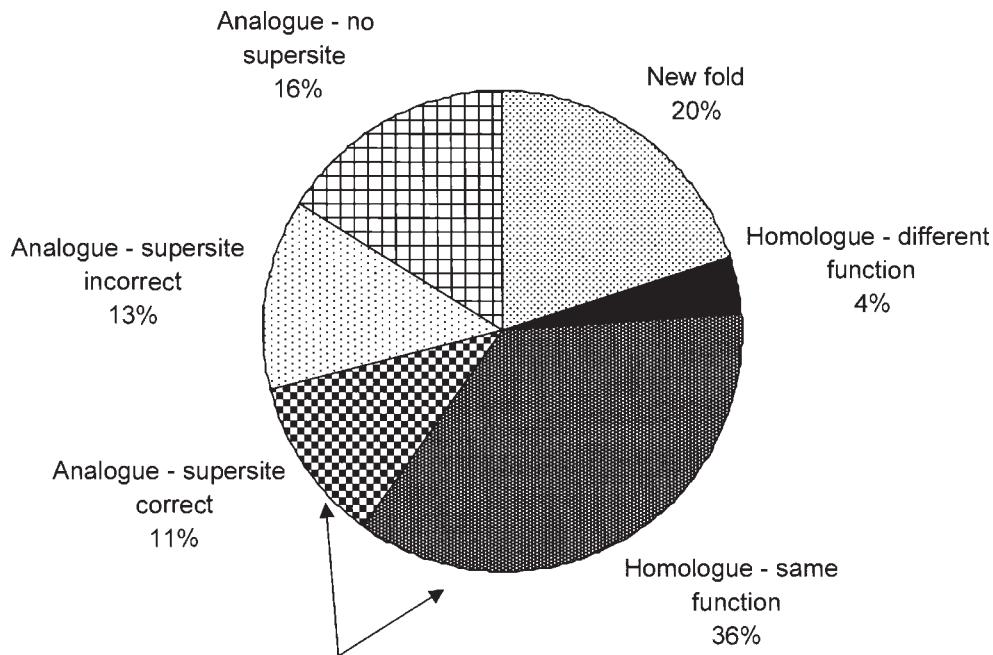


Fig. 3. Pie chart showing how often new structures will have correct binding site information predicted through structure comparison (adapted from ref. 52).

3. Two Examples

Two examples of protein structure similarities are described as below. In both instances the similarity was not reported by the crystallographers. Both similarities had clear biological implications that augmented the understanding of protein function following structure determination.

3.1. Example 1: β -Glucosyltransferase

The structure of β -glucosyltransferase (BGT) was originally reported to contain two domains of similar topology, each reminiscent of a nucleotide binding fold (55). Subsequent comparison of BGT to other known structures during two independent studies revealed a close similarity with glycogen phosphorylase, GPB (56,57). BGT and GPB differ greatly in length: BGT contains only 351 amino acids compared to GPB's 842. Despite this, 13 β -strands and 9 α -helices are equivalent (Fig. 4), and 256 pairs of C_α atoms can be superimposed with a root-mean-square deviation (RMSD) of 3.4 Å (56) (alternatively, 114 C_α atoms can be superimposed with an RMSD of 1.72 Å [57]). The common fold comprises the entire core BGT structure, with GPB containing numerous long insertions, which appear to modulate function (see Fig. 4). Superimposition also reveals striking similarities in the active sites of the two enzymes (despite surprisingly few residue identities). The observations lead to the suggestion that BGT and GPB share an ancient common ancestor.

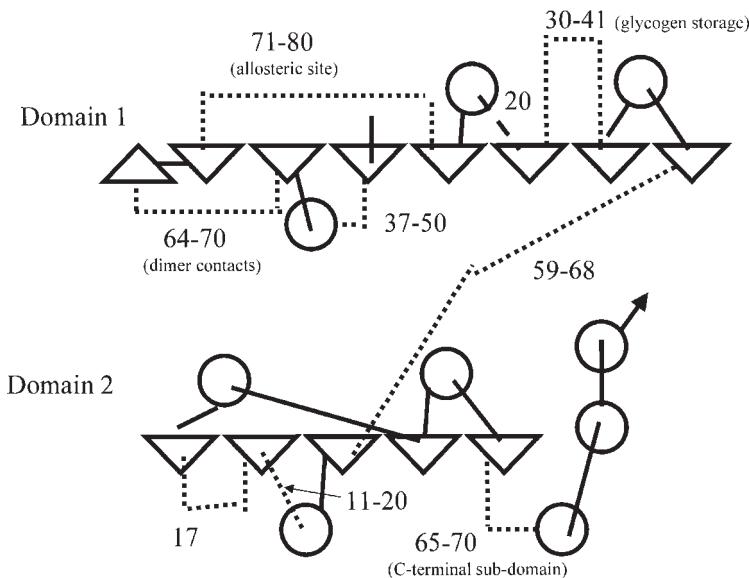


Fig. 4. Molscript (68) figure showing the similarity between adenylyl cyclase (PDB code 1ab8) and the palm domain from DNA polymerase (PDB code 1dpi). The location of key aspartyl residues is shown (N.B.: these are serines in the cyclase domain shown, but are aspartates in the other similar domain forming the active heterodimer). Equivalent regions in the two structures are shown as ribbons/coil, non equivalent regions (including the inserted fingers domain) are shown in C_α trace.

3.2. Example 2: Adenylyl Cyclase

The structure of adenylyl cyclase structure was originally reported to contain a new protein fold (58). However, subsequent comparison of the structure to the database found a striking similarity with DNA/RNA polymerases (20) (see Fig. 5). The core fold adopts the very common ferredoxin-like fold, and although this fold is seen in many proteins, cyclases and polymerases have a similar binding site, a similar reaction mechanism, and both contain key Mg^{2+} binding aspartate residues (59), known to be critical for polymerase function. The similarity thus provides key insights into the mechanism of the less-well-understood cyclases.

4. Protein Structure Classifications

Several protein structure classification schemes have become available via the Internet over the last 5 yr. Below the relative merits of each are discussed. Perhaps the most important general comment is that none of the classifications give a complete picture. Because all have different strengths, it is best to consider as many as possible.

4.1. SCOP

SCOP is maintained by Murzin et al. (17) in Cambridge, and is an entirely manual classification. The class definitions are after Levitt and Chothia (16). Proteins are generally divided into functional domains, and these are grouped into a hierarchy consisting of class, fold, superfamily, family, protein, and species. Proteins are put into the same fold if they have a similar core, which is decided by manual analysis. The fold definitions in SCOP are

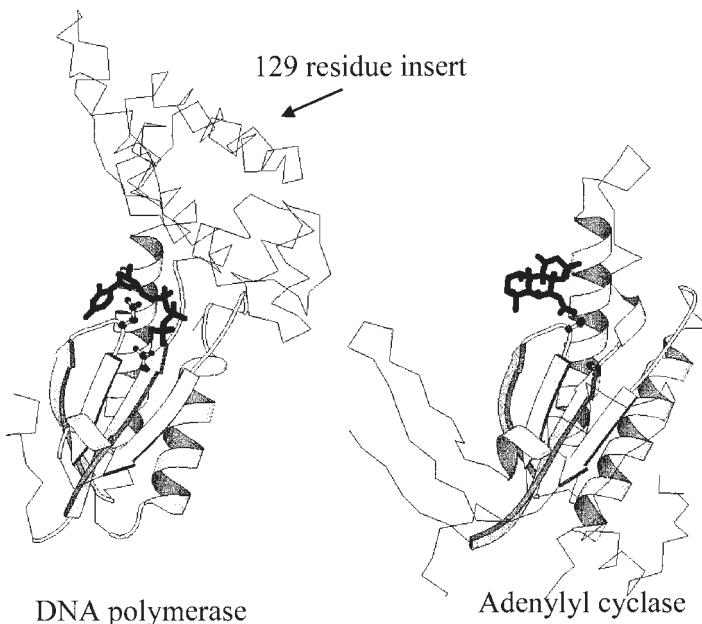


Fig. 5. Topology diagram showing the similarity between bacterial β -glucosyltransferase (BGT; PDB code 1bgu) and glycogen phosphorylase b (GPB; PDB code 1gpb). The figure was adapted from refs. 56 and 57. Dashed lines indicate those regions in the core of the fold where GPB contains very long insertions relative to BGT. Descriptions of the function of each insertion are given where known.

more stringent than the other classifications, and several similar structures are not put into the same fold, sometimes simply to avoid exceedingly large groups of structures (for example, the three layered α - β - α Rossmann-fold like structures; A. G. Murzin, personal communication). The subdivisions within each fold (superfamily, family, protein, and species) group proteins according to the degree of homology. **Figure 6** shows a schematic example of the SCOP classification scheme.

The great strength of SCOP is the very careful assignment of evolutionary relationships, even in the absence sequence similarity. Proteins in the same fold, but in different superfamilies are lacking in evidence for a common ancestor (analogous folds); those in the same fold and the same superfamily show some evidence of a common ancestor, which is often based on the features discussed above.

4.2. CATH

CATH is maintained by Orengo et al. (12) at University College, London. The classification is partly automated and partly manual, although they work toward a mostly automated system. They classify proteins according to a hierarchy that is similar to SCOP, although with some important differences. The class (C) layer is directly equivalent to that found in SCOP, with the exception that no distinction between α / β and α + β domains is made. The Architecture (A) layer, a unique feature of CATH, is an intermediate between class (C) and fold (or topology, T, in CATH). Protein architecture defines the orientation of the secondary structures composing the fold, independent of the connectivity or direction of secondary structure elements. For example, all protein

SCOP

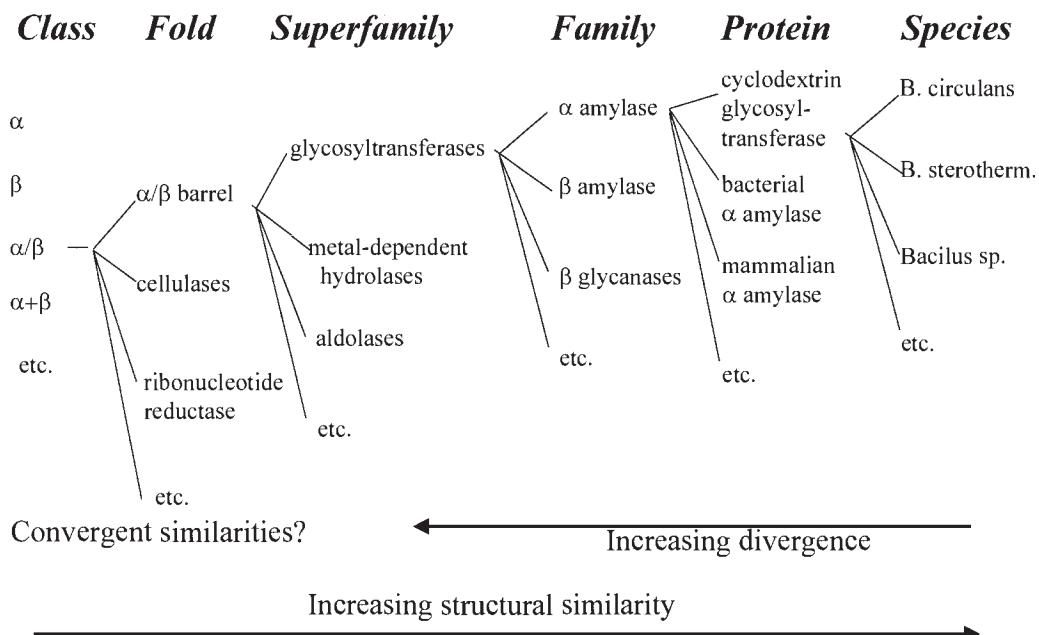


Fig. 6. Example of the classification hierarchy for the SCOP database.

domains containing a β -barrel regardless of the number of strands forming the barrel, or the connectivity are placed in a single architecture. The extra level to the hierarchy makes browsing classifications somewhat easier, and it makes structure space more continuous than in some of the other classifications. Architecture also encapsulates many of the descriptions often given with newly solved 3D structures (e.g., the protein contains a β -barrel structure). [Figure 7](#) shows a schematic example of the CATH classification scheme.

CATH provides excellent peripheral information for every protein structure in the database. Detailed graphical information as to structural motifs (60), bound ligands (61), and cross-references to many other data sources are all available.

4.3. FSSP

The FSSP database is provided by Holm and Sander (29) at the European Bioinformatics Institute (Hinxton, UK). Rather than a classification, FSSP is a list of protein structural neighbors. After each update of the PDB, each new protein is compared to all others using sequence and structure comparison methods. Thus, for each PDB entry, one can obtain a list of sequence and structure neighbors (the results of a search). There are no discrete boundaries discerning similarity from dissimilarity. Frequently, proteins that are not grouped in, e.g., SCOP or CATH, are structural neighbors within FSSP, reflecting weaker matches not always captured by other classifications that may, nevertheless, represent biologically meaningful examples (see [ref. 48](#) for an example). The main drawback to the FSSP database at present is the lack-of-domain definitions.

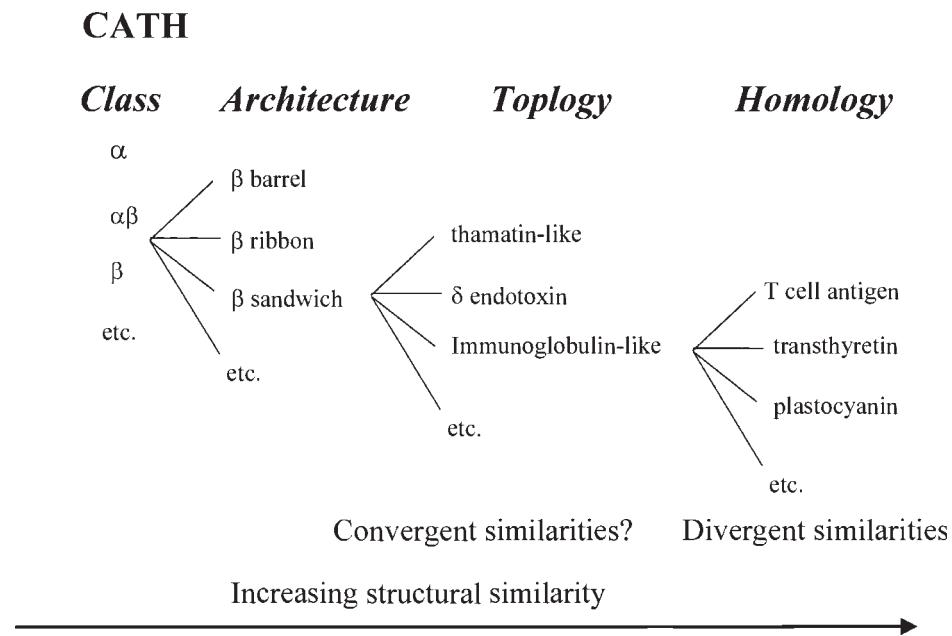


Fig. 7. Example of the classification hierarchy for the CATH database (adapted from **ref. 12**).

Multidomain proteins are compared as a whole to the database, meaning that it may be difficult to see similarities involving a rare domain when it is connected to one occurring frequently.

4.4. VAST

VAST is a program for structure comparison written by Bryant and co-workers (32) at the National Institutes of Health NCBI, and forms the basis for adding structural information to the ENTREZ database facility (62). Like FSSP, this database is more a list of similarities than a classification, with the same neighboring system found throughout the ENTREZ system. It has the advantages of being updated immediately following protein databank updates, and because of its location in ENTREZ, it contains excellent links to protein and nucleotide sequence data, and to Medline literature references.

4.5. 3Dee

The Protein Domains Database (PDB) (Barton et al., European Bioinformatics Institute, in press), provides a set of carefully defined protein domains for the entire protein databank, a structural classification and links to SCOP and other databases. In addition to providing a graphical view of superimpositions, 3DEE also provides the unique ability to view different domain assignments.

5. Notes

1. Domain assignment: Repeating units in a structure need not define individual domains, as many single domain structures possess internal symmetry, e.g., the β -trefoils, which contain three similar trefoil motifs that form a single domain. It is unlikely that the motifs could fold or be functionally active in isolation.

2. Class assignment: Assign class based on the core structure. For example, if a protein contains a β -barrel with numerous helical insertions, then it is usually best classified as all- β .
3. Fold assignment: If you have similarities involving separate domains, attempt to extend them by adding domains. For example, BGT (see **Section 3**), was originally commented to have two Rossmann fold domains (55). However, both structural domains can be superimposed on glycogen phosphorylase (56,57), indicating an ancient common ancestor.
4. Superfamily assignment:
 - a. The structure with the highest degree of structural similarity to a probe structure may not necessarily be the best candidate for superfamily or functional similarity (see adenylyl cyclase and DNA polymerase (20) in the NCBI-VAST database [63]). Partly this can be due to limitations in the structure comparison method.
 - b. Even homologous protein structures can have different functions (e.g., refs. 43,64,65). Consider, e.g., the similarity between sonic hedgehog (a factor) and DD carboxypeptidase (an enzyme) (65).
 - c. A common binding-site location is not sufficient to group proteins into the same superfamily, as some proteins appear to show binding-site similarity in the absence of homology (e.g., the α/β -barrels; see refs. 52,66).
5. Classifications:

SCOP Advantages

- a. Classification is done manually, and with careful consideration of the literature.
- b. Includes classifications for structures for which no coordinates are publicly available.
- c. Evolutionary classification is better than any other system.
- d. Interactive interface to local copy of protein databank (via RasMol [67]).

SCOP Limitations

- a. Groupings at the fold level are fairly stringent, meaning that similar structures are often not grouped together. Note that this means that proteins belonging to different folds in SCOP can still show some degree of structural similarity (e.g., Ig folds and cupredoxins).
- b. Fold/Superfamily definitions are not static. This is also an advantage, as misclassifications are corrected when more information becomes available.
- c. Some folds have been studied in more detail than others.
- d. Updates only occur about twice annually.
- e. No facility for viewing alignments or superimpositions to date.

CATH Advantages

- a. Groupings at the fold level are more lenient than in SCOP, and more useful for tasks like the assessment of protein fold recognition.
- b. Architecture division makes classification easier to follow.
- c. Excellent peripheral resources (e.g., Rasmol, ligand binding, structural characterization, and enzyme-classification annotation).
- d. Careful assignment of proteins into domains.

CATH Limitations

- a. Updates are infrequent.
- b. No facility for viewing alignments or superimpositions to date.
- c. Domains are often structural, which means that some fold/superfamily similarities are missed (e.g., the trypsin-like serine proteases).

FSSP/DALI Advantages

- a. Fully automated, and as up to date as the PDB.
- b. Provides good interactive interface to view both superimpositions and alignments of structures.

- c. Ability to search the PDB with a new structure.
- d. Statistical measure provides reliable significance for each similarity.

FSSP/DALI Limitations

- a. Fully automated, thus can contain some misclassifications owing to lack of human interpretation.
- b. Currently, the lack of domain assignments can make classification of multidomain proteins difficult.

VAST/NCBI Advantages

- a. Fully automated, and as up to date as the protein databank.
- b. Excellent crossreferencing to protein databank, protein/DNA sequence and literature data through the Entrez system (62).
- c. Statistical measure provides reliable significance for each similarity.

VAST/NCBI Limitations

- a. Similarities are detected based on arrangements of secondary-structures, which means some similarities may be missed owing to poor definitions.
- b. No domain definitions at present.

Acknowledgments

The author is grateful to Chris Rawlings, David Searls, and Ford Calhoun (SmithKline Beecham) for encouragement, and Mike Sternberg (Imperial Cancer Research Fund) for helpful discussions. Thanks also go to Richard Copley for a detailed proofreading of the manuscript.

Appendix: URLs

Structural Classifications

SCOP (MRC/LMB Cambridge, UK): <http://scop.mrc-lmb.cam.ac.uk/scop> (mirrors around the world)

CATH (University College, London, UK): <http://www.biochem.ucl.ac.uk/bsm/cath>

FSSP/DALI (European Bioinformatics Institute, Cambridge, UK): <http://www2.ebi.ac.uk/dali/fssp/fssp.html>

NCBI/VAST (NCBI, NIH, Bethesda, MD): <http://www.ncbi.nlm.nih.gov/Structure/vast.html>

DDBASE (Department of Biochemistry, Cambridge University, UK): <http://www-cryst.bioc.cam.ac.uk/~ddbase/>

3DEE (EBI, Cambridge, UK): http://circinus.ebi.ac.uk:8080/3Dee/help/help_intro.html

Algorithms

Data

Protein Databank (PDB): <http://www.pdb.bnl.gov/>

Secondary Structure Assignment

DSSP: <ftp://ftp.ebi.ac.uk/pub/software/unix/dssp/>

STRIDE: <http://www.embl-heidelberg.de/stride/stride.html>

Domain Assignment

DAD algorithm: <http://www.icnet.uk/bmm/domains/assign.html> calculates domains given a set of coordinates

DOMAK program: <http://barton.ebi.ac.uk/> downloadable program for calculating domains

Structure–Database Comparison

DALI: <http://www2.ebi.ac.uk/dali/dali.html> — compares a query set of protein coordinates to a database of known structures

SSAP: <http://www.biochem.ucl.ac.uk/~orengo/ssap.html> — information on downloading the SSAP program for protein structure alignment and superimposition.

SARF: <http://www-lmmb.ncifcrf.gov/~nicka/run2.html> — compare two protein structures from the protein databank.

SARF: <http://www-lmmb.ncifcrf.gov/~nicka/prerun.html> — download SARF2 program for structure comparison.

STAMP: <http://barton.ebi.ac.uk/> download STAMP program for structure comparison.

References

1. Orengo, C. A., Swindells, M. B., Michie, A. D., Zvelebil, M. J., Driscoll, P. C., Waterfield, M. D., and Thornton, J. M. (1995) Structural similarity between the pleckstrin homology domain and verotoxin: the problem of measuring and evaluating structural similarity. *Protein Sci.* **4**, 1977–1983.
2. Colloc'h, N., Etchebest, C., Thoreau, E., Henrissat, B., and Mornon, J. P. (1993) Comparison of three algorithms for the assignment of secondary structure in proteins: the advantages of a consensus assignment. *Protein Eng.* **6**, 377–382.
3. Frishman, D. and Argos, P. (1995) Knowledge-based protein secondary structure assignment. *Proteins: Struct. Funct. Genet.* **23**, 566–579.
4. Taylor, W. R. (1992) Patterns, predictions and problems, in *Patterns in Protein Sequence and Structure* (Taylor, W. R., ed.), Springer-Verlag, Berlin.
5. Sternberg, M. J. E., Hegyi, H., Islam, S. A., Luo, J., and Russell, R. B. (1995) Towards an intelligent system for the automatic assignment of domains in globular proteins. *Ismb* **3**, 376–383.
6. McLachlan, A. D. (1979) Gene duplications in the structural evolution of chymotrypsin. *J. Mol. Biol.* **128**, 49–79.
7. Holm, L. and Sander, C. (1994) Parser for protein folding units. *Proteins: Struct. Funct. Genet.* **19**, 256–268.
8. Siddiqui, A. S. and Barton, G. J. (1995) Continuous and discontinuous domains: an algorithm for the automatic generation of reliable protein domain definitions. *Protein Sci.* **4**, 872–884.
9. Sowdhamini, R. and Blundell, T. L. (1995) An automated method involving cluster analysis of secondary structures for the identification of domains in proteins. *Protein Sci.* **4**, 506–520.
10. Islam, S. A., Luo, J., and Sternberg, M. J. E. (1995) Identification and analysis of domains in proteins. *Protein Eng.* **8**, 513–525.
11. Swindells, M. B. (1995) A procedure for detecting structural domains in proteins, *Protein Sci.* **4**, 103–112
12. Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., and Thornton, J. M. (1997) CATH—a hierachic classification of protein domain structures. *Structure* **5**, 1093–1108.

13. Jones S., Stewart M., Michie A., Swindells M. B., Orengo C., and Thornton, J. M. (1998) Domain assignment for protein structures using a consensus approach: characterization and analysis. *Protein Sci.* **7**, 233–242.
14. Holm, L. and Sander, C. (1998) Dictionary of recurrent domains in protein structures. *Proteins: Struct. Funct. Genet.* **33**, ???–???.
15. Russell, R. B. (1994) Domain insertion. *Protein Eng.* **7**, 1407–1410.
16. Levitt, M. and Chothia, C. (1976) Structural patterns in globular proteins. *Nature* **261**, 552–558.
17. Murzin, A. G., Brenner, S. E., Hubbard, T. J., and Chothia, C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540.
18. Holm, L. and Sander, C. (1997) Mapping the protein universe. *Science* **273**, 595–602.
19. Holm, L. and Sander, C. (1997) An evolutionary treasure: unification of a broad set of amidohydrolases related to urease. *Proteins: Struct. Funct. Genet.* **28**, 72–82.
20. Artymiuk, P. J., Poirette, A. R., Rice, D. W., and Willett, P. (1997) A polymerase I palm domain in adenylyl cyclase? *Nature* **388**, 33–34.
21. Orengo, C. A. (1994) Classification of protein folds. *Curr. Opin. Struct. Biol.* **4**, 429–440.
22. Holm, L. and Sander, C. (1994) Searching protein structure databases has come of age. *Proteins: Struct. Funct. Genet.* **19**, 165–183.
23. Holm, L. and Sander, C. (1997) New structure — novel fold? *Structure* **5**, 165–171.
24. Taylor, W. R. and Orengo, C. A. (1989) Protein structure alignment. *J. Mol. Biol.* **208**, 1–22.
25. Orengo, C. A., Brown, N. P., and Taylor W. R. (1992) Fast structure alignment for protein databank searching. *Proteins: Struct. Funct. Genet.* **14**, 139–167
26. Alexandrov, N. N., Takahashi, K., and Go, N. (1992) Common spatial arrangements of backbone fragments in homologous and non-homologous proteins. *J. Mol. Biol.* **225**, 5–9.
27. Russell, R. B. and Barton, G. J. (1992) Multiple sequence alignment from tertiary structure comparison. Assignment of global and residue confidence levels. *Proteins: Struct. Funct. Genet.* **14**, 309–323.
28. Holm, L. and Sander, C. (1993) Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233**, 123–138.
29. Holm, L. and Sander, C. (1994) The FSSP database of structurally aligned protein fold families. *Nucleic Acids Res.* **22**, 3600–3609. *See also Nucleic Acids Res.* **24**, 206–210.
30. Mitchell, E. M., Artymiuk, P. J., Rice, D. W., and Willett, P. (1990) Use of techniques derived from graph theory to compare secondary structure motifs in proteins. *J. Mol. Biol.* **212**, 151–166.
31. Quanta. Molecular Simulations, San Diego, CA.
32. Gibrat, J.-F., Madej, T., Bryant, S. H. (1996) Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.* **6**, 377–385.
33. Kleywegt, G. J. and Jones, T. A. (1997) Detecting folding motifs and similarities in protein structures. *Methods Enzymol.* **277**, 525–545.
34. Russell, R. B. and Ponting, C. P. (1998) Protein fold irregularities that hinder sequence analysis. *Curr. Opin. Struct. Biol.* **8**, 364.
35. Murzin, A. G. (1998) Probable circular permutation in the flavin-binding domain. *Nat. Struct. Biol.* **5**, 101.
36. Liepinsh, E., Kitamura, M., Murakami, T., Nakaya, T., and Otting, G. (1997) Pathway of chymotrypsin evolution suggested by the structure of the FMN-binding protein from *Desulfovibrio vulgaris*. *Nat. Struct. Biol.* **4**, 975–979.
37. Chothia, C. (1992) One thousand families for the molecular biologist. *Nature* **357**, 543–544.
38. Orengo, C. A., Jones, D. T., and Thornton, J. M. (1994) Protein superfamilies and domain superfolds. *Nature* **372**, 631–634.

39. Blundell, T. L. and Johnson, M. S. (1993) Catching the common fold. *Protein Sci.* **2**, 877–883.
40. Crippen, G. M. and Mariov, V. (1995) How many protein folding motifs are there? *J. Mol. Biol.* **252**, 144–151.
41. Russell, R. B., Saqi, M. A. S., Sayle, R. A., Bates, P. A., and Sternberg, M. J. E. (1997) Recognition of analogous and homologous protein folds. Analysis of sequence and structure conservation. *J. Mol. Biol.* **269**, 423–439.
42. Jones, D. T. (1997) Progress in protein structure prediction. *Curr. Opin. Struct. Biol.* **7**, 377–387.
43. Murzin, A. G. (1993a) Sweet tasting protein monellin is related to the cystatin family of thiol proteinase inhibitors. *J. Mol. Biol.* **230**, 689–694.
44. Russell, R. B. and Barton, G. J. (1994) Structural features can be unconserved in proteins with similar folds. An analysis of side-chain to side-chain contacts, secondary structure and accessibility. *J. Mol. Biol.* **244**, 332–350.
45. Holm, L. and Sander, C. (1997) Decision support system for the evolutionary classification of protein structures. *Intel. Syst. Mol. Biol.* **5**, 140–146.
46. Swindells, M. B. (1993) Classification of doubly wound nucleotide binding topologies using automated loop searches. *Protein Sci.* **2**, 2146–2153.
47. Murzin, A. G. (1995) A ribosomal protein module in EF-G and DNA gyrase, *Nat. Struct. Biol.* **2**, 25–26.
48. Holm, L. and Sander, C. (1995) DNA polymerase β belongs to an ancient nucleotidyl-transferase superfamily. *Trends Biochem. Sci.* **20**, 345–347.
49. Holm, L. and Sander, C. (1997) An evolutionary treasure: unification of a broad set of amidohydrolases related to urease. *Proteins: Struct. Funct. Genet.* **28**, 72–82.
50. Holm, L. and Sander, C. (1997) Enzyme HIT. *Trends Biochem. Sci.* **22**, 116.
51. Park J., Teichmann S. A., Hubbard T., and Chothia C. (1997) Intermediate sequences increase the detection of homology between sequences. *J. Mol. Biol.* **273**, 349–354.
52. Russell, R. B., Saseini, P. D., and Sternberg, M. J. E. (1998) Supersites within superfolds. Binding site similarity in the absence of homology. *J. Mol. Biol.* **28**, 903.
53. Hol, W. G., van Duijnen, P. T., and Berendsen, H. J. (1978) The alpha-helix dipole and the properties of proteins. *Nature* **273**, 443–446.
54. Pennisi, P. (1998) Taking a structured approach to understanding proteins. *Science* **279**, 978–979.
55. Vrielink, A., Ruger, W., Driessens, H. P., Freemont, P. S. (1994) Crystal structure of the DNA modifying enzyme beta-glucosyltransferase in the presence and absence of the substrate uridine diphosphoglucose. *EMBO J.* **13**, 3413–3422.
56. Holm, L. and Sander, C. (1995) Evolutionary link between glycogen phosphorylase and a DNA modifying enzyme. *EMBO J.* **14**, 1287–1293.
57. Artymiuk, P. J., Rice, D. W., Poirette, A. R., and Willett, P. (1995) β -Glucosyltransferase and phosphorylase reveal their common theme, *Nat. Struct. Biol.* **2**, 117–120.
58. Zhang, G., Liu, Y., Ruoho, A. E., and Hurley, J. H. (1997) Structure of the adenylyl cyclase catalytic core. *Nature* **386**, 247–253.
59. Tesmer, J. J., Sunahara, R. K., Gilman, A. G., and Sprang, S. R. (1997) Crystal structure of the catalytic domains of adenylyl cyclase in a complex with Gs α .GTP γ S. *Science* **278**, 1907–1916.
60. Laskowski, R. A., Hutchinson, E. G., Michie, A. D., Wallace, A. C., Jones, M. L., and Thornton, J. M. (19??) PDBsum: a Web-based database of summaries and analyses of all PDB structures. *Trends Biochem. Sci.* **22**, 488–490.
61. Wallace, A. C., Laskowski, R. A., and Thornton, J. M. (1995) LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. *Protein Eng.* **8**, 127–134.

62. Schuler, G. D., Epstein, J. A., Ohkawa, H., and Kans, J. A. (1996) Entrez: molecular biology database and retrieval system *Methods Enzymol.* **266**, 141–162.
63. Bryant, S. H., Madej, T., Janin, J., Liu, Y., Ruoho, A. E., Zhang, G., and Hurley, J. H. (1997) A polymerase I palm in adenylyl cyclase? *Nature* **388**, 34.
64. Murzin, A. G. (1993b) Can homologous proteins evolve different enzymatic activities? *Trends Biochem. Sci.* **18**, 403–405.
65. Murzin, A. G. (1996) Structural classification of proteins: new superfamilies. *Curr. Opin. Struct. Biol.* **6**, 386–394.
66. Farber, G. K. and Petsko, G. A. (19??) The evolution of a/b barrel enzymes. *Trends Biochem Sci* **15**, 228–234.
67. Sayle, R. A. and Milner-White, E. J. (1995) RASMOL Biomolecular Graphics for all. *Trends Biochem. Sci.* **20**, 374.
68. Kraulis, P. J. (1991) Molscript: a program to produced detailed and schematic plots of protein structures. *J. Appl. Cryst.* **24**, 946–950.
69. Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. E., Brice, M. D., Rodgers, M. D., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.

Protein Threading

Andrew E. Torda

1. Introduction

Theoreticians have been trying to predict protein structure based on sequence information for decades. Literally, more than a quarter century ago, there were optimistic reports that one could use simulation methods to calculate the structure of a small protein given only its sequence (1,2). To this day, devotees of this approach persevere and may ultimately win over the problems with force fields and the enormous search space. In the meantime, a class of protein structure methods have developed, traveling under names such as “protein threading” and “fold recognition.”

In the most general case, protein structure prediction is a truly ferocious problem, whose size can be made clear by a model calculation. Imagine that every peptide plane (ω) angle is fixed, planar, and *trans*. At every residue, one still has the phi and psi (ϕ , ψ) backbone angles, and there might be two or three preferred local conformations. Even in this unrealistically simple case, a protein of 100 residues has between 10^{30} (2^{100}) and 10^{47} (3^{100}) conformations to be considered. These numbers come without even considering side chains or the fact that backbone conformations are continuous variables and do not fall into two or three discrete locations. At the risk of being a doomsayer, one could also note that computers double their speed every few years, but the size of the computational problem doubles with every extra amino acid. If you can predict the structure of a 50-residue protein this year, it could be a few more years until you can do 51 residues.

Rather than give up, one can look for a simpler version of the problem or a subset that might be solvable. Proteins probably do not manage to fold into every shape a polymer chemist could imagine (3). Instead, there may be only a finite number of protein folds in nature (4–12), and certain kinds of structure seem to be remarkably popular among apparently unrelated sequences (13–20). This has an important consequence. Even when an experimentalist may not expect any similarity, the structure they are about to solve may be quite similar to one that is already known. In recent years, less than 15% of structures deposited in the Protein Data Bank (PDB) (21) could even be considered new folds. This is the rationale for the entire area of threading or fold recognition. The protein sequence of interest may have no detectable sequence homology to anything of known structure, but there may well be some similar structure waiting to be recognized. If one can recognize the related structure and do a sequence-to-structure alignment calculation, one should produce a useful model. It would be even better if one could detect those cases where the method will fail and the sequence will fold to a

From: *The Proteomics Protocols Handbook*
Edited by: J. M. Walker © Humana Press Inc., Totowa, NJ

new structure. Unlike normal sequence comparison, the alignment method should take advantage of the structural information of each template.

Some of the statements above are poorly quantified, and no two groups may agree on what constitutes a new fold and how often one is found. At the same time, one can quote some findings on when sequence similarity is sufficient to infer similarity. It is often said that if a sequence has more than 30% identity to a known structure, it is possible to build a reasonable model, but at 20 to 25%, the similarity may be purely coincidental. In practice, this rule of thumb should not be used. For example, Brenner et al. give the example of a pair of proteins with 39% sequence identity, but no detectable structural similarity (22). Rather than look for a single number (sequence identity), one must look at the length of the proteins and aligned regions. Intuitively, it is obvious that 40% similarity over a small peptide is much more likely to happen by chance than 40% over 500 residues. This issue has been addressed by comparing large numbers of protein pairs (22,23). Basically, for 50 residues, you would want 40% sequence identity before deeming it reliable, but for 250 residues, 25% might suffice. These numbers are purely statistical, so there is always the distinct possibility that a weak sequence identity does not reflect structural similarity. A better approach is to resist the temptation to concentrate on pairwise numbers. Sequence database searching programs such as FASTA (24) and Basic Local Alignment Search Tool (BLAST) (25) estimate the reliability of a sequence match by looking at it in the context of the whole library of sequence scores. More recent, iterated versions of BLAST (26) render the interpretation of pairwise sequence identity even more meaningless. Because programs such as position specific iterative (PSI)-BLAST work with a sequence profile, a database hit is often statistically reliable, even with less than 20% sequence identity.

With these results in mind, one is left with an unsatisfying but practical way to decide whether or not a threading calculation is of interest. If a simple database search finds a reliable homolog of known structure for a sequence, it is the best way to build a model. If a careful, exhaustive, iterated database search cannot find a statistically reliable homolog, or if you wish for a confirmation of your beliefs, a threading calculation is called for.

2. Threading Overview

For a threading calculation, there are some elements common to most programs. Firstly, you have a sequence of interest and a library of templates or known structures, as shown in [Fig. 1](#). Presumably, these are PDB structures and the library contains all known protein folds. Next, one takes the sequence and “threads” it through each template in the library, as shown in [Fig. 2](#). The word *threading* implies that one drags the sequence (ACDEFG...) step by step through each location on each template, but really one is searching for the best arrangement of the sequence, as measured by some score or quasi-energy function. In the third alignment in [Fig. 2](#), the sequence of interest has been aligned so it skips over part of the template. Finding the best arrangement of residues, including these gaps and insertions, is the problem of sequence-to-structure alignment, discussed later. Finally, all the candidate models with their scores are collected in [Fig. 3](#). The best-scoring (lowest-energy) one is then taken as the structure prediction.

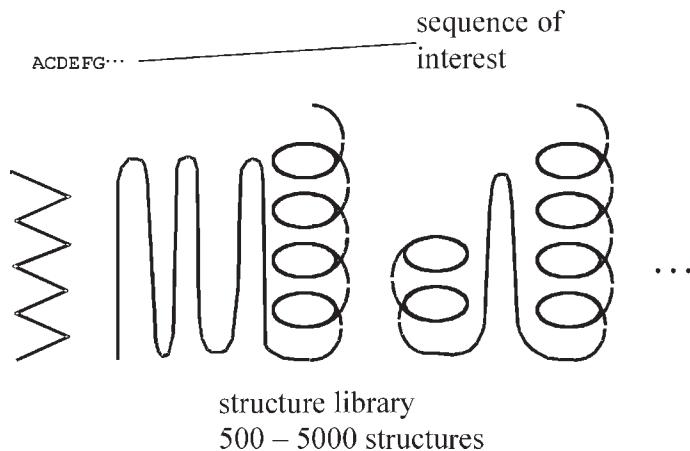


Fig. 1. A sequence of unknown structure and a template library (collection of known structures).

Before considering technical details, this simple picture highlights some problems. First, the result will depend on the size and details of the library at the first step. Typically, different groups will have libraries ranging from 500 to 5000 members, and there is definitely no consensus as to the optimal size. On the one side, the library should be small. Threading calculations are often slow, so one may want to use the smallest possible library. At the same time, threading score functions are far from perfect. The closer a template is to the correct answer, the more likely the sequence is to score well on it. Thus libraries should be large. However, imagine you include 10 small variations on one particular protein fold, but one representative from another. Statistically, the well-represented fold is more likely to score well by chance. Thus, libraries should be small. Finally, there is no agreed-upon way to select the particular members.

One could argue that library members should be single chains or domains. One could argue that from a protein family, one should select the member that has the best-quality coordinates or the one that is in some way most representative of the family of structures. Continuing in this vein, the idea of representative structures implies that one has already clustered all known proteins down to a set of families. This could be based on sequence identity or some measure of structural similarity. Finally, one is not even limited to simple PDB coordinates. Madej et al. used a library based on extracted cores from proteins (27), and some have suggested that the structures in the library could be optimized so as to make the ranking of models statistically more reliable (28).

The simple set of pictures also introduces the next questions. One needs some way to score the sequences and structures, and then one will need a way to find the best alignment of a sequence on a template structure.

3. Score Functions

Much of computational chemistry is centered about finding the best conformation for some molecule. In protein calculations, this usually involves a classical atomistic model for the potential energy of a system (29–32). In the case of protein threading,

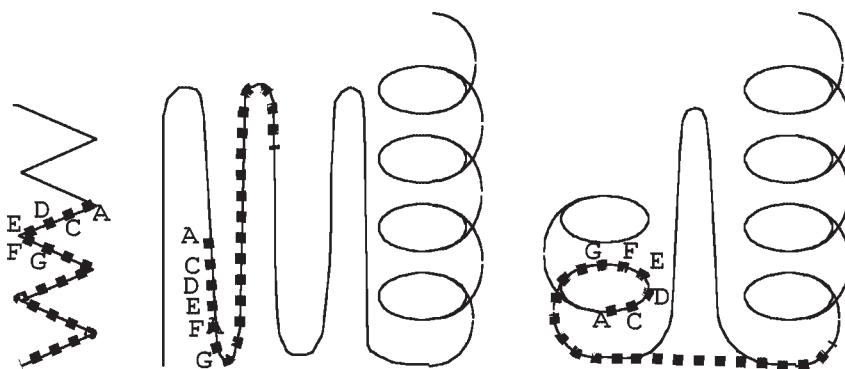


Fig. 2. Threading and aligning a sequence through a template library.

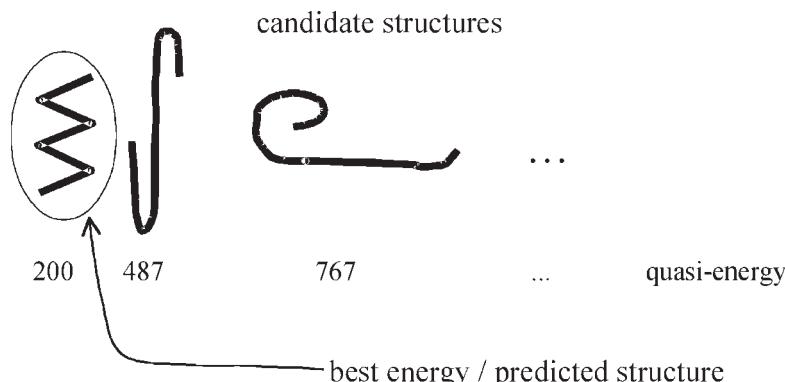


Fig. 3. Set of candidate structures for sequence. Models correspond to alignments from [Fig. 2](#).

one is not bound by this philosophy. One really just wants a score function that is capable of recognizing correct arrangements of protein residues. It need not perform all the feats of a conventional force field or work in the same application areas. For example, in the procedure described so far, one need never take the derivative of score with respect to coordinates, although this would usually be an essential step in many energy minimization schemes or dynamics simulations. Similarly, if a score function is used only in a threading context, it will never be faced with atoms hitting each other, stretched bonds, distorted angles, or any of the other situations that might confront a normal force field. This should mean that it is easier to build a through-space, threading score function than a full force field for molecular mechanics calculations.

Threading score functions are also usually more coarse-grained than those used in a real energy calculation. In a threading calculation, the sequence residues are placed on the backbone of the template structure, and from there, one can calculate ideal coordinates for the C^β atom. One does not know where the rest of the residue is, so it will be extremely difficult to use a score function, which uses the coordinates of all atoms. Consequently, a threading score function usually represents each residue by one or a few interaction sites. Often, most of the chemical identity of a residue comes from an interaction site located at the C^β residue or a point closer to the side-chain center of mass.

With this level of representation, it is not common to rely on pure physics. For example, a threading score function does not usually have a term like Coulomb's law for electrostatics or a Lennard-Jones term for other atomic interactions. Instead, there are two common approaches to building a score function: (1) potentials of mean force, and (2) from an optimization calculation.

Potentials of mean force are described in statistical-mechanics textbooks, often based on the distributions of particles in simulations (33). At a coarse-grained level, suitable for threading, these were parameterized in the 1970s, 1980s (34,35), and repeatedly since (36,37). In the protein literature, they usually travel under names such as statistics- or Boltzmann-based force fields or score functions, and sometimes even knowledge-based force fields. The principle is easily illustrated by example. If we know the concentration of two species (particle types A and B), we can calculate how often they will be observed at a certain distance from each other by chance. If AB pairs are seen more often than expected at 5 Å, the system is behaving as if there is some kind of energy minimum between the particles at 5 Å. If the pairs of particles are seen less often than expected at that distance, it appears that the interaction is unfavorable at that distance. To formalize this, one must remember the words of Herr Boltzmann and refer to some free energy, G , which is a function of the distance r_{AB} between particles of types A and B:

$$G(R_{AB}) = kT \ln \frac{\rho_{r_{AB}}}{\rho_{r_{AB}}^0}$$

k and T have their normal meanings of Boltzmann's constant and temperature. $\rho_{r_{AB}}$ is the observed frequency of AB pairs at distance r . $\rho_{r_{AB}}^0$ is less obvious. It is the frequency of AB pairs at distance r you would expect to see by chance. This formulation is very general and is easy to apply to proteins. Instead of considering particles A and B, you might consider C^β atoms on Ala and Trp residues. Then you could build a potential of mean force for Ala/Trp C^β atoms, and you could do this for every combination of amino acids. One could even parameterize this kind of function in terms of torsion angles or any other property that seems to be important for determining a protein's structure.

This framework relies on measuring $\rho_{r_{AB}}$ and estimating $\rho_{r_{AB}}^0$. With protein structures, the best you can do for $\rho_{r_{AB}}$ is collect statistics from the PDB and pretend this is a statistical mechanical ensemble. For $\rho_{r_{AB}}^0$, the frequency you would expect by chance, one can use an analogy with chemistry, treat A and B as species in solution and consider the concentrations [A] and [B]. For proteins, you could treat the amino acid composition as if it were a mole fraction.

In practice, it may not be valid to treat proteins as if they were disconnected solutions of amino acids (38). There might be artifacts due to packing effects and problems with the fictitious statistical mechanics (39,40). It is hard to see what kind of ensemble a survey of the PDB really is, but it has been argued that the resulting numbers are Helmholtz free energies (41,42). Pragmatically, it may not matter much how close these statistical score functions are to free energies. They definitely do reflect statistical tendencies within proteins, and this may be all one needs for a threading application

(43). Despite the debate over details, the approach is clear. One takes a large set of proteins, collects statistics, and converts them to a score function. One then expects this function to work well for proteins not included in its parameterization.

If one believes the statistical mechanical basis for the statistics-based score functions, then one is dealing with a real energy that is properly calibrated against the rest of the world. There is, however, a quite different school of thought. If one is dealing with protein threading, or perhaps structure prediction in general, than maybe one need not be too concerned with real energies or reproducing the physics of protein folding. It is not important that a score function represent every false minimum or kinetic trap that a protein visits when folding. Instead, one wants a function that can distinguish between a correct and an incorrect structure. This function will usually have some adjustable parameters, and perhaps these can be optimized for protein fold recognition (44–49). The result may be a function that does not look like a conventional model for energy, but formally is still a force field.

Although this approach sounds attractive, it is not so simple to put in place. First, one must select the underlying basis functions. In the literature, these have ranged from quasi-Lennard–Jones terms (44,48) to various kinds of sigmoidal function (47,50). Next, there is a problem with the question as posed. We want to distinguish between correct and incorrect structures. We can say that the correct structure is whatever is given in the PDB, but unfortunately, there is almost an infinity of incorrect structures for a sequence, and one would like the score function to penalize all of them. One way to encode this idea is to adopt a statistical approach and try to consider the distribution of incorrect structures (50–53). Imagine you have some score function that produces an energy, E_X , for your sequence on a template structure X. If you generate a large number of incorrect or alternate structures, you can calculate their energies (E_{alt}). One convenient way is to take a sequence and put its residues onto every template you can find that is larger than the sequence. This guarantees that the alternative structures are protein-like. Next, you could plot out a histogram of the energies of the alternate structures, as shown in **Fig. 4A**. Empirically, the distribution of alternate energies (E_{alt}) looks like a Gaussian curve (50), and it can even be theoretically justified (53). As well as the distribution of E_{alt} , **Fig. 4A** shows E_{nat} , the energy of the native structure. What we would like is to adjust the force-field parameters so that E_{nat} is well separated from the mean energy of all the wrong structures, $\langle E_{alt} \rangle$. In other words, one wants to make $z-score = \frac{E_{nat} - \langle E_{alt} \rangle}{\sigma_{E_{alt}}}$ as large as possible, as shown in **Fig. 4B**. At the same time, one does not want to simply scale the figure. Instead, one should keep the standard deviation of the distribution, as small as possible. This idea is captured by the standard statistical term, the Z-score, given by

$$z-score = \frac{E_{nat} - \langle E_{alt} \rangle}{\sigma_{E_{alt}}}$$

So, with this philosophy, the aim is to find the score function that gives the greatest Z-score. At the same time, the score function should not only work for one protein, it should work (ideally) for every protein it will ever be faced with. Then, the approach usually taken is to select a set of proteins for parameterization and to adjust the force

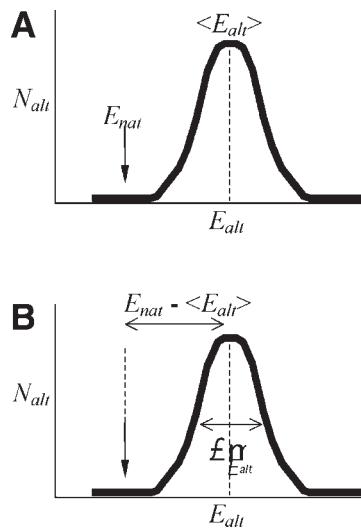


Fig. 4. Z-score optimization for force-field construction. E_{alt} = the energy of an alternative conformation; E_{nat} = the energy of the native (correct) structure; $\langle E_{alt} \rangle$ = the mean of the alternative conformation energies; N_{alt} = the number of alternative conformations of a given energy. In (B), $\sigma_{E_{alt}}$ is the standard deviation of the E_{alt} distribution; $E_{nat} - \langle E_{alt} \rangle$ is the difference between the energy of the native conformation and the average of alternate conformation energies.

field to give the best Z-score over the whole set. Numerically, one has to take the expression for the Z-score, expand it in terms of the underlying energy expressions and parameters, and use a numerical optimization method to adjust the parameters. Hopefully, the score function will then work well, even with proteins it has never faced before.

4. Sequence-to-Structure Alignment

Given some kind of score function, there are two areas where it will be applied. First, one needs the best arrangement of sequence residues on the template (sequence-to-structure alignment). Second, one needs a score function to rank the final structures, discussed under Section 6. As discussed later, this may really lead to two different score functions.

Finding the best alignment of a sequence to a template structure is vital, but perhaps still a problem. In 1995, it was noted that sequence-to-structure alignments were typically error-prone (54). More recently, the problem has been re-examined along with the consequences for fold recognition (55). One can see the severity of the problem even with tiny errors. The average distance between C^α atoms is 3.8 Å, so a single residue misalignment would be enough to render a model useless in an application like drug design, even if the template molecule is close to correct for the unknown structure. Next, a larger misalignment, putting a gap of several residues at the wrong location, could easily send β -strand residues to a piece of α -helix or random coil. In the context of fold recognition, the problem is worse. Looking at Fig. 3, one can see that if the alignments are wrong, the models and calculated scores are wrong, and it makes no sense to rank them.

There are two very clear reasons why sequence-to-structure alignment problems are difficult. The first is that the simple score functions commonly used are far from perfect. It is not practical to use the best atomistic force fields in the literature, and the simpler, more coarse-grained ones cannot work as well. Next, the problem is formally difficult and in the most general case is NP-complete (56). This can be explained by comparison with sequence-to-sequence comparison. **Figure 5A** shows an alignment of “ACDEF” to some template that has both a sequence “QRSTVW” and the structure shown. With the two gaps present, only three residues are aligned. If we consider a classic dynamic programming calculation for the alignment (57,58), we have to construct a scoring matrix as shown in **Fig. 5B**. The elements of the matrix reflect the similarity of amino acids. For example, the element indexed by “DR” comes from looking up the similarity of aspartate and arginine in a literature substitution matrix. The path marked on the score matrix corresponds to the alignment in part A.

In contrast, consider the situation in **Fig. 5C**. The same sequence is to be aligned, but now to a structure. One wants to construct a similar score matrix, as shown on the right, but it is not possible. Looking again at the cell indexed by D2, one wants some kind of compatibility score. This implies the interactions shown on the left. While we can place residue D at position 2, the interactions with the other sites cannot be calculated, since the other residues have not been aligned. For example, if one wants to calculate the interaction between sites 2 and 4, we may say that D is at place 2, but one does not know who it is interacting with at site 4. Clearly, sequence-to-structure alignments are routinely calculated, so the problem is not impossible. It merely requires heuristics and approximations blending optimism, brute force, and cunning.

One approach is to give up on dynamic programming completely. One has a score function and a discrete space, so the score of a trial alignment can always be calculated. In that case, the problem seems well suited to Monte Carlo/simulated annealing (59). This, however, does not alleviate the problem of the huge search space. Allowing gaps and insertions at any position and of any length leads to a combinatorial explosion of possibilities. The calculation can be made tractable by restricting the search space and forbidding gaps except in recognized loops in template structures (27,60).

In contrast, a dynamic programming approach has the advantage that it is deterministic and there are at least three approximations that squeeze a pairwise, through-space calculation into the framework of **Fig. 5**. First, one could use the sequence of the template structure to start a process. In **Fig. 5C**, the template has been drawn without its original sequence. One can, instead, leave the template residues in place. Then, to fill out an element in the score matrix such as the D2 position, one knows the interaction partners. For example, the first interaction could be calculated as D at position 2 with an S at position 3, where the S comes from the template sequence. After calculating a first alignment, the residue identities could be taken from the correct sequence and another alignment calculated. This method, usually known as the frozen approximation, can be iterated a fixed number of times or until convergence (36,61,62).

Another method for approximating the missing information was introduced by Jones et al. in 1992, who used a second level of dynamic programming (37). To continue to concentrate on one matrix element, one could conceptually place residue D at position 2 and then arrange the rest of the sequence to interact as favorably as possible. To score D at position 3, one would again recalculate the best arrangement of its sequence neigh-

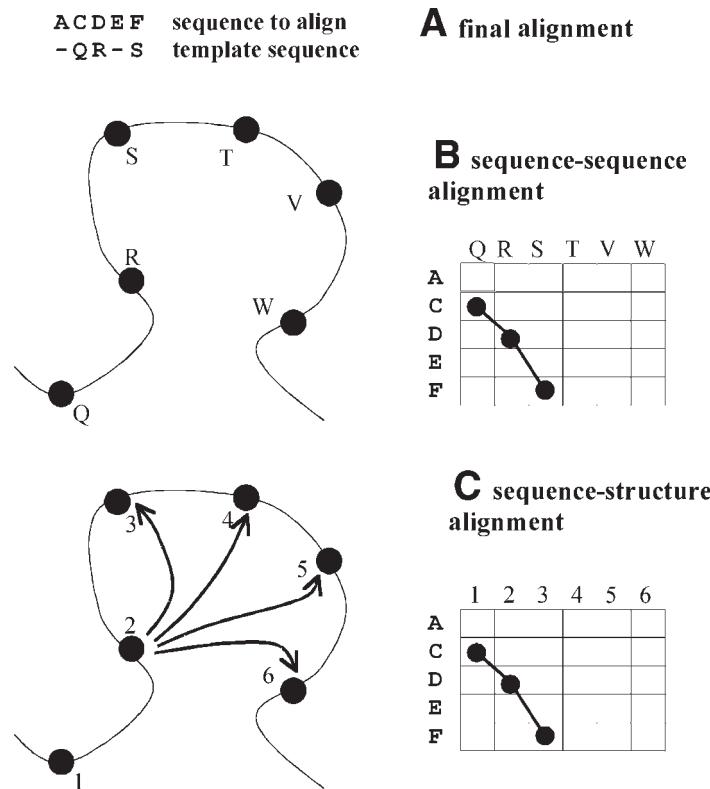


Fig. 5. Comparison of sequence-to-sequence and sequence-to-structure alignments.

bors. This is still only an approximation to the correct answer, and has been described as finding the best arrangement for every residue that it could have (63).

Third, one could modify the score function itself so as to make it suitable for a pairwise calculation. This can be done by using a score function that uses the identity of only one member of each interaction pair, like DX, EX, FX, ... Furthermore, the score function can actually be optimized to work in this mean-field manner (64).

Given the approximations necessary to treat the alignment problem by dynamic programming, several groups have developed branch and bound methods. Working in the phenomenally complex space of possible alignments, they try to successively rule out regions that cannot contribute to the answer (65–68). The only disadvantage is that they tend to be slow and difficult to implement. A recent version appears to be both swift and remarkably effective (69). If more people were capable of programming these approaches, they might displace the ugly approximations in common use.

5. Fold Recognition

If the steps described so far have been successful, one has a library of protein templates that is comprehensive and representative. There is a score function and a fast method for producing the best possible sequence-to-structure alignments and thus the best models possible. Unfortunately, the problem is still not solved. Imagine one has a library of 1000 structures and only one of the templates is close to the correct answer.

It is an act of faith to assume that the most correct model is the one that scored best during the alignment step.

One can introduce the problem of fold recognition by comparing it with a sequence database search. In that case, one assumes that the more similar a sequence, the more similar residues will have been aligned and the higher the score will be. In protein threading, one uses similar reasoning and says that only a similar template will provide a framework that lets the sequence residues interact favorably, so templates will score better, the closer they are to the correct answer. Unfortunately, it has been pointed out that the argument is not strictly valid (70). A sequence-to-structure alignment allows residues to move along non-physical degrees of freedom. In other words, a sequence-to-structure alignment may produce a favorable arrangement of residues in space, but it may not be one that occurs in nature. In practical terms, a good sequence-to-structure alignment method may arrange a sequence so that secondary structure is formed and hydrophobic residues will be close to each other, but on a completely wrong template.

Perhaps one should not even expect that one score function should be best for both arranging residues on a template and then ranking the models (64,71). When calculating alignments, the score function is being asked which parts of a sequence are more suitable than other portions of the sequence for certain parts of a structure. When ranking models, one is asking a score function to rank the same sequence in different conformations. The situation, however, is even more interesting from a statistical point of view.

Imagine a score function that is useful for both sequence-to-structure alignments and recognizing correct models, and has no systematic bias or error. The only failing it has is a susceptibility to some quasi-random noise. In this case you might take the scores of your models, plot out the distribution, and count the number of standard deviations (σ) that separate your best scores from the mean of the distribution. If your best score is 10σ from the mean, it is almost certainly not a chance occurrence. If it is one σ from the mean, there is a significant chance that it is simply coincidental. Unfortunately, this approach, which was popular some years ago, is woefully inaccurate. The scores, especially when gaps and insertions are allowed, will be far from Gaussian distributed.

In pure sequence comparison, a mixture of theory and empiricism has been applied to assessing the significance of scores by estimating p -values (probability of a score occurring by chance) and E values (expected number of times the score will be seen given the size of the database). For ungapped sequence comparisons, one can assume an extreme-value or Gumbel distribution (72–75). For gapped alignments, this is a useful approximation, but may not be absolutely correct (76). For sequence-to-structure alignments, the problem is worse. As you add residues to a sequence, the score does not grow linearly. Instead, each residue you add may interact with its $N - 1$ neighbors, so one might expect scores to grow with N^2 . Unfortunately, the use of interaction cutoffs means that a conventional pairwise interaction score is expected to grow as N_k , where k is between 1 and 2, but varies depending on protein size. This means that the analytical formulae or regression approaches used in sequence comparison will not work with sequence-to-structure alignments.

Sommer et al. actually had some success treating sequence-to-structure scores as if they followed a known distribution (77), but there is a different philosophy available. If one does not know what the most important features of a reliable model are, one could instead take the likely descriptors and use a machine-learning method to see what is useful. These would include the length of the sequence and template, the length of the alignment, and various score function components (78–81). As is common in neural networks, the approach is often effective, but not transparent. It is interesting that this could be interpreted as an example of using one function for sequence-to-structure alignments and a different one for ranking the resulting models.

With all these caveats, it is interesting to note that automatic prediction servers do give estimates of confidence in predictions. One should bear in mind that these are approximations and probably not as accurate as the statistical estimates for pure sequence comparison produced by programs such as BLAST, PSI-BLAST, or FASTA (24–26).

6. Threading Implementations and the Broader Context

To devotees, pure threading has an intellectual appeal. By using structural information, one should be able to detect similarities that are too weak to find by sequence-based methods. With structural information, one would hope to find similarities even when there is no obvious evolutionary connection between a target sequence and close template. In practice, none of this may be true. Only a very brave spectator would name the best method for alignment and fold recognition, but it would be hard to argue that pure sequence-based methods are not among the very best. It is true that a simple sequence comparison method does not work well with weak homology, but current methods are much more advanced. Both PSI-BLAST (26) and the hidden Markov model methods (82–84) use families of related sequences, taking advantage of the site-specific information found in a sequence alignment and the fact that although proteins A and B are not obviously similar, they are both reliably related to some protein C.

This probably does not spell the end of threading approaches. Instead, most threading approaches now incorporate information beyond pure through-space scoring information. For example, consider again **Fig. 5**, which shows a score matrix for some sequence against a template of known sequence and structure. **Figure 5B,C** shows different score matrices from the sequence-sequence and sequence-structure terms. If they offer independent information, there is no reason they cannot be combined. This implies some weighting of the different terms either by trial and error (85) or even by applying a numerical optimization method (86). Rather than simply add in a sequence term, one can take advantage of the profiles of sequences related to the sequence of interest, the library template, or both (81,85,87). Obviously, combining sequence-to-sequence and sequence-to-structure terms is useful only if they contain independent information, but all the proponents would assert that they do.

This idea can be extended to other kinds of information. The secondary structure of a template is easily calculated. If one could reliably predict the secondary structure of a sequence, one could match it to the template. Even without perfect secondary-structure predictions, they certainly provide more signal than noise, and are routinely added to threading calculations (78,81,87–95).

7. Context, Application, and Obsolescence

Given the selection of methods in the literature, threading means different things to different groups. Whatever the definition, is it ever useful and are there places where it should be avoided? For the sake of argument, one can call threading some method that implements a through-space scoring function, combined with some of the terms from Section 6 and performs sequence-to-structure alignments.

First, one can say that threading must produce better alignments than methods using only sequence information. This is true, because structure only adds information. If it is not true in practice, it means that implementations are not optimal or one is not making a fair comparison. Pure sequence-based methods with profiles are extremely sensitive in finding remote homologs. Pure outdated threading methods are not as sensitive. Newer threading methods use sequence profiles and have absorbed many of the methods of sequence analysis. They certainly should not do worse than any other method.

Next, can one define areas where threading should be the technique of choice? The ideal problem for a threading-partisan is:

- A sequence of unknown structure;
- The sequence should have no detectable homology to anything of known structure;
- There should be a known structure that is very similar to the unknown;
- There should be no functional clues as to the structural class, otherwise a biochemist may recognize the similarity.

This situation can occur and it is not always recognized. A more likely scenario is that the borders are blurred and the thresholds uncertain. There may be some functional information about a sequence, but a chemist would like confirmation of beliefs or reassurance from a calculation. Sequence searches may have suggested plausible homologs of known structure, but with too little statistical confidence to be reliable.

One may not be obliged to follow a threading procedure as a fixed recipe. If sequence searches have suggested structural templates, but of very low sequence identity, then a sequence-to-structure alignment may be a useful step in building a model. This would not count as a threading calculation, but would use methods developed under the methodological umbrella of threading.

Changing viewpoint, can one identify times when threading should be avoided? If a sequence has very high homology to something of known structure, then threading should not do any harm, but may be a waste of time. Occasionally, however, the additional information from through-space score functions will not be helpful. If a protein has unusual structural properties, they may not be well modeled in the simple scoring functions commonly used. For example, calculations on a protein that seems to have no structure in the absence of a cofactor or prosthetic group may produce a disaster. Membrane-bound proteins are also a special problem, since most low-resolution force fields implicitly assume water is the solvent. Even simple factors such as size may be important. Small proteins may be disulfide rich or problematic simply because of their large surface-to-volume ratio.

Maybe the question of when to thread is not really a problem. Threading calculations should be cheap, and one does not have to use or believe the results. They are also not difficult to run. One can either find a relevant code and run it locally or use one of

the Web sites that provide an interface to several methods and even an assessment of the different implementations (96–98).

If one is worried about the reliability of answers, one can also look for an area where some errors are tolerated. If you are interested in genome-scale applications, it is a natural consequence that you will accept a finite error rate, perhaps using threading calculations as just part of a larger computational pipeline for screening sequences (81,99,100). Furthermore, there will even be applications where the exact structure is not important. If one wants to pick targets for structural genomics, one may try to find those sequences whose structure is most difficult to predict. Again, protein threading may be one of the tools used (101).

Since threading has already changed since the early implementations, it is also clear that the methods will continue to evolve. Some techniques combine elements of threading with methods for *de novo* structure prediction (102–105). This holds the promise of being able to predict structures unlike any previously solved. Threading may also be applied in new contexts, such as macromolecular interactions and multimolecular assemblies (100,106).

In the absence of any intellectual progress, the simple accumulation of experimental data makes prediction methods work better. The raw bulk of sequence data means that sequence profiles built now are almost always better than those of a year ago. This alone helps threading calculations that use sequence profiles. At the same time, the growth of the PDB means that there is an ever-increasing chance of a structural homolog existing for the sequence of interest.

Probably the most frightening prospect for an advocate of pure threading has been the occasional success of some fragment-assembly methods and their remarkable predictions, even for previously unseen folds (107–110). If methods for *ab initio* or *de novo* structure prediction become reliable, protein threading will be obsoleted without ever really having had a phase of glory.

References

1. Levitt, M. (1975) Computer simulation of protein folding. *Nature* **253**, 694–698.
2. Levitt, M. (1976) A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.* **104**, 59–107.
3. Crippen, G. M. and Maiorov, V. N. (1995) How many protein-folding motifs are there? *J. Mol. Biol.* **252**, 144–151.
4. Leonov, H., Mitchell, J. S. B., and Arkin, I. T. (2003) Monte Carlo estimation of the number of possible protein folds: effects of sampling bias and folds distributions. *Proteins* **51**, 352–359.
5. Wolf, Y. I., Grishin, N. V., and Koonin, E. V. (2000) Estimating the number of protein folds and families from complete genome data. *J. Mol. Biol.* **299**, 897–905.
6. Govindarajan, S., Recabarren, R., and Goldstein, R. K. (1999) Estimating the total number of protein folds. *Proteins* **35**, 408–414.
7. Zhang, C. O. and DeLisi, C. (1998) Estimating the number of protein folds. *J. Mol. Biol.* **284**, 1301–1305.
8. Wang, Z. X. (1998) A re-estimation for the total numbers of protein folds and superfamilies. *Protein Eng.* **11**, 621–626.
9. Zhang, C. T. (1997) Relations of the numbers of protein sequences, families and folds. *Protein Eng.* **10**, 757–761.

10. Wang, Z. X. (1996) How many fold types of protein are there in nature? *Proteins* **26**, 186–191.
11. Orengo, C. A., Jones, D. T., and Thornton, J. M. (1994) Protein superfamilies and domain superfolds. *Nature* **372**, 631–634.
12. Chothia, C. (1992) Proteins—1000 families for the molecular biologist. *Nature* **357**, 543–544.
13. England, J. L., Shakhnovich, B. E., and Shakhnovich, E. I. (2003) Natural selection of more designable folds: a mechanism for thermophilic adaptation. *Proc. Natl. Acad. Sci. USA* **100**, 8727–8731.
14. Li, H., Tang, C., and Wingreen, N. S. (2002) Designability of protein structures: a lattice-model study using the Miyazawa-Jernigan matrix. *Proteins* **49**, 403–412.
15. Miller, J., Zeng, C., Wingreen, N. S., and Tang, C. (2002) Emergence of highly designable protein-backbone conformations in an off-lattice model. *Proteins* **47**, 506–512.
16. Helling, R., Li, H., Melin, R., et al. (2001) The designability of protein structures. *J. Mol. Graph. Mod.* **19**, 157–167.
17. Shahrezaei, V. and Ejtehadi, M. R. (2000) Geometry selects highly designable structures. *J. Chem. Phys.* **113**, 6437–6442.
18. Bornberg-Bauer, E. (1997) How are model protein structures distributed in sequence space? *Biophys. J.* **73**, 2393–2403.
19. Govindarajan, S. and Goldstein, R. A. (1996) Why are some protein structures so common? *Proc. Natl. Acad. Sci. USA* **93**, 3341–3345.
20. Orengo, C. (1994) Classification of protein folds. *Curr. Opin. Struct. Biol.* **4**, 429–440.
21. Berman, H. M., Westbrook, J., Feng, Z., et al. (2000) The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242.
22. Brenner, S. E., Chothia, C., and Hubbard, T. J. P. (1998) Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl. Acad. Sci. USA* **95**, 6073–6078.
23. Rost, B. (1999) Twilight zone of protein sequence alignments. *Protein Eng.* **12**, 85–94.
24. Pearson, W. and Lipman, D. (1988) Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* **85**, 2444–2448.
25. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.
26. Altschul, S. F., Madden, T. L., Schaffer, A. A., et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402.
27. Madej, T., Gibrat, J. F., and Bryant, S. H. (1995) Threading a database of protein cores. *Proteins* **23**, 356–369.
28. Huber, T. and Torda, A. E. (2002) Protein structure prediction by threading: force field philosophy, approaches to alignment. In Tsigelny, I. F. (ed.), *Protein Structure Prediction: A Bioinformatic Approach*, International University Line, La Jolla, pp. 263–298.
29. Cornell, W. D., Cieplak, P., Bayly, C. I., et al. (1995) A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **117**, 5179–5197.
30. van Gunsteren, W. F., Billeter, S. R., Eising, A. A., et al. (1996) Biomolecular simulation: the GROMOS96 manual and user guide, vdf Hochschulverlag AG an der ETH Zurich and BIOMOS b.v., Zurich and Groningen.
31. MacKerell, A. D., Bashford, D., Bellott, M., et al. (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **102**, 3586–3616.
32. Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., and Karplus, M. (1983) CHARMM—a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **4**, 187–217.

33. Chandler, D. (1987) *Introduction to Modern Statistical Mechanics*, Oxford University Press, New York.
34. Miyazawa, S. and Jernigan, R. L. (1985) Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules* **18**, 534–552.
35. Tanaka, S. and Scheraga, H. A. (1976) Statistical mechanical treatment of protein conformation. 1. Conformational properties of amino-acids in proteins. *Macromolecules* **9**, 142–159.
36. Sippl, M. J. (1993) Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. *J. Comput. Aided Mol. Des.* **7**, 473–501.
37. Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992) A new approach to protein fold recognition. *Nature* **358**, 86–99.
38. Skolnick, J., Jaroszewski, L., Kolinski, A., and Godzik, A. (1997) Derivation and testing of pair potentials for protein folding. When is the quasichemical approximation correct? *Protein Sci.* **6**, 676–688.
39. Ben-Naim, A. (1997) Statistical potentials extracted from protein structures: are these meaningful potentials? *J. Chem. Phys.* **107**, 3698–3706.
40. Thomas, P. D. and Dill, K. (1996) Statistical potentials extracted from protein structures: how accurate are they? *J. Mol. Biol.* **257**, 457–469.
41. Sippl, M. J. (1996) Helmholtz free energy of peptide hydrogen bonds in proteins. *J. Mol. Biol.* **260**, 644–8.
42. Sippl, M. J., Ortner, M., Jaritz, M., Lackner, P., and Flockner, H. (1996) Helmholtz free energies of atom pair interactions in proteins. *Fold. Des.* **1**, 289–98.
43. Shortle, D. (2003) Propensities, probabilities, and the Boltzmann hypothesis. *Protein Sci.* **12**, 1298–1302.
44. Crippen, G. M. and Snow, M. E. (1990) A 1.8 angstrom resolution potential function for protein folding. *Biopolymers* **29**, 1479–1489.
45. Crippen, G. M. (1996) Easily searched protein folding potentials. *J. Mol. Biol.* **260**, 467–75.
46. Goldstein, R. A., Luthey-Schulten, Z. A., and Wolynes, P. G. (1992) Protein tertiary structure recognition using optimized Hamiltonians with local interactions. *Proc. Natl. Acad. Sci. USA* **89**, 9029–9033.
47. Maiorov, V. N. and Crippen, G. M. (1992) Contact potential that recognizes the correct folding of globular-proteins. *J. Mol. Biol.* **227**, 876–888.
48. Seetharamulu, P. and Crippen, G. M. (1991) A potential function for protein folding. *J. Math. Chem.* **6**, 91–110.
49. Ulrich, P., Scott, W., van Gunsteren, W. F., and Torda, A. E. (1997) Protein structure prediction force fields—parametrization with quasi-Newtonian dynamics. *Proteins* **27**, 367–384.
50. Huber, T. and Torda, A. E. (1998) Protein fold recognition without Boltzmann statistics or explicit physical basis. *Protein Sci.* **7**, 142–149.
51. Hao, M. H. and Scheraga, H. A. (1996) How optimization of potential functions affects protein folding. *Proc. Natl. Acad. Sci. USA* **93**, 4984–4989.
52. Mirny, L. A., and Shakhnovich, E. I. (1996) How to derive a protein folding potential—a new approach to an old problem. *J. Mol. Biol.* **264**, 1164–1179.
53. Koretke, K. K., Luthey-Schulten, Z., and Wolynes, P. G. (1996) Self-consistently optimized statistical mechanical energy functions for sequence structure alignment. *Protein Sci.* **5**, 1043–1059.
54. Lemer, C. M., Roonan, M. J., and Wodak, S. J. (1995) Protein structure prediction by threading methods: evaluation of current techniques. *Proteins* **23**, 337–355.

55. Chang, J., Carrillo, M. W., Waugh, A., Wei, L. P., and Altman, R. B. (2002) Scoring functions sensitive to alignment error have a more difficult search: a paradox for threading. In: (Eaton, G. R., Wiley, D. C., and Jardetzky, O., eds.) *Structures and Mechanisms: From Ashes to Enzymes*, vol. 827. Oxford University Press, Oxford, UK: 309–320.
56. Lathrop, R. H. (1994) The protein threading problem with sequence amino acid interaction preferences is NP-complete. *Protein Eng.* **7**, 1059–1068.
57. Smith, T. F. and Waterman, M. S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197.
58. Needleman, S. B. and Wunsch, C. D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453.
59. Kirkpatrick, S., Gelatt, Jr., C. D., and Vecchi, M. P. (1983) Optimization by simulated annealing. *Science* **220**, 671–680.
60. Bryant, S. H. and Lawrence, C. E. (1993) An empirical energy function for threading protein-sequence through the folding motif. *Proteins* **16**, 92–112.
61. Wilmanns, M. and Eisenberg, D. (1995) Inverse protein folding by the residue pair preference profile method: estimating the correctness of alignments of structurally compatible sequences. *Protein Eng.* **8**, 627–639.
62. Godzik, A., Kolinski, A., and Skolnick, J. (1992) Topology fingerprint approach to the inverse protein folding problem. *J. Mol. Biol.* **227**, 227–238.
63. Taylor, W. R. (1997) Multiple sequence threading: an analysis of alignment quality and stability. *J. Mol. Biol.* **269**, 902–943.
64. Huber, T. and Torda, A. E. (1999) Protein sequence threading, the alignment problem and a two step strategy. *J. Comput. Chem.* **20**, 1455–1467.
65. Xu, Y. and Xu, D. (2000) Protein threading using prospect: design and evaluation. *Proteins* **40**, 343–354.
66. Lathrop, R. H. (1999) An anytime local-to-global optimization algorithm for protein threading in $o(m^2n^2)$ space. *J. Comput. Biol.* **6**, 405–418.
67. Xu, Y. and Uberbacher, E. C. (1996) A polynomial-time algorithm for a class of protein threading problems. *Comput. Appl. Biosci.* **12**, 511–517.
68. Lathrop, R. H. and Smith, T. F. (1996) Global optimum protein threading with gapped alignment and empirical pair score functions. *J. Mol. Biol.* **255**, 641–665.
69. Xu, J. and Li, M. (2003) Assessment of RAPTOR's linear programming approach in CAFASP3. *Proteins* **53**, 579–584.
70. Crippen, G. M. (1996) Failures of inverse folding and threading with gapped alignment. *Proteins* **26**, 167–171.
71. Park, B. H., Huang, E. S., and Levitt, M. (1997) Factors affecting the ability of energy functions to discriminate correct from incorrect folds. *J. Mol. Biol.* **266**, 831–846.
72. Altschul, S. F., Boguski, M. S., Gish, W., and Wootton, J. C. (1994) Issues in searching molecular sequence databases. *Nat. Genet.* **6**, 119–129.
73. Altschul, S. F. and Gish, W. (1996) Local alignment statistics. *Methods Enzymol.* **266**, 460–480.
74. Karlin, S. and Altschul, S. F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. USA* **87**, 2264–2268.
75. Pearson, W. R. (1998) Empirical statistical estimates for sequence similarity searches. *J. Mol. Biol.* **276**, 71–84.
76. Mott, R. (2000) Accurate formula for p-values of gapped local sequence and profile alignments. *J. Mol. Biol.* **300**, 649–659.
77. Sommer, I., Zien, A., von Ohsen, N., Zimmer, R., and Lengauer, T. (2002) Confidence measures for protein fold recognition. *Bioinformatics* **18**, 802–812.

78. Jones, D. T. (1999) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.* **287**, 797–815.
79. Juan, D., Grana, O., Pazos, F., Fariselli, P., Casadio, R., and Valencia, A. (2003) A neural network approach to evaluate fold recognition results. *Proteins* **50**, 600–608.
80. Xu, Y., Xu, D., and Olman, V. (2002) A practical method for interpretation of threading scores: an application of neural network. *Stat. Sin.* **12**, 159–177.
81. McGuffin, L. J. and Jones, D. T. (2003) Improvement of the GenTHREADER method for genomic fold recognition. *Bioinformatics* **19**, 874–881.
82. Karplus, K., Sjolander, K., Barrett, C., et al. (1997) Predicting protein structure using hidden Markov models. *Proteins*, 134–139.
83. Karplus, K., Barrett, C., and Hughey, R. (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics* **14**, 846–856.
84. Karplus, K., Barrett, C., Cline, M., Diekhans, M., Grate, L., and Hughey, R. (1999) Predicting protein structure using only sequence information. *Proteins*, 121–125.
85. Panchenko, A. R., Marchler-Bauer, A., and Bryant, S. H. (2000) Combination of threading potentials and sequence profiles improves fold recognition. *J. Mol. Biol.* **296**, 1319–1331.
86. Russell, A. and Torda, A. E. (2002) Protein sequence threading—averaging over structures. *Proteins* **47**, 496–505.
87. Kelley, L. A., MacCallum, R. M., and Sternberg, M. J. E. (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.* **299**, 499–520.
88. Fischer, D. and Eisenberg, D. (1996) Protein fold recognition using sequence-derived predictions. *Protein Sci.* **5**, 947–955.
89. Russell, R. B., Copley, R. R., and Barton, G. J. (1996) Protein fold recognition by mapping predicted secondary structures. *J. Mol. Biol.* **259**, 349–365.
90. Rost, B., Schneider, R., and Sander, C. (1997) Protein fold recognition by prediction-based threading. *J. Mol. Biol.* **270**, 471–480.
91. Di Francesco, V., Munson, P. J., and Garnier, J. (1999) FOREST: fold recognition from secondary structure predictions of proteins. *Bioinformatics* **15**, 131–140.
92. Ayers, D. J., Gooley, P. R., Widmer-Cooper, A., and Torda, A. E. (1999) Enhanced protein fold recognition using secondary structure information from NMR. *Protein Sci.* **8**, 1127–1133.
93. Hargbo, J. and Elofsson, A. (1999) Hidden Markov models that use predicted secondary structures for fold recognition. *Proteins* **36**, 68–76.
94. Ota, M., Kawabata, T., Kinjo, A. R., and Nishikawa, K. (1999) Cooperative approach for the protein fold recognition. *Proteins*, 126–132.
95. Koretke, K. K., Russell, R. B., Copley, R. R., and Lupas, A. N. (1999) Fold recognition using sequence and secondary structure information. *Proteins*, 141–148.
96. Rost, B. and Liu, J. F. (2003) The PredictProtein server. *Nucleic Acids Res.* **31**, 3300–3304.
97. Eyrich, V. A. and Rost, B. (2003) META-PP: single interface to crucial prediction servers. *Nucleic Acids Res.* **31**, 3308–3310.
98. Koh, I. Y. Y., Eyrich, V. A., Marti-Renom, M. A., et al. (2003) EVA: evaluation of protein structure prediction servers. *Nucleic Acids Res.* **31**, 3311–3315.
99. Kim, D., Xu, D., Guo, J. T., Ellrott, K., and Xu, Y. (2003) PROSPECT II: protein structure prediction program for genome-scale applications. *Protein Eng.* **16**, 641–650.
100. Lu, L., Lu, H., and Skolnick, J. (2002) MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading. *Proteins* **49**, 350–364.
101. McGuffin, L. J. and Jones, D. T. (2002) Targeting novel folds for structural genomics. *Proteins* **48**, 44–52.

102. Jones, D. T. (2001) Predicting novel protein folds by using FRAGFOLD. *Proteins*, 127–132.
103. Skolnick, J., Kolinski, A., Kihara, D., Betancourt, M., Rotkiewicz, P., and Boniecki, M. (2001) *Ab initio* protein structure prediction via a combination of threading, lattice folding, clustering, and structure refinement. *Proteins*, 149–156.
104. Zhang, Y., Kolinski, A., and Skolnick, J. (2003) TOUCHSTONE II: a new approach to *ab initio* protein structure prediction. *Biophys. J.* **85**, 1145–1164.
105. Kihara, D., Lu, H., Kolinski, A., and Skolnick, J. (2001) TOUCHSTONE: an *ab initio* protein structure prediction method that uses threading-based tertiary restraints. *Proc. Natl. Acad. Sci. USA* **98**, 10,125–10,130.
106. Lu, L., Arakaki, A. K., Lu, H., and Skolnick, J. (2003) Multimeric threading-based prediction of protein-protein interactions on a genomic scale: application to the *Saccharomyces cerevisiae* proteome. *Genome Res.* **13**, 1146–1154.
107. Simons, K. T., Strauss, C., and Baker, D. (2001) Prospects for *ab initio* protein structural genomics. *J. Mol. Biol.* **306**, 1191–1199.
108. Simons, K. T., Kooperberg, C., Huang, E., and Baker, D. (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.* **268**, 209–225.
109. Chivian, D., Robertson, T., Bonneau, R., and Baker, D. (2003) *Ab initio* methods. In: (Bourne, P. E., and Weissig, H., eds.) *Structural Bioinformatics* vol. 44. Wiley-Liss, Hoboken, NJ: 547–548.
110. Bonneau, R. and Baker, D. (2001) *Ab initio* protein structure prediction: progress and reports. *Annu. Rev. Biophys. Biomol. Struct.* **30**, 173–189.

High-Throughput Crystallography for Structural Proteomics

Jeff Yon, Mladen Vinković, and Harren Jhoti

1. Introduction

The last decade has seen the success of large-scale sequence determination projects, and an increasing number of sequenced genomes have become available (<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome>). This in turn has led to efforts to understand the functions of the many newly discovered proteins, and hence to greater interest in high-throughput protein structure determination (1,2). Thus the last 5 yr has seen the rise of structural proteomics efforts, whose goal has been the large-scale determination of protein structures using the techniques of X-ray crystallography and nuclear magnetic resonance (NMR) (the terms *structural proteomics* and *structural genomics* are often used interchangeably, and no distinction is made in this review). Structural proteomics is expected to contribute to functional studies, as the structures of novel proteins may give insight into their functions as well as having a major impact on drug discovery, as some of these new proteins will be therapeutic targets (3,4). An additional level of complexity results from the fact that proteins are often found as part of multiprotein complexes. Although high-throughput structure determination of multiprotein complexes will be even more difficult than for isolated proteins, it will undoubtedly lead to a better understanding of function and mechanism (5).

The first publicly funded structural proteomics initiative was set up in Japan (<http://www.rsg1.riken.go.jp>, *see also* ref. 6), and has since been followed by numerous other programs. In the United States, the National Institute of General Medical Sciences (NIGMS) Protein Structure Initiative (<http://www.nigms.nih.gov/psi>, *see also* ref. 7) embraces diverse projects focused on method development and on target proteins from a wide range of organisms, including thermophilic and pathogenic bacteria, *Caenorhabditis elegans*, *Arabidopsis*, as well as *Homo sapiens*. More recently, the number of projects in Europe has also increased (<http://www.spineurope.org>, *see also* ref. 8), and large projects have begun in Canada (9) and other countries (<http://www.isgo.org>). The initial focus of these high-throughput crystallography programs has been on developing high-throughput methods for the gene-to-structure process. The route from target selection to structure solution is a complex, multidisciplinary one, with many potential bottlenecks to overcome in order to achieve proteome-scale structure determination. This review will focus on the process of structure determina-

tion by X-ray crystallography, and in particular on technical developments over the last 2–3 yr that have resulted in higher throughput and improved efficiency.

2. Protein Production

It is generally accepted that proteins for crystallization studies should be highly purified (typically >95% pure) and chemically and conformationally homogeneous. Homogeneity may be difficult to achieve: for example, it is often necessary to separate different phosphorylated forms of a protein kinase. A heterogeneous sample of a protein can result in aggregation, which is detrimental to crystallization; therefore, sample analysis using size exclusion chromatography or dynamic light scattering (DLS) is a key step. It is widely accepted that monodisperse solution behavior, as measured by DLS, correlates with an improved chance of crystallization (10). Thus the task in protein production for crystallography is to generate large amounts (usually tens of milligrams) of pure, homogeneous protein. It is therefore no surprise that production of suitable protein is a major bottleneck in high-throughput X-ray crystallography. It has been estimated that up to 80% of the activities in a structural biology laboratory focus on protein expression, purification, and characterization (11). Furthermore, one large-scale structural proteomics program, the Protein Structure Factory in Berlin, has found that their capacity to crystallize and solve structures using X-ray crystallography considerably exceeds their capacity for protein production (12).

2.1. Molecular Biology

The production of recombinant proteins necessitates the cloning of a suitable DNA sequence, generation of expression constructs, testing and optimization of expression, scale-up, purification, and characterization of the proteins produced. Many different options exist for each step, but some common themes have emerged for high-throughput approaches (13,14). Uninterrupted DNA coding sequences suitable for use in expression studies can be obtained either from genomic DNA (for prokaryotes) or from cDNA. Many cDNA libraries from different tissues are now available commercially, and collections of individual cDNAs are becoming more comprehensive (e.g., the FANTOM collection of mouse cDNAs [15]), so that obtaining suitable starting material is often straightforward. The initial cloning step is usually based on polymerase chain reaction (PCR) amplification, which has proven to be a robust and automation-friendly approach. This is often coupled with a TA[®] cloning step (Invitrogen, Carlsbad, CA), or a ligase-free cloning step, such as those based on a viral topoisomerase (TOPO[®] cloning, Invitrogen) or ligation-independent cloning (Novagen, Madison, WI), for reliable cloning of large numbers of cDNA constructs. More recently, recombination-based cloning systems have become commercially available (e.g., GatewayTM from Invitrogen or CreatorTM from Clontech [Palo Alto, CA]) that also allow rapid, parallel cloning of PCR fragments.

Alterations in the expression constructs for a given target (for example varying the residues chosen as the N- and C-termini of the construct, or placing an affinity tag at the N- vs the C-terminus) may have significant effects on expression levels and solubility. To explore some of these variables, multiple expression constructs must often be made before the best is identified. A recent trend has been to make and test many constructs in parallel in order to try to identify an optimal construct more quickly. The recombination-based approaches in particular have the advantage of speed and high

fidelity, and are of great value when the same coding sequence must be subcloned into several different vectors—for example, to test the effect of different affinity tags or tag positions, or to switch between different host/vector systems (16). All of the above cloning and subcloning methods consist principally of liquid addition steps, and can be automated using standard commercially available liquid-handling robots (13).

2.2. Expression Optimization

When screening for expression of a target protein in *Escherichia coli*, variables commonly tested for each construct include strain, growth temperature, and medium composition. Structural proteomics programs rely heavily on selenomethionine labeling of target proteins to assist with structure determination, so media formulations suitable for labeling will often be explored (17). Testing many conditions per construct rapidly gives rise to large numbers of samples, and hence has led to the use of microplate formats for both cell growth and subsequent analysis (18). Analysis of expression levels is usually by denaturing polyacrylamide gel analysis or by dot blot using target-specific or tag-specific antibodies. An affinity purification step may be included prior to this analysis. In addition to constructs of the target of interest, it may be valuable to include closely related targets from other species to improve the chance of success. Recently, Savchenko et al. (19) attempted to quantify the benefits of this approach by studying 68 pairs of orthologous proteins from *E. coli* and *Thermotoga maritima*. These authors concluded that the inclusion of a single ortholog nearly doubled the probability of obtaining a sample suitable for structural studies. To date, most high-throughput structural proteomics initiatives have used *E. coli* as the expression host, for reasons of ease of use and speed. However, obtaining soluble, correctly folded protein has been a major challenge (20), and significant efforts have been devoted to finding ways to improve this, as described later.

High-level production of soluble, tagged bacterial target proteins expressed in *E. coli* has a success rate of approx 50% (21). However, proteins from higher organisms, such as *Homo sapiens*, are often more difficult to produce in *E. coli*, so other expression systems are likely to be required (20,12). High-throughput parallel expression has been described in the yeasts *Saccharomyces cerevisiae* (22) and *Pichia pastoris* (23). However, most structural biology groups routinely use the baculovirus/insect cell expression system for production of mammalian proteins (24). Whereas generation and handling of baculovirus is more complex than bacterial or yeast expression, the system can also be used in microplate format (18,25).

Other approaches to protein production currently receiving attention include cell-free expression and directed-evolution methods. Cell-free expression systems have been a major focus in Japan as a production method for structural proteomics targets. Cell-free systems offer potential advantages of speed and reduced sensitivity to toxic effects of recombinant proteins (26). Current systems based on *E. coli* lysates are capable of producing milligram amounts of protein (27), and selenomethionine labeling has also been demonstrated (28). A 96-well *E. coli*–based cell-free expression screen with a dot-blot readout has recently been described (29). Systems based on wheat germ have also been improved (30), and in a recent report achieved a very high success rate in a screen for production of *Arabidopsis* proteins in 96-well format (31), suggesting that the wheat-germ system may also have utility in structural proteomics programs.

In the directed-evolution methods, the problem of insoluble expression is tackled by screening for protein variants that give high-level soluble expression. The protein of interest is fused to a reporter protein in such a way that the readout (via the reporter) is dependent on expression of the fusion as a soluble protein. Typically, a large pool of variants of the target protein is generated by error-prone PCR and gene-shuffling methods, and variants showing increased solubility are selected based on increased signals from the reporter. Perhaps the best-known example involves the use of green fluorescent protein as a fusion partner (32), and use of this technique has resulted in at least two published crystal structures (33,34). It is not yet clear, however, what impact either cell-free or directed-evolution methods will have on current structural proteomics efforts (14).

2.3. Purification and Characterization

Once a promising construct has been identified, expression is scaled up and the sample is purified to a level suitable for crystallization studies. In an ideal situation, purification can be via a standard protocol using a single affinity chromatography step. In single-step affinity protocols, buffers and elution conditions can be standardized, and if required the whole method can be automated on an high-performance liquid chromatography (HPLC) or fast protein liquid chromatography (FPLC) system capable of multidimensional chromatography—for example, AKTA from Amersham Biosciences (<http://www1.amershambiosciences.com>) or BioCAD from Applied Biosystems (<http://home.appliedbiosystems.com>)—to give higher throughputs of purified protein samples. In some cases, however, additional purification steps will be required, which have to be tailored to the individual proteins, and so may reduce throughput (16). As mentioned above, high-quality samples can be important for crystallization, and some labs use stringent checks by DLS, circular dichroism (CD), Fourier transform infrared (FTIR) spectroscopy, and calorimetry (12) to ensure that the protein is of a high standard.

2.4. Protein Production Automation

All of the steps from cloning through expression testing can be carried out in high-density microplate formats that are suitable for automation with standard liquid-handling robots. Most structural proteomics laboratories have applied automation solutions via commercial vendors rather than building and developing robotic systems in-house. Complete automation packages for high-density expression screening comprising a liquid-handling robot, reagents, and consumables are commercially available from Qiagen (<http://www1.qiagen.com/>). Novagen (<http://www.novagen.com>) also offers Robopop™ reagents, which have been validated on Tecan (<http://www.tecan.com>) and Perkin Elmer (<http://las.perkinelmer.com>) workstations. In contrast to using commercially available automation, some laboratories, such as the Genetics Institute of the Novartis Foundation and Syrrx (35), have developed custom-built robotics for 96-sample bacterial fermentation and purification. Although undoubtedly capable of generating enormous throughputs, such an investment in in-house robotics expertise remains a daunting prospect for most groups.

3. Structure Determination

3.1. Crystallization

The crystallization process is considered a major bottleneck in protein crystallography. The availability of limited amounts of sample, combined with the large number of crystallization parameters, makes the discovery of initial crystallization conditions challenging. However, this situation has been dramatically improved recently by developments in automation, miniaturization, and process integration. A variety of different approaches are used to entice protein molecules to come together within a solution to form a protein crystal (36). Vapor diffusion methods are still the most popular; with the sitting drop technique currently more popular than the more traditional hanging drop approach, as it is simpler to automate. In fact, the micro-batch method of crystallization under oil is an alternative approach to the vapor diffusion methods and actually is the most straightforward crystallization method to automate. However, groups that use micro-batch as a method of first choice, such as the Hauptman-Woodward Institute HTP Crystallization Laboratory (37), are still in a minority.

A typical high-throughput crystallization screening campaign starts with automated preparation or reformatting of precipitant solutions, to which the protein samples are added, using liquid-handling systems (Fig. 1). Thousands of crystallization experiments can then be set up using robotic systems in which a wide range of variables are explored, such as pH, temperature, and protein concentration, as well as the effect of including additives such as co-factors, ligands, and metal ions (36). If the amount of protein is limited, a new approach referred to as *nano-crystallogenesis* can be employed, in which very small crystallization drops are used, containing as little as 25–100 nL of protein (38). The resulting large numbers of crystallization drops can be monitored regularly by an imaging robot such as the Robomicroscope III (Robodesign International, <http://www.robodesign.com>) or Rhombix Vision (DataCentric Automation, <http://www.dcacorp.com>), which are capable of scanning hundreds of crystallization drops within minutes. Collected images can either be analyzed manually on a computer screen or by image-recognition software. Although there have been numerous efforts in developing crystal-recognition software (39,40), at present it is mostly used to eliminate clear and other drops that the software can reliably classify as negatives. The rest of the images are then inspected manually.

Some researchers believe that the crystallizability of a protein sample can be assessed by using only approx 100 crystallization conditions (41). However, there are examples of protein samples that produce crystals in only one of thousands of screening conditions (42), and therefore most groups rely on extensive screening. More than 1500 different crystallization-screening solutions are available commercially from different manufacturers (see for example Hampton Research, <http://www.hampton-research.com>), and screens can also be designed in-house. In both cases, liquid-handling systems such as those from Tecan or Hamilton (<http://www.hamilton.ch>) are used either to reformat solutions into crystallization plates or to prepare them from stock solutions. Matrix Maker from Emerald Biostructures (<http://www.decode.com/emeraldbiostructures>) is specifically designed for high-throughput crystallization solution preparation. The same systems are frequently used for dispensing microliter-scale crystallization drops. However, a dedicated nano-dispensing robot is usually used for crystallization drops in the nanoliter range.

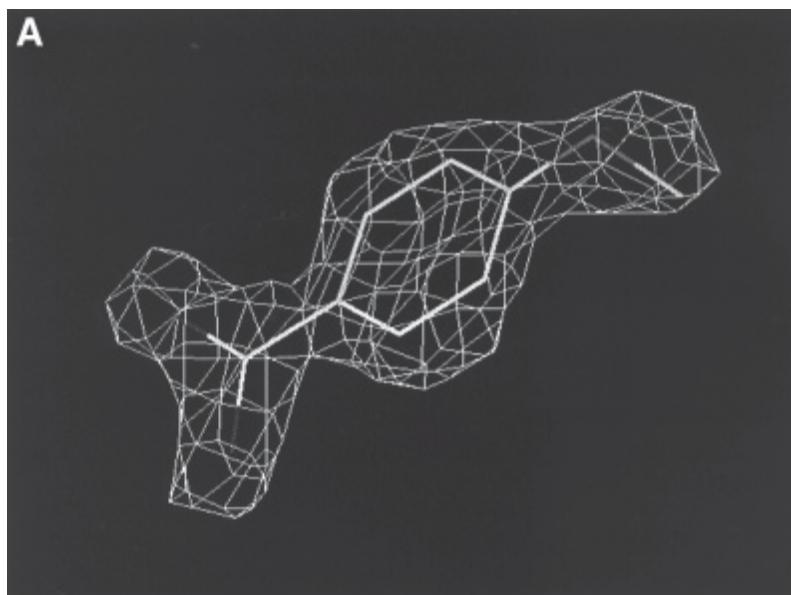


Fig. 1. The PixSys system from Genomic Solutions uses solenoid valve-based SynQuad technology for non-contact dispensing of protein crystallization drops.

Although robots for protein crystallization capable of dispensing volumes as little as 200 nL have been available since the early 1990s (for example the IMPAX and Oryx from Douglas instruments, <http://www.douglas.co.uk>), the wave of interest in nano-crystallogenesis has been initiated with solenoid valve-based nano-dispensers. Some of the pioneering work in nano-crystallogenesis was performed at the University of California, Berkeley (43), and the University of Alabama, Birmingham, while Syrrx (<http://www.syrrx.com>) was the first company to use the technique in an industrial high-throughput environment. The current Cartesian line of products from Genomic Solutions (<http://www.genomicsolutions.com>) is capable of dispensing drops from 20 nL to several μ L in sitting or hanging drop plates (Honeybee systems), as well as to dispense drops through the oil in the micro-batch method. Critically, these machines are able to dispense very viscous crystallization solutions like 30% PEG8K with high accuracy. The other proven technology for nano-dispensing protein crystallization drops uses positive-displacement nano-tips and is implemented in the TTP Labtec Mosquito robot (<http://www.ttplabtech.com>). A recent development in protein crystallization has been the introduction of disposable micro-fluidic crystallization chips with the Topaz system by Fluidigm (<http://www.fluidigm.com>). They employ the free interface diffusion method at low nano-liter scale (25 nL). This method explores more crystallization space in each experiment than vapor diffusion and has claimed several successes (44).

3.2. X-Ray Data Collection

Well-ordered crystals and adequate data-collection equipment are essential for a successful diffraction experiment. The first step in data collection is the positioning of a crystal onto an X-ray data-collection machine, and until recently this has been a

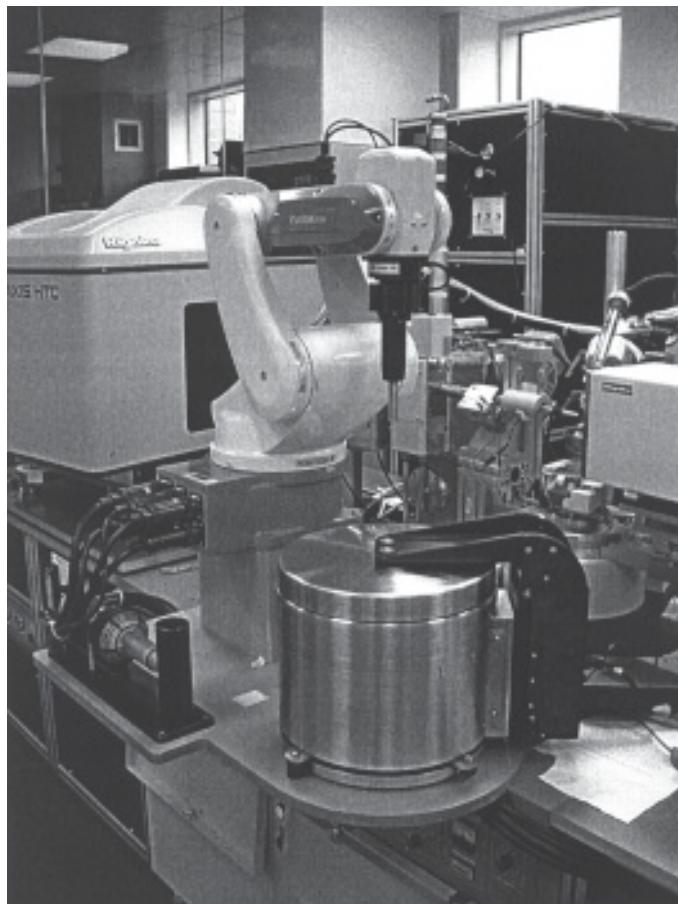


Fig. 2. The Actor robot from Rigaku/MSC holds 60 crystals in its cryo-store, which are mounted sequentially on the goniometer for automated unattended data collection.

manual step. However, new crystal-mounting robots such as ACTOR (Fig. 2), which is manufactured by Rigaku MSC (<http://www.rigakumsc.com>), allow fully automated crystal testing and data collection. These robots have now been installed at many facilities, including high-intensity synchrotron sources as well as on several in-house data-collection systems. The benefits of automated crystal mounting are in more efficient use of X-ray sources by performing unattended data collection from multiple protein crystals. Data-collection software such as Crystal Clear/Director from Rigaku MSC is able to control the experiment by placing the crystal in the correct orientation, initiating and ending the data collection before moving on to the next crystal.

For high-throughput crystallography efforts, high-intensity synchrotron radiation sources have been instrumental in maintaining rapid data-collection throughput. X-ray beams from third-generation synchrotrons are so powerful that data-collection time has typically been reduced by 2- to 10-fold. Another valuable property of synchrotron radiation sources is that, when equipped with a suitable monochromator, they can deliver radiation at different wavelengths. This enables anomalous diffraction experiments, which exploit heavy-atom labeling and have been very effective for novel pro-

tein structure determination (45). In parallel, significant improvements have also been made in laboratory generators, which typically use rotating anodes to produce X-rays. For example, the FRE generator from Rigaku MSC, coupled with exceptionally effective optics (MicroMax from Osmic, <http://www.osmic.com>), generates X-ray flux comparable to some synchrotron beamlines.

High-throughput X-ray crystallography would not be possible without the development of fast and sensitive X-ray detectors. These systems can be grouped into two types: image plate (IP) and charge-coupled device (CCD) detectors. New IP detectors, such as HTC (Rigaku MSC), contain three X-ray-sensitive IPs, which are successively exposed. This ensures continuous data collection by avoiding the need to wait for the reading and erasing of each individual IP, as required by previous models. CCD detectors are a more recent development in X-ray detector technology and allow direct read-out of diffraction data, without the need for scanning. In some cases, CCD detectors can collect X-ray data five times faster than an IP detector.

3.3. Protein Structure Solution and Refinement

X-ray data processing and analysis can be very demanding on the crystallographer's time. This is because there are many separate steps involved, and these are usually performed with different computer programs. They demand a great deal of file and molecular model inspection and editing, contributing to relatively low efficiency. Therefore, it is not surprising that in a high-throughput crystallography environment, this is area in which major time savings have been achieved. Increased throughput can derive from improvements in existing programs, such as those for phase and model refinement; from development of new software for new tasks like automated ligand density interpretation; as well as from integration of individual steps into one continuous automated computing process. There are two different problems that are subjected to automation in specific ways: crystal structure solution of new proteins and of protein-ligand complexes.

A key element in the structure determination process is to obtain phase information for the diffracted X-rays. Historically, this involved immersing protein crystals into solutions of heavy atoms, such as mercuric salts, with the hope that one or more heavy atoms would attach to the protein molecules. These heavy atoms could then be located and used to provide the phase information. This process has now been supplemented by innovative molecular biology techniques that can incorporate atoms such as selenium, which can be exploited using anomalous scattering methods to obtain the phase information. Use of selenomethionine-labeled proteins has transformed the efficiency of structure determination (45). Specialized software has been developed to exploit this approach; for example, programs like SHARP (Statistical Heavy Atom Refinement and Phasing, <http://www.globalphasing.com>) use information about location of anomalous scatterers to refine phases and, linked to solvent flattening programs, which calculate contribution to phases from solvent, are capable of generating exceptionally high-quality electron-density maps. In cases where the protein of interest is highly homologous to a protein of known structure, molecular replacement with the structure of the known homolog as search model may be used to solve the phase problem. Where similarity to proteins of known structure is low, rapid automated molecular replacement by evolutionary search is proving useful (46).

Once the phase information has been obtained, then an electron-density map can be calculated from which a model of the protein is to be built. The interpretation of initial electron-density maps (model building) has traditionally been very laborious. However, automated model-building by software such as ARP/wARP (47,48) can be applied successfully on structures at 2.5-Å resolution or better. Loop regions are particularly difficult to build because they are frequently less ordered in the crystal, giving weak density. Programs like Rapper (49) have algorithms to deal with such challenges. Once an initial model has been built, it then undergoes a process of optimization known as *refinement*. Until recently, refinement of the protein model was performed by least-square-based algorithms, but new maximum likelihood-based programs, like Refmac (50) and CNX (51), are now in routine use. These are especially advantageous in the early stages of refinement, when the model is still far from optimal.

Structure solution and refinement of protein–ligand complexes has also been significantly improved using new tools. The generation of a 3-D model of a ligand, fitting ligand into the electron density in the correct configuration and conformation, defining the tautomeric and ionic state, as well as designing a dictionary for ligand refinement can all be time-consuming, error-prone steps. New software such as AutoSolve® (developed at Astex [52]) can automatically identify ligand density in an active site and build a model of the ligand into it within minutes (Fig. 3). Coupled with other tools including ligand dictionary generation, this has significantly increased the throughput of generating protein/ligand crystal structures. Furthermore, these developments have allowed X-ray crystallography to be used as a screening technology in fragment-based drug discovery (see **Subheading 5.**).

The whole procedure of data processing in a high-throughput approach is integrated in programs such as PHENIX (53) or ACrS (54) for automatic protein structure solution. Another very important aspect is the user interface to automated data processing, enabling simple and quick set-up of calculations and inspection of the structure and electron density, as exemplified by AstexViewer™ incorporated into a World-Wide Web browser (55).

4. Databases

As the numbers of samples at various stages of the structure determination process increases, the need for a laboratory information management system (LIMS) becomes more pressing, both to record the samples and their location, and to allow tracking of target samples. The generation of vast amounts of data during crystallography also makes databases essential. This has led to the development of databases tailored to capturing high-throughput structural information. In some instances, such databases can store information from all stages of the gene-to-structure process (12,56). Indeed, biotechnology companies involved in structure-based drug discovery have developed databases that capture information from target gene all the way to small-molecule inhibitor, as for example at Astex (unpublished) and Structural Genomix (57).

5. Protein Structure in Drug Discovery

The advent of high-throughput X-ray crystallography has also allowed the development of novel approaches to small-molecule lead discovery. In particular, the ability to rapidly obtain the structures of protein–ligand complexes enables X-ray crystallogra-

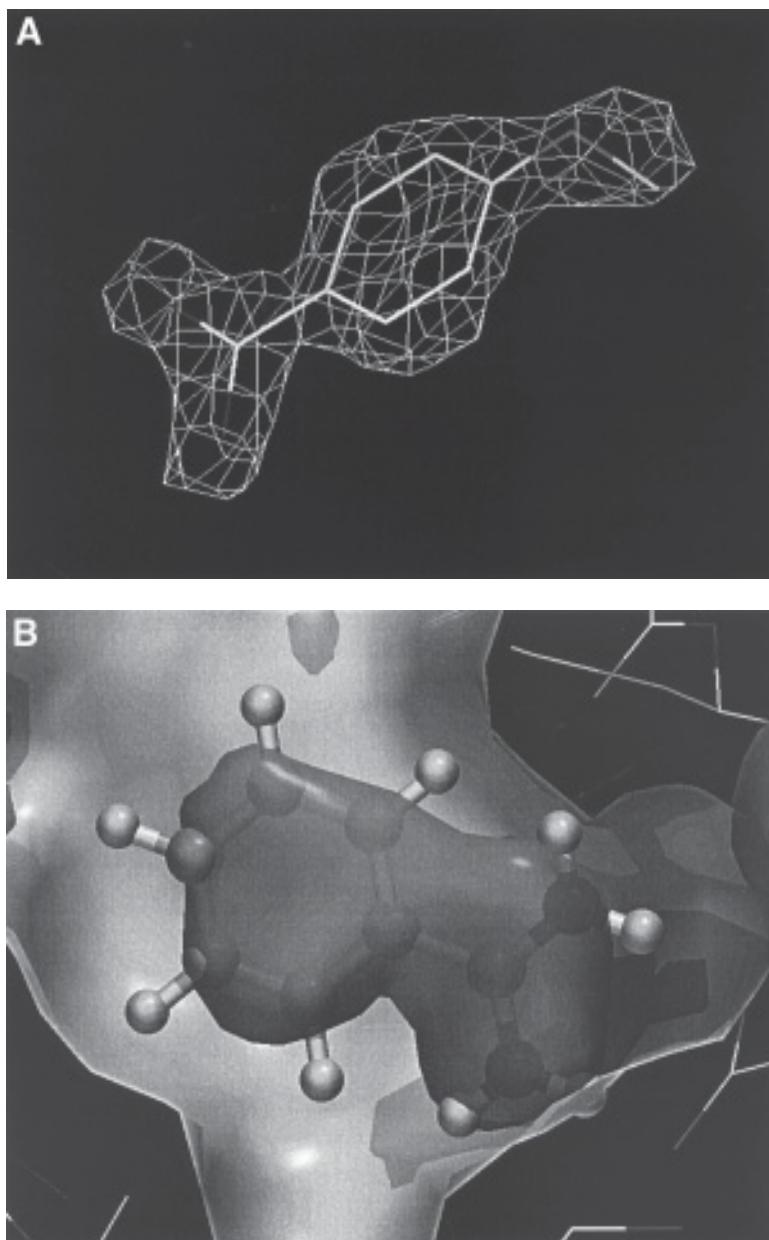


Fig. 3. (A) AutoSolve® software from Astex Technology identifies ligand density in the active site and builds ligand into it. (B) AstexViewer™ visualizes electron density (dark gray) and protein and ligand models. Hydrogen atoms are represented as white spheres for clarity, but are usually not visible in protein crystal structures.

phy to be used as a screening tool for hit identification. This has been very powerful in the area of fragment-based lead discovery, where an initial low-molecular-weight drug fragment hit is evolved into a nanomolar compound (58,59). Fragment binding is difficult to detect reliably by activity-based HTS methods, as the fragments are small (mol. wt. 100–300) and bind weakly, typically in the mM range. X-ray crystallography, on

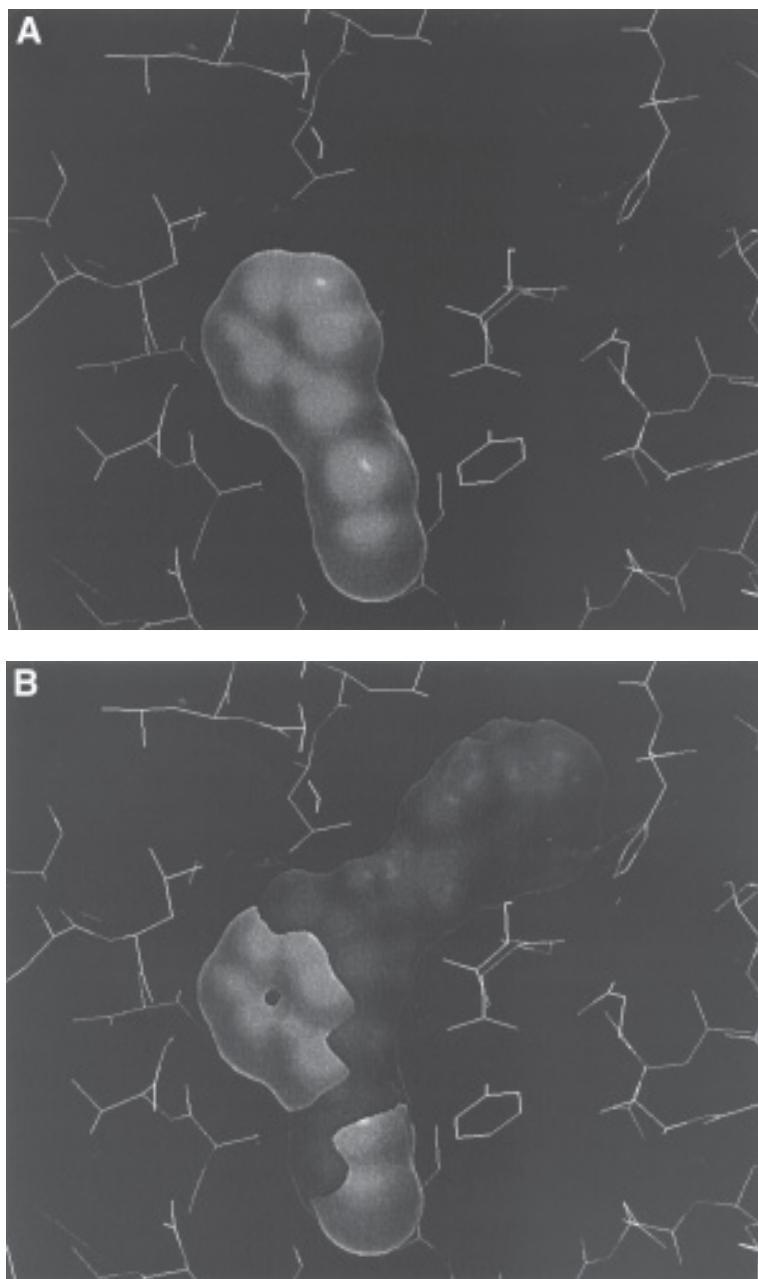


Fig. 4. (A) High-throughput crystallography was used to identify the initial hit fragment binding to p38 MAP kinase (compound is shown in light gray). (B) Iterative structure-based medicinal chemistry was then used to rapidly evolve the initial hit into a potent lead compound (later compound darker gray).

the other hand, can readily detect binding of these fragments (**60**), with the added advantage that the binding mode of the hit is known from the outset, allowing structure-based optimization to begin immediately. Two groups have pioneered this approach, at Abbott (**61**) and Astex (**60**). **Figure 4** gives an example of this process,

showing a hit from a screen against p38 MAP kinase at Astex. The initial hit (mM binding) was rapidly evolved into a nanomolar inhibitor by structure-based medicinal chemistry. In summary, therefore, the ability to determine protein–ligand complex structures at high throughput has opened up the possibility of using X-ray crystallography in hit identification, which may prove to have a major impact on drug discovery.

6. Conclusion

As outlined in this review, the last few years have seen developments in high-throughput methods at all stages of the gene-to-structure process, from the introduction of automation into cloning, expression, purification, and crystallization to the development of software for data analysis and structure solution. These developments should clearly enable greater throughputs in terms of deposited structures, though it should be remembered when analyzing the outputs of structural proteomics programs that the initial focus has largely been on this method development. To assist in this aim, the targets have largely been of bacterial origin, and membrane proteins have been avoided. As of October 2003, close to 40,000 target proteins had been selected and over 500 structures solved and deposited in the Protein Data Bank (PDB) (<http://targetdb.pdb.org>). Attrition rates in the gene-to-structure process have been high, particularly in obtaining soluble expression of the targets (20), but the rate of deposition of structures from these programs is increasing rapidly.

More traditional structural biology labs are now adopting many of the techniques and processes developed in structural proteomics initiatives. Thus it has become increasingly common to test many expression constructs in parallel, or to set up crystallization trays with the help of a robot. This has had an impact in terms of key structures: as examples, the biotechnology companies Syrrx and Astex have used high-throughput methods to enable the solution of medically important kinase (62,63) and cytochrome P450 (64) targets, respectively. It can be expected that the rate of deposition of protein structures in the PDB will continue to increase, both as a direct result of structural proteomics programs and as a result of increased throughputs in more traditional labs. And finally, high-throughput methods for obtaining protein/ligand crystal structures are increasing the impact of rational design in drug discovery, with the advent of fragment-based approaches using X-ray crystallography.

References

1. Burley, S. K., et al. (1999). Structural genomics: beyond the human genome project. *Nature Genet.* **23**, 151–157.
2. Service R. F. (2000). Structural genomics offers high-speed look at proteins. *Science* **287**, 1954–1956.
3. Skolnick, J., et al. (2000). Structural genomics and its importance for gene function analysis. *Nature Biotech.* **18**, 283–287.
4. Zhang, C. and Kim, S.-H. (2003) Overview of structural genomics: from structure to function. *Curr. Opin. Chem. Biol.* **7**, 28–32.
5. Sali, A., et al. (2003). From words to literature in structural proteomics. *Nature* **422**, 216–225.
6. Yokoyama, S., et al. (2000). Structural genomics project in Japan. *Nat. Struct. Biol.* **7**, 943–945.
7. Norvell, J. C. and Machalek, A. Z. (2000). Structural genomics programs at the US National Institute of General Medical Sciences. *Nat. Struct. Biol.* **7**, 931.

8. Heinemann, U. (2000). Structural genomics in Europe: slow start, strong finish? *Nat. Struct. Biol.* **7**, 940–942.
9. Service, R. F. (2003). Canada vaults into drug-oriented protein research. *Science* **300**, 28.
10. Zulauf, M. and D'Arcy, A. (1992) Light scattering of proteins as a criterion for crystallization. *J. Crystal Growth* **122**, 102–106.
11. Blundell, T. L., Jhoti, H., and Abell, C. (2002). High-throughput crystallography for lead discovery in drug design. *Nature Rev. Drug Discov.* **1**, 45–54.
12. Heinemann, U., et al. (2003) Facilities and methods for the high-throughput crystal structural analysis of human proteins. *Acc. Chem. Res.* **36**, 157–163.
13. Dieckman, L., et al. (2002). High-throughput methods for gene cloning and expression. *Prot. Exp. Purif.* **25**, 1–7.
14. Edwards, A. M., et al. (2003) Producing proteins. In: (Chasman, D. I., ed.) *Protein Structure: Determination, Analysis, and Applications for Drug Discovery*. Marcel Dekker, New York, NY: 9–25.
15. Okazaki, Y., et al. (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420**, 563–573.
16. Stevens, R. C. (2000) Design of high-through methods of protein production for structural biology. *Structure* **8**, R177–R185.
17. Doublie, S. (1997) Preparation of selenomethionyl proteins for phase determination. *Methods Enzymol.* **276**, 523–530.
18. Chambers, S. P. (2002). High-throughput protein expression for the post-genomic era. *Drug Disc. Today* **7**, 759–765.
19. Savchenko, A., et al. (2003) Strategies for structural proteomics of prokaryotes: quantifying the advantages of studying orthologous proteins and of using both NMR and X-ray crystallography approaches. *Proteins: Struct. Funct. Genet.* **50**, 392–399.
20. Service, R. F. (2002). Tapping DNA for structures produces a trickle. *Science* **298**, 948–950.
21. Braun, P. and LaBaer, J. (2003) High throughput protein production for functional proteomics. *Trends Biotechnol.* **21**, 383–388.
22. Holz, C., et al. (2003) Establishing the yeast *Saccharomyces cerevisiae* as a system for expression of human proteins on a proteome-scale. *J. Struct. Func. Gen.* **4**, 97–108.
23. Boettner, M., et al. (2002) High-throughput screening for expresion of heterologous proteins in the yeast *Pichia pastoris*. *J. Biotechnol.* **99**, 51–62.
24. Yon, J. and Jhoti, H. (2003) High throughput structural genomics and proteomics: where are we now? *TARGETS* **2**, 201–207.
25. Albala, J. S., et al. (2000). From genes to proteins: high-throughput expression and purification of the human proteome. *J. Cellular Biochem.* **80**, 187–191.
26. Yokoyama, S. (2003) Protein expression systems for structural genomics and proteomics. *Curr. Opin. Chem. Biol.* **7**, 39–43.
27. Kigawa, T., et al. (1999). Cell-free production and stable-isotope labeling of milligram quantities of proteins. *FEBS Lett.* **442**, 15–19.
28. Kigawa, T., et al. (2001). Selenomethionine incorporation into a protein by cell-free synthesis. *J. Struct. Funct. Genomics* **2**, 29–35.
29. Busso, D., Kim, R., and Kim, S-H. (2003) Expression of soluble recombinant proteins in a cell-free system using a 96-well format. *J. Biochem. Biophys. Methods* **55**, 233–240.
30. Endo, Y. and Sawasaki, T. (2003) High-throughput, genome-scale protein production method based on the wheat germ cell-free expression system. *Biotechnol. Adv.* **21**, 695–713.
31. Sawasaki, T., et al. (2002). A cell-free protein synthesis system for high-throughput proteomics. *Proc. Natl. Acad. Sci. USA* **99**, 14,652–14,657.
32. Waldo, G. S. (2003). Genetic screens and directed evolution for protein solubility. *Curr. Opin. Chem. Biol.* **7**, 33–38.

33. Pedelacq, J.-D., et al. (2002) Engineering soluble proteins for structural genomics. *Nature Biotechnol.* **20**, 927–932.
34. Yang, J. K., et al. (2003) Directed evolution approach to a structural genomics project: Rv2002 from *Mycobacterium tuberculosis*. *Proc. Natl. Acad. Sci. USA* **100**, 455–460.
35. Lesley, S. A. (2001) High-throughput proteomics: protein expression and purification in the postgenomic world. *Protein Exp. Purif.* **22**, 159–164.
36. McPherson, A. (1999) *Crystallization of Biological Macromolecules*. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
37. Luft, J. R., et al. (2001) Macromolecular crystallization in a high throughput laboratory - the search phase. *J. Cryst. Growth* **232**, 591–595.
38. Stevens, R. C. (2000) High-throughput protein crystallization. *Curr. Opin. Struct. Biol.* **10**, 558–563.
39. Wilson, J. (2002) Towards the automated evaluation of crystallization trials. *Acta Crystallogr.* **D58**, 1907–1914.
40. Spraggan, G., et al. (2002) Computational analysis of crystallization trials. *Acta Crystallogr.* **D58**, 1915–1923.
41. Dale, G. E. et al. (2003) The protein as a viable in protein crystallization. *J. Struct. Biol.* **142**, 88–97.
42. Luft, J. R., et al. (2003) A deliberate approach to screening for initial crystallization conditions of biological macromolecules. *J. Struct. Biol.* **142**, 170–179.
43. Santarsiero, B. D., et al. (2002) An approach to rapid protein crystallization using nanodroplets. *J. Appl. Cryst.* **35**, 278–281.
44. Hansen, C. L., et al. (2002) A robust and scalable microfluidic metering method that allows protein crystal growth by free interface diffusion. *PNAS* **99**, 16,531–16,536.
45. Boggon, T. J. and Shapiro, L. (2000) Screening for phasing atoms in protein crystallography *Structure* **8**, R143–R149.
46. Kissinger, C. R., Gehlhaar, D. K., and Fogel, D. B. (1999) Rapid automated molecular replacement by evolutionary search. *Acta Crystallogr.* **55**, 484–491.
47. Perrakis, A., Morris, R., and Lamzin, V. S. (1999) Automated protein model building combined with iterative structure refinement. *Nat. Struct. Biol.* **6**, 458–463.
48. Morris, R. J., Perrakis, A., and Lamzin, V. S. (2002) ARP/wARP's model-building algorithms. I. The main chain. *Acta Crystallogr.* **D58**, 968–975.
49. DePristo, M. A., de Bakker, P. I. W., Shetty, R. P., and Blundell, T. L. (2003) Discrete restraint-based protein modeling and the Ca-trace problem. *Protein Sci.* **12**, 2032–2046.
50. Murshudov, G. N., Vagin, A. A., and Dodson, E. J. (1997) Refinement of Macromolecular Structures by the Maximum-Likelihood Method. *Acta Crystallogr.* **D53**, 240–255.
51. Adams, P. D., et al. (1997) Cross-validated maximum likelihood enhances crystallographic simulated annealing refinement. *PNAS* **94**, 5018–5023.
52. Blundell, T. L., et al. (2002) High-throughput X-ray crystallography for drug discovery. In: (Flower, D. R., ed.) *Drug Design: Special Publication*, Royal Society of Chemistry, Cambridge, UK **279**, 53–59.
53. Adams, P. D., et al. (2002) PHENIX: building new software for automated crystallographic structure determination. *Acta Cryst.* **D58**, 1948–1954.
54. Brunzelle, J. S., et al. (2003) Automated crystallographic system for high-throughput protein structure determination. *Acta Cryst.* **D59**, 1138–1144.
55. Hartshorn, M. J. (2002) AstexViewer: a visualisation aid for structure-based drug design. *J. Comput. Aided Mol. Des.* **16**, 871–881.
56. Goh, C.-S., et al. (2003) SPINE 2: a system for collaborative structural proteomics within a federated database framework. *Nucleic Acids Res.* **31**, 2833–2838.
57. Peat, T., et al. (2002) From information management to protein annotation: preparing protein structures for drug discovery. *Acta Cryst.* **D58**, 1968–1970.

58. Boehm, H.-J., et al. (2000). Novel inhibitors of DNA gyrase: 3D structure based biased needle screening, hit validation by biophysical methods, and 3D guided optimization. A promising alternative to random screening. *J. Med. Chem.* **43**, 2664–2674.
59. Fejzo, J., et al. (1999). The SHAPES strategy: an NMR-based approach for lead generation in drug discovery. *Chem. Biol.* **6**, 755–769.
60. Carr, R. and Jhoti, H. (2002). Structure-based screening of low-affinity compounds. *Drug Discovery Today* **7**, 522–527.
61. Nienaber, V. L., et al. (2000). Discovering novel ligands for macromolecules using X-ray crystallographic screening. *Nat. Biotechnol.* **18**, 1105–1107.
62. Nowakowski, J., et al. (2002). Structures of the cancer-related Aurora-A, FAK and EphA2 protein kinases from nanovolume crystallography. *Structure* **10**, 1659–1667.
63. Mol, C. D., et al. (2003) Structure of a c-kit product complex reveals the basis for kinase transactivation. *J. Biol. Chem.* **278**, 31,461–31,464.
64. Williams, P. A., et al. (2003) Crystal structure of human cytochrome P450 with bound warfarin. *Nature* **424**, 464–468.

Automated High-Throughput Protein Crystallization

Arezou Azarani

1. Introduction

X-ray crystallography provides key biological insights into the three-dimensional structure and function of proteins, as well as essential information on protein–protein and protein–ligand interactions, therefore facilitating the design of more effective clinical drugs. The three most popular protein crystallization methods—vapor-diffusion sitting drop, hanging drop, and microbatch (1–6), most commonly used by investigators—convey major economical disadvantages for the setup of rapid large-scale crystallization of new proteins. Lack of suitable automation for the numerous lengthy and labor-intensive setup steps, irreproducibility as a result of manual intervention, waste of precious and scarce protein samples caused by the absence of precise low-volume dispensers and appropriate plate technology, exorbitant consumption of time, and cost (4,5) are among the most common drawbacks of high-throughput crystallization.

Crystallization of a protein sample occurs when a concentrated solution of that protein is brought to a state of supersaturation. In this state, called the crystal phase, the protein starts to nucleate and form crystals. Supersaturation is induced by the addition of a precipitating (crystallization or screen) buffer, which regulates different thermodynamic variables, such as pH, salt concentration, dielectric constant, and protein concentration of the reaction (3). This translates into setting up hundreds of different conditions, by trial and error, to determine the most favorable milieu(s).

Until most recently, crystallographers used nonstandard 24-well plates to manually set up crystallization reactions. However, screening of a large number of crystallization conditions, specifically for scarce proteins, was not realizable in this platform. As a result, many companies, such as Corning (www.corning.com, Acton, MA), Greiner Bio-One (www.greinerbioone.com, Germany), Emerald Biostructures (www.emerald-biostructures.com, Bainbridge Island, WA), and Art Robbins Enterprises (Mountain View, CA) came up with a new high-throughput screening plate technology for vapor-diffusion sitting-drop and hanging-drop crystallization as well as microbatch techniques. These new screening plates (96-, 192-, 288-, or 384-well plate format) are Society for Biomolecular Screening (SBS)-standard, which means that they are compatible with automated robotic equipment. This new plate technology facilitates the setup of hundreds of crystallization reactions without manual intervention (therefore resulting in higher precision and speed). In addition, these plates allow the setting up of nanoliter volumes of crystallization reagents and protein samples, therefore reducing waste (miniaturization).

From: *The Proteomics Protocols Handbook*
Edited by: J. M. Walker © Humana Press Inc., Totowa, NJ

Table 1
Companies Providing Automated Dispensers for Setting Up High-Throughput Crystallization Plates

Company	Instrument/technology	Lowest dispensing volume (nL)
Tecan www.tecan.com	Genesis -Piezo technology -1-8 pipetting devices called ActiveTip M	500
Douglas Instruments www.douglas.co.uk	Oryx6 -Syringe-based (6 channel) Impax -Syringe-based (5 channel)	100
Gilson www.gilson.com	926 PC Workstation -Syringe-based (1, 8, or 96 channel)	1000
Beckman www.beckman.com	Biomek® FX Laboratory Workstation -Syringe-based (8, 96, or 384 channel)	500
CyBio-AG www.cybio-ag.com	CyBi™ HTPC -Syringe-based (8, 96, or 384 channel)	500
Robbins Scientific/Matrix www.matrixcorp.com	Hydra®-Plus-One Hydra® II-Plus-One -Syringe-based (96 or 384) plus a noncontact microsolenoid dispenser called the NanoFill®	100
Cartesian Technology www.cartesiantech.com	ProSys™ 4950 -96 or 384 glass capillaries plus 8 noncontact microsolenoid dispensers	25

Many automated dispensers are now available in different formats (different dispensing technology, dispensing range, size, and price range) through companies such as Gilson, Beckman, Douglas Instrument, CyBio, Tecan, Cartesian, Matrix, and so on (Table 1). Setting up crystallization reactions in the SBS-footprint (96 to 1536) plates using many of these existing automated dispensers can take hours. The slow dispensing speed results in sample evaporation (a major issue when nanoliter dispenses are performed) causing protein denaturation and crystallization failure. To prevent sample evaporation, many dispensers need expensive humidifying chambers, the result being an increase in the cost of the dispenser.

Another major disadvantage of setting up crystallization reactions with the existing automated high-throughput dispensing technologies is the waste of scarce and expensive proteins as well as crystallization reagents. The commonly used 96- or 384-channel dispensers (tip- or syringe-based technology) generally aspirate samples from a

reservoir with a minimal working volume (the minimum volume needed to cover the floor of the reservoir to allow the filling of each channel) of up to 15 mL to dispense nanoliters of reagents. Therefore, even though the dispensers can be used to scale down the volumes of costly reagents (by providing down to nanoliter-dispensing capabilities), there is still a waste of expensive material. In addition, many of these dispensers can not dispense nanoliter volumes of samples (their lowest dispensing range being in the microliter range).

The aim of this chapter is to introduce a new technology, which circumvents many of the discussed difficulties of high-throughput crystallization by providing miniaturization (thus reducing cost) and automation (reducing manual intervention and therefore increasing precision and speed). This novel crystallization platform combines SBS-footprint high-throughput screening-plate technology with a highly accurate nano-dispensing robot. This robot is designed to dispense nanoliters of very viscous protein samples in a noncontact fashion, using a microsolenoid technology, while simultaneously dispensing many different crystallization screens (with different viscosity) using precision glass syringes by positive displacement technology (**Fig. 1**). A few different biotechnology companies provide this technology, and one such system called the Hydra®-Plus-One dispenser (Robbins Scientific-Matrix) (7,8) is described in this chapter.

1.1. The Hydra-Plus-One System

The Hydra-Plus-One system is composed of a single-channel noncontact microsolenoid dispenser, called the NanoFill® dispenser (7,8), and a 96- or 384-channel, nondisposable precision glass syringes system called the Hydra-PP (9,10) (**Fig. 2**). The NanoFill system uses a microsolenoid and positive pressure (created by helium gas) to dispense samples. It has a dispensing nozzle connected to a tube filled with system fluid (water). The sample is aspirated from a microtube (0.6 mL) into the nozzle and is separated from the system fluid by an air gap of 0.75 μ L. The NanoFill system has a hybrid valve architecture, in which the microsolenoid actuator is isolated from the sample flow, therefore preventing clogging issues (a major difficulty observed with the existing microsolenoid technology used for dispensing protein samples).

The microsolenoid dispenser transfers as low as 100 nL of viscous protein solutions (such as BSA solutions of up to 100 mg/mL) with a dispense-precision variation of less than 10% (7,8). A major advantage of this system is its dispensing speed (**Table 2**). The speed of the noncontact dispenser is 58 s per 96-well plate of the same solution (or 0.6 second per spot). As a result, protein sample evaporation is prevented at nanoliter volumes.

When using the single-channel dispenser to dispense protein sample, there is no waste, due to full material recovery from the noncontact dispenser. This means the unused protein sample (left in the dispenser) can be fully retrieved (back into the microtube source). Furthermore, because sample can be aspirated directly from the source tube (0.6 mL microtube), the single-channel dispenser eliminates the dead volume associated with aspirating from reservoirs, and thereby reduces sample waste.

Another advantage of noncontact dispensing is the ability to dispense samples without any wash requirement between dispenses of the same sample. Washes are necessary only when switching from one protein sample to another. This results in higher throughput and speed.

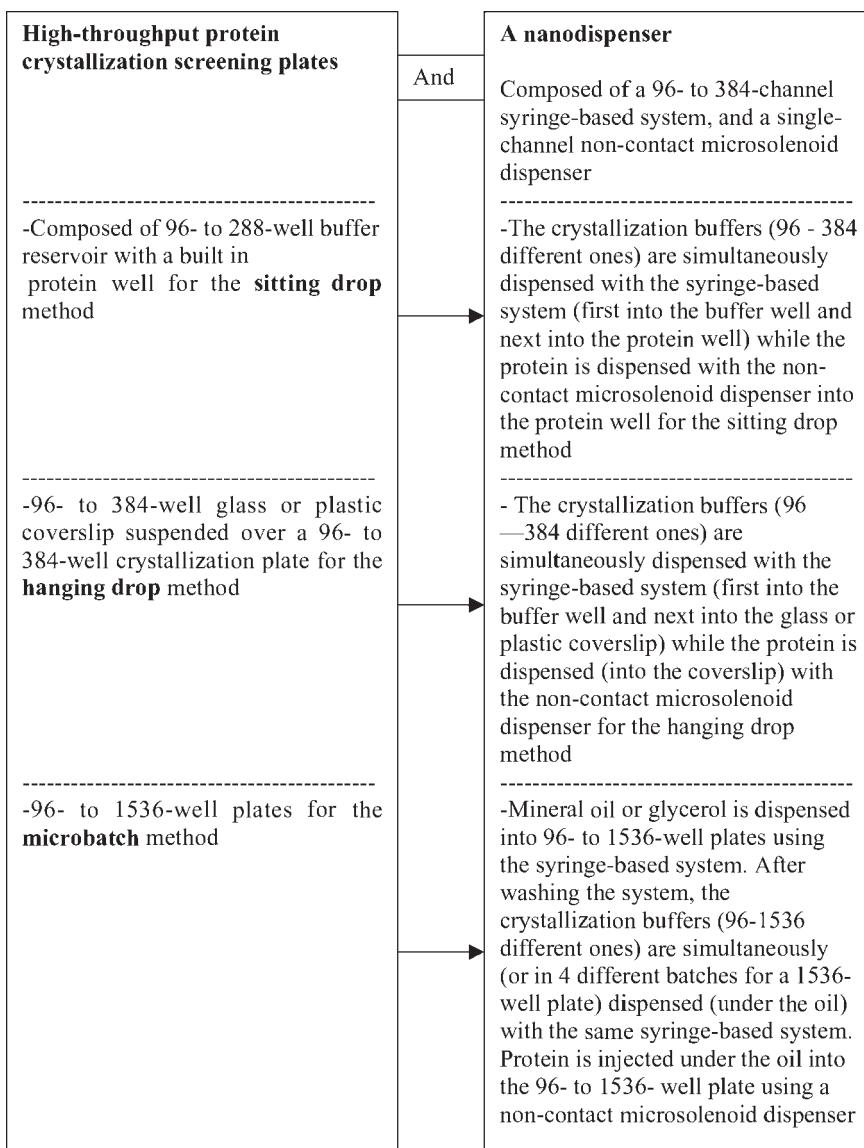


Fig. 1. An automated protein crystallization platform composed of high-throughput crystallization plates and a nanodispenser.

The Hydra-PP part of the Hydra-Plus-One system is equipped with 96 or 384 nondisposable precision glass syringes (with Teflon®-coated stainless steel needles) arrayed in standard microplate spacing, a computer-controlled plate-positioning stage composed of two nests, and an automated syringe washing station (9,10). For full automation capabilities, this system can be equipped with robotic plate handlers.

While the noncontact dispenser part of the Hydra-Plus-One system can be used for the precise low-volume dispensing of valuable protein samples, the Hydra-PP part of the system can dispense 96 or 384 different screening buffers (with different viscosities) simultaneously. This system can dry-dispense viscous samples as low as 200 nL,



Fig. 2. The Hydra-Plus-One system (reproduced from **ref. 4**).

Table 2
Setting Up High-Throughput Crystallization Reactions Using
the Hydra-Plus-One System

Function performed	Speed of the process
<i>For the sitting drop or the hanging drop:</i>	
Aspirating 96 crystallization buffers (from a deep-well plate) and simultaneously dispensing 50 μ L of samples, first into the buffer wells and then into the protein wells or a coverslip (as low as 200nL), using the Hydra-PP component of the Hydra-Plus-One system	20 s
<i>For the microbatch:</i>	
-Aspirating oil and dispensing into a 96-384 well plate using the Hydra-PP component of the Hydra-Plus-One system equipped with 96-384 syringes. The syringes are then washed before the next step	3 min (including a 2 min was after the oil dispense)
-Using the same system, after washing the syringes, and aspirating crystallization buffers (96-384) and simultaneously dispensing under the oil	20 s
<i>For the sitting drop, hanging drop, or the microbatch:</i>	
Aspirating the protein sample from a microtube and dispensing the protein sample (as low as 100nL) into the 96-protein wells or a coverslip or under the oil using the NanoFill microsolenoid dispenser	58 s
Total setup time not including the washes for the hanging drop or the sitting drop	2 min
Total setup time including washes (performed after the setup of plates) for the hanging drop or the sitting drop	4 min
Total setup time including the in-between washes for the microbatch	4.5 min

Table 3
High-Throughput Crystallization Plates for Sitting-Drop, Hanging-Drop, and Microbatch Technology

Company	Plates
Hampton Research www.hamptonresearch.com	96-, 192-, 288-, 384-, and 1536-well plates for the sitting drop, hanging drop, and microbatch technologies
Corning www.corning.com	96-, 192-, 384-, 1536-well plates for the sitting drop and microbatch technologies (CrystalEX™ Protein Crystallization Microplate)
Greiner Bio-One www.greinerbioone.com	96-well plates for the sitting drop technology (CrysalQuick Protein CRYSTAL Plate)
Emerald Biostructures www.emeraldbiostructures.com	96-well plates for the sitting drop technology (CompactClover Crystallization Plates)
Art Robbins Enterprises	96-well plates for the sitting drop technology (Intelli-Plate)

containing up to 30% polyethylene glycol (PEG) 8000 and 50% glycerol, or as low as 100 nL of aqueous solutions with CVs of less than 10%.

The use of fixed needles instead of disposable tips in dispensing systems provides critical advantages. In addition to facilitating the creation of a cost-efficient (inexpensive and reusable) (9,10) and environmentally friendly (creates less waste) method of liquid dispensing, precision glass syringes are simple to operate, robust, and easily replaceable. Since samples do not evaporate or dry once inside the syringes, there is no need for costly humidifying chambers. Other difficulties typically encountered in dispensing with existing inkjet or pin technologies (11)—such as difficulties in thoroughly cleaning the spotting devices, clogging, and inconsistencies in the transfer volume due to the difference in sample viscosity—are not observed when precision glass syringes are used for dispensing.

Another advantage of using syringes for dispensing is that they can be siliconized to prevent the buildup of air bubbles. Bubbles have been reported to cause protein denaturation and therefore crystallization failure.

The use of fixed needles in the Hydra-PP system dictates the need for an efficient syringe—cleaning procedure for the inhibition of carryover contamination and clogging issues. Wash procedures for this system are described in **Subheading 3**.

1.2. High-Throughput Crystallization Plates

High-throughput protein crystallization is now possible by the new SBS-standard plate technology. This new technology facilitates the setup of hundreds of protein crystallization reactions in the nanoliter range. A list of companies offering these plates is shown in **Table 3**. These plates can be used for the vapor-diffusion sitting-drop and hanging-drop crystallization techniques or the microbatch technique. The sitting-drop plates come in 96-, 192-, 288-, or 384-well format, while the hanging-drop plates come in 96- or 384-well format. The microbatch reactions can be set up in regular 96-, 384-, or 1536-well plates.

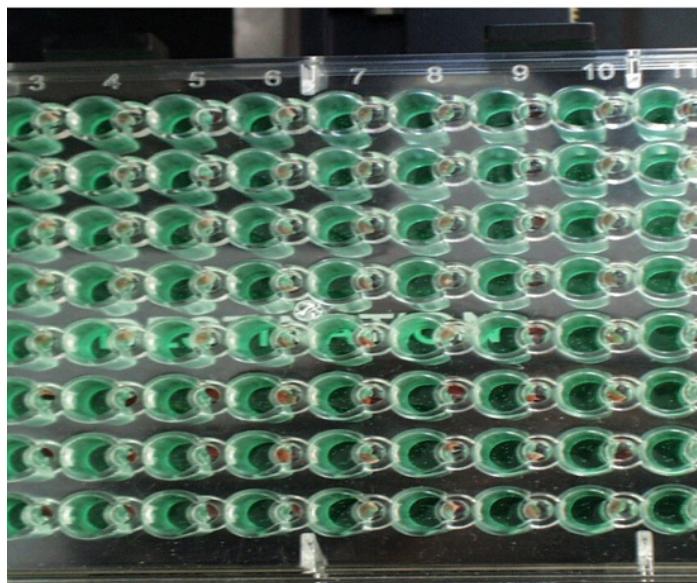
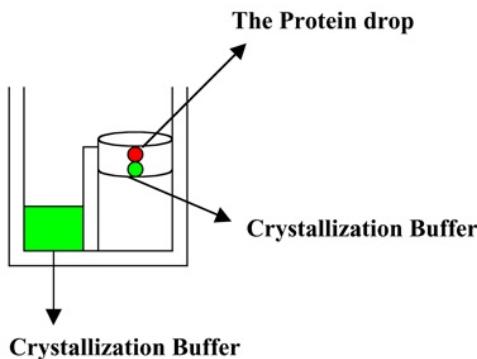


Fig. 3. Crystallization by the sitting-drop technology.

The high-throughput sitting-drop plates have buffer wells, each with a built-in (conical, square, or circular) flat- or round-bottomed protein well. The protein sample is combined with the different crystallization buffers (mostly in a 1:1 ratio) in each of the protein wells. The buffer well contains the same buffer as that present in its protein well. However, the concentration of the buffer in each of the buffer wells is at least twice the concentration of the same buffer in its protein well. Therefore, vapor diffusion occurs from the protein well to the buffer well, resulting in protein nucleation or formation of crystals (Fig. 3).

The hanging-drop plates have a glass or plastic coverslip where the protein sample is combined with the crystallizing agent (mostly in a 1:1 ratio). This cover slip is then inverted over a 96- or 384-well plate containing crystallization screens, allowing the coverslip protein samples to equilibrate with the crystallization buffers in each well by vapor diffusion. Since the concentration of the buffer in each well is at least twice the

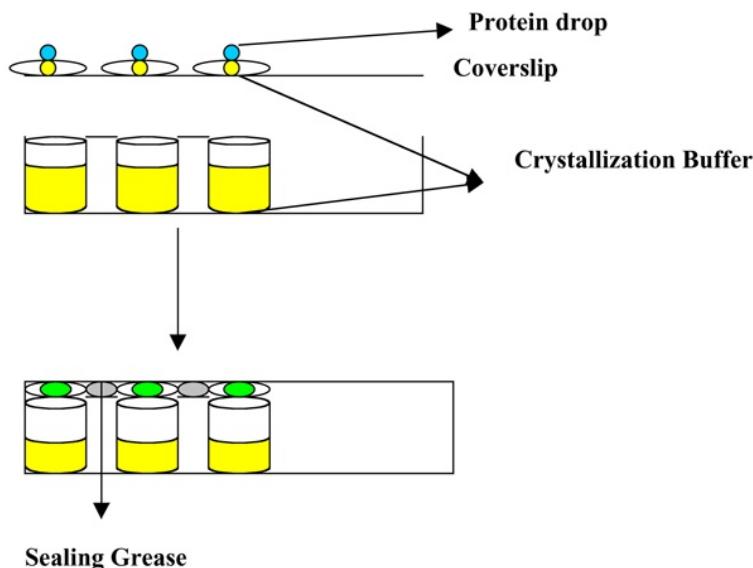


Fig. 4. Crystallization by the hanging-drop technology.

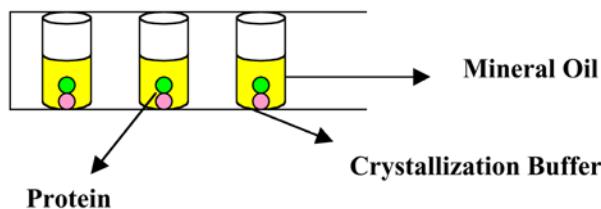


Fig. 5. Protein crystallization by the microbatch technique.

concentration of the buffer on the coverslip, vapor diffusion occurs from the coverslip toward each well, initiating protein crystallization. To create a seal between the wells and the coverslip, grease or Vaseline is used (Fig. 4).

For the microbatch experiments, mineral oil or glycerol is dispensed into 96- to 1536-well plates. Next, the crystallization buffers are dispensed under the oil, followed by dispensing the protein. The buffer and the protein drop (mostly in a 1:1 ratio) must come together under the oil (Fig. 5).

These new crystallography plates offer numerous advantages, such as allowing the setting up of high-throughput nanoliter volumes of crystallization experiments, thus preventing protein waste. They are designed with SBS-footprint, making them compatible with high-throughput, automated imaging and liquid-handling equipment, therefore minimizing user intervention. These plates have been designed to have high chemical resistance toward the most commonly used crystallization buffers (such as dimethyl sulfoxide [DMSO], acetone, acids, alcohol, and ammonia). In addition, the newly designed plates provide high optical clarity, with minimal background interference, enabling easy detection of protein crystals under polarized light.

Depending on the plastic material the plates are made of, the plastic might cause the creation of static forces, separating the protein and the buffer drops in the protein wells upon dispense. This protein-buffer-separation problem, very pronounced in polystyrene and polypropylene plates, leads to protein denaturation (specifically at lower protein volumes (<1 μ L). As a result, a mandatory, difficult-to-automate centrifugation step is required to prevent the separation of the buffer and protein drops in these plates.

Another disadvantage associated with numerous plates on the market is evaporation issues for the nanoliter volumes of protein dispensed, and as a result protein denaturation. However, many companies are coming up with proprietary hydrophilic coating technologies to create plates with low water-absorption characteristics, preventing very low volumes of the protein drops from evaporating.

2. Materials

1. The Hydra-Plus-One system can be purchased from Matrix. Other noncontact dispensers or syringe-based dispensers can be purchased from companies listed in **Table 1**.
2. The protein crystallization plates were purchased from Hampton Research, Laguna Niguel, CA; Corning, Acton, MA; Greiner Bio-One, Germany; Emerald Biostructures, Bainbridge Island, WA; or Art Robbins Enterprises, Mountain View, CA.
3. Crystallization screens were purchased from Hampton Research, Laguna Niguel, CA, or Emerald Biostructures, Bainbridge Island, WA.
4. An image-acquisition system such as VersaScan can be purchased from Velocity 11, Palo Alto, CA.
5. Fluorescein (Molecular Probes, Eugene, Oregon, Cat. no. F-1300) was dissolved to a concentration of 10 μ g/mL solution using a 0.1 M Tris-HCl (pH 8.0) buffer.

3. Methods

3.1. Preparation of the Hydra-Plus-One System

Prior to the setup of experiments, in between different dispenses, and after the last use of the instrument for the day, the Hydra syringes as well as the noncontact dispenser components of the Hydra-Plus-One system should be washed thoroughly to prevent sample-to-sample carryover, cross-contamination, and clogging. These complete procedures have been published elsewhere (7,8,11–15). If other nano-instruments with a similar technology are being used for the setting up of the crystallization plates, refer to the manual provided by the distributor for the recommended washing procedures.

3.2. Measuring the Dispensing Precision of the Hydra-Plus-One System

To ensure reproducibility, prior to the use of the dispensing system for the setting up of crystallization experiments, it is crucial to determine the uniformity and consistency of the protein and buffer volumes dispensed. Precision for volumes dispensed across the array of syringes as well as by the noncontact dispenser can be determined by the coefficient of variance for specific dispensing volumes. These procedures have been reported in detail in other publications (7,8,11–13). In short, different volumes of a 10 μ g/mL fluorescein solution are dispensed into each well of a 96-well plate containing 0.1 M Tris buffer. The final volume in each well is 100 μ L. Each plate is incubated for a period of 1 h, then read in a TECAN SpectraFluorTM fluorescence plate reader, and the CVs are determined across each plate for each of the dispensed volumes. A high

uniformity for dispensing volumes of ≥ 100 nL used in the crystallization experiments is evident, with CVs of less than 10%.

3.3. Setting Up the Hydra-Plus-One System for Preparing the Crystallization Plates

The Hydra stage is composed of two nests. For setting up crystallization plates for the sitting-drop or hanging-drop experiments, one nest is dedicated to the 96-deep-well buffer plate and the other nest to the crystallization plate or a coverslip. While the 96–384 different crystallization screens were dispensed using the Hydra-PP syringe-based component of the system, the protein was dispensed using the noncontact microsolenoid dispenser component.

For the sitting-drop experiments, using the syringes, 5 μL of air followed by the appropriate volume of the crystallization buffers (for example, 52 μL of mother liquors to dispense 50 μL of buffer) were aspirated from the 96-deep-well screen plates. After a triple trial dispense (500 nL dispensed back into the deep-well screen plate, to increase the accuracy of dispense), the crystallization buffers were dispensed into the buffer reservoirs of high-throughput crystallization plates, followed by a nano- to microliter dispense of the same buffer into the protein wells (anywhere from 200 nL to 2 μL). Next, 1 μL of air followed by the appropriate volume of the protein (for example, 20 μL of protein to dispense 200 nL for each well of a 96-well plate) was aspirated by the microsolenoid dispenser from a 0.5-mL microtube and dispensed into the protein wells after a triple 200-nL trial dispense (dispensed back into the protein tube).

For the hanging-drop experiments, the same procedures were followed as for the sitting-drop, with the exception that the crystallization buffers were dispensed into the buffer reservoirs first, and next onto the coverslip. The protein was dispensed onto the coverslip by the microsolenoid dispenser.

For the microbatch experiments, mineral oil or glycerol is dispensed into 96- to 1536-well plates using the syringe-based Hydra-PP system. After washing the syringes, the crystallization buffers (96–1536 different ones) are simultaneously (or in four different batches for a 1536-well plate) dispensed with the same syringe-based system. Protein is then injected under the oil into the 96- to 1536-well plate using the noncontact microsolenoid dispenser.

Plates were sealed manually or automatically and stored at room temperature. Depending on the plate used, a centrifugation step might be essential to ensure that the protein and the buffer drop come together, specifically if low nanoliter volumes are used in plastic plates with high-static problems.

3.4. Protein Crystallization

This melange of the syringe- and the microsolenoid-based technology in automated dispensers, such as Hydra-Plus One system, can be used to set up crystallization by all three techniques. Crystals grown by high-throughput plate and automated nano-dispensing technologies have been published in many journals (7,8,16,17). An example is shown in **Fig. 6**.

4. Notes

1. The advantages associated with the use of a nano-dispenser composed of a 96–384 syringe-based component and a noncontact microsolenoid component, as well as the use of high-



Fig. 6. Crystal of an unknown protein (LBNL, Berkeley, CA), called protein 1139, was obtained by the sitting-drop vapor-diffusion method in 1 μ L drops (500 nL of the protein and 500 nL of the mother liquor in each protein well) suspended over 50- μ L buffer reservoirs at room temperature. All dispensing was performed using a Hydra-Plus-One system. Plate used was CrystalEX Protein Crystallization Plate from Corning.

throughput plate technology in setting up protein crystallization trials, are reproducibility, accuracy, precision, speed, reduction in protein waste, the ability to dispense samples without any wash requirements between dispenses of the same buffer or protein samples, and cost and time reduction.

2. Complications such as difficulties in thoroughly cleaning the spotting devices, the drying of samples during the process of delivery, the buildup of bubbles, and clogging issues are circumvented when precision glass syringes and the noncontact dispensers are used for setting up crystallization plates.
3. Following the proper syringe and noncontact microsolenoid dispenser washing procedures is essential for the prevention of crossover contamination.

References

1. McPherson, A. (1982) Preparation and analysis of protein crystals. Krieger, Malabar, FL.
2. Rhodes, G. (2000) Crystallography made crystal clear: a guide for users of macromolecular models. Academic, San Diego, CA.
3. Hansen, C. L., Skordalakes, E., Berger, J. M., and Quake, S. R. (2002) A robust and scalable microfluidic metering method that allows protein crystal growth by free interface diffusion. *Proc. Natl. Acad. Sci. USA* **99**, 16,531–16,536.
4. DeLucas, L. J., Bray, T. L., Nagy, L., et al. (2003) Efficient protein crystallization. *J. Struct. Biol.* **142**, 188–206.
5. Villasenor, A., Sha, M., Thana, P., and Browner, M. (2002) Fast drops: a high-throughput approach for setting up protein crystal screens. *BioTechniques* **32**, 184–189.
6. Segelke, B. W. (2001) *J. Crystal Growth* **232**, 553–562.
7. Krupka, H. I., Rupp, B., Segelke, B. W., et al. (2002) The high-speed Hydra®-Plus-One system for automated high-throughput protein crystallography. *Acta. Cryst. D* **53**, 1523–1526.

8. James, A., Wu, H.-C., Braunthal, N., Shieh, J., and Azarani, A. (2003) Setting Up high-throughput, low-volume sequencing and PCR reactions using an automated system equipped with precision glass syringes and a non-contact microsolenoid dispenser. *JALA* **8**, 37–40.
9. Stanchfield, J., Wright, D., Hsu, S., Lamsa, M., and Robbins, A. (1996) Precision 96-channel dispenser for microchemical techniques. *BioTechniques* **20**, 292–296.
10. Mardis, E. R., Weinstock, L., Simonyan, A., and Stanchfield, J. E. (1997) M13 DNA preparations for a large scale sequencing project using the Hydra-96 microdispenser. Product Application Note 5, Apogent Discoveries (Web site at <http://www.apogentdiscoveries.com>).
11. Shieh, J., Carramao, J., Nishimura, N., Maruta, Y., Hashimoto, Y., Wright, D., Wu, H.-C., and Azarani A. (2002) High-throughput array production using precision glass syringes. *BioTechniques* **32**, 1360–1365.
12. To, C., Todd, P., Wright, D., and Azarani, A. (2001) Prevention of carry-over contamination from organic compounds, DNA and protein samples when using the Robbins Tango Liquid Handling System. Product Application Note 10, Apogent Discoveries (web site at <http://www.apogentdiscoveries.com>).
13. Wu, H.-C., Shieh, J., Wright, D., and Azarani, A. (2003) DNA sequencing using Rolling Circle Amplification and precision glass syringes in a high-throughput liquid handling system. *BioTechniques*, **34**, 204–207.
14. Azarani, A. (2004) The use of precision glass syringes and a noncontact microsolenoid dispenser for the production of high-throughput low-density arrays. In: (Fung, E. T., ed.) *Protein Arrays: Methods and Protocols*. Humana, Totowa, NJ.
15. James, A. T., Wu, H.-C., Braunthal, N., and Azarani, A. (2003) A study of a high-throughput plasmid DNA purification system, *JALA* **8**, 36–39.
16. Jurisica, I., Rogers, P., Glasgow, J. I., et al. (2001) *IBM Systems Journal* **40**, 394–409.
17. Mueller, U., Nyarsik, L., Horn, M., et al. (2001) Development of a technology for automation and miniaturization of protein crystallization. *J. Biotech.* **85**, 7–14.

NMR-Based Structure Determination of Proteins in Solution

Andrzej Ejchart and Igor Zhukov

1. Introduction

Nuclear magnetic resonance (NMR) spectroscopy is well suited to play an important part in proteomics programs because this method provides structural information at the atomic level. Nuclei of isotopes of biologically important elements display narrow resonance lines. Internuclear interactions, modulated by even small structural and conformational changes, influence line position, line shape, and intensity of signals in NMR spectra. Last but not least, NMR provides high-resolution structures in solution, allowing the study of proteins that fail to crystallize or comparison of differences between their crystal and solution structure (1). The potential of the NMR method, however, has not been reflected by the present number of deposited structures; less than 14% of the protein structures in the Protein Data Bank (PDB) have been determined by NMR spectroscopy (2).

The following factors have hampered a broad use of NMR in the determination of three-dimensional structures of proteins:

- Low sensitivity of NMR spectroscopy. A typical amount of protein used in NMR studies is approx $0.5 \mu M$, several orders of magnitude more than in mass spectrometry or optical spectroscopy.
- Because NMR measurements are time-consuming, protein stability in solution at room temperature is required for an extended period.
- Signals in spectra of individual isotopes of proteins are usually strongly superposed. For instance, in a protein built up of 200 amino acid residues, one can expect approx 1200 1H signals, 1000 ^{13}C signals, and more than 200 ^{15}N signals placed in narrow spectral windows.
- Strong solvent signal generates dynamic range problems in 1H spectra ($S_{\text{water}}/S_{\text{protein}} = \text{approx } 10^5$) and obscures a diagnostically important spectral region of H_a signals.
- In large proteins, fast transverse nuclear relaxation brings about line broadening that aggravates superposition of signals and eliminates their fine structure.
- NMR-derived data determine ambiguous structural constraints, leaving a number of possible solutions.
- Manual signal assignment and constraint identification is laborious and time-consuming.

Recently, most of these obstacles have been overcome, because of the progress in spectrometer design, isotope labeling, new spectral techniques, and new, often automated, computational procedures for spectral assignment and structure calculation (3)

(see **Subheading 3.**). In the following sections, the most important NMR techniques used in the structure determination of proteins in solution are presented.

2. Basic NMR Concepts

Isotopes possessing nuclei with spins, and thus magnetic moments, are magnetically active, and their NMR spectra can be measured. The resonance frequency of a nucleus is given by a formula:

$$f = \gamma B_0 (1 - \sigma) / 2\pi$$

where B_0 is the strength of the spectrometer magnetic field and the magnetogyric ratio, γ , is characteristic for a given isotope. The shielding constant, σ , strongly depends on the distribution of the electronic charge in the vicinity of the nucleus; therefore, for local molecular structure, typically it is of the order of 10^{-4} – 10^{-5} . Since the use of shielding constants is not convenient, the related parameter *chemical shift* has been introduced. It is given by the difference in resonance frequencies between the nucleus of interest and a reference nucleus, and is expressed as a dimensionless parameter Δ :

$$\Delta = 10^6(f - f_{ref})/f_{ref} = \text{approx } 10^6(\sigma_{ref} - \sigma)$$

A factor of 10^6 is introduced for convenience, and the chemical shift is expressed in parts per million (ppm).

The sensitivity of NMR measurements expressed as a signal-to-noise ratio is given by:

$$S/N = \text{approx } \gamma_i \gamma_d^{3/2} B_0^{3/2} N T^{1/2} / T$$

where γ_i and γ_d are the magnetogyric ratios for initially perturbed and detected nuclei, respectively, provided they are different. Thus, measurements that start and end with large-magnetogyric-ratio nuclei are most sensitive. NT is the number of coherently added experiments. In order to increase signal-to-noise ratio n times, the total experimental time has to be increased n^2 times. Magnetic field strength and temperature also influence the sensitivity.

Nuclear spins can mutually interact in two ways—through chemical bonds and through space. The interaction transmitted through chemical bonds, the spin-spin (scalar) interaction, is generally observed between nuclei separated by four or fewer chemical bonds. It depends on the type of elements and chemical bonds in the pathway. The spin-spin interaction manifests itself in the spectrum as a signal splitting, with components separated by the spin-spin coupling constant, J , which contains structural information. Three-bond, vicinal coupling constants, 3J , are particularly important from this standpoint, because they depend on the dihedral angle defined by the intervening bonds. Besides the structural information, spin-spin coupling constants are utilized for polarization transfer—the technique used for sensitivity enhancement in NMR spectroscopy. For this purpose, heteronuclear one-bond couplings are especially useful,

$$^1J(^1H, ^{13}C) > 120 \text{ Hz and } ^1J(^1H, ^{15}N) < -90 \text{ Hz.}$$

Nuclear spins interact through space as magnetic dipoles. Static dipolar coupling constants, D , depend on the distance between interacting nuclei and their magnetogyric ratios; $D_{IJ} = \text{approx } \gamma_i \gamma_j r_{ij}^{-3}$. Typical D values are several orders of magnitude larger than spin-spin coupling constants. In proteins, for instance, $D(C_\alpha, H_\alpha) = \text{approx } -24$

kHz, or $D(N,C') = \text{approx } 1 \text{ kHz}$. However, dipolar interactions, which depend on the orientation of internuclear vectors relative to B_0 , do not split NMR signals in isotropic solutions, because they are averaged to zero by the fast diffusional tumbling of solute molecules. Even so, dipolar interactions manifest themselves in isotropic solutions and generate the nuclear Overhauser effect (NOE), which is the change in the intensity of the NMR signal of a nuclear spin when the thermodynamic equilibrium of another nuclear spin interacting with a given one is perturbed. The NOE arises due to cross-relaxation taking place during an appropriately designed and performed experiment, which results in the transfer of magnetization between dipolarly interacting protons. The most important dependence of the cross-relaxation rate σ_{ij} between two protons i and j from a structural standpoint is that on the inverse sixth power of the internuclear distance, $\sigma_{ij} = \text{approx } D^2 = \text{approx } r_{ij}^{-6}$. Therefore, NOEs bear structural information about interatomic distances in molecules. If solute molecules are placed in anisotropic solution, the averaging of dipolar interactions will be incomplete, and so-called residual dipolar couplings (RDC) will appear. Their magnitudes depend on the degree of alignment of solute molecules and the orientation of the internuclear vector relative to the principal axes of the molecular orientation tensor, giving another type of structural information.

A perturbed system of nuclear spins regains its thermodynamic equilibrium owing to the process termed *relaxation*. Longitudinal relaxation restores the static magnetization along the B_0 , whereas transverse relaxation causes the decay of transverse magnetization, which at equilibrium is equal to zero. Transverse relaxation rate is the main factor determining line width in NMR spectroscopy: both these parameters increase with the mass of protein and viscosity of solution. Two relaxation mechanisms are of importance in the relaxation of nuclei with spin $I = 1/2$ in macromolecules: the dipolar (DD) mechanism and chemical shift anisotropy (CSA). Nuclei with spin $I > 1/2$ possess quadrupolar moments besides the magnetic ones. Interaction of nuclear quadrupolar moment with molecular electric field gradients is usually much stronger than the DD and CSA interactions. For this reason transverse relaxation rates of quadrupolar nuclei are much faster than appropriate rates of 1/2-spin nuclei lacking quadrupolar moment. Fast relaxation causes significant broadening of NMR signals and often makes their observation impossible. Therefore, the main interest of NMR spectroscopists is focused on displaying narrow signals of one-half-spin nuclei. Luckily, for all biologically important elements except oxygen, isotopes with one-half-spin nuclei exist. ^1H and ^{31}P are isotopes with large natural abundance (99.98% and 100%, respectively), whereas the opposite is true for ^{13}C and ^{15}N isotopes (1.1% and 0.37%, respectively). Therefore, the enrichment of the latter isotopes can be necessary for successful NMR study. In NMR spectroscopy, resonance frequencies of nuclei fall in the radio frequency range: e.g.,

$$f(^1\text{H}) = \text{approx } 500 \text{ MHz}, f(^{13}\text{C}) = \text{approx } 125.7 \text{ MHz}, \text{ and } f(^{15}\text{N}) = \text{approx } 50.7 \text{ MHz} \text{ at } B_0 = 11.7 \text{ T.}$$

In the conventional, one-dimensional (1-D) experiment, generation of the initial nonequilibrium state of the spin system (preparation) is followed by the optional magnetization transfer between different spins (mixing) and the detection of the response of a spin system (acquisition). Such a signal of free induction decay (FID) has to be

Fourier transformed in order to change the time domain of the FID to the frequency domain of the spectrum.

In two-dimensional (2-D) spectroscopy after the preparation period the perturbed spin system is allowed to evolve at its characteristic frequencies (evolution). Owing to mixing, these frequencies can modulate the oscillations detected during acquisition. If the evolution period is systematically incremented, a set of 1-D data, each differently modulated, will be obtained, forming a second time domain. Two subsequent Fourier transformations applied to both time domains move the data to the two-dimensional frequency domain. If evolved and detected spins belong to the same isotope, the resulting homonuclear 2-D spectrum will display two types of signals: diagonal peaks essentially corresponding to those appearing in a 1-D spectrum, and off-diagonal cross peaks, whose coordinates $f_1 \neq f_2$ represent chemical shifts of interacting nuclear spins. The type of interaction (spin-spin or dipolar) is chosen by the appropriate sequence of mixing periods. If evolved and detected spins belong to different isotopes, a corresponding heteronuclear 2-D spectrum will contain only cross peaks. The procedure can be generalized for third and further dimensions by adding additional sets of evolution and mixing periods.

One has to be aware that evolution time(s) should be long enough to provide sufficiently good spectral resolution. This means that the number of individual measurements becomes equal to the product of the number of increments of all evolution times resulting in the significant increase of a total experimental time, sometimes beyond the sensitivity requirement.

NMR spectroscopy is a method that is sensitive to dynamic processes of a broad range of frequencies. Therefore, the definition of the NMR timescale is crucial for understanding the influence of molecular motions on the NMR parameters. If a given motion is fast in the NMR timescale, it will average an NMR parameter, giving a weighted average of the individual species. This takes place when the lifetimes of those species, τ , fulfill the condition: $\tau \ll 1/\Delta P$, where ΔP represents the parameter difference between species. Otherwise, separate spectra are observed for each species. Chemical shifts are typically averaged by processes in which the lifetimes are 10^{-3} s or shorter. Since the major part of conformational motions is faster, usually conformationally averaged NMR spectra are observed. This means that usually NMR-derived structural information is conformationally averaged.

3. Tools

3.1. Isotope Labeling

The key aims of isotope labeling are to increase sensitivity, to eliminate signal overlap, and to narrow linewidths. ^{15}N and ^{13}C labeling simplifies the spectral assignments and provides new types of structural constraints. On the other hand, deuteration allows reduction of linewidths, to partially or totally eliminate ^1H signals and to reduce the parasitical spin diffusion effect. Uniform $^{15}\text{N}/^{13}\text{C}$ double labeling has been successfully used in the identification of signals in the spectra of ^1H , ^{13}C , and ^{15}N isotopes. Selective $^1\text{H}/^{13}\text{C}$ labeling of methyl groups of Ala, Val, Leu, and Ile or aromatic rings of Phe, Tyr, and Trp in otherwise deuterated proteins, facilitates identification of side chain-side chain dipolar interactions crucial in the determination of hydrophobic core conformation. Segmental isotope labeling allows decreasing of the number of signals

in NMR spectra, thus reducing signal overlap. Such proteins are produced through the ligation of labeled and unlabeled polypeptide chains in a self-catalytic protein splicing process (4,5).

3.2. Polarization Transfer

The purpose of the insensitive nuclei enhanced by polarization transfer (INEPT) technique is to transfer magnetization from nuclei with large magnetogyric ratio (^1H) to nuclei with small magnetogyric ratio (^{13}C , ^{15}N) by means of the spin–spin interaction. A sensitivity gain of $\gamma_{\text{H}}/\gamma_{\text{C}} = \text{approx } 4$ or $\gamma_{\text{H}}/\gamma_{\text{N}} = \text{approx } 10$ is significant. If this technique is combined with reverse INEPT, when the magnetization is transferred back and forth, the sensitivity gain is equal to $(\gamma_{\text{H}}/\gamma_{\text{X}})^{5/2}$, giving an enormous sensitivity enhancement, 32 and 306 for ^{13}C and ^{15}N nuclei, respectively (6).

During the time required for the completion of INEPT—of the order of $1/J_{\text{HX}}$ (7–10 ms)—relaxation processes take place, defocusing the magnetization. For large proteins ($MW > 30 \text{ kDa}$), a very fast transverse relaxation leads to the significant sensitivity loss of the technique. Another transfer technique, the cross relaxation-enhanced polarization transfer (CRINEPT), which is much less sensitive to transverse relaxation effects, allows one to obtain an efficient magnetization transfer in molecules far beyond $MW = 30 \text{ kDa}$ (7).

3.3. Transverse Relaxation-Optimized Spectroscopy

Fast transverse nuclear relaxation caused by DD and CSA mechanisms is one of the major factors limiting the size of proteins that can be investigated by NMR spectroscopy. These two mechanisms can interfere, modifying the apparent relaxation rate. Such interference is a source of structural information, as discussed in **Subheading 5.5**. It also results in the differentiation of relaxation broadening of the spin–spin coupled multiplet components.

In a two-spin scalar-coupled system, wherein the DD and CSA mechanisms effectively interfere, the line widths of doublet components are different, owing to the mutual compensation of the relaxation mechanisms for one component and their enhancement for another one. The transverse relaxation-optimized spectroscopy (TROSY) approach retains solely the narrow component, thus improving spectral resolution and sensitivity. The TROSY technique can be incorporated in most NMR experiments used in protein structure determination (8).

3.4. Multidimensional NMR Spectroscopy and Reduced Dimensionality

The introduction of multidimensional spectroscopy exerted a strong impact on biomolecular NMR. In 1-D spectra, only small molecules with a very limited number of nonequivalent nuclei do not show signal overlap. In proteins containing numerous repeating subunits, extensive superposition of spectral lines precludes their unequivocal identification and structural assignment. In multidimensional spectroscopy, each nucleus is identified not only by its own resonance frequency but also by the frequency or frequencies of nuclei interacting with it. Since usually resonance frequencies of interacting nuclei are weakly correlated, a probability of occurrence of identical sets of resonance frequencies characterizing a given nucleus is virtually eliminated. This feature is particularly noticeable in heteronuclear multidimensional spectra (9). For instance, the signal overlap in one-dimensional ^1H and ^{15}N spectra of proteins usually

is mostly removed in the 2-D $^1\text{H}/^{15}\text{N}$ correlation. Several clusters of cross-peaks in such spectra can be further resolved in a three-dimensional HNCO spectrum correlating chemical shifts of ^1H and ^{15}N amide group nuclei with the chemical shift of the carbonyl

^{13}C nucleus of the preceding amino acid residue (**Fig. 1**).

An extended experimental time required for the collection of multidimensional spectra is the main drawback of this approach. The method of reduced dimensionality (RD) significantly reduces experimental time in comparison with the standard multidimensional methods (10–12). It relies on the simultaneous incrementation of two or more evolution times, resulting in the encoding of several chemical shifts in one of the retained spectral dimensions. An appropriate data processing allows one to obtain a set of spectra wherein signal positions are linear combinations of the chemical shifts determined from the compressed dimensions (**Fig. 2**). As a result, NMR spectra with good spectral resolution and satisfactory sensitivity are obtained in acceptable total experimental time.

3.5. Automated Assignment of NMR Spectra and Ambiguous Constraints

Automated or computer-assisted assignment procedures of protein backbone resonances can significantly speed up this laborious process. A number of methods utilizing data from multidimensional heteronuclear spectra have become available, allowing for faster and more routine analyses in the first step of structure determination by NMR (13,14).

Identification of a large number of unambiguous structural constraints is a key point in the process of evaluation of high-quality protein structure. In practice, a major part of NMR constraints is ambiguous due to intensive cross-peak overlap, limited data accuracy, and spectral artifacts. Several attempts have been made to automate the process of constraint identification and verification (15–17). It should be stressed, however, that automatic procedures require high accuracy and complete input data.

4. Strategies for NMR-Based Protein Structure Determination

An NMR-based procedure of structure determination comprises three stages: assignment of as many signals as possible in the spectra of NMR-active isotopes, identification of structural constraints, and calculation of a family of three-dimensional structures fulfilling experimental constraints. Approaches used at the first two stages depend on the size of protein studied, whereas the approach applied at the third stage depends on the type and number of identified constraints.

Small proteins with molecular weight $MW < 10$ kDa are usually studied using solely two-dimensional (2-D) ^1H NMR spectra. In the first step, spin systems of similar topology are identified from the correlations utilizing ^1H - ^1H scalar couplings. Sequential assignment of spin systems is based on the short-range dipolar interactions identified in 2-D nuclear Overhauser effect (NOESY) spectra. Next, constraints characterizing secondary-structure elements are elucidated from medium-range NOEs and vicinal scalar couplings, $^3\text{J}(\text{H}_\text{N}\text{H}_\text{o})$. Finally, long-range NOEs and hydrogen bonds are used for tertiary-structure determination (18).

For proteins with $MW > 10$ kDa, two simultaneous limitations occur: progressive overlap of signals and their broadening owing to fast transverse nuclear relaxation.

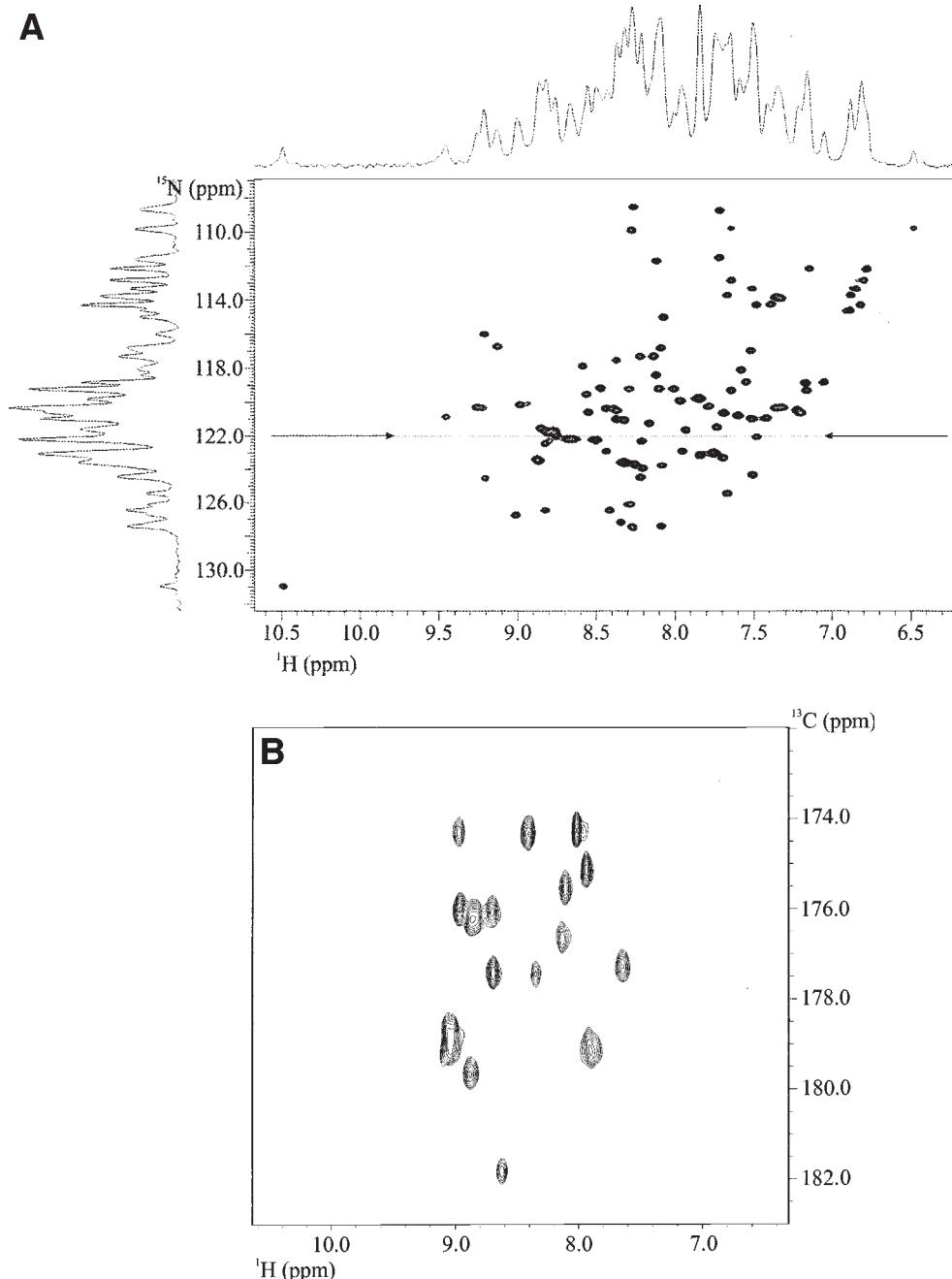


Fig. 1. (A) Two-dimensional nuclear magnetic resonance (NMR) spectrum of homodimeric $^{15}\text{N}/^{13}\text{C}$ -labeled S100A1 protein (93 a.a. in subunit) correlating ^1H (horizontal axis) and ^{15}N (vertical axis) chemical shifts in amide groups shows a good spectral dispersion. On the other hand, signals in both one-dimensional spectra (traces above and to the left of the spectrum) are strongly overlapped. (B) In the cross-section of 3-D HNCO spectrum at the place shown by arrows in the A part, each cross-peak is characterized by the chemical shifts of the ^1H and ^{15}N amide group nuclei and the carbonyl ^{13}C nucleus of the preceding amino acid residue. The cross-section shows dispersion of cross-peaks clustered in spectrum A along the ^{13}C chemical shift.

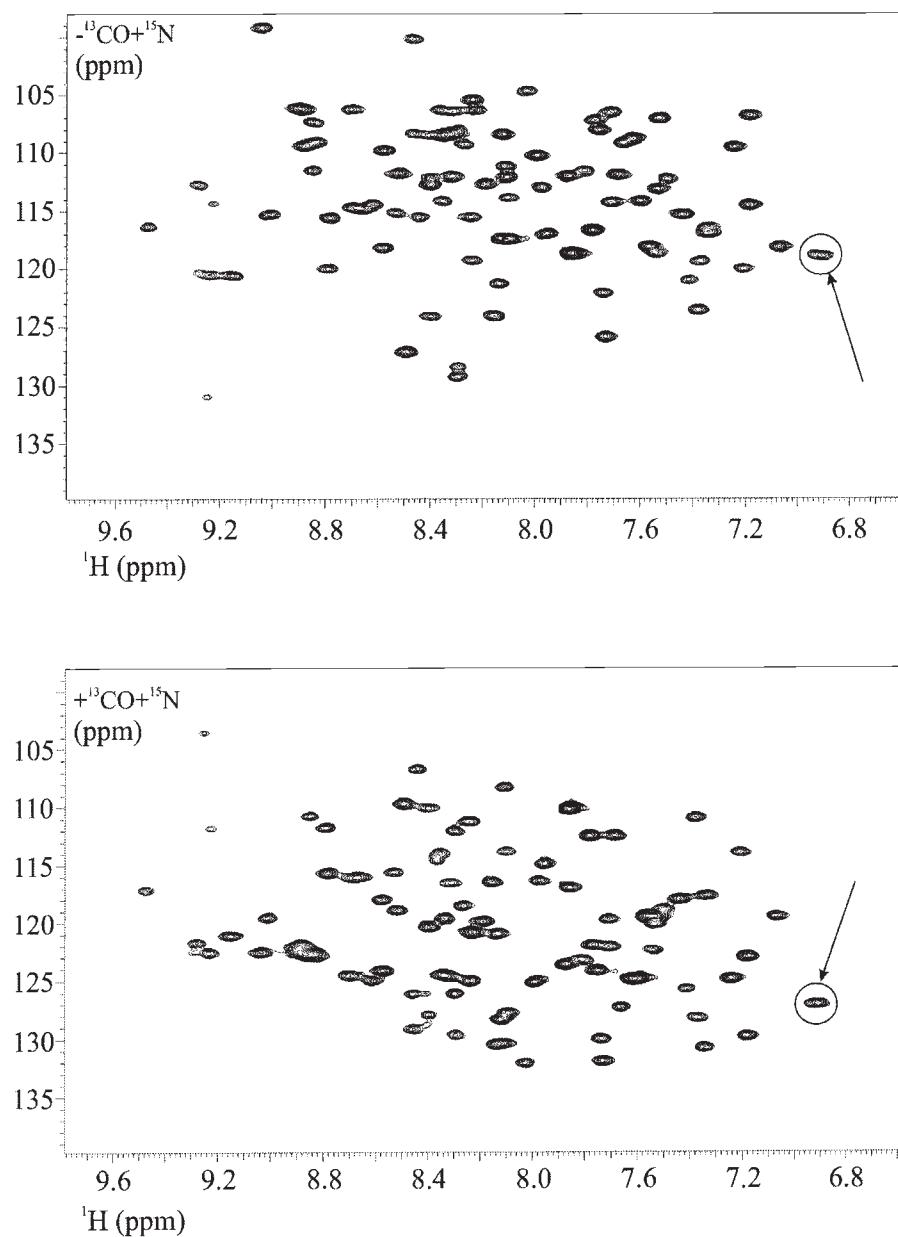


Fig. 2. Two-dimensional HNCO spectrum with reduced dimensionality measured for ^{15}N / ^{13}C -labeled S100A1 protein at $B_0 = 11.7$ T. Vertical coordinate contains information on ^{15}N and ^{13}C chemical shifts. $\Delta^{(15)\text{N}}$ is calculated as an average of coordinates in two spectra denoted as $[-^{13}\text{CO} + ^{15}\text{N}]$ and $[+^{13}\text{CO} + ^{15}\text{N}]$. For a marked cross-peak, it gives the value 122.95 ppm. $\Delta^{(13)\text{C}}$ displacement from ^{13}C frequency (in this measurement set to 176 ppm) is obtained from the frequency difference of signals relative to $\Delta^{(15)\text{N}}$ and its recalculation to the ^{13}C chemical shift scale, $176.0 + (122.95 - 118.87) \times 50.7/125.7$, giving $\Delta^{(13)\text{C}} = 177.64$. Experimental time for 2-D HNCO-RD was equal to 95 min, whereas for the corresponding 3-D HNCO, the time was 215 min.

Usually proteins with $MW < 30$ kDa are uniformly $^{15}\text{N}/^{13}\text{C}$ double labeled in order to take advantage of heteronuclear, multidimensional techniques. They allow replacement of NOE-based sequential assignment with correlations transmitted through heteronuclear scalar couplings, which are more sensitive and display better signal dispersion. Moreover, correlations in NOESY spectra can be spread out due to ^{15}N and/or ^{13}C editing (9,19). Additionally, combined information on ^1H , ^{13}C , and ^{15}N resonance frequencies of individual backbone nuclei can be used in such statistical methods of secondary-structure determination as chemical shift index (20) or TALOS software (21). These methods become insufficient for proteins with $MW > 30$ kDa. In very large proteins and protein assemblies, signal overlap can be further diminished by selective or segmental isotope labeling. On the other hand, relaxation-based line broadening and loss of sensitivity can be limited by deuteration and by application of TROSY and CRINEPT techniques (3).

5. NMR-Derived Structural Constraints

Three-dimensional structure of any molecule built up of N atoms is unequivocally determined by $3N - 6$ internal coordinates—interatomic distances, valence angles, and dihedral angles. It is usually assumed that distances between directly bound atoms are well represented by bond lengths and the valence angles by their standard values. On the other hand, experimentally derived structural constraints are required for the determination of dihedral angles. Experimental constraints are also important for the verification of other internal coordinates. NMR spectroscopy is a source of several types of structural constraints.

5.1. $^1\text{H} - ^1\text{H}$ Distances

$^1\text{H} - ^1\text{H}$ distances are determined quantitatively or semiquantitatively from the nuclear Overhauser effect, which depends on the inverse sixth power of the internuclear distance. Therefore, the relative intensities of NOE cross-peaks in NOE spectra can be used to quantify internuclear distances, providing the cross-peak(s) between protons of fixed, known separation (e.g., geminal H_β protons or *ortho* protons H_Δ and H_ϵ in Phe and Tyr) were identified and used for the distance calibration. In a more conservative approach, NOE cross-peaks are divided according to intensity into three groups—strong ($0.18 \text{ nm} < r < 0.25 \text{ nm}$), medium ($0.18 \text{ nm} < r < 0.35 \text{ nm}$), and weak ($0.18 \text{ nm} < r < 0.50 \text{ nm}$). In practice, the maximum distance so available is approx 0.5 nm (22). Larger distances are usually influenced by spin diffusion, a multistep magnetization transfer, which can lead to incorrect internuclear distances and hence to imposition of tighter interproton distance constraints than is justified. These problems can be circumvented when complete relaxation matrix methods allowing for the spin diffusion are used. One should realize that a single internuclear distance determines a sphere of radius r , and as many as four different distances are required to remove the spatial ambiguity.

In order to disperse superposed signals in the crowded NMR spectra of large proteins, multidimensional NOE spectroscopy is routinely used. Two-dimensional NOE (NOESY) spectra of medium and large proteins usually show strong signal overlap. To overcome this problem, ^{15}N and/or ^{13}C edited 3-D/4-D NOESY spectra of respectively labeled proteins should be measured (Fig. 3). Selective deuteration is especially useful for this purpose.

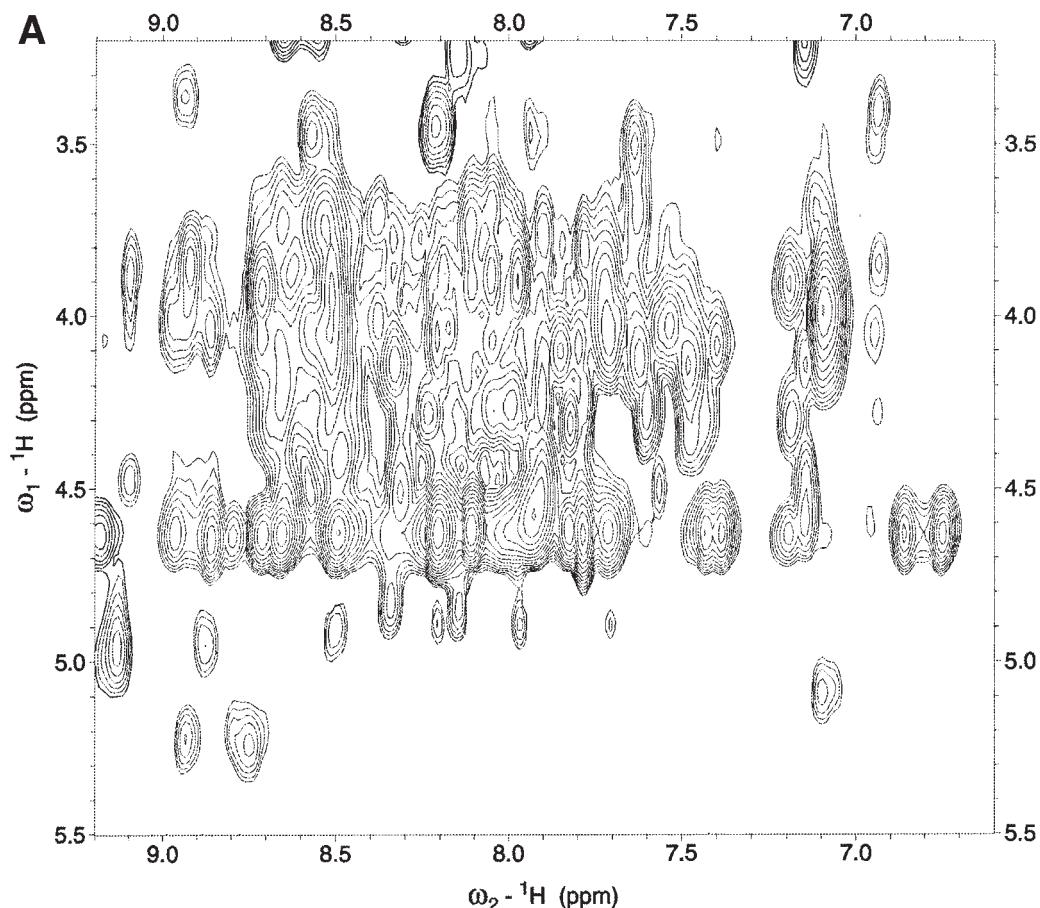
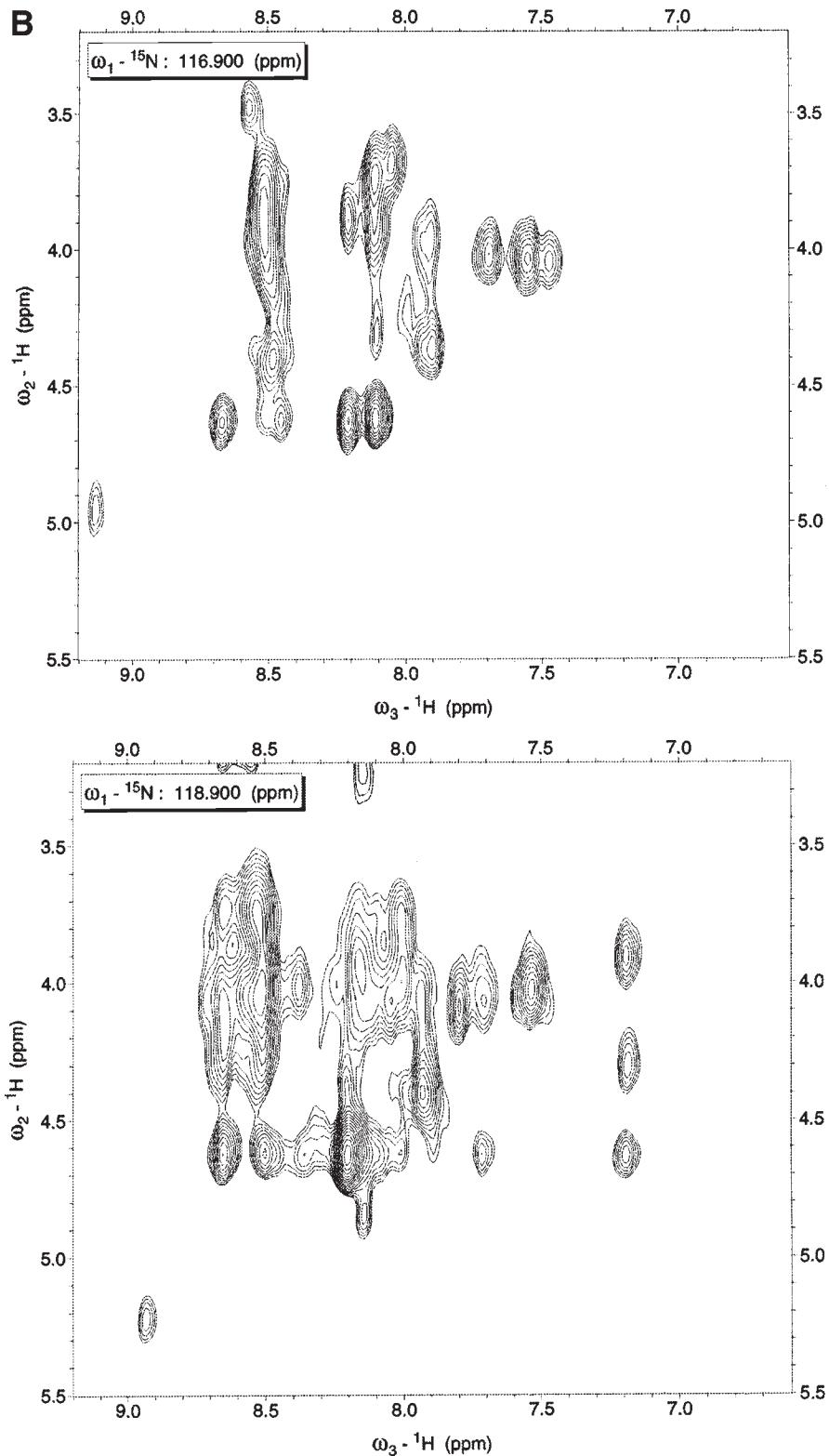


Fig. 3. (A) Part of 2-D NOESY spectrum shows severely superposed $\text{H}_\text{N}/\text{H}_\alpha$ correlations for S100A1 protein, making full signal identification impossible. In the 3-D ^{15}N -edited NOESY spectrum, signals are spread along third axis, representing ^{15}N chemical shifts of the same amide group as a corresponding H_N proton. (B) Two cross-sections (^{15}N chemical shifts given in the figures) show that the signal superposition has been mostly removed.

5.2. Hydrogen Bond Donor-Acceptor Distances

Hydrogen bonds are of key importance for stabilizing protein structures. The presence of hydrogen bonds indicates the spatial proximity and relative arrangement of the atoms involved. Direct evidence for the existence of hydrogen bonds can be established in proteins by the observation of scalar couplings between an amide ^{15}N and carbonyl ^{13}C nuclei of two residues, $^{15}\text{N} - \text{H} \cdots \text{O} = ^{13}\text{C}$ across a hydrogen bond (23). Thus, the detection of scalar coupling through a hydrogen bond unambiguously imposes a valuable distance constraint at the stage of the backbone assignment in a protein; $0.18 \text{ nm} < d(\text{H}_\text{N} \cdots \text{O}) < 0.20 \text{ nm}$ and $0.27 \text{ nm} < d(\text{N} \cdots \text{O}) < 0.30 \text{ nm}$. This method requires the use of $^{15}\text{N}/^{13}\text{C}$ double-labeled proteins.



5.3. Dihedral Angles Evaluated From Vicinal Scalar Coupling Constants

In conformational studies, the Karplus relation between vicinal (through three bonds) scalar couplings, 3J , and dihedral angles, ϕ , is of great importance. This relation can be represented by the general formula: $^3J = A \cdot \cos^2\phi + B \cdot \cos\phi + C$. Coefficients A , B , and C depend on a variety of molecular parameters. Of these, the type of elements forming the central bond as well as electronegativity and relative position of their substituents are most important. Therefore, applications for a given class of molecules have so far used empirical calibration of the coefficients derived from measurements of model compounds. In structural studies of proteins, scalar couplings determining the Φ backbone angle have been the most widely used. Six homo- and heteronuclear scalar couplings are related to this dihedral angle and corresponding experimental Φ -dependent Karplus curves. Owing to the periodicity of the Karplus equation, a single value of vicinal coupling constant can correspond to as many as four different dihedral angles, introducing an ambiguity to the scalar coupling-based constraints. However, an appropriate dihedral angle can be derived from a combination of several scalar coupling constants. Isotopic labeling, which is required when heteronuclear scalar couplings are measured, facilitates the determination of homonuclear scalar couplings as well (Fig. 4) (24).

5.4. Orientations of Internuclear Vectors in the Molecular Reference Frame Derived From Residual Dipolar Couplings

Orientational information relative to a common molecular reference frame can be obtained from residual dipolar couplings, owing to partial molecular alignment. The partial alignment of proteins can be induced by solvation in dilute anisotropic media such as phospholipid bicelles, filamentous phage, or strained gels (25). It prevents complete averaging of dipolar interactions as in isotropic solution. The direct measurement of the RDCs provides long-range orientational information for internuclear vectors positioned throughout the studied macromolecule (25–29). Similar information is provided by heteronuclear relaxation parameters in anisotropically tumbling molecules. Geometric dependence of RDC values on the orientation of internuclear vectors determining dipolar interaction relative to the order matrix is similar to the dependence of relaxation parameters on the orientation of specific relaxation vectors relative to the diffusion tensor. Identical experimental values are distributed on two elliptic cone surfaces. Measurements of RDCs as well as relaxation parameters require labeled proteins. An example of the determination of RDC values is shown in Fig. 5. It is noteworthy that a multitude of different vectors in proteins whose orientations are available from RDCs allows one to greatly improve both the precision and accuracy of solution structures for proteins and their complexes. Recently, the determination of protein backbone conformation using only RDCs constraints has been reported (30).

As pointed out in (27), the degree of alignment should be small enough to scale down dipolar couplings approximately three orders of magnitude. Stronger alignment results in broadening of 1H multiplets as an effect of many unresolvable homonuclear dipolar couplings. It also yields heteronuclear 1H – ^{15}N and 1H – ^{13}C couplings that become of a magnitude comparable to the one-bond 1J coupling. The latter effect can preclude the use of the INEPT technique in cases where the dipolar and scalar contributions to the splitting have opposite signs.

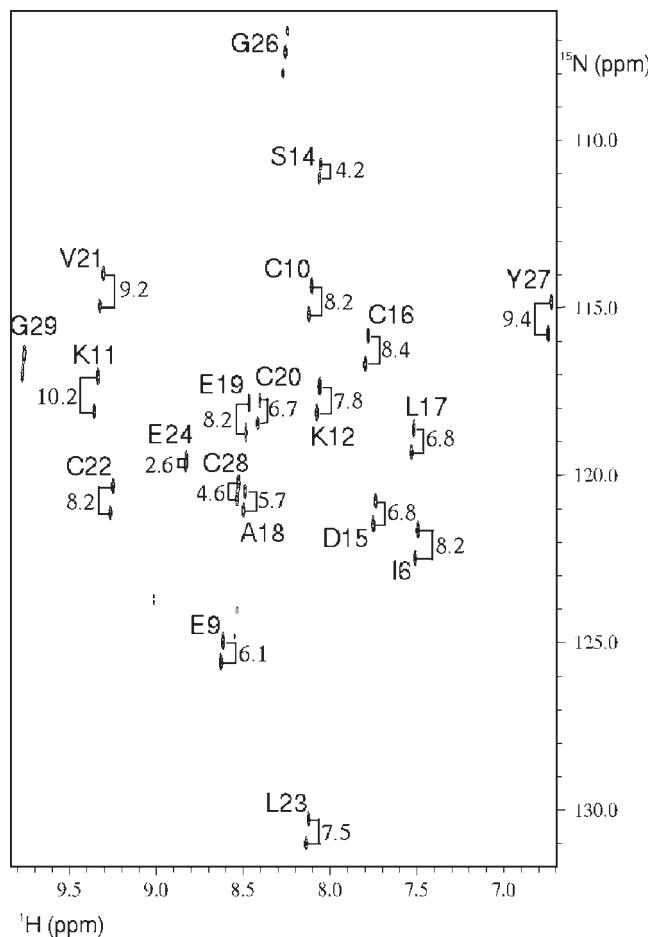


Fig. 4. PPJ-HMQC spectrum (19) measured for the ^{15}N -labeled CMTI-I(M8L) protein. Homonuclear $^3\text{J}(\text{H}_\text{NH}_\alpha)$ scalar couplings result in the splitting of correlation signals along vertical axis. Sequential assignments and scalar coupling values are given in spectrum.

5.5. Relative Orientations of Internuclear Vectors Evaluated From the Interference of Nuclear Relaxation Mechanisms

Two mechanisms dominating the relaxation of heteronuclei in proteins, namely dipolar (DD) and chemical shift anisotropy (CSA), can interfere with one another. The resulting cross-correlation terms, together with auto-correlation terms, contribute to the total nuclear magnetic relaxation. Recently, interference effects of different DD mechanisms, or DD and CSA mechanisms, or different CSA mechanisms in $^{15}\text{N}/^{13}\text{C}$ double-labeled proteins have been employed to determine angles between vectors characterizing cross-correlated mechanisms. In turn, these angles can be related to the backbone dihedral angles (31). For instance, determination of the backbone dihedral angle Ψ_i was obtained from the interference of $\text{DD}(\text{C}_\alpha\text{H}_\alpha)$ and $\text{DD}(\text{N}_{i+1}\text{H})$ or $\text{DD}(\text{C}_\alpha\text{H}_\alpha)$ and $\text{CSA}(\text{C}'_i)$ mechanisms. In general, as many as four dihedral angle values can correspond to a single interference relaxation rate. When more than one interference rate related to a given dihedral angle is available, the ambiguity can be reduced or removed.

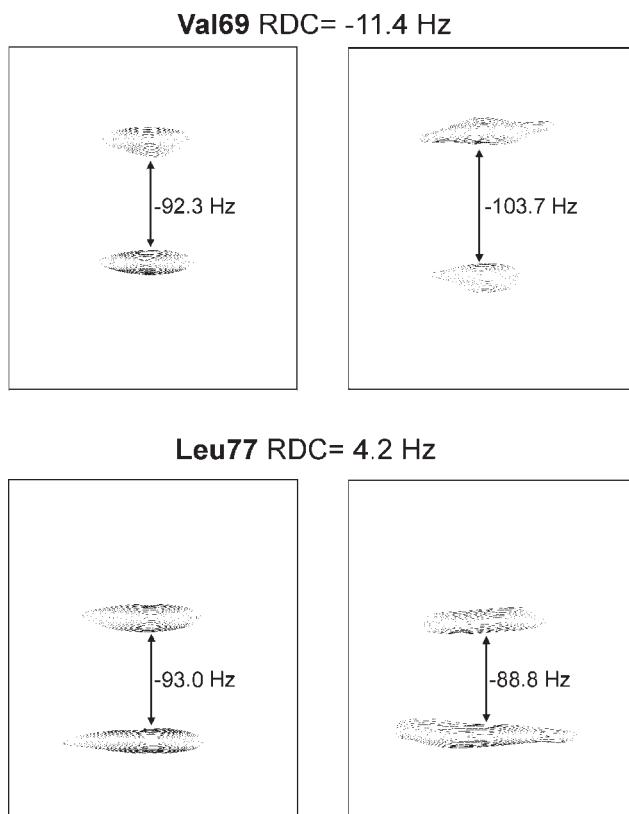


Fig. 5. Two examples of the determination of residual dipolar couplings in the $^{15}\text{N}/^{13}\text{C}$ -labeled S100A1 protein. Left-side figures display fragments of $^1\text{H}/^{15}\text{N}$ correlation spectrum measured in isotropic solution: vertical splittings in the ^{15}N dimension correspond to $^1\text{J}(\text{NH})$ scalar couplings. Right-side figures display the same fragments of spectrum obtained in the anisotropic bicelle solution. Residual dipolar couplings are calculated as the differences of two splittings.

6. Structure Calculations

A thorough discussion of computational methods used for the calculation of protein structure from NMR-derived constraints is outside of the scope of this review. The most successful method is based on restrained molecular dynamics simulations combined with the simulated annealing approach. A detailed analysis of those methods can be found in the dedicated publications (22,32–34).

References

1. Prestegard, J. H., Valafar, H., Glushka, J., and Tian, F. (2001) Nuclear magnetic resonance in the era of structural genomics. *Biochemistry* **40**, 8677–8685.
2. <http://www.rcsb.org/pdb/>
3. Wider, G. and Wüthrich, K. (1999) NMR spectroscopy of large molecules and multimolecular assemblies in solution. *Curr. Opin. Struct. Biol.* **9**, 594–601.
4. Lian, L.Y. and Middleton, D. A. (2001) Labelling approaches for protein structured studies by solution-state and solid-state NMR. *Prog. NMR Spectrosc.* **39**, 171–190.

5. Rajesh, S., Nietlispach, D., Nakayama, H., et al. (2003) A novel method for the biosynthesis of deuterated proteins with selective protonation at the aromatic rings of Phe, Tyr and Trp. *J. Biomol. NMR* **27**, 81–86.
6. Griesinger, C., Schwalbe, H., Schleucher, J., and Sattler, M. (1994) Proton-detected heteronuclear and multidimensional NMR. In: (Croasmun, W. R. and Carlson R. M. K., eds.) *Two-Dimensional NMR Spectroscopy*, VCH, New York, NY: 457–580.
7. Riek, R., Wider, G., Pervushin, K., and Wüthrich, K. (1999) Polarization transfer by cross-correlated relaxation in solution NMR with very large molecules. *Proc. Natl. Acad. Sci. USA* **96**, 4918–4923.
8. Pervushin, K., Riek, R., Wider, G., and Wüthrich, K. (1997) Attenuated T_2 relaxation by mutual cancellation of dipole-dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution. *Proc. Natl. Acad. Sci. USA* **94**, 12,366–12,371.
9. Clore, G. M. and Gronenborn, A. M. (1994) Structures of larger proteins, protein-ligand and protein-DNA complexes by multidimensional heteronuclear NMR. *Protein Sci.* **3**, 372–390.
10. Szyperski, T., Wider, G., Bushweller, J. H., and Wüthrich, K. (1993) Reduced dimensionality in triple-resonance NMR experiments. *J. Am. Chem. Soc.* **115**, 9307–9308.
11. Kozminski, W. and Zhukov I. (2003) Multiple quadrature detection in reduced dimensionality experiments. *J. Biomol. NMR* **26**, 157–166.
12. Bersch, B., Rossy, E., Covès, J., and Brutscher, B. (2003) Optimized set of two-dimensional experiments for fast sequential assignment, secondary structure determination, and backbone fold validation of $^{13}\text{C}/^{15}\text{N}$ -labelled proteins. *J. Biomol. NMR* **27**, 57–67.
13. Moseley, H. N. B. and Montelione, G. T. (1999) Automated analysis of NMR assignments and structures for proteins. *Curr. Opin. Struct. Biol.* **9**, 635–642.
14. Moseley, H. N. B., Monleon, D., and Montelione, G. T. (2001) Automated determination of protein backbone resonance assignments from triple resonance nuclear magnetic resonance data. In: (James, T. L., Dötsch, V., and Smitz, U., eds.) *Methods in Enzymology*, vol. 339 Academic, San Diego, CA: 91–108.
15. Mumenthaler, C., Güntert, P., Braun ,W., and Wüthrich, K. (1997) Automated combined assignment of NOESY spectra and three-dimensional protein structure determination. *J. Biomol. NMR* **10**, 351–362.
16. Linge, J. P., O'Donoghue, S. I., and Nilges, M. (2001) Automated assignment of ambiguous nuclear Overhauser effects with ARIA. In: (James, T. L., Dötsch, V., and Smitz, U., eds.) *Methods in Enzymology*, vol. 339 Academic, San Diego, CA: 71–90.
17. Herrmann, T., Güntert, P., and Wüthrich, K. (2002) Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. *J. Mol. Biol.* **319**, 209–227.
18. Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids*. Wiley, New York.
19. Cavanagh, J., Fairbrother, W. J., Palmer III, A. G., and Skelton, N. J. (1996) *Protein NMR Spectroscopy*. Academic Press, New York.
20. Wishart, D. S., Sykes, B. D., and Richards, F. M. (1992) The Chemical Shift Index: a fast and simple method for the assignment of protein secondary structure through NMR spectroscopy. *Biochemistry* **31**, 1647–1651.
21. Cornilescu, G., Delaglio, F., and Bax, A. (1999) Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J. Biomol. NMR* **13**, 289–302.
22. Dyson H. J. and Wright P. E. (1994) Protein structure calculation using NMR restraints. In: (Croasmun, W. R. and Carlson, R. M. K., eds.) *Two-Dimensional NMR Spectroscopy*, VCH, New York, NY: 655–698.

23. Grzesiek, S., Cordier, F., and Dingley, A. J. (2001) Scalar couplings across hydrogen bonds. In: (James, T. L., Dötsch, V., and Smitz, U., eds.) *Methods in Enzymology*, vol. 338 Academic, San Diego, CA: 111–133.
24. Kozminski, W. (1999) A pure-phase homonuclear J-modulated HMQC experiment with tilted cross-peak patterns for an accurate determination of homonuclear coupling constants. *J. Magn. Reson.* **141**, 185–190.
25. Bax, A. (2003) Weak alignment offers new NMR opportunities to study protein structure and dynamics. *Protein Sci.* **12**, 1–16.
26. Prestegard, J. H., Tolman, J. R., Al-Hashimi, H. M., and Andrec, M. (1999) Protein structure and dynamics from field-induced residual dipolar couplings. In: (Rama Krishna, N. and Berliner, L. J., eds.) *Biological Magnetic Resonance*, vol. 17 Kluwer/Plenum, New York, NY: 311–355.
27. Bax, A., Kontaxis, G., and Tjandra N. (2001) Dipolar couplings in macromolecular structure determination. In: (James, T. L., Dötsch, V., and Smitz, U., eds.) *Methods in Enzymology*, vol. 339 Academic, San Diego, CA: 127–174.
28. Brunner E. (2001) Residual dipolar couplings in protein NMR. *Concepts Magn. Reson.* **13**, 238–259.
29. de Alba, E. and Tjandra N. (2002) NMR dipolar coupling for the structure determination of biopolymers in solution. *Prog. NMR Spectrosc.* **40**, 175–197.
30. Chou, J. J., Li, S., Klee, C. B., and Bax, A. (2001) Solution structure of Ca^{2+} -calmodulin reveals flexible hand-like properties of its domains. *Nature Struct. Biol.* **8**, 990–997.
31. Schwalbe, H., Carlomagno, T., Hennig, M., et al. (2001) Cross-correlated relaxation for measurement of angles between tensorial interactions. In: (James, T. L., Dötsch, V., and Smitz, U., eds.) *Methods in Enzymology*, vol. 338 Academic, San Diego, CA: 35–81.
32. Clore, G. M. and Gronenborn, A. M. (1998) Determining structures of large proteins and protein complexes by NMR. In: (Krishna, N. R. and Berliner, L. J., eds.) *Biological Magnetic Resonance*, vol. 16 Kluwer Academic/Plenum, New York, NY: 3–26.
33. van Gunsteren, W. F., Bonvin, A. M. J. J., Daura, X., and Smith, L. J. (1999) Aspects of modelling biomolecular structure on the basis of spectroscopic or diffraction data. In: (Krishna, N. R. and Berliner, L. J., eds.) *Biological Magnetic Resonance*, vol. 17 Kluwer Academic/Plenum, New York, NY: 3–35.
34. O'Donoghue, S. I. and Nilges, M. (1999) Calculation of symmetric oligomer structures from NMR data. In: (Krishna, N. R. and Berliner, L. J., eds.) *Biological Magnetic Resonance*, vol. 17 Kluwer Academic/Plenum, New York, NY: 131–161.

Index

A

- AACompIdent tool, 585–588
AACompSim tool, 588
Adaptive evolution, 547–548
Adenylyl cyclase, 910
Agarose gels (IEF), 146–148
Alanine scanning energetics database, 780–782
Albumin—removal from serum, 50–52
Aldente tool, 590–592
Amidosulfobetaine 14 (ASB 14), 167, 174
Amino acid-coded mass tagging, 393–403
Ammoniacal silver nitrate stain, 173
Ampholytes, 32
ANTHERPROT, 486, 524
Antibody-affinity purification for interacting proteins, 683–687
Antibody–antigen interactions, 683–687, 699–705
Arabidopsis thaliana, 9, 334
ASB 14, *see* Amidosulfobetaine

B

- Bacteria, sample preparation for 2D-PAGE, 19–25
BIND, *see* Biomolecular interaction network database
Binding database (Binding DB), 770–774
Binding interface database, 782–783
BioGraph tool, 597–599
Biomolecular interaction network database (BIND), 763–767
Biotinylation, 746–747
BLAST, 529, 533–534, 555–565
Blood cells
 preparation of, 70–71
 sorting, 72
BRENDA, *see* Comprehensive enzyme information system
Brij, 3

C

- CATH database, 806, 835, 911–912, 914
Cell membrane microarrays, 729–730
Cell signaling networks database, 756
Cell sorting, 72–77
Cells (mammalian)
 differential detergent fractionation of, 42
 preparation of for 2D-PAGE, 7, 33–34, 154
 purification by FACS, 67–75
Chaotropes, 2
CHAPS, 3, 24, 28, 167
Chemically assisted fragmentation, 325–331
Chlamydia pneumoniae, 20–21
Cleland’s reagent, *see* Dithiothreitol
CluSTr analysis, 624
Clustal programs, 493–501
COILS, 529
Colloidal Coomassie blue, 161–162
Co-localization method for protein interactions, 633–634, 638–639
Comprehensive enzyme information system (BRENDA), 774–775
Comprehensive yeast genomic database, 756, 774–775
Computer pI/Mw tool, 575–576
ConA, *see* Concanavalin A
Concanavalin A, 197
Coomassie fluor orange, 219–220
Crystallization methods, 955–965
Crystallography, 939–950, 959–965
Cyanine dyes, 5–6, 62, 224, 228, 230–233

D

- Database of interacting domains, 757
Database of interacting proteins, 760–763
DeCyder software, 232–233
Deglycosylation, 195, 345–348
Delta 2D software, 199–200
Deltamass, 436
Deoxycholate (DOC), 45

- Dephosphorylation, 206, 464
- Detergents, 3, 24, 28, 37–43, 45, 167, 174
- Differential detergent fractionation, 37–46
- Difference in-gel electrophoresis, 5–6, 109–111, 223–236
- DIGE, *see* Difference in-gel electrophoresis
- Digitonin, 42, 45
- DIP, *see* Database of interacting proteins
- Disulfide linkages, 433
- Dithioerythritol (DTE), 31
- Dithiothreitol (DTT), 4–5, 23
- DNase treatment, 4, 23
- DOC, *see* Deoxycholate
- Domain identification, 527, 903–905
- 2D-PAGE, *see* Two dimensional PAGE
- Drosophila* protein–protein interaction map, 756
- DTE, *see* Dithioerythritol
- DTT, *see* Dithiothreitol
- Dynamic range of proteins, 2
- E**
- Electrospray MS, 313, 315–318, 379–381, 445–447, 463–464
- ENDscript, 486
- Entrez Proteins database, 610
- Ettan CAF MALDI sequencing kit, 325
- ExPASy server, 571–605
- Extraction of proteins, 1–6
- Evolutionary genomics, 543–552, 815–816
- F**
- FACS, *see* Fluorescence-activated cell sorting
- FASTA, 503–524, 565
- FFF, *see* Fuzzy Functional Folds
- FindMod tool, 436, 592–595
- FindPept tool, 437, 595–597
- Flicker, 281–303
- Fluorescence-activated cell sorting, 67–76
- Folds, classification of, 903–915
- Fold
- similarity, 810, 822, 823
 - assignment, 836–838, 906–908
 - recognition, 929–931
- FSSP database, 912, 914
- Function prediction, 629–648, 801–823, 908
- Fuzzy Functional Folds, 813
- G**
- Gel staining, *see* Stains
- Gene duplication, 802
- Gene fusion prediction of protein interactions, 641–642, 802–803
- Gene neighbor-based prediction of protein interactions, 638–639
- GenPept database, 610
- Glucosyltransferase, 909
- GlycanMass tool, 436, 595
- GlycoMod tool, 436, 595
- Glycoprotein
- analysis, 389–391, 432–433, 439–454
 - detection with lectins, 197
 - deglycosylation, 195, 345–348
 - sequencing, 444–446, 448–450
 - staining, 196–197
- GlycoSuiteDB, 346
- GO Slim analysis, 624–625
- Gradient gels, 126
- Grand average hydrophobicity values (GRAVY) 140
- GRAVY, *see* Grand average hydrophobicity values
- GW FASTA, 511–512, 521–523
- H**
- Hepatocyte sorting, 72
- Hidden Markov model, 532, 536, 538
- High MW proteins (2D-PAGE), 145–149
- Homoarginine, 408
- Human protein reference database (HPRD), 756, 775
- Human unidentified gene encoded protein–protein interaction database, 775–778
- Hydra-plus-one system, 957–965
- HydroGel slides, 469, 713–718
- Hydrophobically modified pullans (HMCMPs), 3
- Hydrophobicity scores, 140
- Hydrophobic lipid balance, 3
- Hydroxyethyl disulfide, 157, 161

I

- ICAT, *see* Isotope-coded affinity tags
IEF, *see* Isoelectric focusing
Igepal CA-630, 31
IMAC, *see* Metal affinity chromatography
Image analysis software
 DeCyder, 232–233
 Delta2D, 197–199
 Flicker, 281–303
 Melanie, 267–277
 Z3, 197–199
Image Master 2D Platinum software, *see*
 Melanie
Immobilized pH gradients, 101–104,
 122–123, 154–160, 171–172
Immunoglobulins, removal from serum, 2,
 49–52
In-gel digestion, 211, 312
InterPro database, 478, 619–626
IonIQ, 349
IPG-Dalt, *see* Immobilized pH gradients
Isoelectric focusing, 8, 371, 418
 in agarose gels, 146–148
 in 2D gels, 101–104, 122–123, 138–139,
 154–160, 171–172
 prefractionation by, 97–115
Isotope-coded affinity tags (ICAT),
 385–392
Isotope labeling, 393–404, 970–971

J

- JPRED, 529

K

- Kinases, 206–207, 470
Kinase pathway domain database, 757

L

- Large format 2D gels, 119–130
Laser assisted microdissection, 6, 59–66
Laser capture microscopy, *see* Laser
 assisted microdissection
LC-MS, 375–383, 411–412
LC-MS/MS, 73–75, 111–112, 367–372
LC/LC-MS/MS, 111–112
Lectins, 197

Liver cells - preparation of, 71

Low MW proteins

 2D-PAGE, 140

 nanospray MS, 313

Lysine, modification of, 408

M

- MALDI-PSD, 325–339
MALDI-TOF MS, 319–324, 328, 348–349,
 398–401, 417–427
Maltoside *n*-dodecyl- β -D-maltoside, 24
Mammalian tissue
 preparation for 2D-PAGE, 31–34
 preparation for IEF, 369–370
MASCOT, 349, 356–359
Mass-coded abundance tagging, 407–414
Mass spectrometry
 compatible silver stains, 181–182, 189
 MALDI-TOF, 319–324, 328, 348–349,
 398–401, 417–427
 nanospray, 313, 315–318, 379–381,
 445–447, 463–464
 of glycoproteins, 432–433
 on-plate washing, 321–322
 peptide mass fingerprinting, 355–363
 phosphorylation sites, 459–465
 post source decay MALDI, 322, 325–
 229, 444–445
 posttranslational modifications, 431–436,
 459–465, 574
 quantitative, 393–403, 407–414
 sample clean-up, 307–309
 SELDI-TOF, 699–707
MCAT, *see* Mass-coded abundance tagging
Melanie software, 267–277
Membrane proteins
 2D-PAGE of, 133–143, 167–174
 hydrophobicity values, 140
 preparation of, 136
 solubilization of, 3–4, 136–137
Metal affinity chromatography, 462
Methylisourea, 408
MHC peptide interaction database, 790–794
Minnesota Biodegradation and Biocatalyst
 database, 787
MINT, *see* Molecular interactions database

- Microarrays, 679–704, 709–720, 723–731, 735–741, 743–751
 analysis of, 718–719
 antibody based, 699–704, 709–720, 727–729
 cell membrane arrays, 729–730
 HydroGel slides, 469, 713–718
in situ, 735–741
 phosphorylation detection, 467–473
 robotic pin printing, 723–730
 staining for phosphoproteins, 408, 470–472
- Mitochondria sorting, 72
- Modeling, 831–849
- Molecular interactions database (MINT), 767–770
- Molecular weight determination, 319–324
- MudPIT, 379–380, 411–412
- MultiAlin, 529
- Multident tool, 588–590
- Multiple sequence alignment, 475–487, 493–501, 503–524, 534–537, 838–839
- Myelin proteins, 170–171
- N**
- Nanoelectrospray, *see* Nanospray
- Nanospray MS, 313, 315–318, 379–381, 445–447, 463–464
- NCBI BLAST, *see* BLAST
- NEpHGE, *see* Nonequilibrium pH gradient electrophoresis
- NetOGlyc, 529
- NMR, *see* Nuclear magnetic resonance
- Nonequilibrium pH gradient electrophoresis, 167–174
- Nonidet NP40, 24
- Nuclear magnetic resonance, 967–980
- Nucleic acid-protein interactions database, 785
- Nucleic acids, removal from 2-D samples, 4, 23
- Nucleoli, 79–84
- O**
- On-membrane digestion, 348
- On-plate washing (MS), 321–322
- Organism-specific databases, 616
- P**
- PCMA, 529
- PCR-directed *in situ* microarrays, 735–741
- PDB, *see* Protein data bank
- Peptide Cutter, 579–581
- PeptideMass tool, 437, 578–579
- Peptide
- analysis by LC-MS/MS, 75
- extraction, 312
- mass fingerprinting, 348–349, 355–363
- microarrays, 714
- production, 312
- sequence determination (MS) 315–316, 325–359, 336, 412–413
- Pharmalyte, 24
- PHD, 529
- Phenylmethylsulphonyl fluoride, 24, 41
- Phosphopeptides
- analysis, 459–464
- sequencing, 336
- Phosphoproteins
- analysis by MS, 207, 459–464
- dephosphorylation, 206
- staining, 203, 467–473
- Phosphorylation (in vitro), 206–207, 470
- Phosphorylation—detection of, 459–464, 467–473
- Phosphoserine—modification of, 463
- Phylogenetic analysis, 518–519
- Phylogenetic profile-based prediction of protein interactions, 634, 639–641
- Piezoelectric arraying, 344, 713–714
- PINTS, 813
- PipeAlign, 486
- PIR-PSD database, 611–612
- Plant tissue
- protein extraction, 5
- removal of phenolics, 4
- sample preparation for 2D-PAGE, 55–57
- Plasma (*see also* Serum)
- electrophoresis of, 344–345
- sample preparation for 2D PAGE, 49–54
- proteins, 350
- PNGase F, 195, 345–348
- Polyphenols, 4
- Polyvinylpyrrolidone, 4
- Post-source-decay MALDI, 322, 325–339, 444–445

- Posttranslational modifications, 431–436, 459–465, 574
- Prefractionation methods, 87–94, 97–115
- Prions, 881–883
- ProFound, 349
- PROFphd, 875–893
- Pro-Q Diamond, 203, 467–473
- Pro-Q Emerald, 196
- Protease inhibitors, 24, 41
- Protein A agarose, 685
- Protein blotting, *see* Western blotting
- Protein data bank (PDB), 778
- Protein family databases, 616
- Protein G, 51–52
- Protein-ligand databases, 785–787
- Protein prospector, 356–361
- Protein–protein interactions
- antibody affinity purification method, 683–687
 - by surface plasmon resonance, 689–697
 - by yeast two-hybrid system, 636–637, 653–678
 - computer prediction of, 629–645
 - databases, 632, 645–648, 753–796
- Protein
- classification of sequences and structure, 861–870
 - folds, 903–916
 - function prediction, 629–648, 801–823, 908
 - sequence analysis, 527–539
 - sequence databases, 609–617
 - signature databases, 620–621
 - solubilization (for gels), 1–12
 - structure prediction, 831–849, 875–903, 921–933
 - threading, 921–933
- ProtParam tool, 576–578
- ProtScale, 581–582
- PROWL, 365
- PSI-BLAST, 564–565
- PSORT, 529
- PVP, *see* Polyvinylpyrrolidone
- Q**
- Quantitation
- by MS, 393–403
 - using Coomassie blue, 229
- R**
- Reducing agents, 4–5, 23, 157, 161
- Reflection TOF, 22
- RefSeq database, 610
- Reversed phase HPLC for sample prefractionation, 87–94
- RNAse treatment, 4, 23
- Robotic pin printing, 723–730
- RP-HPLC, *see* Reversed phase HPLC
- S**
- Saccharomyces cerevisiae*, (*see also* Yeast) 27–29, 641
- SAM, 529
- Sample clean-up for MS, 307–309
- SCOP database, 806, 910–911, 914
- Secondary structure, 903–915
- SELDI, 699–705
- Selected ion monitoring (SIM), 434
- Sequence databases, 572–575, 609–617
- Sequence determination
- MALDI-PSD, 325–339
 - peptide (nanospray), 315–318, 412–413
- Sequences
- classification of, 865–870
- Sequence similarity searching, 475–487, 493–501, 503–524, 527–538
- SEQUEST, 361–363
- Serum (*see also* Plasma)
- albumin depletion, 50–52
 - immunoglobulin depletion, 2, 49–52
 - preparation for 2D PAGE, 7, 49–54
- SignalP, 529
- Silver staining, 161–162, 173, 177–183
- SIM, *see* Selected ion monitoring
- SMD, *see* Stanford microarray database
- Solubilization of proteins, 1–12
- SPR, *see* Surface plasmon resonance
- SSEARCH, 565
- Stains
- Colloidal Coomassie blue, 161–162
 - Coomassie Fluor orange, 219–220
 - multiple, 193–200
- Pro-Q Diamond, 203, 467–473
- Pro-Q Emerald, 196
- silver, 161–162, 173, 177–183
- SYPRO orange, 220–221
- SYPRO red, 220–221

- SYPRO ruby, 196–197, 204–205, 209–214, 234–235
 SYPRO tangerine, 221–222
 zinc reverse, 185–190
 Stanford microarray database (SMD), 642
 Statistical analysis of 2D gel patterns, 239–257
 Structural motifs, 811
 Structural/mutational databases, 778
 Structure–function relationships, 629–648, 801–823
 Structure prediction, 831–849, 875–903, 921–933
 Structures
 classification of, 865–870
 Sulphobetaine, 24
 Superfamilies, 907–908
 Superfolds, 810, 908
 Surface plasmon resonance (SPR), 689–697
 SWISS-2D PAGE, 260–264
 Swiss-Prot database, 572–575, 612–614
 SYPRO orange, 220–221
 SYPRO red, 220–221
 SYPRO ruby, 162, 196–197, 204–205, 209–214, 234–235
 SYPRO tangerine, 221–222
- T**
 TagIdent tool, 582–585
 TCA precipitation, 22, 687
 T-coffee, 529
 TESS, 813
 TFE buffer, 10
 Thiourea, 23, 28
 Threading, 921–933
 TMAP, 529
 TMHMM, 529
 TMPred, 529
 TrEMBL, 572–574, 614–615
 Tributyl phosphine, 24, 31
 Trifluorethanol, *see* TFE buffer
 Triton X-100, 24, 28, 45
 TRIZOL, 370
 Trypan blue, 80
 Trypsin, 363, 411
 in gel digestion with, 211, 312
 Two-dimensional PAGE, 62–64, 119–130, 137–140, 145–149
 agarose, 146–148
 comparison of gels, 279–303
 databases, 259–264
 high molecular mass proteins, 145–149
 image analysis, 197–199, 267–277, 281–303
 low molecular mass proteins, 140
 membrane proteins, 133–143
 prefractionation of samples, 2, 49–52, 87–94, 97–115
 sample preparation, 1–12, 19–25, 27–29, 31–34, 49–54, 55–57, 79–84
 statistical analysis, 239–257
 ultrazoom gels, 151–163
- U**
 Ultra-zoom gels, 151–163
 Urea, 23
 Urine, 8
- V**
 VAST, 913, 915
 Virtual 2D gel electrophoresis, 417–427
- W**
 Wavelets, 245
 Western blotting, 344–355
 Wise2, 529
 WU BLAST, 529, 555
- X**
 X-ray crystallography, 939–950, 955–965
- Y**
 Yeast 2-hybrid system, 653–678
 in silico, 636–637
 Yeast interacting protein database (YIPD), 756
 Yeast—preparation for 2D-PAGE, 27–29
 YIPD, *see* Yeast interacting protein database
- Z**
 Z3 software, 197–199
 Zinc reverse staining, 185–190
 ZipTip, 307–309
 Zoom IEF fractionator, 97–115