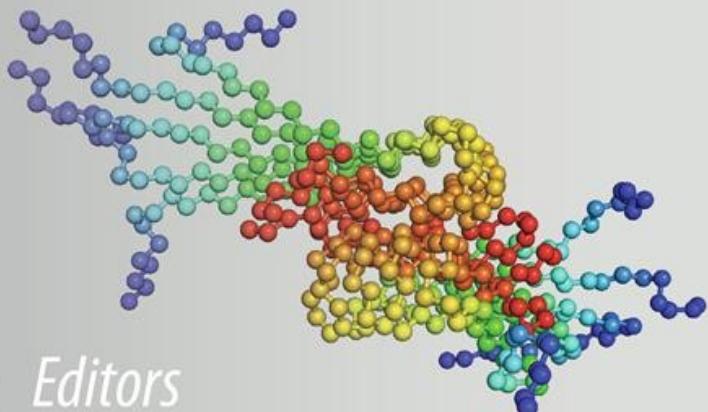


Vladimir Voynov
Justin A. Caravella *Editors*



Therapeutic Proteins

Methods and Protocols

Second Edition

METHODS IN MOLECULAR BIOLOGY™

Series Editor
John M. Walker
School of Life Sciences
University of Hertfordshire
Hatfield, Hertfordshire, AL10 9AB, UK

For further volumes:
<http://www.springer.com/series/7651>

Therapeutic Proteins

Methods and Protocols

Second Edition

Edited by

Vladimir Voynov

MedImmune, LLC, Gaithersburg, Maryland, USA

Justin A. Caravella

*Department of Physical Biochemistry, Biogen Idec
Cambridge, Massachusetts, USA*



Editors

Vladimir Voynov
MedImmune, LLC, One MedImmune Way
Gaithersburg, Maryland
USA

Justin A. Caravella
Department of Physical Biochemistry
Biogen Idec, 14 Cambridge Center
Cambridge, Massachusetts
USA

ISSN 1064-3745 ISSN 1940-6029 (electronic)
ISBN 978-1-61779-920-4 ISBN 978-1-61779-921-1 (eBook)
DOI 10.1007/978-1-61779-921-1
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2012940190

© Springer Science+Business Media, LLC 2012

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Humana Press is a brand of Springer
Springer is part of Springer Science+Business Media (www.springer.com)

Preface

The first proteins developed as therapeutic agents were purified naturally occurring proteins or recombinant versions thereof. In recent years, the field of therapeutic proteins has expanded, and more effort has been put into increasingly sophisticated protein design and engineering efforts. However, design, discovery, and engineering of a new protein are only the first steps towards the development of a therapeutic molecule. Once a potential protein drug has been identified, the next critical step is the production of sufficient authentic material for testing, characterization, and clinical trials. If a candidate protein drug makes it through this lengthy and costly process, methodology that allows the production of the protein on a scale large enough to meet demand must be implemented. It is also necessary to have robust methods for the purification, characterization, viral inactivation, and continued testing of the final protein product.

The aim of this second edition of *Therapeutic Proteins: Methods and Protocols* is to cover each of these key aspects of protein drug production. As in the first edition, attention is given to production, purification, and characterization of protein therapeutics. This second edition emphasizes new developments since the publication of the first edition in 2005. This second edition also includes additional emphasis on discovery, including new display and screening methods as well as the design and engineering of new types of therapeutic proteins. There is also discussion of computational and bioinformatics methods, and chapters on safety aspects of therapeutic protein development. All contributing authors are based at highly esteemed industrial and academic institutions from around the world.

This book contains complete protocols set out in a simple step-by-step manner. It opens with an introductory chapter that reviews the history of the field and contains thoughts on the direction of future developments. A number of other chapters provide an overview of a key area of therapeutic protein development. Most chapters are experimental protocols that contain a useful introduction describing the theory and background to a particular method, which is then followed by a list of all equipment and materials required to complete the protocol. The Methods section describes every step of the protocol and is cross-referenced to a Notes section that describes possible difficulties or problems that may arise, alternative methods, and invaluable hints.

A large number of people have helped organize this book so that it ultimately provides a very useful resource to all those working in the field of therapeutic proteins. We would especially like to thank all contributors for their excellent chapters based on expert knowledge and experience. We are also grateful to John Walker, the series editor, for asking us to edit this book, and for his help and advice in preparing the final product. Thanks also to David Casey and his colleagues at Springer, Humana Press, who have helped put this together.

*Gaithersburg, MD, USA
Cambridge, MA, USA*

*Vladimir Voynov
Justin A. Caravella*

Contents

Preface	v
Contributors	ix
1 Therapeutic Proteins	1
<i>Dimitar S. Dimitrov</i>	
2 Synthetic Antibody Libraries	27
<i>Bryce Nelson and Sachdev S. Sidhu</i>	
3 The Construction of “Phylomer” Peptide Libraries as a Rich Source of Potent Inhibitors of Protein/Protein Interactions	43
<i>Nadia Milech and Paul Watt</i>	
4 Ribosome Display and Screening for Protein Therapeutics	61
<i>Damjana Kastelic and Mingyue He</i>	
5 Yeast Display of Engineered Antibody Domains	73
<i>Qi Zhao, Zhongyu Zhu, and Dimitar S. Dimitrov</i>	
6 Expression, Purification, and Characterization of Engineered Antibody CH2 and VH Domains	85
<i>Rui Gong, Weizao Chen, and Dimitar S. Dimitrov</i>	
7 Engineering of Affibody Molecules for Therapy and Diagnostics	103
<i>Joachim Feldwisch and Vladimir Tolmachev</i>	
8 Protein Design for Diversity of Sequences and Conformations Using Dead-End Elimination	127
<i>Karl J.M. Hanf</i>	
9 Design and Generation of DVD-Ig TM Molecules for Dual-Specific Targeting	145
<i>Enrico DiGiammarino, Tariq Ghayur, and Junjian Liu</i>	
10 Engineering and Expression of Bibody and Tribody Constructs in Mammalian Cells and in the Yeast <i>Pichia pastoris</i>	157
<i>Steve Schoonooghe</i>	
11 Use of <i>E. coli</i> for the Production of a Single Protein	177
<i>Lili Mao and Masayori Inouye</i>	
12 Folding Engineering Strategies for Efficient Membrane Protein Production in <i>E. coli</i>	187
<i>Brent L. Nannenga and François Baneyx</i>	
13 Transient Expression Technologies: Past, Present, and Future	203
<i>Sabine Geisse and Bernd Voedisch</i>	
14 Stable Transfection Pools for Large Quantity of Protein Production	221
<i>Jianxin Ye</i>	
15 Mammalian Stable Expression of Biotherapeutics	227
<i>Thomas Jostock and Hans-Peter Knopf</i>	

16	Transgenic Expression of Therapeutic Proteins in <i>Arabidopsis thaliana</i> Seed	239
	<i>Cory L. Nykiforuk and Joseph G. Boothe</i>	
17	Methods for Chromatographic Removal of Endotoxin	265
	<i>Adam J. Lowe, Cameron L. Bardliving, and Carl A. Batt</i>	
18	Effectiveness of Various Processing Steps for Viral Clearance of Therapeutic Proteins: Database Analyses of Commonly Used Steps	277
	<i>Dana Cipriano, Michael Burnham, and Joseph V. Hughes</i>	
19	High-Throughput Quantitative N-Glycan Analysis of Glycoproteins	293
	<i>Margaret Doherty, Ciara A. McManus, Rebecca Duke, and Pauline M. Rudd</i>	
20	High-Throughput Multimodal Strong Anion Exchange Purification and N-Glycan Characterization of Endogenous Glycoprotein Expressed in Glycoengineered <i>Pichia pastoris</i>	315
	<i>Sujatha Gomathinayagam, Erik Hoyt, Alissa M. Thompson, Eric Brown, Khanita Karaveg, Stephen R. Hamilton, and Huijuan Li</i>	
21	Databases and Tools in Glycobiology	325
	<i>Natalia V. Artemenko, Andrew G. McDonald, Gavin P. Davey, and Pauline M. Rudd</i>	
22	Characterization of PEGylated Biopharmaceutical Products by LC/MS and LC/MS/MS	351
	<i>Lihua Huang and P. Clayton Gough</i>	
23	Identification of Asp Isomerization in Proteins by ¹⁸ O Labeling and Tandem Mass Spectrometry	365
	<i>Jennifer Zhang and Viswanatham Katta</i>	
24	Monitoring of Subvisible Particles in Therapeutic Proteins	379
	<i>Satish K. Singh and Maria R. Toler</i>	
25	Size-Exclusion Chromatography with Multi-angle Light Scattering for Elucidating Protein Aggregation Mechanisms	403
	<i>Erinc Sahin and Christopher J. Roberts</i>	
26	Computational Methods to Predict Therapeutic Protein Aggregation	425
	<i>Patrick M. Buck, Sandeep Kumar, Xiaoling Wang, Neeraj J. Agrawal, Bernhardt L. Trout, and Satish K. Singh</i>	
27	Coarse-Grained Simulations of Protein Aggregation	453
	<i>Troy Cellmer and Nicolas L. Fawzi</i>	
28	Chitosan-Based Nanoparticles as Delivery Systems of Therapeutic Proteins	471
	<i>Pedro Fonte, José Carlos Andrade, Vítor Seabra, and Bruno Sarmento</i>	
29	Challenges in the Development and Manufacturing of Antibody-Drug Conjugates.	489
	<i>Laurent Ducry</i>	
	<i>Index</i>	499

Contributors

- NEERAJ J. AGRAWAL • *Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA*
- JOSÉ CARLOS ANDRADE • *Department of Pharmaceutical Sciences, CICS, Health Sciences Research Center, Instituto Superior de Ciências da Saúde, Gandra, Portugal*
- NATALIA V. ARTEMENKO • *NIBRT Glycobiology Laboratory, The National Institute for Bioprocessing Research and Training, Dublin, Ireland*
- FRANÇOIS BANEYX • *Department of Chemical Engineering, University of Washington, Seattle, WA, USA*
- CAMERON L. BARDLIVING • *Biomedical Engineering, Cornell University, Ithaca, NY, USA*
- CARL A. BATT • *Department of Food Science, Cornell University, Ithaca, NY, USA*
- JOSEPH G. BOOTHE • *SemiBioSys Genetics Inc, Calgary, AB, Canada*
- ERIC BROWN • *GlycoFi Inc., A wholly owned subsidiary of Merck & Co Inc, Lebanon, NH, USA*
- PATRICK M. BUCK • *Biotherapeutics Pharmaceutical Research and Development, Pfizer, Inc, St. Louis, MO, USA*
- MICHAEL BURNHAM • *WuXi AppTec, Inc, Philadelphia, PA, USA*
- TROY CELLMER • *Laboratory of Chemical Physics, National Institute of Digestive and Diabetes and Kidney Diseases, National Institutes of Health, Bethesda, MD, USA*
- WEIZAO CHEN • *Protein Interactions Group, Frederick National Laboratory for Cancer Research, National Cancer Institute, National Institutes of Health, Frederick, MD, USA*
- DANA CIPRIANO • *WuXi AppTec, Inc, Philadelphia, PA, USA*
- GAVIN P. DAVEY • *School of Biochemistry and Immunology, Trinity College Dublin, Dublin, Ireland*
- ENRICO DiGAMMARINO • *Abbott Laboratories, Abbott Bioresearch Center, Worcester, MA, USA*
- DIMITER S. DIMITROV • *Protein Interactions Group, Frederick National Laboratory for Cancer Research, National Cancer Institute, National Institutes of Health, Frederick, MD, USA*
- MARGARET DOHERTY • *NIBRT Glycobiology Laboratory, The National Institute for Bioprocessing Research and Training, Dublin, Ireland*
- LAURENT DUCRY • *Lonza Ltd, Visp, Switzerland*
- REBECCA DUKE • *NIBRT Glycobiology Laboratory, The National Institute for Bioprocessing Research and Training, Dublin, Ireland*
- NICOLAS L. FAWZI • *Laboratory of Chemical Physics, National Institute of Digestive and Diabetes and Kidney Diseases, National Institutes of Health, Bethesda, MD, USA*
- JOACHIM FELDWISCH • *Affibody AB, Solna, Sweden*
- PEDRO FONTE • *Department of Pharmaceutical Sciences, CICS, Health Sciences Research Center, Instituto Superior de Ciências da Saúde, Gandra, Portugal*
- SABINE GEISSE • *Novartis Institutes for BioMedical Research, Basel, Switzerland*
- TARIQ GHAYUR • *Abbott Bioresearch Center, Worcester, MA, USA*
- SUJATHA GOMATHINAYAGAM • *GlycoFi Inc., A wholly owned subsidiary of Merck & Co Inc, Lebanon, NH, USA*

- RUI GONG • *Protein Interactions Group, Frederick National Laboratory for Cancer Research, National Cancer Institute, National Institutes of Health, Frederick, MD, USA*
- P. CLAYTON GOUGH • *Bioproduct Research & Development, Lilly Research Laboratories, Lilly Corporate Center, Eli Lilly and Company, Indianapolis, IN, USA*
- STEPHEN R. HAMILTON • *GlycoFi Inc., A wholly owned subsidiary of Merck & Co Inc, Lebanon, NH, USA*
- KARL J.M. HANF • *Department of Physical Biochemistry, Biogen Idec, Cambridge, MA, USA*
- MINGYUE HE • *The Inositide Laboratory, Babraham Institute, Cambridge, UK*
- ERIK HOYT • *GlycoFi Inc., A wholly owned subsidiary of Merck & Co Inc, Lebanon, NH, USA*
- LIHUA HUANG • *Bioproduct Research & Development, Lilly Research Laboratories, Lilly Corporate Center, Eli Lilly and Company, Indianapolis, IN, USA*
- JOSEPH V. HUGHES • *WuXi AppTec, Inc, Philadelphia, PA, USA*
- MASAYORI INOUYE • *Department of Biochemistry, Center for Advanced Biotechnology and Medicine, Robert Wood Johnson Medical School, Piscataway, NJ, USA*
- THOMAS JOSTOCK • *Novartis Pharma AG, Basel, Switzerland*
- KHANITA KARAVEG • *GlycoFi Inc., A wholly owned subsidiary of Merck & Co Inc, Lebanon, NH, USA*
- DAMJANA KASTELIC • *University of Ljubljana, Ljubljana, Slovenia*
- VISWANATHAM KATTA • *Protein Analytical Chemistry, Genentech Inc, South San Francisco, CA, USA*
- HANS-PETER KNOPF • *Novartis Pharma AG, Basel, Switzerland*
- SANDEEP KUMAR • *Biotherapeutics Pharmaceutical Research and Development, Pfizer, Inc, St. Louis, MO, USA*
- HUIJUAN LI • *GlycoFi Inc., A wholly owned subsidiary of Merck & Co Inc, Lebanon, NH, USA*
- JUNJIAN LIU • *Abbott Laboratories, Abbott Bioresearch Center, Worcester, MA, USA*
- ADAM J. LOWE • *Graduate Field of Microbiology, Cornell University, Ithaca, NY, USA*
- SRC Inc, Syracuse, NY, USA
- LILI MAO • *Department of Biochemistry, Center for Advanced Biotechnology and Medicine, Robert Wood Johnson Medical School, Piscataway, NJ, USA*
- ANDREW G. McDONALD • *School of Biochemistry and Immunology, Trinity College Dublin, Dublin, Ireland*
- CIARA A. McMANUS • *NIBRT Glycobiology Laboratory, The National Institute for Bioprocessing Research and Training, Dublin, Ireland*
- NADIA MILECH • *Telethon Institute for Child Health Research and Centre for Child Health, University of Western Australia, Subiaco, WA, Australia; Phylogica Limited, Subiaco, WA, Australia*
- BRENT L. NANNENGA • *Department of Chemical Engineering, University of Washington, Seattle, WA, USA*
- BRYCE NELSON • *Department of Molecular Genetics, Banting and Best Department of Medical Research, Terrence Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, ON, Canada*
- CORY L. NYKIFORUK • *SemBioSys Genetics Inc, Calgary, AB, Canada*
- CHRISTOPHER J. ROBERTS • *Department of Chemical Engineering, University of Delaware, Newark, DE, USA*
- PAULINE M. RUDD • *NIBRT Glycobiology Laboratory, The National Institute for Bioprocessing Research and Training, Dublin, Ireland*

ERINC SAHIN • *Drug Product Science & Technology R&D, Bristol-Myers Squibb, New Brunswick, NJ, USA*

BRUNO SARMENTO • *Department of Pharmaceutical Sciences, CICS, Health Sciences Research Center, Instituto Superior de Ciências da Saúde, Gandra, Portugal; Department of Pharmaceutical Technology, University of Porto, Porto, Portugal; INEB - Instituto de Engenharia Biomédica Organization, University of Porto, Porto, Portugal*

STEVE SCHOONOOGHE • *Cellular and Molecular Immunology Lab, Vrije Universiteit Brussel, Brussel, Belgium*

VÍTOR SEABRA • *Department of Pharmaceutical Sciences, CICS, Health Sciences Research Center, Instituto Superior de Ciências da Saúde, Gandra, Portugal*

SACHDEV S. SIDHU • *Banting and Best Department of Medical Research, Department of Molecular Genetics, Terrence Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, ON, Canada*

SATISH K. SINGH • *Biotherapeutics Pharmaceutical Sciences, Pfizer, Inc, Chesterfield, MO, USA*

ALISSA M. THOMPSON • *GlycoFi Inc., A wholly owned subsidiary of Merck & Co Inc, Lebanon, NH, USA*

MARIA R. TOLER • *Biotherapeutics Pharmaceutical Sciences, Pfizer, Inc, Chesterfield, MO, USA*

VLADIMIR TOLMACHEV • *Rudbeck Laboratory, Division of Biomedical Radiation Sciences, Department of Radiology, Oncology, and Clinical Immunology, Uppsala University, Uppsala, Sweden*

BERNHARDT L. TROUT • *Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA*

BERND VOEDISCH • *Novartis Institutes for BioMedical Research, Basel, Switzerland*

XIAOLING WANG • *Biotherapeutics Pharmaceutical Research and Development, Pfizer, Inc, St. Louis, MO, USA*

PAUL WATT • *Phylogica Limited, Subiaco, WA, Australia*

JIANXIN YE • *Bioprocess Research & Development, Merck & Co., Inc, Rahway, NJ, USA*

JENNIFER ZHANG • *Protein Analytical Chemistry, Genentech Inc, South San Francisco, CA, USA*

QI ZHAO • *Protein Interactions Group, Center for Cancer Research Nanobiology Program, National Cancer Institute, National Institutes of Health, Frederick, MD, USA*

ZHONGYU ZHU • *Protein Interactions Group, Center for Cancer Research Nanobiology Program, National Cancer Institute, National Institutes of Health, Frederick, MD, USA*

Chapter 1

Therapeutic Proteins

Dimitar S. Dimitrov

Abstract

Protein-based therapeutics are highly successful in clinic and currently enjoy unprecedented recognition of their potential. More than 100 genuine and similar number of modified therapeutic proteins are approved for clinical use in the European Union and the USA with 2010 sales of US\$108 bln; monoclonal antibodies (mAbs) accounted for almost half (48%) of the sales. Based on their pharmacological activity, they can be divided into five groups: (a) replacing a protein that is deficient or abnormal; (b) augmenting an existing pathway; (c) providing a novel function or activity; (d) interfering with a molecule or organism; and (e) delivering other compounds or proteins, such as a radionuclide, cytotoxic drug, or effector proteins. Therapeutic proteins can also be grouped based on their molecular types that include antibody-based drugs, Fc fusion proteins, anticoagulants, blood factors, bone morphogenetic proteins, engineered protein scaffolds, enzymes, growth factors, hormones, interferons, interleukins, and thrombolytics. They can also be classified based on their molecular mechanism of activity as (a) binding non-covalently to target, e.g., mAbs; (b) affecting covalent bonds, e.g., enzymes; and (c) exerting activity without specific interactions, e.g., serum albumin. Most protein therapeutics currently on the market are recombinant and hundreds of them are in clinical trials for therapy of cancers, immune disorders, infections, and other diseases. New engineered proteins, including bispecific mAbs and multispecific fusion proteins, mAbs conjugated with small molecule drugs, and proteins with optimized pharmacokinetics, are currently under development. However, in the last several decades, there are no conceptually new methodological developments comparable, e.g., to genetic engineering leading to the development of recombinant therapeutic proteins. It appears that a paradigm change in methodologies and understanding of mechanisms is needed to overcome major challenges, including resistance to therapy, access to targets, complexity of biological systems, and individual variations.

Key words: Therapeutics, Proteins, Antibodies, Vaccines, Cancer, Immune diseases, Immunogenicity, Safety, Efficacy

1. Introduction

Proteins, e.g., albumin from egg whites, blood serum albumin, fibrin, and wheat gluten, were recognized in the eighteenth century as biological molecules with distinct properties mostly by their

ability to coagulate under treatments with heat or acid. The term “protein” to describe these molecules was proposed in 1838 by Jöns Jakob Berzelius—from French protéine and German Protein originated from Greek πρώτεῖος (primary) derived from πρῶτος (first, foremost, in time, place, order, or importance). Although one can argue which molecules are most important for life and they all are, currently the proteins as therapeutics are the most important biologicals in terms of their clinical utility. Close to 100 genuine unmodified therapeutic proteins have been approved for clinical use in the European Union (EU) and the USA by July 2011; in contrast, we still have to wait for the first approved DNA-based therapeutic. In 2010, sales of mainly recombinant therapeutic proteins and antibodies exceeded US\$ 100 bln (from US\$ 92 bln in 2009 to US\$ 108 bln in 2010); therapeutic monoclonal antibodies (mAbs) accounted for almost half (48%) of the sales—the top five biologics in 2010 sales are four mAbs and one antibody (IgG1 Fc)-derived fusion protein (Table 1) (<http://www.lamerie.com/press-room/biologics-sales-2010-exceeded-us-100-bln.html>).

Therapeutic mAbs and other therapeutic proteins have been reviewed previously (see recent reviews (1–13) and articles cited there). Therefore, here I will only update and briefly overview currently approved therapeutic proteins, and will describe comparatively therapeutic mAbs as the fastest growing groups of protein therapeutics to illustrate properties, challenges, and future directions that are common for all therapeutic proteins. I will also describe a new classification based on their molecular mechanism of activity.

We have suggested that there were two major paradigm changes in the development of therapeutic antibodies (9). The first occurred more than a century ago and resulted in the development of the serum therapy which saved thousands of lives; von Behring who in the 1880s developed an antitoxin that did not kill the bacteria but neutralized the toxin that the bacteria released into the body was awarded the first Nobel Prize in Medicine in 1901 for his role in the discovery and development of a serum therapy for diphtheria. The second major paradigm change began in the 1970s with the discovery of the hybridoma technology (14) which can provide unlimited quantities of mAbs with predefined specificity. The use of a number of molecular biology techniques, mostly recombinant DNA technology, and the increased understanding of the antibody structure and function led to the development of chimeric and humanized mAbs. Finally, phage display techniques and other techniques based on the progress of molecular biology, including the generation of transgenic animals, allowed the development of fully human antibodies which completed the paradigm change which occurred mostly during a period of two to three decades beginning in the 1970s and ending in the 1990s.

Table 1
The 30 top-selling therapeutic proteins in 2010 (in bln US\$) (modified from LaMerie Business Intelligence, Barcelona)

# (09)	Name	Target/mechanism	Type	Company	Indication	Sales
1 (1)	Erlotinib	TNF α	Fc fusion TNFR2 ECD	Amgen Wyeth	Immune diseases	7.287
2 (3)	Bevacizumab	VEGF	Humanized IgG	Genentech Roche Chugai	Cancer	6.973
3 (4)	Rituximab	CD20	Chimeric IgG	Genentech Biogen-IDEC Roche	Cancer	6.859
4 (5)	Adalimumab	TNF α	Human IgG	Abbott Eisai	Immune diseases	6.548
5 (2)	Infliximab	TNF α	Chimeric IgG	Centocor (J&J) Schering-Plough Mitsubishi Tanabe	Immune diseases	6.520
6 (7)	Trastuzumab	Her2	Humanized IgG	Genentech Chugai Roche	Cancer	5.859
7 (8)	Insulin glargine	Insulin receptor	Modified insulin	Sanofi-Aventis	Diabetes	4.834
8 (6)	Epoetin alfa	EPO-R	Human EPO	Amgen Ortho Biotech Kyowa Hakko Kirin	Anemia	4.590
9 (9)	Pegfilgrastim	G-CSF receptor	PEGhuman G-CSF	Amgen	Neutropenia	3.558
10 (11)	Ranibizumab	VEGF	Humanized Fab	Genentech Novartis	AMD	3.106
11 (10)	Darbepoetin alfa	EPO-R	Modified human EPO	Amgen Kyowa Hakko Kirin	Anemia	2.995
12 (12)	Interferon beta-1a (Avonex)	Interferon beta receptor	Human interferon beta-1a	Biogen Idec	Multiple sclerosis	2.518
13 (13)	Interferon beta-1a (Rebif)	Interferon beta receptor	Human interferon beta-1a	Merck Serono	Multiple sclerosis	2.297
14 (17)	Insulin aspart	Insulin receptor	Modified insulin	Novo Nordisk	Diabetes	2.198

(continued)

**Table 1
(continued)**

# (09)	Name	Target/mechanism	Type	Company	Indication	Sales
15 (14)	Rhu insulin	Insulin receptor	Modified insulin	Novo Nordisk	Diabetes	2.185
16 (15)	Octocog alfa	Factor VIII replacement	Factor VIII	Baxter Healthcare	Hemophilia A	2.095
17 (16)	Insulin lispro	Insulin receptor	Modified insulin	Eli Lilly	Diabetes	2.054
18 (19)	Cetuximab	EGF-R	Chimeric IgG	Eli Lilly BMS Merck Serono	Cancer	1.791
19 (20)	Peginterferon alfa-2a	Interferon alfa receptor	PEGhuman protein	Roche	Hepatitis C	1.775
20 (18)	Interferon beta-1b	Interferon beta receptor	Human protein	Berlex Bayer Schering	Multiple sclerosis	1.661
21 (23)	Eptacog alfa	Initiate coagulation	Human factor VIIa	Novo Nordisk	Hemophilia	1.483
22 (25)	Insulin aspart	Insulin receptor	Modified insulin	Eli Lilly	Diabetes	1.445
23 (-)	OnabotulinumtoxinA	SNAP-25 cleavage	Botulinum toxin type A	Allergan GSK	Medical and esthetic	1.414
24	Epoetin beta	EPO-R	Human EPO	Roche	Anemia	1.387
25	Rec antihemophilic factor	F VIII substitution	Human protein	Bayer Schering	Hemophilia A	1.383
26	Filgrastin	G-CSF receptor	Human protein	Amgen	Neutropenia	1.286
27	Insulin detemir	Insulin receptor	Modified insulin	Novo Nordisk	Diabetes	1.271
28	Natalizumab	α 4 / β 1 / 7 integrin	Humanized IgG	Biogen Idec Elan	Multiple sclerosis	1.230
29	Insulin (humulin)	Insulin receptor	Human insulin	Eli Lilly	Diabetes	1.089
30	Palivizumab	RSV	Humanized IgG	MedImmune	RSV	1.038

The numbers in parentheses are for year 2009
Currencies as of March 2, 2011: 1 €= 1.37726 US\$; 1 CHF = 1.07917 US\$, 1 Yen = 0.0121955 US\$; 1 DKK = 0.184739 US\$; 1 SEK = 0.157766 US\$

The revolutionary changes in science that resulted in the development of antibody therapeutics had broader implications for protein therapeutics in general. The first protein therapeutic other than antibodies, insulin, was purified from animal pancreases and administered to patients with diabetes mellitus in 1922 following the first paradigm change. The availability, cost, and immunogenicity of animal-derived insulin limited its use. It took 60 years and the second paradigm change in the 1970s to produce the first recombinant protein therapeutic, humulin (human insulin) (15). The second paradigm change for therapeutic proteins other than mAbs began with the development of recombinant DNA technologies in the 1970s, PCR, and other advancement in molecular biology, and ended in the 1990s. Interestingly, mAbs are unique among therapeutic proteins in that the hybridoma technology independently on recombinant DNA and other molecular biology methodologies has been capable to provide single well-characterized species that have therapeutic potential. Indeed, the first therapeutic mAb approved for clinical use (1986, Table 1), Muromonab-CD3 (OKT3), is a murine mAb produced by hybridoma technology; it was withdrawn from the market and supplies were exhausted in 2010. Currently, therapeutic proteins are being gradually improved for efficacy, safety, quality, and cost and new targets are being explored, but there are no new concepts similar to those leading to the development of recombinant proteins and identification of unique high-affinity binders out of billions of different molecules.

Excluding protein-based vaccines and diagnostics which will not be discussed here, currently, there are more than 100 approved for clinical use genuine therapeutic proteins of which 29 are mAbs, 22 are enzymes, and the rest are of various structures and function. Based on their pharmacological activity, they can be divided into five groups: (a) replacing a protein that is deficient or abnormal; (b) augmenting an existing pathway; (c) providing a novel function or activity; (d) interfering with a molecule or organism; and (e) delivering other compounds or proteins, such as a radionuclide, cytotoxic drug, or effector proteins (8). Protein therapeutics can be also grouped based on their molecular types that include antibody-based drugs, anticoagulants, blood factors, bone morphogenetic proteins, engineered protein scaffolds, enzymes, Fc fusion proteins, growth factors, hormones, interferons, interleukins, and thrombolytics (5). Based on their molecular mechanism of function, they can be divided into three groups: (a) specific noncovalent binders, (b) proteins affecting covalent bonds, and (c) others. The first group of proteins approved by the European Union or the USA for clinical use by July 2011 consists of 73 genuine unmodified proteins including 29 mAbs (Tables 2, 3, and 4), the second—22 including 21 enzymes (Table 5); and the third group has only one representative, human serum albumin, which is used

Table 2
Therapeutic monoclonal antibodies approved or in review in the European Union or the USA

Name	Trade name	Type	Indication first approved	First EU (US) approval year
Muromonab-CD3	Orthoclone OKt3	Anti-CD3; Murine IgG2a	Reversal of kidney transplant rejection	1986 ^a (1986#2010@)
Abciximab	Reopro	Anti-GPIIb/IIIa; Chimeric IgG1 Fab	Prevention of blood clots in angioplasty	1995 ^a (1994)
Rituximab	MabThera, RituXan	Anti-CD20; Chimeric IgG1	Non-Hodgkin's lymphoma	1998 (1997)
Basiliximab	Simulect	Anti-IL2R; Chimeric IgG1	Prevention of kidney transplant rejection	1998 (1998)
Daclizumab	Zenapax	Anti-IL2R; Humanized IgG1	Prevention of kidney transplant rejection	1999 (1997); #2009
Palivizumab	Synagis	Anti-RSV; Humanized IgG1	Prevention of respiratory syncytial virus infection	1999 (1998)
Infliximab	Remicade	Anti-TNF; Chimeric IgG1	Crohn's disease	1999 (1998)
Trastuzumab	Herceptin	Anti-HER2; Humanized IgG1	Breast cancer	2000 (1998)
Gemtuzumab ozogamicin	Mylotarg	Anti-CD33; Humanized IgG4	Acute myeloid leukemia	NA (2000#2010)
Alemtuzumab	MabCampath, Campath-1H	Anti-CD52; Humanized IgG1	Chronic myeloid leukemia	2001 (2001)
Adalimumab	Humira	Anti-TNF; Human IgG1	Rheumatoid arthritis	2003 (2002)
Tositumomab + ¹³¹ I-Tositumomab	Bexxar	Anti-CD20; Murine IgG2a	Non-Hodgkin lymphoma	NA (2003)
Efalizumab	Raptiva	Anti-CD11a; humanized IgG1	Psoriasis	2004 (2003); #2009
Cetuximab	Erbitux	Anti-EGFR; chimeric IgG1	Colorectal cancer	2004 (2004)
Ibritumomab tiuxetan	Zevalin	Anti-CD20; murine IgG1	Non-Hodgkin's lymphoma	2004 (2002)

Omalizumab	Xolair	Anti-IgE; humanized IgG1	Asthma	2005 (2003)
Bevacizumab	Avastin	Anti-VEGF; humanized IgG1	Colorectal cancer	2005 (2004)
Natalizumab	Tysabri	Anti-a4 integrin; humanized IgG4	Multiple sclerosis	2006 (2004)
Ranibizumab	Lucentis	Anti-VEGF; humanized IgG1 Fab	Macular degeneration	2007 (2006)
Panitumumab	Vectibix	Anti-EGFR; human IgG2	Colorectal cancer	2007 (2006)
Eculizumab	Soliris	Anti-C5; humanized IgG2/4	Paroxysmal nocturnal hemoglobinuria	2007 (2007)
Certolizumab pegol	Cimzia	Anti-TNF; humanized Fab, pegylated	Crohn disease	2009 (2008)
Golimumab	Simponi	Anti-TNF; human IgG1	Rheumatoid and psoriatic arthritis, ankylosing spondylitis	2009 (2009)
Canakinumab	Ilaris	Anti-IL1b; human IgG1	Muckle-Wells syndrome	2009 (2009)
Catumaxomab	Removab	Anti-EPCAM-/CD3; rat/mouse bispecific mAb	Malignant ascites	2009 (NA)
Ustekinumab	Stelara	Anti-IL12/23; human IgG1	Psoriasis	2009 (2009)
Tocilizumab	RoActemra, Actemra	Anti-IL6R; humanized IgG1	Rheumatoid arthritis	2009 (2010)
Ofatumumab	Azerra	Anti-CD20; human IgG1	Chronic lymphocytic leukemia	2010 (2009)
Denosumab	Prolia	Anti-RANK-L; human IgG2	Bone loss	2010 (2010)
Belimumab	Benlysta	Anti-BLYS; human IgG1	Systemic lupus erythematosus	(2011)
Raxibacumab	(Pending)	Anti- <i>B. anthracis</i> PA; human IgG1	Anthrax infection	NA (in review)
Ipilimumab	Yervoy	Anti-CTLA-4; human IgG1	Metastatic melanoma	(2011)
Brentuximab vedotin	(Pending)	Anti-CD30; chimeric IgG1; immunoconjugate	Hodgkin lymphoma, systemic ALCL	NA (application submitted)

Information current as of July 2011. Updated and modified from Janice M. Reichert, Editor-in-Chief, mAbs

*Country-specific approval; approved under concordance procedure; #Voluntarily withdrawn from the market. @Supplies are exhausted in 2010. BLyS B lymphocyte stimulator, C5 complement 5, CD cluster of differentiation, CTLA-4 cytotoxic T lymphocyte antigen 4, EGFR epidermal growth factor receptor, EPCAM epithelial cell adhesion molecule, GP glycoprotein, IL interleukin, NA not approved, PA protective antigen, RANT-L receptor activator of NFκB ligand, RSV respiratory syncytial virus, TNF tumor necrosis factor, VEGF vascular endothelial growth factor. Not included here are polyclonal antibodies against infectious diseases and toxins

Table 3
List of Fc fusion proteins, the year denotes date of approval

1. Etanercept (TNFR2 ECD, 1998)
2. Alefacept (LFA3 ECD, 2003)
3. Abatacept (CTLA4 ECD, 2005)
4. Rilonacept (IL-1RI/IL-1RacP ECD, 2008)
5. Romiplostim (41aa thrombopoietin (TPO) analogue peptide, 2008)
6. Belatacept (CTLA4 ECD, 2011)

ECD extracellular domain

Table 4
List of genuine noncovalent binders other than mAbs, Fc fusion proteins, and polyclonal immunoglobulins approved for clinical use

1. Insulin (blood glucose regulator)
2. Pramlintide acetate (glucose control)
3. Growth hormone GH (growth failure)
4. Pegvisoman (growth hormone receptor antagonist)
5. Mecasermin (IGF1, growth failure)
6. Factor VIII (coagulation factor)
7. Factor IX (coagulation factor)
8. Protein C concentrate (anti-coagulation)
9. α 1-proteinase inhibitor (anti-trypsin inhibitor)
10. Erythropoietin (stimulates erythropoiesis)
11. Filgrastim (granulocyte colony-stimulating factor, G-CSF; stimulates neutrophil proliferation)
12. Sargramostim36, 37 (granulocytemacrophage colony-stimulating factor, GM-CSF)
13. Oprelvekin (interleukin11, IL11)
14. Human follicle-stimulating hormone (FSH)
15. Human chorionic gonadotropin (HCG)

(continued)

Table 4
(continued)

16. Lutropin- α (human luteinizing hormone)
17. Interleukin 2 (IL2)
18. Denileukin diftitox (fusion of IL2 and Diphtheria toxin)
19. Interferon alfacon 1 (consensus interferon)
20. Interferon- α 2a (IFN α 2a)
21. Interferon- α 2b (IFN α 2b)
22. Interferon- α n3 (IFN α n3)
23. Interferon- β 1a (rIFN- β)
24. Interferon- β 1b (rIFN- β)
25. Interferon- γ 1b (IFN γ)
26. Salmon calcitonin (32-amino acid linear polypeptide hormone)
27. Teriparatide (part of human parathyroid hormone 1–34 residues)
28. Exenatide (Incretin mimetic with actions similar to glucagon-like peptide 1)
29. Octreotide (octapeptide that mimics natural somatostatin)
30. Dibotermín- α (recombinant human bone morphogenic protein 2)
31. Recombinant human bone morphogenic protein 7
32. Histrelin acetate (gonadotropin-releasing hormone; GnRH)
33. Palifermin (keratinocyte growth factor, KGF)
34. Bepaclermin (platelet-derived growth factor, PDGF)
35. Nesiritide (recombinant human B-type natriuretic peptide)
36. Lepirudin (recombinant variant of hirudin, another variant is Bivalirudin)
37. Anakinra (interleukin 1 (IL1) receptor antagonist)
38. Enfuvirtide (an HIV-1 gp41-derived peptide)

Information for the protein and/or abbreviations used is provided in parentheses.
Modified and updated from (8)

Table 5
List of genuine therapeutic proteins affecting covalent bonds—enzymes and antithrombin III—approved for clinical use

1. β -Glucocerebrosidase (hydrolyzes to glucose and ceramide)
2. Alglucosidase- α (degrades glycogen)
3. Laronidase (digests glycosaminoglycans within lysosomes)
4. Idursulfase (cleaves O-sulfate preventing GAGs accumulation)
5. Galsulfase (cleaves terminal sulphate from GAGs)
6. Agalsidase- β (human α -galactosidase A, hydrolyzes glycosphingolipids)
7. Lactase (digest lactose)
8. Pancreatic enzymes (lipase, amylase, protease; digest food)
9. Adenosine deaminase (metabolizes adenosine)
10. Tissue plasminogen activator (tPA, serine protease involved in the breakdown of blood clots)
11. Factor VIIa (serine protease, causes blood to clot)
12. Drotrecogin- α (serine protease, human activated protein C)
13. Trypsin (serine protease, hydrolyzes proteins)
14. Botulinum toxin type A (protease, inactivates SNAP-25 which is involved in synaptic vesicle fusion)
15. Botulinum toxin type B (protease that inactivates SNAP-25 which is involved in synaptic vesicle fusion)
16. Collagenase (endopeptidase, digest native collagen)
17. Human deoxyribonuclease I (endonuclease, DNase I, cleaves DNA)
18. Hyaluronidase (hydrolyzes hyaluronan)
19. Papain (cysteine protease, hydrolyzes proteins)
20. L-Asparaginase (catalyzes the conversion of L-asparagine to aspartic acid and ammonia)
21. Rasburicase (urate oxidase, catalyzes the conversion of uric acid to allantoin)
22. Streptokinase (Anistreplase is anisoylated plasminogen streptokinase activator complex (APSAC))
23. Antithrombin III (serine protease inhibitor)

Information for the protein and/or abbreviations used is provided in parentheses.
Modified and updated from (8)

to increase plasma osmolarity. One should note that there is a group of polyclonal antibodies (either nonspecific pooled human immunoglobulin (Ig) or specific Ig) which are still in clinical use against toxins (diphtheria, tetanus, botulism), viruses (hepatitis A, hepatitis B, cytomegalovirus, varicella zoster, rabies, measles, vaccinia), venom toxins, and toxic drugs (digoxin); nonspecific pooled human IgG is also approved for use against idiopathic thrombocytopenic purpura, Kawasaki disease, and IgG deficiency. If these IgGs are included as therapeutic proteins, then the total number of genuine therapeutic proteins exceeds 100. The largest and currently most selling therapeutic protein group is the first one and especially mAbs and Fc fusion proteins, including the top six-selling protein therapeutics in 2010 (Table 1).

2. mAbs and Fc Fusion Proteins

2.1. mAbs and Fc Fusion Proteins

Approved for Clinical Use or in Clinical Trials

Currently (as of July 2011) 29 mAbs are approved for clinical use in the European Union or the USA (Table 2) and 6 Fc fusion proteins (Table 3). Four of the approved mAbs were withdrawn from the market for safety or utility reasons. One of the mAbs, Synagis, is for prevention and not for therapy but traditionally is included as a therapeutic mAb. Worldwide sales of mAb-based drugs in 2010 were \$52 billion of \$108 billion total for all protein biopharmaceuticals; six of the ten top-selling therapeutic proteins in 2010 were mAbs (infliximab, bevacizumab, rituximab, adalimumab, trastuzumab, and ranibizumab) and one (etanercept)—Fc fusion protein (Table 1).

One antibody drug conjugate (ADC), Brentuximab vedotin, is pending approval; if approved, it will be the first in class. It comprises an anti-CD30 mAb attached by a protease-cleavable linker to a potent, synthetic drug, monomethyl auristatin E (MMAE) utilizing Seattle Genetics' proprietary technology. The ADC employs a novel linker system that is designed to be stable in the bloodstream but to release MMAE upon internalization into CD30-expressing tumor cells. This approach is intended to spare non-targeted cells, which may help minimize the potential toxic effects of traditional chemotherapy while allowing for the selective targeting of CD30-expressing cancer cells, thus potentially enhancing the antitumor activity.

Hundreds of mAbs are in thousands of clinical trials including 25 (12) in phase 3 trials—2,909 entries for planned, ongoing or completed clinical trials were retrieved from <http://www.clinical-trials.gov> by searching with therapy AND mAbs as of July 2011 of which 484 are in phase 3. Significant number of all new medicines are mAbs (see also <http://www.phrma.org/research/new-medicines>). More than 200 different antibody-based candidate

therapeutics are in clinical trials targeting more than 70 different molecules. At least one to three different antibodies are being developed at different companies for each relevant therapeutic target. However, some molecules are targeted by many more mAbs, e.g., the insulin-like growth factor receptor type I (IGF-IR) is targeted by more than ten different mAbs (16). During the last decade and especially in the last several years, the number of clinical trials with therapeutic antibodies has increased dramatically. However, this increase has been largely due to an increase in the number of indications for the same antibodies, especially in combination with other therapeutics. The number of targets and corresponding antibodies in preclinical development and in the discovery phase has also increased significantly during the past decade.

Second- and third-generation mAbs are being developed against already validated targets. For example, based on Synagis, an antibody (motavizumab—MEDI-524; NuMax) was developed with much higher affinity to the F protein of the RSV (17). The improvement of already existing antibodies also includes an increase (to a certain extent) of their binding to Fc receptors for enhancement of antibody-dependent cell-mediated cytotoxicity (ADCC) and half-life, selection of appropriate frameworks to increase stability and yield, decrease of immunogenicity by using *in silico* and *in vitro* methods, and conjugation to small molecules and various fusion proteins to enhance cytotoxicity. A major lesson from the current state of antibody-based therapeutics is that gradual improvement in the properties of existing antibodies and identification of novel antibodies and novel targets are likely to continue in the foreseeable future. This is likely to be a major driving force of the field until saturation is reached presumably in the next decade or two, and various combinations of antibodies and other drugs may dominate unless a major change in the current paradigm occurs. Currently, research and development (R&D) of mAbs as potential therapeutics is growing, and therapeutic mAbs are the fastest growing group of therapeutic proteins.

Antibodies are used not only as antigen binders but also as contributing effector functions through their Fc to already known binders. Currently approved fusion proteins are mostly Fc fusions (Table 3). The rationale to develop Fc fusion proteins is that Fc can confer effector functions and long half-life because of binding to the neonatal Fc receptor (FcRn) and increase in size. In addition, because Fc is dimeric, such fusion proteins could exhibit activity due to increase in valency. An Fc fusion protein (etanercept) continues to be the top-selling therapeutic protein in 2010 with worldwide sales of \$7.287 bln. It is a TNF α inhibitor and is effective only in its dimeric form due to the Fc which also confers long half-life in the circulation. Interestingly, two other TNF α inhibitors (adalimumab and infliximab) are also among the top five selling protein therapeutics; the total sales for these three therapeutics

exceed \$20 bln which makes TNF α a molecule targeted by the highest-selling therapeutic proteins. Romiplostim is the first peptide-Fc fusion (“peptibody”) approved as a human therapeutic; it is a thrombopoietin (TPO) receptor agonist for treatment for thrombocytopenia. Two Fc fusion proteins (Abatacept and Belatacept) are fused to the same molecule (CTLA4) but are used for different indications—rheumatoid arthritis and for the prevention of acute rejection in adult patients who have had a kidney transplant, respectively.

2.2. Beyond Traditional Antibodies: Engineered Antibody Domains

One question is whether a new paradigm change could trigger a new dramatic expansion of some novel, still unknown, types of therapeutics as it happened several decades ago. We do not know the answer to this question and surprises are always possible, but currently there are no indications that another paradigm change in the discovery of biological therapeutics is coming anytime soon (one methodology which could contribute to such paradigm change is the revolution in high-throughput sequencing but time will show how useful it is for development of therapeutic proteins). It rather appears that there will be gradual improvements of existing antibodies and identification of antibodies to novel targets using currently available methodologies. However, one area where one could expect conceptually novel antibody-based candidate therapeutics even though within the current paradigm is going beyond traditional antibody structures.

Currently, almost all FDA-approved therapeutic antibodies (Table 2) and the vast majority of those in clinical trials are full-size antibodies mostly in IgG1 format of about 150 kDa size. A fundamental problem for such large molecules is their poor penetration into tissues (e.g., solid tumors) and poor or absent binding to regions on the surface of some molecules (e.g., on the HIV envelope glycoprotein) which are accessible by molecules of smaller size. Therefore, a large amount of work especially during the last decade has been aimed at developing novel scaffolds of much smaller size and higher stability (see, e.g., recent reviews (9, 18, 19)). Such scaffolds are based on various human and nonhuman molecules of high stability and could be divided into two major groups for the purposes of this review—antibody derived and others. Here, I will briefly discuss advantages of antibody-derived scaffolds, specifically those derived from antibody domains, and binders selected from libraries based on engineered antibody domains (eAds); an excellent recent review describes the second group (18).

Firstly, their size (12–15 kDa) is about an order of magnitude smaller than the size of an IgG1 (about 150 kDa). The small size leads to relatively good penetration into tissues and the ability to bind into cavities or active sites of protein targets which may not be accessible to full-size antibodies. This could be particularly important for the development of therapeutics against rapidly mutating

viruses, e.g., HIV. Because these viruses have evolved in humans to escape naturally occurring antibodies of large size, some of their surface regions which are critical for the viral life cycle may be vulnerable for targeting by molecules of smaller size including eAds. Secondly, eAds may be more stable than full-size antibodies in the circulation and can be relatively easily engineered to further increase their stability. For example, some eAds with increased stability could be taken orally or delivered via the pulmonary route or may even penetrate the blood–brain barrier, and retain activity even after being subjected to harsh conditions, such as freeze-drying or heat denaturation. In addition, eAds are typically monomeric, of high solubility, and do not significantly aggregate or can be engineered to reduce aggregation. Their half-life in the circulation can be relatively easily adjusted from minutes or hours to weeks by making fusion proteins of varying size and changing binding to the FcRn. In contrast to conventional antibodies, eAds are well expressed in bacterial, yeast, and mammalian cell systems. Finally, the small size of eAds allows for higher molar quantities per gram of product, which should provide a significant increase in potency per dose and reduction in overall manufacturing cost. However, in spite of all these advantages, there is still no candidate therapeutic based on such scaffolds in phase III clinical trial as of July 2011.

Research on novel antibody-derived scaffold continues. We identified a VH-based scaffold which is stable and highly soluble (20). It was used for construction of a large-size (20 billion clones) eAd phage library by grafting CDR3s and CDR2s from five of our other Fab libraries and randomly mutagenizing CDR1. Panning of this library with an HIV Env complexed with CD4 resulted in the identification of a very potent broadly cross-reactive eAd against HIV, m36, which neutralized primary HIV isolates from different clades with IC50s and IC90s in the low µg/ml range (21). Fc fusion proteins of m36 were even more potent and neutralized all tested isolates (22). I also proposed to use engineered antibody constant domains (CH2 of IgG, IgA, and IgD, and CH3 of IgE and IgM) as scaffolds for construction of libraries (23). Because of their small size and the domain role in antibody effector functions, these have been termed nanoantibodies, the smallest fragments that could be engineered to exhibit simultaneously antigen binding and effector functions. Several large libraries (up to 50 billion clones) were constructed and antigen-specific binders successfully identified (24). We have recently engineered CH2-based scaffolds with high stability by introducing an additional disulfide bond (25) and by shortening CH2 (26). It is possible that these and other novel scaffolds under development could provide new opportunities for identification of potentially useful therapeutics.

3. Therapeutic Proteins Other Than Antibody-Based

3.1. Therapeutic Proteins Approved for Clinical Use

Therapeutic proteins other than mAbs and Fc fusion proteins approved for clinical use by the US FDA include noncovalent binders (Table 4), proteins that affect covalent bonds which are almost all enzymes (Table 5), and albumin. Based on their molecular type and similarity in function, they can be divided into anticoagulants, blood factors, bone morphogenetic proteins, engineered protein scaffolds, enzymes, growth factors, hormones, interferons, interleukins, and thrombolytics (5). There are several major differences and similarities between these proteins and antibody-based therapeutics which may explain why on average per therapeutic the approved mAbs and Fc fusion proteins are more successful in terms of sales. The Fc portion of the antibody confers them with relatively long half-life by binding to the FcRn and effector functions including ADCC and complement (27). There is no other protein capable of performing those functions simultaneously. A second major advantage is the ability of the Fabs to bind to large number of targets keeping their Ig-based scaffold. Attempts to design similar protein scaffolds based on other proteins are ongoing, but so far none has reached approval for clinical use. A third one is that antibodies have evolved to fight diseases and are in high concentrations (tens of mgs/ml) in the blood without significant side effects. Therefore, mAbs and Fc fusion proteins on average should be less toxic than other protein therapeutics.

3.2. Engineered Proteins to Resemble Fc Properties

Various methodologies have been used to engineer therapeutic proteins to resemble some of the properties which mAbs already have mostly through their Fc. To increase their half-lives, they are PEGylated (conjugated with PEG) which is currently the most successful, clinically and commercially, conjugation of proteins. An example of such protein is peginterferon- α 2b. There are currently several alternative competing conjugations with conformationally flexible stretches of amino acid residues, e.g., G, that lead to an increase in the hydrodynamic radius and correspondingly to an increase in the half-life of protein therapeutics. To confer cytotoxic functions, proteins can be conjugated to small molecule drugs and radionuclides similarly to antibodies.

3.3. Enzymes

Enzymes are a special class of protein therapeutics (Table 5). Although there are known antibodies with catalytic properties, typically they are not that active as highly specialized enzymes and currently all enzymes approved for clinical use are proteins different than antibodies. Interestingly, there is one protein approved for clinical use which changes a covalent bond but is not an enzyme—antithrombin III (AT-III). It inactivates thrombin by forming a covalent bond between the catalytic serine residue of thrombin and an arginine reactive site on AT-III.

4. Therapeutic Proteins: Successes and Challenges

The success of protein-based therapeutics is mostly due to the use of concepts and methodologies developed during the second paradigm change decades ago that resulted in dramatic improvement of three key features in candidate therapeutics required for FDA approval: safety, efficacy, and quality. They are critical for the success of any drug and are discussed in more detail below mostly with examples for antibody-based therapeutics which share similar challenges with other therapeutic proteins.

4.1. Safety

Side effects due to therapeutic proteins could be divided into two large groups: (a) interactions with intended targets and (b) interactions with unintended targets. Binding to an intended target can lead to undesirable side effects, e.g., by immunomodulatory antibodies that could be suppressory or stimulatory. Administration of suppressory therapeutic proteins could lead to wide range of side effects related to decreased function of the immune system. An important example is the use of the best-selling antibody-based protein therapeutics targeting TNF α (etanercept, infliximab, certolizumab pegol, and adalimumab) which can lead to infectious complications (28). The overstimulation of the immune system can also produce life-threatening illness. In one case which gained wide publicity, administration of a single dose of the stimulatory anti-CD28 mAb TGN1412 resulted in induction of a systemic inflammatory response characterized by a rapid induction of pro-inflammatory cytokines in all six volunteers, leading to critical illness in 12–16 h (29). One important difference between antibody-based therapeutic containing Fc and other therapeutic proteins (not conjugated with toxic molecules) is that the antibody effector functions including ADCC and complement-dependent cytotoxicity (CDC) could lead to toxicities after binding to intended target molecules but on tissues other than those intended. An example of this is the trastuzumab-associated cardiotoxicity that is potentiated when the antibody is used concurrently or sequentially with an anthracycline (30).

Interactions with unintended targets can lead to a wide range of side effects in many cases with poorly understood mechanisms. An important example is the adverse acute infusion reactions after administration of proteins, where cytokine release plays a pivotal role, but other not fully explained mechanisms could be involved; such reactions were reported for many proteins, including infliximab, rituximab, cetuximab, alemtuzumab, trastuzumab and panitumumab (31), insulin, and interferon. Infusion side effects for rituximab can result from release of cellular contents from lysed malignant B cells (32). Administration of proteins can also lead to hypersensitivity reactions, including anaphylactic shock and serum

sickness (28). Preexisting IgEs that cross-react with therapeutic proteins can increase the number and severity of such reactions, which can occur even with the first protein infusion. A notable example of this occurred with administration of cetuximab (31). Hypersensitivity is frequently associated with immunogenicity.

4.2. Immunogenicity

Immunogenicity of proteins can be a significant safety and efficacy issue (28, 33–38). For example, the success of the mAb-based therapeutics was critically related to the development of less immunogenic proteins. Murine mAbs were used initially as candidate therapeutics in the 1980s, but their high immunogenicity resulted in high titers of human anti-mouse antibodies (HAMAs), and related toxicities and low potency. Development of the less immunogenic chimeric mAbs, which contain human Fc fragments, and humanized mAbs, which contain mouse complementarity determining regions (CDRs) grafted into human antibody framework, was critical for the clinical success of the products. Human antibodies exhibit low immunogenicity on average, and are currently the favored type of antibody in development, although most of the therapeutic antibodies approved for clinical use are still chimeric and humanized mAbs.

Immunogenicity can be influenced by factors related to protein structure, composition, posttranslational modifications, impurities, heterogeneity, aggregate formation, degradation, formulation, storage conditions, as well as properties of its interacting partner, the patient's immune system and disease status, concomitant medications, dose, route, and time and frequency of administration especially when administered as multiple doses over prolonged periods (34). Even human proteins can elicit human anti-human antibodies. In one of the most studied cases of anti-TNF α mAbs, treatment with the human mAb adalimumab resulted in antibodies against the therapeutic that varied from <1% to up to 87% for different cohorts of patients, protocols, disease, and methods of measurement (39).

A likely mechanism for the immunogenicity of human mAbs involves the unique antibody sequences that confer antigen binding and specificity, but may appear foreign. Human therapeutic proteins can also break immune tolerance and aggregation can be a major determinant of antibody elicitation (34). Aggregation can result in repetitive structures that may not require T cell help (40). Protein immunogenicity may also affect efficacy through either the pharmacokinetic or neutralizing effects of the antibody responses that are dependent on a number of factors, including the affinity, specificity, and concentration of the induced antibodies (33). Because immunogenicity is an important factor in both safety and efficacy, significant efforts to predict and reduce immunogenicity of therapeutic proteins are ongoing (35–38).

Individual immune responses to therapeutic proteins vary widely. A key, and largely unanswered, question is what determines these variations. Despite extensive laboratory and clinical studies that were instrumental in delineating general concepts about critical factors involved in immunogenicity, it is impossible to predict the extent to which a novel therapeutic protein will be immunogenic in human patients. Little is known about the individual antibodies composing the polyclonal response to therapeutic proteins. The germline antibody repertoire at any given time could be a major determinant of individual differences, and so knowledge of large portions of antibodies generated by the human immune system, preferably the complete set, i.e., the antibodyome (6), could ultimately help to predict individual immune responses to therapeutic proteins.

In spite of the possibility for immunogenicity and other side effects, protein therapeutics are relatively safe due primarily to their high specificity. This is a fundamental advantage compared to small molecule drugs which on average are less specific and can bind nonspecifically to large number of molecules. However, in some cases, there are significant side effects, and safety concerns can lead to the withdrawal of therapeutic proteins from the market. The psoriasis drug efalizumab was withdrawn because of a potential risk of patients developing progressive multifocal leukoencephalopathy (PML), which is a rare, serious, progressive neurologic disease caused by the JC virus (JCV). More than 80% of the general population is infected with JCV. Why the virus becomes activated and causes disease only in minority of the treated patients is unknown, although typically PML occurs in people whose immune systems have been severely weakened. Thus, choosing the most appropriate animal model for toxicity testing is very important and species cross-reactivity should be included when identifying new candidate mAb therapeutics. If such a model does not exist, transgenic animals expressing the human target and surrogate protein that is cross-reactive with the human homologous target in relevant animals can be used (41).

4.3. Efficacy

After safety, efficacy is the most important parameter considered by FDA for approval. Many therapeutic proteins are highly effective in vivo and have revolutionized treatment of diseases, e.g., insulin for diabetes, epoetin for anemia, and rituximab for non-Hodgkin lymphoma (32), to name a few. Alemtuzumab plays an important role in the therapy of hematological malignancies (42). Another example is trastuzumab as adjuvant systemic therapy for human epidermal growth factor receptor type 2 (HER2)-positive breast cancer (43). Results from six trials randomizing more than 14,000 women with HER2-positive early breast cancer to trastuzumab versus non-trastuzumab-based adjuvant chemotherapy demonstrate that the

addition of trastuzumab reduces recurrence by approximately 50% and improves overall survival by 30% (44).

On average, the efficacy of therapeutic mAbs and some other therapeutic proteins is not high and there is substantial individual variability. One prominent example is trastuzumab (Herceptin) which has clearly revolutionized the treatment of HER2-positive patients; however, half of the patients still have non-responding tumors, and disease progression occurs within a year in the majority of cases (45). For patients with disease progression, combination with small molecules could be useful, e.g., the addition of a dual tyrosine kinase inhibitor of epidermal growth factor receptor (EGFR) and HER2 lapatinib to capecitabine was shown to provide superior efficacy for women with HER2-positive, advanced breast cancer progressing after treatment with anthracycline-, taxane-, and trastuzumab-based therapy (46). Current data do not support the use of trastuzumab for more than 1 year; the appropriate length of treatment, optimum timing, and administration schedule are not known (43). Like other therapeutic proteins trastuzumab does not appear to efficiently cross the blood–brain barrier, and it is unclear if the current practice of local therapy of the central nervous system and continued trastuzumab is optimal (45).

Anti-angiogenic therapies that target the vascular endothelial growth factor (VEGF), e.g., bevacizumab, and the VEGF receptor (VEGFR) are effective adjuncts for treatment of solid tumors, and are commonly administered in combination with cytotoxic chemotherapy. However, at least half of patients fail to respond to anti-angiogenic treatment of gliomas, and the response duration is modest and variable (47). The use of bevacizumab plus paclitaxel as a first-line treatment of patients with metastatic breast cancer doubled median progression-free survival (PFS; 11.8 months versus 5.9 months; hazard ratio = 0.60; $P < .001$) compared with paclitaxel alone; however, a statistically significant improvement in overall survival was not provided by the addition of bevacizumab, although a post hoc analysis demonstrated a significant increase in 1-year survival for the combination arm (48).

The anti-EGFR mAbs cetuximab and panitumumab, either as single agents or in combination with chemotherapy, have demonstrated clinical activity against metastatic colorectal cancer, but appear to benefit only select patients with predictive markers of efficacy, including EGFR overexpression, development of skin rash, and the absence of a K-ras mutation (49). In general, as single agents or in combination, therapeutic mAbs and other proteins have produced only modest clinical responses in solid tumors (50). There are no mAbs approved for treatment of a number of tumors, e.g., prostate cancer. However, for prostate cancer, there are 30 candidates in the pipeline (16 vaccines and 14 antibodies), and one FDA-approved prostate cancer vaccine (Provenge); of these

candidates, 19 are in phases II and III (9 vaccines and ten antibodies) and 8 are in phase I clinical trials.

The mechanisms underlying the relatively low efficacy of some therapeutic proteins and the high variability of responses to treatment are not well known, but are likely to involve multiple factors. Preexisting resistance or development of resistance is a fundamental problem for any therapeutic. Various mechanisms, including mutations, activation of multidrug transporters, and overexpression or activation of signaling proteins, are operating as exemplified for EGFR-targeted therapies (51). Another major problem is poor penetration into tissues, e.g., solid tumors. A related issue for full-size mAbs is poor or absent binding to regions on the surface of some molecules, i.e., existence of “steric barriers,” e.g., on the HIV envelope glycoprotein (Env) (22).

New approaches are being developed to increase efficacy of mAb and other therapeutic proteins, including enhanced effector functions, improved half-life, increased tumor and tissue accessibility, and greater stability; the methods used involve both protein- and glyco-engineering, and results to date are encouraging (52, 53). mAbs that do not engage the innate immune system’s effector functions are being developed when binding is sufficient (54). Multi-targeted antibodies are being developed and tested in clinical trials, e.g., an antibody targeting HER2/neu and CD3 with preferential binding to activating Fc γ type I/III-receptors, resulting in the formation of tri-cell complexes among tumor cells, T cells, and accessory cells (55). Similar bispecific (targeting CD3 and epithelial cell adhesion molecule, EpCAM) trifunctional mAb, catumaxomab, was approved in the European Union for therapy of malignant ascites in 2009 (Table 2): the first bispecific mAb approved for clinical use. This antibody binds to cancer cells expressing EpCAM on their surface via one arm; to a T lymphocyte expressing CD3 via the other arm; and to an antigen-presenting cell like a macrophage, a natural killer cell, or a dendritic cell via the Fc. This initiates an immunological reaction leading to the removal of cancer cells from the abdominal cavity, thus reducing the tumor burden which is seen as the cause for ascites in cancer patients. Bispecific and multispecific mAbs and other therapeutic proteins are currently being developed to a number of targets.

A promising direction is the modulation of immune responses by mAbs targeting regulators of T cell immune responses. The cytotoxic T lymphocyte antigen 4 (CTLA-4) present on activated T cells is an inhibitory regulator of such responses. Human antibodies and Fc fusion proteins that abrogate the function of CTLA-4 have been tested in the clinic and found to have clinical activity against melanoma (56, 57). It appears that CTLA-4 blockade also enhanced the cancer-testis antigen NY-ESO-1-specific B cell and T cell immune responses in patients with durable objective clinical responses and stable disease suggesting immunotherapeutic designs

that combine NY-ESO-1 vaccination with CTLA-4 blockade (57). Ipilimumab which targets CTLA-4 was approved by the US FDA in 2011 for therapy of metastatic melanoma (Table 2). Therapeutic mAbs that mimic the natural ligand, e.g., the tumor necrosis factor-related apoptosis inducing ligand (TRAIL), have also been developed (58, 59).

Currently, second- and third-generation mAbs against already validated targets, e.g., HER2, CD20, and TNF α , are in clinical studies or already approved. Various approaches have been used to discover novel, relevant targets, but progress has been slow. Modifications of the standard panning procedures have been reported, including enhanced selection of cross-reactive antibodies by sequential antigen panning (60) and competitive antigen panning for focused selection of antibodies targeting a specific protein domain or subunit (61, 62). To ensure better tissue penetration and hidden epitope access, a variety of small engineered antibody domains (about tenfold smaller than IgG) are being developed (19, 20). Knowledge of antibodyomes could be used for generation of semisynthetic libraries for selection of high-affinity binders of small size and minimal immunogenicity (6).

A major lesson from the current state of antibody-based therapeutics is that gradual improvement in the properties of existing therapeutic proteins and identification of novel proteins and targets are likely to continue in the foreseeable future. A fundamental challenge has been to increase dramatically the efficacy of therapeutic antibodies and to apply them to many more diseases. Other major challenges are the development of effective personalized antibody-based therapeutics, and prediction of toxicity or potentially low efficacy *in vivo*.

4.4. Quality

Quality is a very important parameter for approval of any drug by FDA. A specific fundamental feature that distinguishes mAb and other biologics from small molecule drugs is their heterogeneity. Heterogeneity of mAbs is due to modifications, such as incomplete disulfide bond formation, glycosylation, N-terminal pyroglutamine cyclization, C-terminal lysine processing, deamidation, isomerization, oxidation, amidation of the C-terminal amino acid, and modification of the N-terminal amino acids by maleic acid, as well as noncovalent associations with other molecules, conformational diversity, and aggregation (63). Tens of thousands of variants with the same sequence may coexist.

Development of high-quality protein therapeutics with minimal heterogeneity and contamination is essential for their safety and approval by FDA. Process development for production of therapeutic proteins is a very complex operation involving recombinant DNA technologies, verification of a strong expression system, gene amplification, characterization of a stable host cell expression system, optimization and design of the mammalian cell culture

fermentation system, and development of an efficient recovery process resulting in high yields and product quality (64). Titers in the range of 5–10 g/L or even higher, cell densities of more than 20 million cells/ml, and specific productivity of over 20 pg/cell/day (even up to 100 pg/cell per day) have been achieved (65).

Genetic delivery of therapeutic proteins by in vivo production offers a new direction to increase quality and reduce cost; three approaches can be used for the stable long-term expression and secretion of therapeutic proteins in vivo: (1) direct in vivo administration of integrating vectors carrying the gene, (2) grafting of ex vivo genetically modified autologous cells, and (3) implantation of an encapsulated antibody producing heterologous or autologous cells. Another promising direction is the prospects for using molecular farming methods to create relatively low-cost therapeutic proteins in plants, e.g., in genetically engineered tobacco leaves.

5. Biosimilar and Biobetter Therapeutic Proteins

A major direction of current activity is to develop therapeutic proteins that are similar but cheaper than the currently existing or are better in terms of efficacy and safety. By 2015, biologics worth \$60 billion in annual sales will lose patent protection, bolstering hopes for the rapid growth of the biosimilars as generics companies elbow their way into a big new market. Rituxan/MabThera, Remicade, and Enbrel are on the top of the list for biosimilars. Sandoz, e.g., which is leading the pack of generic companies angling to get into the market, expects to see biosimilar revenue jump from \$250 million in 2011 to \$20 billion by 2020. Over the next 5 years, the market for biosimilars will increase to \$10 billion, but only a handful of big pharmaceutical companies and world-class R&D facilities will be able to take part. And that means that most small- and medium-size drug developers will never have a chance of getting into the new market for follow-on biologics.

The niche for most small biotech companies is taking a pre-clinical- or very-early-stage candidate to proof of concept, at which point they can make sale to bigger companies. With biosimilars, the developer will start with proof of concept data and then ramp up the most expensive stage of clinical development, with the added charge of running a likely comparison study to the marketed therapeutic. That will not be cheap. It could take 8 years to run a biosimilar program with development costs sliding from \$100 million to \$150 million. With that much time and money at stake, most biotech companies may never be competitive.

6. Conclusions

The rapid progress made in the last few decades toward the development of potent therapeutic proteins raises a number of questions for the future directions of this field. A key question is whether there are any indications of a paradigm change that could lead to radically different therapeutics as occurred two to three decades ago and which resulted in an explosion of protein therapeutics approved for clinical use during the last decades. If history provides an answer and such a paradigm shift occurs, it will probably take decades before we witness the fruition of such a shift in terms of new licensed protein therapeutics. Meanwhile, gradual improvements in the characteristics of existing protein therapeutics, discovery of novel protein-based drugs and novel targets, combining therapeutics, conjugating them with drugs, nanoparticles, and other reagents using integrative approaches based on cell biology, bioengineering and genetic profiling, as well as predictive tools to narrow down which candidate molecules could be successfully developed as therapeutics, and developing novel protein-based scaffolds with superior properties to those already in use will be major areas of research and development in the coming decades. A decade from now, it is likely that we will see many protein-based therapeutics based on different scaffolds approved for clinical use and hundreds more in preclinical and clinical development.

Acknowledgments

I would like to thank members of the group Protein Interactions, especially W. Chen, X. Xiao, Z. Zhu, Y. Feng, E. Streaker, and J. Owens for discussions, experiments and help, and the Editor V. Voynov for helpful suggestions which improved this article. This study was supported by the NIH NCI CCR intramural program, the NIH intramural AIDS program (IATAP), and the NIAID intramural biodefense program.

References

1. Carter PJ (2006) Potent antibody therapeutics by design. *Nat Rev Immunol* 6:343–357
2. Schrama D, Reisfeld RA, Becker JC (2006) Antibody targeted drugs as cancer therapeutics. *Nat Rev Drug Discov* 5:147–159
3. Waldmann TA (2003) Immunotherapy: past, present and future. *Nat Med* 9:269–277
4. Casadevall A, Dadachova E, Pirofski LA (2004) Passive antibody therapy for infectious diseases. *Nat Rev Microbiol* 2:695–703
5. Carter PJ (2011) Introduction to current and future protein therapeutics: a protein engineering perspective. *Exp Cell Res* 317: 1261–1269

6. Dimitrov DS (2010) Therapeutic antibodies, vaccines and antibodyomes. *MAbs* 2:347–356
7. Walsh G (2010) Biopharmaceutical benchmarks 2010. *Nat Biotechnol* 28:917–924
8. Leader B, Baca QJ, Golan DE (2008) Protein therapeutics: a summary and pharmacological classification. *Nat Rev Drug Discov* 7:21–39
9. Dimitrov DS, Marks JD (2009) Therapeutic antibodies: current state and future trends—is a paradigm change coming soon? *Methods Mol Biol* 525:1–27
10. Ashkenazi A (2008) Directing cancer cells to self-destruct with pro-apoptotic receptor agonists. *Nat Rev Drug Discov* 7:1001–1012
11. Beck A, Reichert JM (2011) Therapeutic Fc-fusion proteins and peptides as successful alternatives to antibodies. *MAbs* 3(5):415–416
12. Reichert JM (2011) Antibody-based therapeutics to watch in 2011. *MAbs* 3:76–99
13. Reichert JM (2010) Metrics for antibody therapeutics development. *MAbs* 2:695–700
14. Kohler G, Milstein C (1975) Continuous cultures of fused cells secreting antibody of predefined specificity. *Nature* 256:495–497
15. Goeddel DV, Kleid DG, Bolivar F, Heyneker HL, Yansura DG, Crea R, Hirose T, Kraszewski A, Itakura K, Riggs AD (1979) Expression in *Escherichia coli* of chemically synthesized genes for human insulin. *Proc Natl Acad Sci USA* 76:106–110
16. Feng Y, Dimitrov DS (2008) Monoclonal antibodies against components of the IGF system for cancer treatment. *Curr Opin Drug Discov Devel* 11:178–185
17. Wu H, Pfarr DS, Tang Y, An LL, Patel NK, Watkins JD, Huse WD, Kiener PA, Young JF (2005) Ultra-potent antibodies against respiratory syncytial virus: effects of binding kinetics and binding valence on viral neutralization. *J Mol Biol* 350:126–144
18. Skerra A (2007) Alternative non-antibody scaffolds for molecular recognition. *Curr Opin Biotechnol* 18:295–304
19. Chen W, Dimitrov DS (2009) Human monoclonal antibodies and engineered antibody domains as HIV-1 entry inhibitors. *Curr Opin HIV AIDS* 4:112–117
20. Chen W, Zhu Z, Feng Y, Xiao X, Dimitrov DS (2008) Construction of a large phage-displayed human antibody domain library with a scaffold based on a newly identified highly soluble, stable heavy chain variable domain. *J Mol Biol* 382:779–789
21. Chen W, Zhu Z, Feng Y, Dimitrov DS (2008) Human domain antibodies to conserved sterically restricted regions on gp120 as exceptionally potent cross-reactive HIV-1 neutralizers. *Proc Natl Acad Sci USA* 105:17121–17126
22. Chen W, Xiao X, Wang Y, Zhu Z, Dimitrov DS (2010) Bifunctional fusion proteins of the human engineered antibody domain m36 with human soluble CD4 are potent inhibitors of diverse HIV-1 isolates. *Antiviral Res* 88: 107–115
23. Dimitrov DS (2009) Engineered CH2 domains (nanoantibodies). *MAbs* 1:26–28
24. Xiao X, Feng Y, Vu BK, Ishima R, Dimitrov DS (2009) A large library based on a novel (CH2) scaffold: identification of HIV-1 inhibitors. *Biochem Biophys Res Commun* 387: 387–392
25. Gong R, Vu BK, Feng Y, Prieto DA, Dyba MA, Walsh JD, Prabakaran P, Veenstra TD, Tarasov SG, Ishima R, Dimitrov DS (2009) Engineered human antibody constant domains with increased stability. *J Biol Chem* 284: 14203–14210
26. Gong R, Wang Y, Feng Y, Zhao Q, Dimitrov DS (2011) Shortened engineered human antibody CH2 domains: increased stability and binding to the human neonatal receptor. *J Biol Chem* 286(3):27288–27293
27. Jiang XR, Song A, Bergelson S, Arroll T, Parekh B, May K, Chung S, Strouse R, Mire-Sluis A, Scheneman M (2011) Advances in the assessment and control of the effector functions of therapeutic antibodies. *Nat Rev Drug Discov* 10:101–111
28. Descotes J, Gouraud A (2008) Clinical immunotoxicity of therapeutic proteins. *Expert Opin Drug Metab Toxicol* 4:1537–1549
29. Suntharalingam G, Perry MR, Ward S, Brett SJ, Castello-Cortes A, Brunner MD, Panoskaltsis N (2006) Cytokine storm in a phase 1 trial of the anti-CD28 monoclonal antibody TGN1412. *N Engl J Med* 355: 1018–1028
30. Ewer SM, Ewer MS (2008) Cardiotoxicity profile of trastuzumab. *Drug Saf* 31:459–467
31. Chung CH (2008) Managing premedications and the risk for reactions to infusional monoclonal antibody therapy. *Oncologist* 13:725–732
32. Winter MC, Hancock BW (2009) Ten years of rituximab in NHL. *Expert Opin Drug Saf* 8:223–235
33. Pendley C, Schantz A, Wagner C (2003) Immunogenicity of therapeutic monoclonal antibodies. *Curr Opin Mol Ther* 5:172–179
34. Schellekens H (2008) How to predict and prevent the immunogenicity of therapeutic proteins. *Biotechnol Annu Rev* 14:191–202

35. Onda M (2009) Reducing the immunogenicity of protein therapeutics. *Curr Drug Targets* 10:131–139
36. Baker MP, Jones TD (2007) Identification and removal of immunogenicity in therapeutic proteins. *Curr Opin Drug Discov Devel* 10:219–227
37. Stas P, Lasters I (2009) Strategies for preclinical immunogenicity assessment of protein therapeutics. *IDrugs* 12:169–173
38. De Groot AS, McMurry J, Moise L (2008) Prediction of immunogenicity: in silico paradigms, ex vivo and in vivo correlates. *Curr Opin Pharmacol* 8:620–626
39. Emi AN, de Carvalho JF, Artur Almeida SC, Bonfa E (2010) Immunogenicity of Anti-TNF-alpha agents in autoimmune diseases. *Clin Rev Allergy Immunol* 38(2–3):82–89
40. Hangartner L, Zinkernagel RM, Hengartner H (2006) Antiviral antibody responses: the two extremes of a wide spectrum. *Nat Rev Immunol* 6:231–243
41. Dixit R, Coats S (2009) Preclinical efficacy and safety models for mAbs: the challenge of developing effective model systems. *IDrugs* 12:103–108
42. Castillo J, Winer E, Quesenberry P (2008) Newer monoclonal antibodies for hematological malignancies. *Exp Hematol* 36:755–768
43. Mariani G, Fasolo A, De BE, Gianni L (2009) Trastuzumab as adjuvant systemic therapy for HER2-positive breast cancer. *Nat Clin Pract Oncol* 6:93–104
44. Bedard PL, Piccart-Gebhart MJ (2008) Current paradigms for the use of HER2-targeted therapy in early-stage breast cancer. *Clin Breast Cancer* 8(Suppl 4):S157–S165
45. Hall PS, Cameron DA (2009) Current perspective—trastuzumab. *Eur J Cancer* 45:12–18
46. Cameron D, Casey M, Press M, Lindquist D, Pienkowski T, Romieu CG, Chan S, Jagiello-Grusfeld A, Kaufman B, Crown J, Chan A, Campone M, Viens P, Davidson N, Gorbounova V, Raats JI, Skarlos D, Newstat B, Roychowdhury D, Paoletti P, Oliva C, Rubin S, Stein S, Geyer CE (2008) A phase III randomized comparison of lapatinib plus capecitabine versus capecitabine alone in women with advanced breast cancer that has progressed on trastuzumab: updated efficacy and biomarker analyses. *Breast Cancer Res Treat* 112:533–543
47. Norden AD, Drappatz J, Wen PY (2008) Novel anti-angiogenic therapies for malignant gliomas. *Lancet Neurol* 7:1152–1160
48. Sachdev JC, Jahanzeb M (2008) Evolution of bevacizumab-based therapy in the management of breast cancer. *Clin Breast Cancer* 8: 402–410
49. Patel DK (2008) Clinical use of anti-epidermal growth factor receptor monoclonal antibodies in metastatic colorectal cancer. *Pharmacotherapy* 28:31S–41S
50. Tashev DV, Cheung NK (2009) Monoclonal antibody therapies for solid tumors. *Expert Opin Biol Ther* 9:341–353
51. Hopper-Borge EA, Nasto RE, Ratushny V, Weiner LM, Golemis EA, Astsaturov I (2009) Mechanisms of tumor resistance to EGFR-targeted therapies. *Expert Opin Ther Targets* 13:339–362
52. Presta LG (2008) Molecular engineering and design of therapeutic antibodies. *Curr Opin Immunol* 20:460–470
53. Jefferis R (2009) Glycosylation as a strategy to improve antibody-based therapeutics. *Nat Rev Drug Discov* 8:226–234
54. Labrijn AF, Aalberse RC, Schuurman J (2008) When binding is enough: nonactivating antibody formats. *Curr Opin Immunol* 20: 479–485
55. Kiewe P, Thiel E (2008) Ertumaxomab: a trifunctional antibody for breast cancer treatment. *Expert Opin Investig Drugs* 17:1553–1558
56. Weber J (2009) Ipilimumab: controversies in its development, utility and autoimmune adverse events. *Cancer Immunol Immunother* 58:823–830
57. Yuan J, Gnjatic S, Li H, Powel S, Gallardo HF, Ritter E, Ku GY, Jungbluth AA, Segal NH, Rasalan TS, Manukian G, Xu Y, Roman RA, Terzulli SL, Heywood M, Pogoriler E, Ritter G, Old LJ, Allison JP, Wolchok JD (2008) CTLA-4 blockade enhances polyfunctional NY-ESO-1 specific T cell responses in metastatic melanoma patients with clinical benefit. *Proc Natl Acad Sci USA* 105:20410–20415
58. Bellail AC, Qi L, Mulligan P, Chhabra V, Hao C (2009) TRAIL agonists on clinical trials for cancer therapy: the promises and the challenges. *Rev Recent Clin Trials* 4:34–41
59. Feng Y, Xiao X, Zhu Z, Dimitrov DS (2010) Identification and characterization of a novel agonistic anti-DR4 human monoclonal antibody. *MAbs* 2:565–570
60. Zhang MY, Shu Y, Phogat S, Xiao X, Cham F, Bouma P, Choudhary A, Feng YR, Sanz I, Rybak S, Broder CC, Quinnan GV, Evans T, Dimitrov DS (2003) Broadly cross-reactive HIV neutralizing human monoclonal antibody Fab selected by sequential antigen panning of a phage display library. *J Immunol Methods* 283:17–25

61. Choudhry V, Zhang MY, Sidorov IA, Louis JM, Harris I, Dimitrov AS, Bouma P, Cham F, Choudhary A, Rybak SM, Fouts T, Montefiori DC, Broder CC, Quinnan GV Jr, Dimitrov DS (2007) Cross-reactive HIV-1 neutralizing monoclonal antibodies selected by screening of an immune human phage library against an envelope glycoprotein (gp140) isolated from a patient (R2) with broadly HIV-1 neutralizing antibodies. *Virology* 363:79–90
62. Zhang MY, Dimitrov DS (2007) Novel approaches for identification of broadly cross-reactive HIV-1 neutralizing human monoclonal antibodies and improvement of their potency. *Curr Pharm Des* 13:203–212
63. Liu H, Gaza-Bulseco G, Faldu D, Chumsae C, Sun J (2008) Heterogeneity of monoclonal antibodies. *J Pharm Sci* 97:2426–2447
64. Birch JR, Racher AJ (2006) Antibody production. *Adv Drug Deliv Rev* 58:671–685
65. Zhou JX, Tressel T, Yang X, Seewoester T (2008) Implementation of advanced technologies in commercial monoclonal antibody production. *Biotechnol J* 3:1185–1200

Chapter 2

Synthetic Antibody Libraries

Bryce Nelson and Sachdev S. Sidhu

Abstract

Synthetic antibody libraries are constructed using designed synthetic DNA that facilitates the use of highly optimized human frameworks and enables the introduction of defined chemical diversity at positions that are most likely to contribute to antigen recognition. Using a relatively simple design based on a single human framework into which diversity is restricted to four complementarity-determining regions and two amino acids (tyrosine and serine), these synthetic antibody libraries are capable of generating specific antibodies against a diverse range of protein antigens. Moreover, by using the methods described here, more complex libraries can be constructed that are able to produce synthetic antibodies with affinities and specificities beyond the capacity of natural antibodies. Since these methods rely entirely upon standard supplies, equipment, and methods, construction of such libraries can be performed by any molecular biology laboratory.

Key words: Synthetic antibodies, Phage display, Synthetic libraries, Protein engineering

1. Introduction

Phage display has been in existence for more than 20 years with a strong track record of generating and improving upon antibodies that would be difficult, if not impossible, to isolate using traditional methods (1–3). This is due in large part to the fact that standard molecular biology techniques can be used to build highly diverse libraries consisting of billions of members that can then be used en masse to screen antigens under controlled in vitro conditions to select antibodies with particular binding specificities. Importantly, since the sequence of each antibody can be decoded from the cognate DNA, affinities and specificities can be later adjusted for particular applications. Furthermore, the recombinant nature of the technology allows for facile reformatting of phage-displayed antibody fragments for the production of full-length

immunoglobulins. Thus, phage display circumvents the need for animal immunization and allows for the rapid generation of recombinant antibodies with specificity to virtually any protein of interest.

Initial phage display libraries consisted of natural antibody repertoires amplified from human immune tissues transferred into phage display vectors. While this approach to library construction remains common, advances in our knowledge of antibody structure and function have enabled the construction of “synthetic” antibody libraries with diversities rivaling or exceeding those of natural immune repertoires (4). These synthetic antibody libraries are constructed through the introduction of degenerate DNA precisely into regions encoding the complementarity-determining regions (CDRs) that make up the antigen-binding site within defined variable domain frameworks. There are numerous advantages offered by synthetic libraries beyond vast diversity. Natural repertoires lack antibodies against self-antigens, whereas synthetic antibodies are completely naïve and are not biased against natural proteins, regardless of source or sequence. Frameworks for synthetic antibodies can be chosen for particular properties including, for example, high stability and expression. Similarly, for therapeutic applications, optimized human frameworks can be used to minimize the risk of immunogenicity, thus obviating the need for humanization. Lastly, design features can be incorporated to allow facile affinity maturation and adaptation to a high-throughput pipeline.

In order to construct a phage display library, specialized vectors such as phagemids that can be easily modified to allow production of protein for purification and analysis are used (Fig. 1). Two origins of replication are contained within a phagemid: a double-stranded DNA origin (dsDNA ori) allows replication as a plasmid in *E. coli* while a filamentous phage origin (f1 ori) enables packaging of single-stranded DNA (ssDNA) into phage particles. Our optimized phagemid has been used for the display of Fabs (5), scFvs (6), V_H domains (7), as well as peptides and a variety of other proteins (8–10). The C-terminal domain of the M13 bacteriophage gene-3 minor coat protein (Protein-3, P3) is fused directly to monomeric scFv or V_H domain to achieve phage display, whereas heterodimeric Fabs require bicistronic expression. Secretion signals direct P3-fused heavy chain and independently expressed light chain to the periplasm, where spontaneous assembly of Fabs occurs. In our Fab system, open reading frames are under the control of the inducible alkaline phosphatase promoter (PhoA) due to the ease of downstream Fab expression and purification while our scFv and V_H domain libraries are controlled by the IPTG-inducible P_{tac} promoter.

We have sought to develop simplified synthetic antibody libraries because we believe that simple designs can enhance our understanding of antibody function and facilitate uptake of the technology

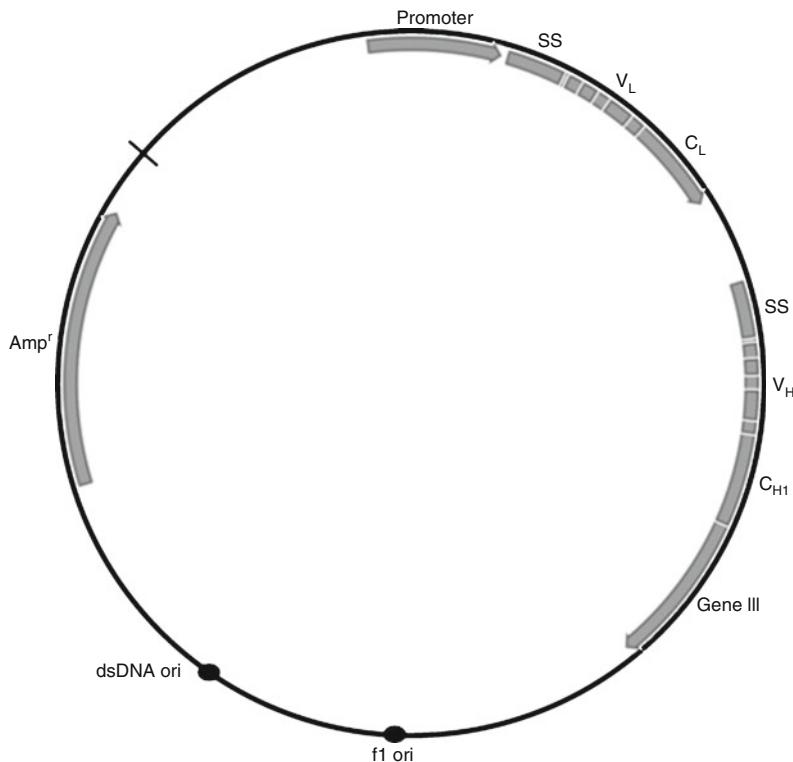


Fig. 1. Phagemid design for phage display. Expression of a bi-cistronic message encoding the light chain (V_L - C_L) and the variable and first constant domains of the heavy chain (V_H - C_{H1}) fused to a truncated gene III coat protein is required for Fab-phage display. N-terminal secretion signals direct the proteins to the periplasm, where light and heavy chains associate to form Fabs. The phagemid also contains origins of single-stranded (f1 ori) and double-stranded (dsDNA ori) DNA replication, as well as a selectable marker (Amp^r) that confers resistance to carbenicillin.

by other researchers. Our work has shown that remarkably diverse antibody functions can be supported by a single framework based on the highly stable therapeutic antibody humanized 4D5 (Fig. 2) (9–13). Moreover, high-affinity antibodies can be generated against most antigens by introducing diversity into only a subset of positions within four of the six CDRs. Most surprisingly, we have been able to restrict chemical diversity without compromising function (14–17), and in the extreme case, we have shown that a binary code of tyrosine and serine is sufficient for generating antigen-binding sites capable of recognizing diverse proteins (18).

This reductionist approach has led to minimalist methodologies that can be readily replicated in any molecular biology laboratory using standard equipment, supplies, and techniques. Here, we provide a complete set of methods that will enable the construction of a simplified synthetic antibody library with binary diversity (Y/S) introduced into four CDRs of a single human framework. Importantly, these methods enable even nonexpert researchers to develop functional synthetic libraries that can yield antibodies

a**V_L**

gatatccagatgaccaggactcccccggagctccctgtccgcctctgtggcgatagggtcacc
 D I Q M T Q S P S S L S A S V G D R V T

atcacctgcgtgcaggatgtgtccactgtctgtggatcaacagaaaacca
 I T C R A S Q D V S T A V A W Y Q Q K P

ggaaaagctccgaagcttctgatttaactcggcatccttcctactctggagttcccttct
 G K A P K L L I Y S A S F L Y S G V P S

cgcttcctggtagcggttccggacggatttactgtccactgtccactcggatcagcgtctgcagccg
 R F S G S G S G T D F T L T I S S L Q P

CDR-L3

gaagacttcgcaacttattactgtcagaacttataactactcctccacgttccggacag
 E D F A T Y Y C Q Q S Y T T P P T F G Q

ggtacccaagggtggagatcaa
 G T K V E I K

V_H

gaggttcagctggggagtctggcggtggcctgggtcagccaggggctactccgttg
 E V Q L V E S G G G L V Q P G G S L R L

CDR-H1

tccctgtcagttctggcttcaacatataagacacactataactgggtgcgtcaggcc
 S C A A S G F N I K D T Y I H W V R Q A

CDR-H2

ccgggttaaggggccttggaaatgggttgcaggatttatcatacgaaatggttataacttagat
 P G K G L E W V A R I Y P T N G Y T R Y

CDR-H3

ctacaatgaacagcttaagagctgaggacactgccgtctattattgttagccgtggggaa
 L Q M N S L R A E D T A V Y Y C S R W G

ggggacggcttctatgtatggactactgggtcaagggaaactagtccacgtctcc
 G D G F Y A M D Y W G Q G T L V T V S S

b

CDR-L3: tattactgtcagcaatmttmttmttmttccacgttcgga
 Y Y C Q Q X X X X P P T F G

CDR-H1: tctggcttcaacattttttttttttttatacactgggtgcgt
 S G F N I X X X X I H W V R

CDR-H2: ctggaatgggttgcattatgccatagcgtc
 L E W V A X I X P X X G X T X Y A D S V

CDR-H3: tattattgtagccgc(tmt)_ngctatggactactgg (n=3-16)
 Y Y C S R (X)_n A M D Y W

Fig. 2. Synthetic antibody library design with a binary code. (a) Nucleotide (*lower case*) and protein sequences (*upper case*) for the light (V_L) and heavy (V_H) variable domains of the humanized 4D5 antibody. Diversified CDRs are denoted by boxes and diversified positions are shaded grey. (b) Sequences of mutagenic oligonucleotides for library construction. Diversified positions (Tyr/Ser in equal proportions) encoded by degenerate "tmt" codons ($m = a/c$ in equal proportions) are denoted by "X" and flanked on either side by 15 bases that anneal perfectly to the surrounding sequences. To introduce length diversity into CDR-H3, a mixture of 14 oligonucleotides is used to obtain an equimolar mixture of all possible lengths with 3–16 "tmt" degenerate codons. DNA sequences are shown in the 5'-3' orientation.

against diverse antigens. Once mastered, the same methods can be used to augment the minimalist design with additional diversity to produce libraries capable of generating synthetic antibodies that rival or even exceed the functional capacity of natural antibodies.

2. Materials

1. 0.2-cm Gap electroporation cuvette.
2. 1.0 M H₃PO₄.
3. 1.0 M Tris-base, pH 8.0.
4. 1.0 mM Hepes, pH 7.4 (4.0 ml of 1.0 M Hepes, pH 7.4 in 4.0 L of ultrapure irrigation USP water, filter sterilize).
5. 3,3',5,5'-Tetramethylbenzidine/H₂O₂ peroxidase (TMB) substrate.
6. 10% (v/v) Ultrapure glycerol (100 ml ultrapure glycerol in 900 ml ultrapure irrigation USP water, filter sterilize).
7. 10 mM ATP.
8. 10× PCR buffer (600 mM Tris-HCl, pH 8.3, 250 mM KCl, 15 mM MgCl₂, 1% Triton X-100, 100 mM β-mercaptoethanol).
9. 10× TM buffer (0.1 M MgCl₂, 0.5 M Tris, pH 7.5).
10. 100 mM dNTP mix (solution containing 25 mM each of dATP, dCTP, dGTP, dTTP).
11. 96-Well Maxisorp immunoplates.
12. 96-Well microtubes.
13. 100 mM HCl.
14. 100 mM Dithiothreitol (DTT).
15. 2YT medium. (10 g bacto-yeast extract, 16 g bacto-trypitone, 5 g NaCl. Add water to 1.0 l; adjust pH to 7.0 with NaOH; and autoclave.)
16. 2YT/carb/cmp medium (2YT, 100 µg/ml carbenicillin, 5 µg/ml chloramphenicol).
17. 2YT/carb/kan medium (2YT, 100 µg/ml carbenicillin, 25 µg/ml kanamycin).
18. 2YT/carb/kan/uridine medium (2YT, 100 µg/ml carbenicillin, 25 µg/ml kanamycin, 0.25 µg/ml uridine).
19. 2YT/carb/KO7 medium (2YT, 100 µg/ml carbenicillin, 10¹⁰ M13KO7-phage/ml).
20. 2YT/tet medium (2YT, 10 µg/ml tetracycline).
21. 2YT/carb/tet medium (2YT, 100 µg/ml carbenicillin, 10 µg/ml tetracycline).

22. 2YT/carb/tet/KO7 (2YT, 100 µg/ml carbenicillin, 10 µg/ml tetracycline, 10¹⁰ M13KO7-phage/ml).
23. 2YT/kan medium (2YT, 25 µg/ml kanamycin).
24. 2YT/kan/tet medium (2YT, 25 µg/ml kanamycin, 10 µg/ml tetracycline).
25. 2YT top agar. (16 g tryptone, 10 g yeast extract, 5 g NaCl, 7.5 g granulated agar. Add water to 1.0 L and adjust pH to 7.0 with NaOH, heat to dissolve, and autoclave.)
26. AmpliTaq DNA polymerase.
27. ECM-600 electroporator.
28. *E. coli* CJ236 (New England Biolabs, Beverly, MA).
29. *E. coli* SS320 (Lucigen, Middleton, WI).
30. *E. coli* XL1-blue (Agilent Technologies, Santa Clara, CA).
31. Horseradish peroxidase/anti-M13 antibody conjugate.
32. LB/carb plates (LB agar, 50 µg/ml carbenicillin).
33. LB/tet plates (LB agar, 5 µg/ml tetracycline).
34. M13KO7 helper phage (New England Biolabs, Ipswich, MA).
35. Magnetic stir bars (2 in.) soaked in ethanol.
36. Phosphate-buffered saline (PBS). (137 mM NaCl, 3 mM KCl, 8 mM Na₂HPO₄, 1.5 mM KH₂PO₄. Adjust pH to 7.2 with HCl, and autoclave.)
37. PBS, 0.2% bovine serum albumin (BSA).
38. PEG/NaCl. (20% PEG-8000 (w/v), 2.5 M NaCl. Mix and filter sterilize.)
39. QIAprep Spin M13 Kit (Qiagen, Valencia, CA).
40. QIAquick Gel Extraction Kit (Qiagen, Valencia, CA).
41. SOC medium. (5 g bacto-yeast extract, 20 g bacto-tryptone, 0.5 g NaCl, 0.2 g KCl. Add water to 1.0 L, adjust pH to 7.0 with NaOH, and autoclave; add 5.0 ml of autoclaved 2.0 M MgCl₂ and 20 ml of filter-sterilized 1.0 M glucose).
42. Superbroth medium (12 g tryptone, 24 g yeast extract, 5 ml glycerol; add water to 900 ml, autoclave, and add 100 ml of autoclaved 0.17 M KH₂PO₄, 0.72 M K₂HPO₄).
43. Superbroth/tet/kan medium (Superbroth medium, 10 µg/ml tetracycline, 25 µg/ml kanamycin).
44. T4 polynucleotide kinase.
45. T4 DNA ligase.
46. T7 DNA polymerase.
47. TAE buffer (40 mM Tris-acetate, 1.0 mM EDTA; adjust pH to 8.0; autoclave).

48. TAE/agarose gel (TAE buffer, 1.0% (w/v) agarose, 1:5,000 (v/v) 10% ethidium bromide).
49. Ultrapure irrigation USP water.
50. Uridine (0.25 mg/ml in water, filter sterilize).
51. Ultrapure glycerol.

3. Methods

The following sections feature optimized protocols for the construction of phage-displayed libraries containing over 10^{10} unique antibodies. Here, a parental antibody framework is displayed using a phagemid that can be modified subsequently to introduce appropriate genetic diversity into the CDRs. This genetic library is then passed through an *E. coli* host, thereby converting it to a phage-displayed protein library that can be used in selections to isolate antigen-specific clones (see Note 1).

3.1. Purification of dU-ssDNA Template (see Note 2)

1. From a fresh LB/carb plate, pick a single colony of *E. coli* CJ236 (or another *dut/unq* strain) containing the appropriate phagemid into 1 ml of 2YT medium supplemented with M13KO7 helper phage (10^{10} pfu/ml) and appropriate antibiotics to maintain the host F' episome and the phagemid. For example, 2YT/carb/cmp medium contains carbenicillin to select for phagemids and chloramphenicol to select for the F' episome of *E. coli* CJ236.
2. Shake at 200 rpm and 37°C for 2 h before addition of kanamycin (25 µg/ml) to select for clones co-infected with M13KO7 (which carries a kanamycin resistance gene).
3. Shake at 200 rpm and 37°C for another 6 h and transfer the culture to 30 ml of 2YT/carb/kan/uridine medium.
4. Shake for 20 h at 200 rpm and 37°C.
5. Pellet bacterial cells by centrifuging for 10 min at $27,000 \times g$ and 4°C in a Sorvall SS-34 rotor. Transfer the phage-containing supernatant to a new tube containing 1/5 final volume of PEG/NaCl and incubate for 5 min at room temperature.
6. Centrifuge 15 min at $27,000 \times g$ and 4°C in an SS-34 rotor. Decant the supernatant; centrifuge briefly at $2,000 \times g$ and aspirate the remaining supernatant. Be sure to use barrier-tips when handling phage to avoid pipetman contamination.
7. Resuspend the phage pellet in 0.5 ml of PBS and transfer to a 1.5-ml microcentrifuge tube.
8. Centrifuge for 5 min at $15,000 \times g$ in a microcentrifuge, and transfer the supernatant to a fresh 1.5-ml microcentrifuge tube.

9. Add 7.0 μ l of buffer MP (Qiagen) and mix. Incubate at room temperature for at least 2 min.
10. Apply the sample to a QIAprep spin column (Qiagen) in a 2-ml microcentrifuge tube. Centrifuge for 30 s at $6,000 \times g$ in a microcentrifuge. Discard the flow through. The phage particles remain bound to the column matrix.
11. Add 0.7 ml of buffer MLB (Qiagen) to the column. Centrifuge for 30 s at $6,000 \times g$ and discard the flow through.
12. Add another 0.7 ml of buffer MLB and incubate at room temperature for at least 1 min.
13. Centrifuge at $6,000 \times g$ for 30 s. Discard the flow through. The DNA is separated from the protein coat and remains adsorbed to the matrix.
14. Add 0.7 ml of buffer PE (Qiagen). Centrifuge at $6,000 \times g$ for 30 s and discard the flow through.
15. Repeat step 14. Residual proteins and salt are removed.
16. Centrifuge the column at $6,000 \times g$ for 30 s to remove residual PE buffer.
17. Transfer the column to a fresh 1.5-ml microcentrifuge tube.
18. Add 100 μ l of buffer EB (Qiagen; 10 mM Tris–Cl, pH 8.5) to the center of the column membrane. Incubate at room temperature for 10 min.
19. Centrifuge for 30 s at $6,000 \times g$. Save the eluant, which contains the purified dU-ssDNA.
20. Analyze the DNA by electrophoresing 1.0 μ l on a TAE/agarose gel. The DNA should appear as a predominant single band, but faint bands with lower electrophoretic mobility are often visible. These are likely caused by secondary structure in the dU-ssDNA.
21. Determine the DNA concentration by measuring absorbance at 260 nm ($A_{260} = 1.0$ for 33 ng/ μ l of ssDNA). Typical DNA concentrations range from 200 to 500 ng/ μ l.

3.2. In Vitro Synthesis of Heteroduplex CCC-dsDNA

3.2.1. Oligonucleotide Phosphorylation with T4 Polynucleotide Kinase (see Note 4)

A three-step procedure is used to incorporate the mutagenic oligonucleotides into heteroduplex covalently closed, circular, double-stranded DNA (CCC-dsDNA), using dU-ssDNA as a template (Fig. 3 and see Note 3).

1. Combine 0.6 μ g of mutagenic oligonucleotide designed to mutate a CDR (Fig. 1b) with 2.0 μ l 10x TM buffer, 2.0 μ l 10 mM ATP, and 1.0 μ l 100 mM DTT. Add water to a total volume of 20 μ l in a 1.5-ml microcentrifuge tube. Each mutagenic oligonucleotide requires a separate phosphorylation reaction.
2. Add 20 U of T4 polynucleotide kinase to each tube and incubate for 1.0 h at 37°C. Use immediately for annealing.

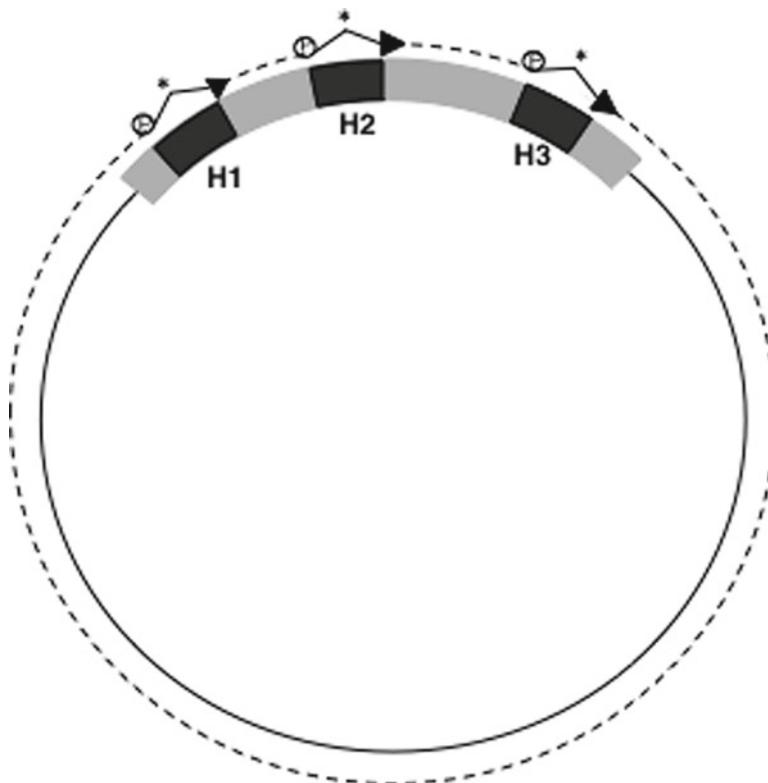


Fig. 3. Library construction by oligonucleotide-directed mutagenesis. Mutagenic phosphorylated oligonucleotides (arrows) containing mutated regions (asterisk) anneal to the three heavy chain CDRs (H1, H2, and H3). Multiple oligonucleotides can be simultaneously annealed to a dU-ssDNA template and used to prime enzymatic synthesis of heteroduplex CCC-dsDNA (dashed circle) subsequently introduced into a *dut⁺/ung⁺* *E. coli* host, where the mutated strand is preferentially amplified.

3.2.2. Annealing of the Oligonucleotides to the Template

1. To 20 µg of dU-ssDNA template, add 25 µl 10× TM buffer, 20 µl of each phosphorylated oligonucleotide, and water to a final volume of 250 µl. These DNA quantities provide an oligonucleotide:template molar ratio of 3:1, assuming that the oligonucleotide:template length ratio is 1:100.
2. Incubate at 90°C for 3 min, 50°C for 3 min, and 20°C for 5 min.

3.2.3. Enzymatic Synthesis of CCC-dsDNA

1. To the annealed oligonucleotide/template mixture, add 10 µl 10 mM ATP, 10 µl 10 mM dNTP mix, 15 µl 100 mM DTT, 30 Weiss units T4 DNA ligase, and 30 U T7 DNA polymerase.
2. Incubate overnight at 20°C.
3. Purify and desalt the DNA using the Qiagen QIAquick DNA purification kit.
4. Add 1.0 ml of buffer QG (Qiagen) and mix.

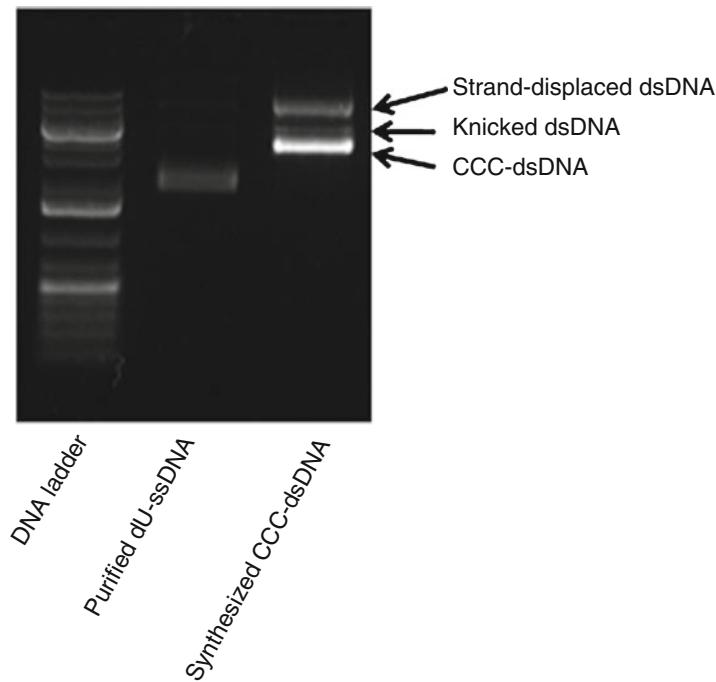


Fig. 4. In vitro synthesis of heteroduplex CCC-dsDNA. The higher mobility band in the synthesized CCC-dsDNA lane represents correctly extended and ligated CCC-dsDNA and should constitute the major product of the reaction.

5. Apply the sample over two QIAquick spin columns placed in 2-ml microcentrifuge tubes. Centrifuge at $15,000 \times g$ for 1 min in a microcentrifuge. Discard the flow through.
6. Add 750 μ l buffer PE (Qiagen) to each column, and centrifuge at $15,000 \times g$ for 1 min. Discard the flow through.
7. Centrifuge the column at $15,000 \times g$ for 1 min to remove excess buffer PE.
8. Transfer the column to a fresh 1.5-ml microcentrifuge tube and add 35 μ l of ultrapure irrigation USP water to the center of the membrane. Incubate for 2 min at room temperature.
9. Centrifuge at $15,000 \times g$ for 1 min to elute the DNA. Combine the eluants from the two columns. The DNA can be used immediately for *E. coli* electroporation or frozen for later use.
10. Electrophorese 1.0 μ l of the eluted reaction product alongside the ssDNA template (Fig. 4 and see Note 5).

3.3. Conversion of CCC-dsDNA into a Phage-Displayed Library

To complete library construction, the heteroduplex CCC-dsDNA requires introduction into an *E. coli* host containing an F' episome to enable M13 bacteriophage infection and propagation (see Note 6).

Preparation of Electrocompetent *E. coli* SS320 (see Note 7)

1. Inoculate 25 ml 2YT/tet medium with a single colony of *E. coli* SS320 from a fresh LB/tet plate. Incubate at 37°C with shaking at 200 rpm to mid-log phase ($OD_{550}=0.8$).
2. Make tenfold serial dilutions of M13K07 by diluting 20 μ l into 180 μ l of PBS (use a new pipette tip for each dilution).
3. Mix 500- μ l aliquots of mid-log phase *E. coli* SS320 with 200 μ l of each M13K07 dilution and 4 ml of 2YT top agar.
4. Pour the mixtures onto prewarmed LB/tet plates and grow overnight at 37°C.
5. Pick an average-sized single plaque and place in 1 ml of 2YT/kan/tet medium. Grow for 8 h at 37°C with shaking at 200 rpm.
6. Transfer the culture to 250 ml of 2YT/kan medium in a 2-L baffled flask. Grow overnight at 37°C with shaking at 200 rpm.
7. Inoculate six 2-L baffled flasks containing 900 ml of superbroth/tet/kan medium with 5 ml of the overnight culture. Incubate at 37°C with shaking at 200 rpm to mid-log phase ($OD_{600}=0.8$).
8. Chill three of the flasks on ice for 5 min with occasional swirling. The following steps should be done in a cold room, on ice, with prechilled solutions and equipment.
9. Centrifuge at $5,000 \times g$ and 4°C for 10 min in a Sorvall GS-3 rotor.
10. Decant the supernatant and add culture from the remaining flasks (these should be chilled while the first set is centrifuging) to the same tubes.
11. Repeat the centrifugation and decant the supernatant.
12. Fill the tubes with 1.0 mM Hepes, pH 7.4, and add sterile magnetic stir bars to facilitate pellet resuspension. Swirl to dislodge the pellet from the tube wall and stir at a moderate rate to resuspend the pellet completely.
13. Centrifuge at $5,000 \times g$ and 4°C for 10 min in a Sorvall GS-3 rotor. Decant the supernatant, being careful to retain the stir bar. To avoid disturbing the pellet, maintain the position of the centrifuge tube when removing from the rotor.
14. Repeat steps 12 and 13.
15. Resuspend each pellet in 150 ml of 10% ultrapure glycerol. Use stirbars and do not combine the pellets.
16. Centrifuge at $5,000 \times g$ and 4°C for 15 min in a Sorvall GS-3 rotor. Decant the supernatant and remove the stir bar. Remove remaining traces of supernatant with a pipette.

17. Add 3.0 ml of 10% ultrapure glycerol to one tube and resuspend the pellet by pipetting. Transfer the suspension to the next tube and repeat until all of the pellets are resuspended.
18. Transfer 350- μ l aliquots into 1.5-ml microcentrifuge tubes.
19. Flash freeze with liquid nitrogen and store at -70°C.

3.4. *E. coli* Electroporation and Phage Propagation

1. Chill the purified, desalted CCC-dsDNA (20 μ g in a maximum volume of 100 μ l) and a 0.2-cm-gap electroporation cuvette on ice.
2. Thaw a 350- μ l aliquot of electrocompetent *E. coli* SS320 on ice. Add the cells to the DNA and mix by pipetting several times (avoid introducing bubbles).
3. Transfer the mixture to the cuvette and electroporate. For electroporation, follow the manufacturer's instructions, preferably using a BTX ECM-600 electroporation system with the following settings: 2.5 kV field strength, 125 Ω resistance, and 50 μ F capacitance.
4. Immediately rescue the electroporated cells by adding 1 ml SOC medium and transferring to 10 ml SOC medium in a 250-ml baffled flask. Rinse the cuvette twice with 1 ml SOC medium. Add SOC medium to a final volume of 25 ml.
5. Incubate for 30 min at 37°C with shaking at 200 rpm.
6. To determine the library diversity, plate serial dilutions on LB/carb plates to select for the phagemid.
7. Transfer the culture to a 2-L baffled flask containing 500 ml 2YT medium, supplemented with antibiotics for phagemid and M13KO7 helper phage selection (e.g., 2YT/carb/kan medium).
8. Incubate overnight at 37°C with shaking at 200 rpm.
9. Centrifuge the culture for 10 min at 16,000 $\times g$ and 4°C in a Sorvall GSA rotor.
10. Transfer the supernatant to a fresh tube and add 1/5 final volume of PEG/NaCl solution to precipitate the phage. Incubate for 5 min at room temperature.
11. Centrifuge for 10 min at 16,000 $\times g$ and 4°C in a GSA rotor. Decant the supernatant. Spin briefly and remove the remaining supernatant with a pipette.
12. Resuspend the phage pellet in 20 ml of PBT buffer.
13. Pellet insoluble matter by centrifuging for 5 min at 27,000 $\times g$ and 4°C in an SS-34 rotor. Transfer the supernatant to a clean tube.
14. Estimate the phage concentration spectrophotometrically ($OD_{268} = 1.0$ for a solution of 5×10^{12} phage/ml).

15. The library can be used immediately for selection experiments. Alternatively, the library can be frozen and stored at -80°C, following the addition of glycerol to a final concentration of 10%.

4. Notes

1. By using optimized procedures ([5](#), [6](#), [19](#)) that are based on the classical oligonucleotide-directed mutagenesis method of Kunkel et al. (Fig. 3) ([20](#)), very large phage-displayed antibody repertoires ($>10^{10}$ members) can be constructed quite rapidly. Importantly, the method is scalable and can be used to mutate up to four independent regions concurrently with very high efficiency. First, a *dut⁻/ung* *E. coli* host is used to propagate uracil-containing ssDNA (dU-ssDNA) template to which mutagenic oligonucleotides are annealed. “Stop templates” contain stop codons in the CDRs intended to be randomized and ensure that only mutated antibodies are displayed; non-mutated parental “stop template” will fail to express functional fusion polypeptides leading to clearing from the pool during selections.
2. Library construction mutagenesis efficiency depends on template purity; therefore, the use of highly pure dU-ssDNA is essential for successful library construction. By using a modified Qiagen QIAprep Spin M13 Kit protocol for dU-ssDNA purification, at least 20 µg of dU-ssDNA for a phagemid with medium copy number (e.g., pBR322 backbone) will be isolated, which is sufficient for the construction of one library.
3. The protocol described here is an optimized, large-scale version of a published method ([20](#)). The first step involves phosphorylation of the oligonucleotide that is subsequently annealed to the dU-ssDNA template. This annealed, phosphorylated oligonucleotide primes the template for enzymatic extension of the entire template sequence. Ligation results in the formation of ~20 µg of highly pure, low-conductance, heteroduplex CCC-dsDNA that will require purification and desalting before electroporation (Fig. 3).
4. Precise control over library design can be introduced through design of the mutagenic oligonucleotides that can contain degenerate codons to introduce complex diversity that is biased in favor of amino acids that are common in natural antibodies ([5](#), [6](#)) or are particularly well suited for antigen recognition ([17](#), [18](#)). With the mutagenic oligonucleotides annealed and serving as primers, the synthesis of a complementary DNA strand subsequently ligated to form a CCC-dsDNA heteroduplex

can occur. Lastly, electroporation is used to introduce this CCC-dsDNA heteroduplex into a *dut⁻/ung⁺* *E. coli* host that preferentially inactivates the uracil-containing template strand resulting in efficient mutagenesis (>80%).

5. A successful reaction results in the complete conversion of ssDNA to dsDNA, which has a lower electrophoretic mobility. No ssDNA should remain and usually at least two product bands are visible. The product band with higher electrophoretic mobility represents the desired product: correctly extended and ligated CCC-dsDNA that transforms *E. coli* efficiently and provides a high mutation frequency (~80%). The product band with lower electrophoretic mobility is a strand-displaced product resulting from intrinsic, unwanted activity of T7 DNA polymerase (21) and provides a low mutation frequency (~20%) that transforms *E. coli* at least 30-fold less efficiently than CCC-dsDNA. If a significant proportion of the single-stranded template is converted to CCC-dsDNA, a highly diverse library with high mutation frequency will result. Sometimes, a third band is visible, with an electrophoretic mobility between the other two product bands. This intermediate band is correctly extended but contains unligated dsDNA (knicked dsDNA) resulting from either insufficient T4 DNA ligase activity or from incomplete oligonucleotide phosphorylation.
6. The limiting factor for phage-displayed library diversities are the methods for introducing DNA into *E. coli*, with the most efficient method being high-voltage electroporation. We have constructed an *E. coli* strain (SS320) that is ideal for both high-efficiency electroporation and phage production (8). The following optimized protocols enable the production of high-diversity libraries by the large-scale electroporation of CCC-dsDNA into specially prepared electrocompetent *E. coli* SS320 infected by M13KO7 helper phage. Transformation into an *E. coli* host results in phagemid replication as a double-stranded plasmid. Upon coinfection with helper phage, ssDNA replication is initiated and phagemid ssDNA is packaged into phage particles containing phagemid-encoded protein, thereby providing linkage to their encoding DNA. Helper phage, such as M13KO7, provides all proteins necessary for assembly of phage virions also containing phagemid-encoded coat protein. These phage particles can then be used for library selections.
7. The following protocol yields approximately 12 ml of highly concentrated, electrocompetent *E. coli* SS320 (~3 × 10¹¹ cfu/ml) infected by M13KO7 helper phage that can be stored indefinitely at -70°C in 10% glycerol. The use of *E. coli* infected by helper phage ensures that, once transformed with a phagemid, each cell will be able to produce phage particles without the need for further helper phage infection.

References

1. Bradbury AR, Sidhu S, Dubel S, McCafferty J (2011) Beyond natural antibodies: the power of in vitro display technologies. *Nat Biotechnol* 29:245–254
2. Hoogenboom HR (2005) Selecting and screening recombinant antibody libraries. *Nat Biotechnol* 23:1105–1116
3. Bradbury AR, Marks JD (2004) Antibodies from phage antibody libraries. *J Immunol Methods* 290:29–49
4. Sidhu SS, Fellouse FA (2006) Synthetic therapeutic antibodies. *Nat Chem Biol* 2:682–688
5. Lee CV, Liang WC, Dennis MS, Eigenbrot C, Sidhu SS, Fuh G (2004) High-affinity human antibodies from phage-displayed synthetic Fab libraries with a single framework scaffold. *J Mol Biol* 340:1073–1093
6. Sidhu SS, Li B, Chen Y, Fellouse FA, Eigenbrot C, Fuh G (2004) Phage-displayed antibody libraries of synthetic heavy chain complementarity determining regions. *J Mol Biol* 338:299–310
7. Bond CJ, Wiesmann C, Marsters JC Jr, Sidhu SS (2005) A structure-based database of antibody variable domain diversity. *J Mol Biol* 348:699–709
8. Sidhu SS, Lowman HB, Cunningham BC, Wells JA (2000) Phage display for selection of novel binding peptides. *Methods Enzymol* 328:333–363
9. Gao J, Sidhu SS, Wells JA (2009) Two-state selection of conformation-specific antibodies. *Proc Natl Acad Sci USA* 106:3071–3076
10. Newton K, Matsumoto ML, Wertz IE, Kirkpatrick DS, Lill JR, Tan J, Dugger D, Gordon N, Sidhu SS, Fellouse FA, Komuves L, French DM, Ferrando RE, Lam C, Compaan D, Yu C, Bosanac I, Hymowitz SG, Kelley RF, Dixit VM (2008) Ubiquitin chain editing revealed by polyubiquitin linkage-specific antibodies. *Cell* 134:668–678
11. Ye JD, Tereshko V, Frederiksen JK, Koide A, Fellouse FA, Sidhu SS, Koide S, Kossiakoff AA, Piccirilli JA (2008) Synthetic antibodies for specific recognition and crystallization of structured RNA. *Proc Natl Acad Sci USA* 105:82–87
12. Uysal S, Vasquez V, Tereshko V, Esaki K, Fellouse FA, Sidhu SS, Koide S, Perozo E, Kossiakoff A (2009) Crystal structure of full-length KcsA in its closed conformation. *Proc Natl Acad Sci USA* 106:6644–6649
13. Fellouse FA, Esaki K, Birtalan S, Raptis D, Cancasci VJ, Koide A, Jhurani P, Vasser M, Wiesmann C, Kossiakoff AA, Koide S, Sidhu SS (2007) High-throughput generation of synthetic antibodies from highly functional minimalist phage-displayed libraries. *J Mol Biol* 373:924–940
14. Birtalan S, Zhang Y, Fellouse FA, Shao L, Schaefer G, Sidhu SS (2008) The intrinsic contributions of tyrosine, serine, glycine and arginine to the affinity and specificity of antibodies. *J Mol Biol* 377:1518–1528
15. Fisher RD, Ultsch M, Lingel A, Schaefer G, Shao L, Birtalan S, Sidhu SS, Eigenbrot C (2010) Structure of the complex between HER2 and an antibody paratope formed by side chains from tryptophan and serine. *J Mol Biol* 402:217–229
16. Birtalan S, Fisher RD, Sidhu SS (2010) The functional capacity of the natural amino acids for molecular recognition. *Mol Biosyst* 6:1186–1194
17. Fellouse FA, Wiesmann C, Sidhu SS (2004) Synthetic antibodies from a four-amino-acid code: a dominant role for tyrosine in antigen recognition. *Proc Natl Acad Sci USA* 101:12467–12472
18. Fellouse FA, Li B, Compaan DM, Peden AA, Hymowitz SG, Sidhu SS (2005) Molecular recognition by a binary code. *J Mol Biol* 348:1153–1162
19. Sidhu SS (2000) Phage display in pharmaceutical biotechnology. *Curr Opin Biotechnol* 11:610–616
20. Kunkel TA, Roberts JD, Zakour RA (1987) Rapid and efficient site-specific mutagenesis without phenotypic selection. *Methods Enzymol* 154:367–382
21. Lechner RL, Engler MJ, Richardson CC (1983) Characterization of strand displacement synthesis catalyzed by bacteriophage T7 DNA polymerase. *J Biol Chem* 258:11174–11184

Chapter 3

The Construction of “Phylomer” Peptide Libraries as a Rich Source of Potent Inhibitors of Protein/Protein Interactions

Nadia Milech and Paul Watt

Abstract

Phylomer libraries are made from random overlapping genome fragments of biodiverse bacteria and *Archaea*. They provide a rich source of high-affinity binders to protein interfaces, and can be used both for target-directed screening approaches and for phenotypic screens to discover new targets. Here, we describe methods used for the construction of a Phylomer library, illustrated by examples of construction in both a yeast two-hybrid vector and a phage display vector.

Key words: Peptide, Library, Phylomer, Inhibitor

1. Introduction

Peptides can be a useful tool for blocking protein–protein interactions (PPIs). However, the quality and quantity of hits derived from libraries of randomly encoded sequences are typically low due to the rarity of appropriate structures in such random libraries, even when conformationally constrained. An alternative approach is to exploit evolution as a source of raw material from which to capture structure-rich peptides capable of binding protein interfaces with high affinity.

Protein folds in prokaryotes and eukaryotes may have evolved by assembly of ancient peptide modules known as “antecedent domain segments,” found in all three current reading frames which are encoded by proto-exon sequences distributed throughout eubacterial and archaeal genomes ([1](#), [2](#)). These minimal structural units (~15–30 amino acids) are thought to specify pro-selective functions, including propensity to interact with other proteins, sometimes even in the absence of intrinsic tertiary structure (see ([3–6](#))). We have

hypothesized that screening phylogenetically diverse gene fragments could enrich for the common secondary structures and subdomains required for interaction with protein targets (7).

We constructed peptide libraries from such biodiverse genomic material which we refer to as “Phylomer libraries” (7). These libraries can indeed provide a rich source of target binders, presumably through containing more privileged structures (8), with evidence from circular dichroism experiments suggesting that at least the secondary structural elements from synthetic Phylomer peptides are typically maintained out of the context of their parent proteins. Since peptide structures on the surface of protein targets often recapitulate supersecondary structures found within monomeric proteins (9), this may explain why Phylomer libraries can be such a rich source of inhibitors of protein interactions.

Many Phylomer peptides lack disulfide bonds, including those derived from thermostable proteins. This enables those Phylomers to be functional in reducing environments, such as inside cells. Primary Phylomer hits can show potent biological activity *in vitro* and *in vivo*, against intracellular or extracellular targets prior to affinity maturation of their sequences, suggesting that they may prove useful in the capture of putative targets using phenotypic screening. These target candidates can then be identified via affinity purification and mass spectrometry, and also used as “protein interference” probes in functional target validation experiments (8).

As described here, Phylomer libraries are constructed by “shotgun cloning” of small genomic or cDNA fragments from biodiverse microorganisms into expression vectors to allow genetic screening against extracellular or intracellular targets using standard techniques (phage display, yeast two-hybrid, or also phenotypic screening). These Phylomer peptide modules (typically 15–50 amino acids) can be joined synthetically or by recombinant fusion to further enhance target binding through avidity or local concentration effects. The fact that many of the genes encoding Phylomer peptides are encoded by thermophilic *Archaea* and bacteria which have evolved to live in extreme environments, such as deep sea volcanic vents and geysers, allows the capture of thermally stable structures which may enhance their biological efficacy. This chapter describes methods for the construction of Phylomer libraries intended for screening against intracellular targets using yeast two-hybrid or reverse two-hybrid approaches, or against extracellular targets using phage display. Screening protocols themselves are described more fully in various patent applications—e.g., (10–12). Using these methods for library construction, we have isolated multiple primary Phylomer peptides, at a high hit rate, with high target affinities and selectivity before the application of any affinity maturation of the sequence.

2. Materials

2.1. Genomic Material

We recommend using fully sequenced genomic material to assist in bioinformatic interpretation of screen results and modeling of structure–activity relationships to inform lead optimization strategies (for example, information on the flanking sequences of a peptide hit or knowledge of other related sequences in the library that proved successful or unsuccessful in the selection). We maximized library structural diversity by using widely phylogenetically-diverse genomes to ensure representation of most of the basic structural elements found in evolution. For sequenced genomes, the fold diversity has been modeled and can be used as a guide (see for example (5, 13)). An example of library composition is given in Fig. 1 and Table 1.

Typically these microbial genomes comprise approximately 80% coding sequence, and given that the library is shotgun cloned, one-sixth of the library clones are expected to be in the natural reading frame, resulting in approximately 13% of the clones on average being in frame with coding genes.

2.2. Plasmids

pJG45_v2 (Fig. 2) is a derivative of the LexA-based yeast two-hybrid prey vector pJG4-5 (14), modified by inserting stop codons in all three reading frames immediately 3' of the *Xba*I restriction enzyme site; Phylomers are expressed as C' fusions.

pNp3 (Fig. 3) is a derivative of the pJuFo M13 p3 phagemid vector (15), modified for N' phage display by expressing the Phylomer peptides as N'-fusions to the phage p3 protein.

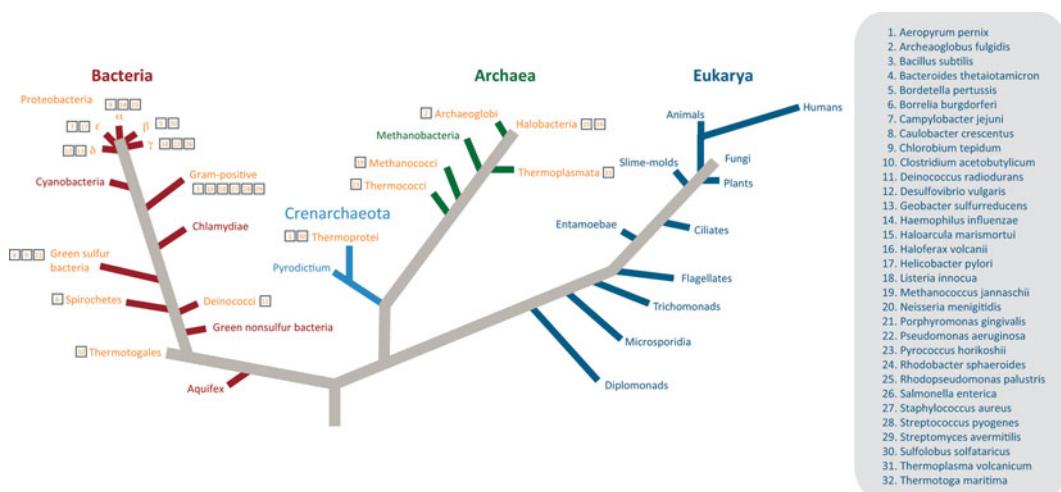


Fig. 1. Unrooted phylogenetic tree displaying sequenced genomes that can be included in Phylomer libraries.

Table 1
Sequenced genomes for Phylomer library creation

Genus species	Bacteria (B)/ Archae (A)	Genome size (kb)	Genome size multiplier (for mixing)
<i>Aeropyrum pernix</i>	A	1,670	1.16
<i>Archeaoglobus fulgidis</i>	A	2,178	1.51
<i>Bacillus subtilis</i>	B	4,214	2.92
<i>Bordetella pertussis</i>	B	4,086	2.83
<i>Borrelia burgdorferi</i>	B	1,444	1.00
<i>Campylobacter jejuni</i> subsp. <i>jejuni</i>	B	1,641	1.14
<i>Caulobacter vibrioides (crescentus)</i>	B	4,017	2.78
<i>Chlorobium tepidum</i>	B	2,155	1.49
<i>Clostridium acetobutylicum</i>	B	4,132	2.86
<i>Deinococcus radiodurans</i>	B	3,284	2.27
<i>Desulfovibrio vulgaris</i>	B	3,571	2.47
<i>Geobacter sulfurreducens</i>	B	3,814	2.64
<i>Haemophilus influenzae</i>	B	1,830	1.27
<i>Haloarcula marismortui</i>	A	4,275	2.96
<i>Haloferax volcanii</i>	A	4,010	2.78
<i>Helicobacter pylori</i>	B	1,667	1.15
<i>Listeria innocua</i>	B	3,011	2.09
<i>Methanococcus jannaschii</i>	A	1,734	1.20
<i>Neisseria meningitidis</i>	B	2,195	1.52
<i>Porphyromonas gingivalis</i>	B	2,343	1.62
<i>Pseudomonas aeruginosa</i>	B	6,264	4.34
<i>Pyrococcus horikoshii</i>	A	1,738	1.20
<i>Rhodobacter sphaeroides</i>	B	4,132	2.86
<i>Rhodopseudomonas palustris</i>	B	5,468	3.79
<i>Salmonella enterica</i> subsp. <i>enterica</i> serovar <i>Thyphimurium</i>	B	4,951	3.43
<i>Staphylococcus aureus</i>	B	2,903	2.01
<i>Streptococcus pyogenes</i>	B	1,852	1.28
<i>Streptomyces avermitilis</i>	B	9,026	6.25
<i>Sulfolobus solfataricus</i>	A	2,992	2.07
<i>Thermoplasma volcanicum</i>	A	1,585	1.10
<i>Thermotoga maritima</i>	B	1,861	1.29
<i>Bacteroides thetaiotamicron</i>	B	6,260	4.34

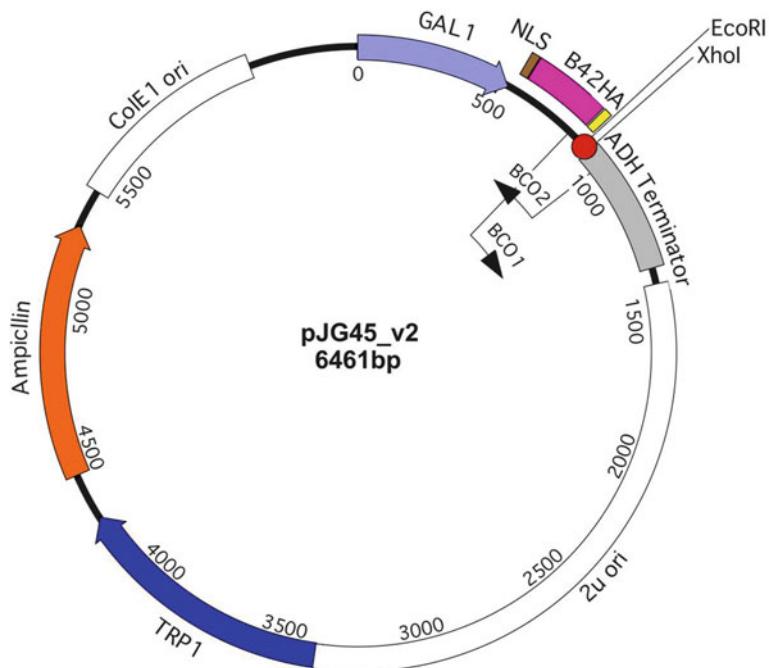


Fig. 2. pJG45_v2: a plasmid to construct a yeast two-hybrid Phylomer library in.

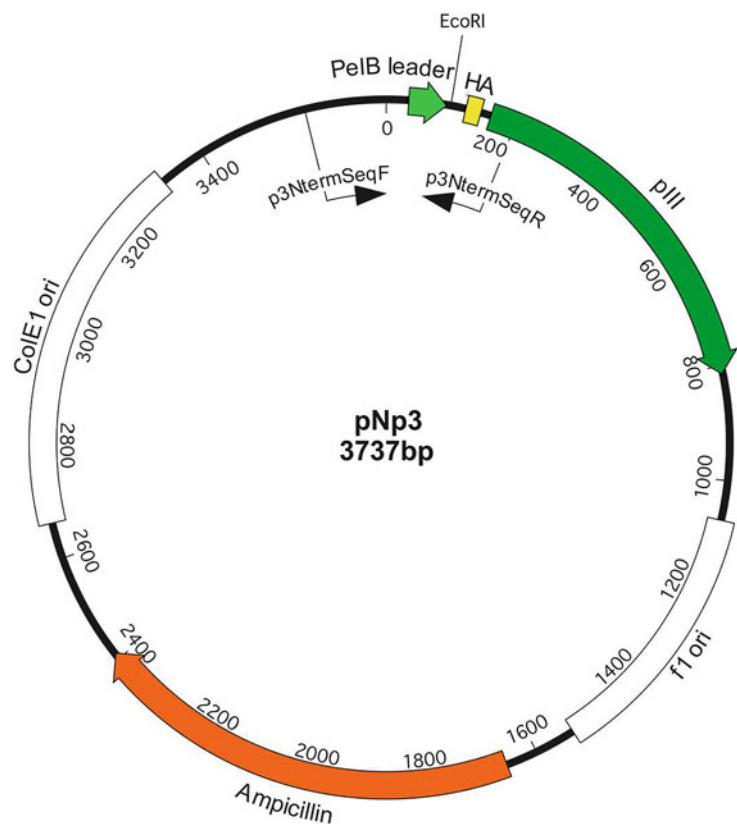


Fig. 3. pNp3: a phagemid Phylomer library in.

2.3. Oligonucleotides

Oligo	Sequence
T7MfeN6	5' GTAATACGACTCATACAATTGC NNNNNN 3'
T7Mfe	5' GTAATACGACTCATACAATTGC 3'
	<i>For yeast two-hybrid libraries with vector pJG45_v2</i>
Forward plasmid oligo (BCO1)	5' CCAGCCTCTTGCTGAGTGGAGATG 3' 5' GACAAGCCGACAACCTTGATTGGAG 3'
Forward plasmid oligo (BCO2)	
	<i>For phagemid libraries with vector pNp3</i>
Forward plasmid oligo (p3NtermSeqF)	5' CAGGCTTTACACTTATGCTTCCG 3' 5' TTGAGGCAGGTCAGACGATTGG 3'
Forward plasmid oligo (p3NtermSeqR)	

2.4. Enzymes and Enzyme Buffers

Enzyme	10× Buffer
DNA polymerase I large (Klenow) fragment	500 mM Tris–HCl (pH 7.2 at 25°C), 100 mM MgSO ₄ , and 1 mM DTT
Taq DNA polymerase	670 mM Tris–HCl (pH 8.8 at 25°C), 166 mM (NH ₄) ₂ SO ₄ , 4.5% Triton X-100, and 2 mg/ml gelatin
EcoRI	900 mM Tris–HCl pH 7.5, 100 mM MgCl ₂ , and 500 mM NaCl
MfeI	200 mM Tris–acetate, 500 mM potassium acetate, 100 mM magnesium acetate, 10 mM dithiothreitol (DTT), pH 7.9 at 25°C
Antarctic shrimp phosphatase	500 mM Bis–Tris–Propane–HCl, 10 mM MgCl ₂ , 1 mM ZnCl ₂ , pH 6.0 at 25°C
T4 DNA ligase	200 mM Tris–HCl pH 7.6, 100 mM MgCl ₂ , and 250 µg/ml acetylated BSA

2.5. Consumable and Chemicals

dNTPs: 2 mM.
 DTT: 100 mM.
 ATP: 10 mM.
 BSA: 1 mg/ml.
 Carbenicillin.
 Kanamycin.
 Buffer exchange column: for example, S200 Microspin column (Amersham).
 Silica-membrane-based DNA purification column: for example, QIAquick cleanup or QIAprep column (QIAGEN).

Agarose.

0.2-cm Electroporation cuvettes.

Optional: glass slides (sterile) and scalpel blades.

Optional: M13K07 Helper Phage.

2.6. Buffers and Solutions

NaCl: 1 M NaCl, autoclave.

PEG8000: 50% PEG8000 (w/v), autoclave.

HEPES: 1 mM HEPES-NaOH, pH 7.4; filter-sterilize.

PEG/NaCl: 20% PEG8000 (w/v), 2.5 M NaCl.

PBS (phosphate-buffered saline): 137 mM NaCl, 2.7 mM KCl, 4.3 mM Na₂HPO₄, 1.47 mM KH₂PO₄, pH 7.4, at 25°C; filter-sterilize.

PBS/Tween: PBS, 0.05% Tween-20 (v/v); filter-sterilize.

PBS/glycerol: PBS, 10% glycerol (v/v); filter-sterilize.

LB/30% Glycerol: LB, 30% glycerol (v/v); filter-sterilize.

2.7. Media

Low-salt LB: 5 g bacto-yeast extract, 10 g bacto-trypotone, 5 g NaCl, ddH₂O to 1 l; for plates: add 20 g agar. Adjust pH to 7.0 with NaOH. Autoclave.

2× YT: 10 g yeast extract, 16 g tryptone, 5 g NaCl, ddH₂O to 1 l. Adjust pH to 7.0 with NaOH. Autoclave.

SOC: 5 g tryptone, 1.25 g yeast extract, 0.125 g NaCl in 230 ml ddH₂O. Add 2.5 ml of 250 mM KCl, adjust pH to 7.0, and make to 250 ml with ddH₂O. Autoclave. Just before use, add: 1.25 ml 2 M MgCl₂ solution (sterile), 5 ml 1 M glucose solution (sterile).

LB/Carb₅₀: LB (broth or agar) supplemented with carbenicillin (50 µg/ml final concentration).

2.8. E. coli Strains

For non-phagemid libraries (e.g., yeast two-hybrid)

DH10B	F- <i>mcrA</i> Δ(<i>mrr-hsdRMS-mcrBC</i>) φ80 <i>lacZ</i> ΔM15 Δ <i>lacX74 recA1 endA1 araD139</i> Δ(<i>ara, leu</i>)7697 <i>galU galK λ-rpsL nupG</i>
-------	---

For phagemid libraries (M13)

SS320 (MC1061F')	[F' <i>proAB</i> ⁺ <i>lacI</i> ^q <i>lacZ</i> ΔM15 <i>Tn10</i> (tet ^r)] <i>hsdR mcrB araD139</i> Δ(<i>araABC-leu</i>)7679Δ <i>lacX74 galU galK rpsL thi</i>
------------------	--

3. Methods

DNA from biodiverse bacterial and *Archaeal* genomes is amplified in a two-stage process: firstly by low temperature extension from oligos ending in N6 random nucleotides using Klenow enzyme, and secondly by specific PCR amplification using a standard *Taq* polymerase. Amplified genomic material is pooled, based on genome size, and digested with *MfeI* in preparation for non-directional cloning into a yeast two-hybrid or phage display cloning plasmid.

3.1. Klenow Amplification of Individual Genomes

1. For each genome, mix 200–300 ng of genomic DNA and 100 pmol T7MfeN6 oligo in a PCR microfuge tube; make the volume to 10 µl with ddH₂O.
2. *Round #1 amplification*
 - Denature by boiling for 5 min (100°C, thermocycler) and place immediately on ice.
 - Add (on ice): 3 µl 10x Klenow buffer, 6 µl 50% PEG8000, 3 µl 2 mM dNTPs, 3 µl 1 M NaCl, 4.4 µl ddH₂O, and 0.6 µl (3 U) Klenow polymerase.
 - Amplify in a thermocycler under the following conditions: 15 min at 10°C and then ramp up to 22°C at 0.5°C/min; 50 min at 22°C; and 15 min at 37°C (see Note 1).
3. *Round #2 amplification*
 - Denature reaction from round #1 by boiling for 5 min (100°C, thermocycler) and place immediately on ice.
 - Add (on ice): 0.5 µl 10x Klenow buffer, 0.5 µl 2 mM dNTPs, 0.5 µl 1 M NaCl, 1 µl ddH₂O, 2 µl (50 pmol) oligo T7MfeN6, 0.5 µl (2.5 U) Klenow polymerase.
 - Amplify on a thermocycler under the following conditions: 15 min at 10°C and then ramp up to 22°C at 0.5°C/min, 50 min at 22°C, 15 min at 37°C.
4. *Round #3–#4 amplification*
 - Repeat steps from Round #2 amplification.
5. Total volume after four rounds of Klenow amplification is 45 µl.
6. Purify amplified DNA product over a buffer exchange column according to manufacturer's instructions (see Note 2).

3.2. PCR Amplification of Klenow-Amplified Genomic DNA for Individual Genomes

1. In a thin-walled PCR plate (on ice), set up the following reaction: 10 µl 10x PCR buffer, 8 µl 25 mM MgCl₂, 10 µl 2 mM dNTPs, 240 pmol T7Mfe oligo, 8 µl Klenow-amplified/purified genome, 4 U Taq DNA polymerase, and ddH₂O to 100 µl.

2. Amplify on a thermocycler under the following conditions: denaturation for 5 min at 94°C; 26 cycles of 30 s at 94°C, 30 s at 60°C, and 1 min at 72°C; and final extension for 2 min at 72°C.
3. Purify 95 µl of the PCR-amplified DNA product over a silica-membrane-based DNA purification column according to manufacturer’s instructions (see Note 3); keep the remaining 5 µl for gel analysis.
4. Assess quality and comparability of the amplified genomes by running 5 µl of unpurified and 5 µl purified PCR-amplified DNA, for each genome, on a 2% agarose gel. Compare the size range and median of the genomic smears for each genome. For best library results, the smears should have similar size ranges and, ideally, similar medians.

**3.3. Preparation of Amplified Genomes:
DNA Quantitation,
Pooling, and Digestion**

1. Determine DNA concentration of each amplified genome using standard spectrophotometry or fluorometry techniques.
2. Pool amplified DNA (see Note 4) according to concentration and genome size (i.e., add proportionally larger amounts of DNA for bigger genomes); see Table 1 for example of ratio for mixing genomes. Keep 3 µl of pooled DNA for later gel electrophoresis analysis (a).
3. Digest the rest of the pooled amplified genomic DNA as follows, making multiple reactions if necessary: 8–12 µg pooled PCR-amplified genomic DNA, 10 µl 10× *Mfe*I reaction buffer, 10 µl 1 mg/ml BSA, 5 U/µg genomic DNA enzyme *Mfe*I, and ddH₂O to 100 µl total volume.
4. Incubate for 2 h at 37°C, then combine digestions, and mix; keep 5 µl of digested unpurified genomic DNA for gel analysis (b).
5. Purify the remainder of digested product over a silica-membrane-based DNA purification column according to manufacturer’s instructions (see Note 3). If purifying digests over multiple columns, combine eluates; keep 5 µl sample for gel analysis (c).
6. Check quality of genomic material by gel electrophoresis, running the following samples on a 2% agarose gel: (a) 3 µl of purified pooled genomic DNA; (b) 5 µl digested, pooled genomic DNA; and (c) 5 µl digested, purified, pooled genomic DNA. The size range of the smear should decrease with digestion (compare samples (b) to (a)) and the cleaved ends of the DNA fragments (seen in sample (b)) should be removed with purification (sample (c)).
7. Determine concentration of purified, digested, amplified genomic DNA using standard spectrophotometry or fluorometry techniques (see Note 5).

Prepare large-scale plasmid preparations using four different conditions: a small fraction of each is used to optimize the cloning conditions. If the cloning frequency is good, this will quality-check the plasmid preparation and the remainder of the digested plasmid can be used in creating the library. Suggested conditions are as follows:

- (A) Digestion: 37°C for 1 h; dephosphorylation: 1 U phosphatase/μg digested plasmid.
- (B) Digestion: 37°C for 1 h; dephosphorylation: 2 U phosphatase/μg digested plasmid.
- (C) Digestion: 37°C for 2 h; dephosphorylation: 1 U phosphatase/μg digested plasmid.
- (D) Digestion: 37°C for 2 h; dephosphorylation: 2 U phosphatase/μg digested plasmid.

3.4. Optimization of Library Ligation Conditions

1. For each digestion condition (A–D), digest 80 μg plasmid DNA (pJG45_v2 or pNp3; see Note 6) with *Eco*RI, digesting 10 μg DNA per 100 μl reaction as follows: 10 μg of plasmid, 10 μl 10× *Eco*RI digestion buffer, 10 μl 1 mg/ml BSA, 60 U *Eco*RI, and ddH₂O to 100 μl.
2. Incubate reactions at 37°C for 1 or 2 h as appropriate.
3. Purify digested plasmid over silica-membrane-based DNA purification columns according to manufacturer's instructions (see Note 3).
4. Combine reaction samples for each condition, and quantitate DNA concentration of combined digests using standard spectrophotometry or fluorometry techniques.
5. Desphosphorylate digested plasmid using 1 or 2 U phosphatase per μg DNA as appropriate; reactions are recommended to be divided as follows: 50 μl linearized plasmid, 6 μl 10× Antarctic Phosphatase buffer, X μl Antarctic phosphatase (80U or 160U, depending on the digestion condition), and ddH₂O to 60 μl total volume.
6. Incubate at 37°C for 15 min, then heat-inactivate enzyme at 65°C for 5 min, or according to manufacturer's instructions, and combine plasmid digests/dephosphorylations appropriately.
7. Assess quality and completion of plasmid digestion by gel electrophoresis, running 5-μl aliquots of each sample on a 0.7–1% agarose gel.
8. Calculate the concentration of linearized vector and insert in pmol/μl (see Note 7).
9. For each plasmid preparation, set up ligation-test reactions (V-L and V+L are ligation controls for uncut plasmid and religands, respectively) in a 96-well PCR plate or equivalent (see Note 8):

Reagent	V-L	V+L	3:1	5:1
Linear, dephos plasmid (pmol)	0.02	0.02	0.02	0.02
Phylomer genomic PCR (pmol)	-	-	0.06	0.10
1/10 Diluted T4 ligase (Novagen ligase) or equivalent (μ l)	-	1	1	1
10x Ligase buffer (μ l)	1	1	1	1
10 mM ATP (μ l)	1	1	1	1
100 mM DTT (μ l)	0.5	0.5	0.5	0.5
ddH ₂ O (μ l)	To 10 μ l			

10. Incubate ligations overnight (~16 h) in a thermocycler under the following conditions: repeated cycles of 10°C for 1 min, ramp 0.1°C/s, and 30°C for 1 min (modified from (16)).
11. After incubation, briefly spin down the plate, and “cut-away” undigested or religated plasmid by digestion with *Eco*RI, adding to each ligation: 1.15 μ l 10x *Eco*RI digestion buffer and 2 U *Eco*RI.
12. Incubate reaction at 37°C for 30 min and then heat-inactivate enzyme at 70°C for 10 min.
13. Transform 1 μ l of each ligation into 50- μ l aliquots of high-efficiency electrocompetent cells (see Subheading 3.10) or equivalent.
14. After transformation outgrowth, make tenfold serial dilutions and for each ligation testing group, plate 100 μ l of each “Phylomer+plasmid” dilution from 10⁻¹ to 10⁻³ and plate 100 μ l of each “V-L” and “V+L” dilution from Neat to 10⁻² onto LB/Carb₅₀ agar plates.
15. For the transformation efficiency test, plate 100 μ l of each dilution from 10⁻² to 10⁻⁶ onto LB/Carb₅₀ agar plates.
16. Incubate plates overnight at 37°C, count the number of colonies on the next day, and calculate the stimulation over background based on colony #s (Phylomer+plasmid/V+L control).
17. From this, determine the optimal ligation conditions with respect to dephosphorylation and insert:vector ratios based on the highest number of clones (Phylomer+plasmid) with the lowest religand background (V+L).

Select the two best ligation conditions from the four conditions tested (two different dephosphorylations multiplied by two

different insert-to-vector ratios) and assess cloning efficiency (insert frequency) by colony PCR.

3.5. Colony PCR to Assess Cloning Efficiency (Insert Frequency)

- For a colony PCR plate, keep one well without template (negative control), one well with vector plasmid (positive/vector control), and pick 40 colonies from each of the 2 best ligation conditions along with 3 colonies each of their corresponding V-L and V+L controls to screen inserts by colony PCR (see Subheading 3.12).
- Check insert frequency (# of colonies with insert compared to total number of colonies) and the range of insert sizes by gel electrophoresis, running 10 μ l samples of the PCRs on a 2% agarose gel.
- The optimal cloning condition should have a good insert size range (the majority of fragments between 200 and 500 bp is desirable) and high insert frequency.

3.6. Constructing the Phylomer Library

- Using the plasmid preparation and vector-to-insert ratio from the optimization experiments, set up the following large-volume ligation:

Reagent	V+L	Optimal ligation
Linear, dephos plasmid (pmol)	0.08	8.0
Phylomer genomic PCR	–	Opt insert pmol
1/10 Diluted T4 ligase (Novagen ligase) or equivalent (μ l)	4	400
10× Ligase buffer (μ l)	4	400
10 mM ATP (μ l)	4	400
100 mM DTT (μ l)	2	200
ddH ₂ O (μ l)	To 40 μ l	To 4 ml

- Set up two 96-well plates. Into each first well (A1) of each, add 20 μ l of the V+L control mix.
- Aliquot 20 μ l of ligation mix per well to all the remaining wells in both plates.
- Incubate overnight and digest post-ligation by adding 2.3 μ l 10× *Eco*RI digestion buffer and 4 U *Eco*RI to each well, incubating digestions as described above (see step 11 in Subheading 3.4).
- Centrifuge plates to collect digested ligations, transfer each V+L ligations to microfuge tubes and keep on ice. For each plate, combine library samples (~2.154 ml each), and set aside 2 μ l of each combined, digested library ligation in a microfuge tube and keep on ice.

6. Purify each of the combined library ligations over a silica-membrane-based DNA purification column according to manufacturer’s instructions (see Note 3). Elute from the column by applying the smallest volume of ddH₂O recommended (likely ~35 µl), and then applying a second volume of ddH₂O to the same column to maximize DNA recovery.
7. Combine purified library eluates and keep on ice. Take ~2 µl of purified digested ligation to quantitate using standard spectrophotometry techniques.
8. Make fresh electrocompetent cells (see Subheading 3.10); calculate the aliquot volume of purified ligation mix to evenly divide the DNA over the number of 350-µl competent cells’ aliquots.
9. Electroporate a 350-µl aliquot of cells with one aliquot volume of purified Phylomer ligation mix (see Subheading 3.11), transferring each electroporated cells/SOC resuspension to a 250-ml conical flask containing 22 ml pre-warmed SOC. Rinse cuvette twice with 1 ml SOC, adding rinses to the flask, and incubate for 30 min at 37°C with shaking at 200 rpm.
10. Repeat this procedure until all of purified ligation mix is transformed into cells.
11. Electroporate 1 µl of each control (unpurified, digested V+L controls (2); unpurified, digested Phylomer ligation (2); and 100 ng plasmid as a transformation efficiency control) in 50µl-aliquots of competent cells.
12. Following transformation outgrowth, combine all library-ligation transformations into one flask. “Rinse” the remaining flasks with 25 ml SOC, and combine this with the rest of the library suspension; record the total volume.
13. In triplicate, make tenfold serial dilutions of the library transformation mix (20 µl of transformed cells into 180 µl LB) and plate 100 µl from dilutions 10⁻² to 10⁻⁶ onto LB/Carb₅₀ plates.
14. Repeat similar tenfold serial dilutions for both Phylomer and V+L control cultures, plating 100 µl of each dilution from Neat to 10⁻³ for the Phylomer and Neat to 10⁻² for V+L controls.
15. Repeat similar tenfold serial dilutions for the plasmid transformation control culture, plating 100 µl of each dilution from 10⁻⁴ to 10⁻⁷.
16. Incubate plated cells overnight at 37°C, and count the number of colonies on the next day.
17. Determine the library complexity (calculated total colony #), library transformation efficiency (colony #/µg transformed for purified Phylomer library), % background based on colony numbers (unpurified Phylomer control/V+L control), and

cell transformation efficiency (colony #/ μ g transformed for plasmid control).

18. See Subheading 3.7 or 3.8 for growth and harvesting of the Phylomer library, depending on the type of library being constructed.

3.7. Growing and Harvesting the Phylomer Library Constructed in Non-phagemid Vectors

1. After making serial dilutions, concentrate combined library transformation culture (from Subheading 3.6) by centrifugation at 2706 $\times g$ for 15 min.
2. Resuspend cells in LB and spread evenly over ~75 large square LB/Carb₅₀ plates (25 \times 25 cm), 1 ml per plate (see Note 9). Allow cell suspension to completely soak in and then incubate plates overnight at 30°C.
3. The next day, examine the growth of the bacteria on the library plates; incubate at 37°C until colonies are as grown as possible without merging.
4. Scrape and collect bacterial cells from plates, using sterile glass slide(s), add an equal volume of LB, and resuspend cells gently but thoroughly.
5. Make about twenty 1 ml frozen stocks of the primary plate library by mixing 10 ml of this cell suspension and 10 ml LB/30% glycerol, and store in 1-ml aliquots at -80°C.
6. Take the remaining cell suspension (the majority of the library), and extract the plasmid DNA over a CsCl gradient using standard laboratory techniques. For the lysis of the cells and the CsCl protocol, treat the cell suspension as if it were (at least) 2 l of culture to ensure sufficient lysis of the cells and optimal recovery of supercoiled plasmid.
7. The purified library DNA is stored at -20°C for future use, for example, transformation into yeast for yeast two-hybrid screening.

3.8. Growing and Harvesting the Phylomer Library Constructed in Phagemid Vectors

1. After making serial dilutions, divide combined library transformation culture (from Subheading 3.6) between four 2-l flasks. Make the volume up to 500 ml in each flask with 2 \times YT broth and add carbenicillin to a final concentration of 50 μ g/ml.
2. Add ~50 μ l M13K07 helper phage to each flask, and incubate for 2 h at 37°C, 220 rpm.
3. Add kanamycin to a final concentration of 25 μ g/ml and incubate overnight at 37°C, 220 rpm.
4. The following day, make ten frozen stocks: 0.5 ml culture + 0.5 ml LB/30% glycerol; store at -80°C.
5. Harvest the expressed phage by standard phage/PEG precipitation techniques (see Subheading 3.13 for a sample protocol), and store at -80°C for future phage display screening.

3.9. Primary Characterization of Phylomer Library by Colony PCR and Sequencing

1. Examine the size diversity of the inserts and the insert frequency by colony PCR (see Subheading 3.12), screening 3× 96-well plates of colonies, picked from the dilution plates spread for library complexity calculations.
2. Sequence these PCR products; these DNA sequences are used for bioinformatic characterization of the Phylomer library (see Note 10).

3.10. Additional Protocol: Electrocompetent Cells (Modified from (17))

1. Inoculate a starter culture: 20 ml 2× YT with a single colony of *E. coli* from a fresh streak-plate; incubate overnight at 37°C with shaking.
2. Inoculate four 2-l flasks of 500 ml of low-salt LB with 5 ml of starter culture each; incubate flasks at 37°C, with shaking, until the OD₆₀₀ reaches ~0.8.
3. Place cultures on ice for 5 min (see Note 11); then divide 1 l of culture over four centrifuge bottles, leaving the remaining cultures on ice, and collect cells by centrifugation at 4,229 ×g, 4°C for 10 min.
4. Divide the remaining 1 l of culture over the same four bottles, and again collect cells by centrifugation at 4,229 ×g, 4°C for 10 min.
5. Wash cell pellets twice in an equal volume of 1 mM HEPES-NaOH, pH 7.4, collecting cells by centrifugation as before.
6. Wash each pellet in 150 ml of 10% glycerol, collecting cells by centrifugation as before.
7. Add 500 µl of 10% glycerol to two tubes, and resuspend the pellets with gentle pipetting. Transfer the suspension to the next tube and resuspend the second pellet. Combine cell suspensions (see Note 12).
8. Aliquot in 350-µl aliquots; include at least five 50-µl aliquots to determine competency and control transformations. Keep cells on ice; transformation should proceed immediately.

3.11. Additional Protocol: Electroporation of Bacteria

1. Mix cells with chilled ligation DNA gently with a pipette, transfer cell/DNA mixture to a chilled 0.2-cm cuvette, and electroporate immediately: 2.5 kV, 125 Ω, and 50 µF.
2. Immediately add 1 ml pre-warmed SOC medium and transfer cell suspension to an appropriate tube for culture growth.
3. Parallel a transformation with undigested plasmid to confirm cell electroporation efficiency.
4. Incubate transformations for 1 h at 37°C with gentle shaking.

3.12. Additional Protocol: Colony PCR

1. Make up a colony PCR master mix: 500 µl 10× PCR buffer, 150 µl 25 mM MgCl₂, 250 µl 2 mM dNTPs, 250 pmol Plasmid forward oligo, 250 pmol Plasmid reverse oligo, 50 U Taq DNA polymerase, and ddH₂O to 2,500 µl total volume.

2. Aliquot 25 μ l master mix per well into a 96-well plate (on ice).
3. Keep one well without template (negative control), to one well add ~10 pg plasmid DNA (positive/vector control), and for the remaining wells “pick” a bacterial colony (see Note 13).
4. Amplify on a thermocycler as follows: denaturation for 10 min at 94°C; 30 cycles of 30 s at 94°C, 30 s at 56°C, and 1 min at 72°C; and final extension for 5 min at 72°C.

3.13. Additional Protocol: PEG Precipitation of M13 Phage

1. Collect cells from the overnight outgrown library culture (from Subheading 3.8) by centrifugation for 20 min at 2,706 $\times g$.
2. Precipitate phage, transferring supernatant equally over eight fresh tubes and adding 1/5 volume of PEG/NaCl. Vortex thoroughly and let stand at room temperature for at least 20 min.
3. Collect phage by centrifugation for 30 min at 14,475 $\times g$; without disturbing pellet(s), discard supernatant and briefly spin again.
4. Remove supernatant and resuspend each phage pellet in 15 ml PBS/Tween.
5. Repeat the precipitation and resuspend each phage pellet in 1 ml PBS/glycerol.
6. Vortex thoroughly, and transfer supernatants into microfuge tubes; let them stand for 10–15 min at room temperature.
7. Centrifuge these samples for 10 min at 14,475 $\times g$, and combine supernatants.
8. Store the collected M13 phage at -80°C in 100- μ l aliquots; ~2 μ l may be used to titrate the phage preparation before freezing (by standard phage infection and titration methods).

4. Notes

1. Recommended: Try different thermocyclers to optimize the amount of DNA produced and to produce a consistent, low-molecular-weight genomic smear.
2. For example, an S200 column.
3. For example, a QIAquick cleanup or QIAprep column. Check maximum amount of DNA that can be cleaned up per column, and use multiple columns if necessary, combining samples after purification if/as appropriate. Tip: Warming the DNA elution buffer to 65°C can increase DNA recovery.

4. Recommended: Pool 15 µg (total) amplified DNA (based on ~50–60% recovery from purification). If necessary, adjust based on the calculated amount of DNA required for the ligation.
5. Recommended: Aliquot *MfeI*-digested pooled amplified genomic DNA into ~5 µg aliquots for library construction and some 1/10 dilutions for ligation testing, and store at -20°C.
6. Recommended: Using cesium-gradient purified plasmid DNA improves cloning efficiency.
7. Calculation of the number of moles of insert depends on average size of *MfeI*-digested Phylomer fragments, as determined by agarose gel electrophoresis (Subheading 3.3).
8. Including an additional control with no ligase and no insert will allow calculation of the uncut vector (inefficiency of digestion).
9. Adjust plate number depending on incubator space.
10. Not all PCR products need to be sequenced immediately; colony PCRs can be stored at -20°C for later sequencing if desired.
11. Competent cell preparation should be done in a 4°C cold room, with prechilled solutions and equipment, and cells kept on ice throughout the process.
12. Optional: Remove a small aliquot (~5 µl), make serial dilutions (10⁻² to 10⁻⁴), and determine density using spectrophotometer OD₆₀₀; preferably, this should be >0.3 for the 10⁻³ dilution.
13. With a disposable tip or similar tool, touch a bacterial colony and transfer the tiny amount of cells into a well, mixing thoroughly to disperse the cells.

References

1. Gilbert W, de Souza SJ, Long M (1997) Origin of genes. Proc Natl Acad Sci USA 94:7698–7703
2. Bogard LD, Deem MW (1999) A hierarchical approach to protein molecular evolution. Proc Natl Acad Sci USA 96:2591–2595
3. Riechmann L, Winter G (2006) Early protein evolution: building domains from ligand-binding polypeptide segments. J Mol Biol 363:460–468
4. Riechmann L, Winter G (2000) Novel folded protein domains generated by combinatorial shuffling of polypeptide segments. Proc Natl Acad Sci USA 97:10068–10073
5. Söding J, Lupas AN (2003) More than the sum of their parts: on the evolution of proteins from peptides. Bioessays 25:837–846
6. Lupas AN, Ponting CP, Russell RB (2001) On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? J Struct Biol 134:191–203
7. Watt P (2006) Screening for peptide drugs from the natural repertoire of biodiverse protein folds. Nat Biotechnol 24:177–183
8. Watt PM (2009) Phenotypic screening of peptide libraries derived from genome fragments (phylomers) for therapeutics discovery and as a tool to identify and validate new targets. Future Med Chem 1:1–8
9. Vanhee P et al (2009) Protein-peptide interactions adopt the same structural motifs as monomeric protein folds. Structure 17:1128–1136
10. Watt P, Hopkins R, Fear M, Milech N (2006) Modulators of biochemical characteristics. International Patent Application (WO 2005/119244 A1R6).

11. Watt P, Hopkins R, Thomas W (2010) Isolating biological modulators from biodiverse gene fragment libraries. Granted European Patent EP 1179059 B1 (PCT/AU00/00414).
12. Watt P, Kees U (2007) Peptide detection method. Granted European Patent EP 1 053 347 B1 (PCT/AU1999/000018).
13. Marsden RL, Lewis TA, Orengo CA (2007) Towards a comprehensive structural coverage of completed genomes: a structural genomics viewpoint. *BMC Bioinformatics* 8(86):1471–2105
14. Gyuris J, Golemis E, Chertkov H, Brent R (1993) Cdk1, a human G1 and S phase protein phosphatase that associates with Cdk2. *Cell* 75:791–803
15. Crameri R, Hemmann S, Blaser K (1996) pJuFo: a phagemid for display of cDNA libraries on phage surface suitable for selective isolation of clones expressing allergens. *Adv Exp Med Biol* 409:103–110
16. Han CS, Buckingham J, Meincke LJ, Doggett NA (2011) Vector for high-throughput sequencing: construction and preparation with cyclic cut-ligation. *Biotechniques* 30:1208–1210
17. Tonikian R, Zhang Y, Boone C, Sidhu SS (2007) Identifying specificity profiles for peptide recognition modules from phage-displayed peptide libraries. *Nat Protoc* 2: 1368–1386

Chapter 4

Ribosome Display and Screening for Protein Therapeutics

Damjana Kastelic and Mingyue He

Abstract

Ribosome display is a cell-free display technology which enables in vitro selection of antibodies from large recombinant DNA libraries. It also allows continuous introduction of mutations into the selected DNA pool by PCR-based mutagenesis in each cycle, enabling selection of antibody variants with improved affinity, specificity, and stability, thus providing a powerful “protein evolution” tool for optimizing antibody therapeutics. Ribosome display selects required molecules by linking individual proteins (phenotype) with their corresponding mRNAs (genotype) through the formation of stable *Protein–Ribosome–mRNA* (PRM) complexes. By affinity interaction with an immobilized ligand, the captured PRM complexes are recovered as cDNA using RT-PCR from the ribosome-attached mRNA. The DNA is then subjected to subsequent ribosome display cycles for further enrichment of rare species or cloning, expression, and sequencing to identify wanted candidates. Both prokaryotic and eukaryotic cell-free systems have been developed for ribosome display of different proteins. In this chapter, we describe ribosome display of antibodies using the eukaryotic rabbit reticulocyte system with an *in situ* single-primer DNA recovery method. A high-throughput *Escherichia coli* expression format is also described for screening of individual antibody binders from the ribosome-selected population.

Key words: Ribosome display, Single-chain antibody, Rabbit reticulocytes

1. Introduction

Ribosome display is a cell-free based technology for in vitro selection and evolution of proteins from recombinant libraries where individually expressed polypeptides (phenotype) are physically linked to their corresponding mRNA (genotype) through the formation of stable *Protein–Ribosome–mRNA* (PRM) complexes as the selection particles. The PRM complex is generated by deleting the stop codon from the DNA construct, which causes stalling of the translating ribosome at the end of mRNA together with the nascent polypeptide not released ([1–3](#)). These stable PRM complexes are then subjected to a selection cycle that consists of

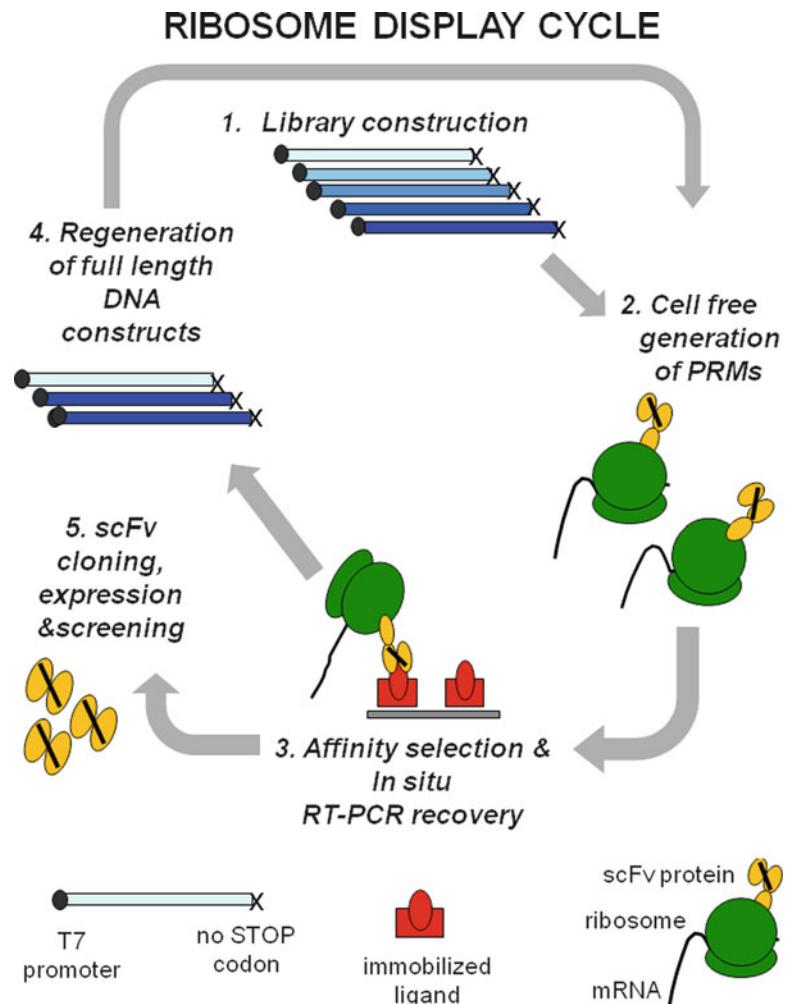


Fig. 1. The ribosome display cycle. The cycle comprises (1) library construction, (2) cell-free generation of PRM complexes, (3) affinity selection on immobilized ligand, followed by in situ RT-PCR recovery and (4) regeneration of the full-length PCR construct for the subsequent selection cycle, or (5) scFv cloning, expression, and screening.

exposing to an immobilized ligand (panning), washing off non-binding complexes, and recovering the selected genetic information as DNA by RT-PCR from the ligand-binding PRM complexes (Fig. 1). The recovered DNA can be either subjected to continuation/repetition of the display cycles to enrich ligand-specific binding molecules from a very large population, or *Escherichia coli* (*E. coli*) cloning, expression, and screening for desirable binders (1–3).

Since the whole process in ribosome display is performed entirely *in vitro* without the need for *E. coli* cloning, the library size is not restricted by *E. coli* transformation efficiency. Thus, ribosome

display screens larger libraries and increases the possibility of discovery of novel molecules (2). Moreover, the use of a PCR library also allows easy introduction of new diversity into the selected DNA pool by random or/and site-directed mutagenesis at each selection cycle, providing an efficient tool for protein evolution *in vitro* (4).

Transcription and translation can be performed in a cell-free system either in a coupled manner, where transcription and translation takes place in the same reaction mixture, or uncoupled, in which the transcription and translation are carried out by separate reactions. Both prokaryotic (*E. coli*) and eukaryotic (wheat germ and rabbit reticulocyte) cell-free systems (1, 5, 6) have been explored for ribosome display of different proteins, including antibody fragments, peptides, scaffolds, novel tags, enzymes, DNA-binding proteins, receptors, membrane proteins, and vaccine candidates (2). A “pure” *E. coli* cell-free system composed of isolated components and enzymes has also been developed for ribosome display (7). In this chapter, we describe the eukaryotic rabbit reticulocyte ribosome display system and its application for selection of single-chain antibody fragments.

2. Materials

Precautions should be taken to avoid any contamination. Always wear gloves and use nuclease-free water in all steps of ribosome display selection. Solutions should be in aliquots and stored at either -20°C or 4°C. All reagents used are of highest purity grade.

2.1. Kits and Molecular Biology Reagents

1. Primers used for PCR and RT-PCR were chemically synthesised (Sigma) (Table 1).
2. GenElute™ Gel Extraction Kit (Sigma).
3. Rabbit Reticulocyte TNT T7 Quick for PCR DNA (Promega).
4. DreamTaq™ Green PCR Master Mix (Fermentas).
5. SuperScript® III RT Kit (Invitrogen).
6. RNase-free DNase I (Roche).
7. RiboLock™ RNase inhibitor (Fermentas).
8. Glutathione; reduced and oxidized form (Sigma).
9. TopYield Strips (NUNC).
10. BSA (Sigma).
11. 100 mM Magnesium acetate (MgAc).
12. *E. coli* BL21(DE3) (Bioline).

Table 1
Primers for PCR and RT-PCR recovery

Primer	Sequence (from 5' to 3')
N-Ab/B	GGAACAGACCACCATGSARGTNSARCTBGWRSAGTCYGG
scFv-link/F ^a	<u>GCTACCGCCACCCTCGAGAGATGGTGCAGCCACAG</u>
Link-C κ /B	<u>CTCGAGGGTGGCGGTAGCACTGTGGCTGCACCATCTGTC</u>
C κ /F	GCACTCTCCCCCTGTTGAAGCT
T7Ab/B	GCAGCTAATACGACTCACTATAGGGAACAGACCACCATGSARGTN SARCTBGWRSAGTCYGG
RTKz1	GAACAGACCACCATGACTTCGCAGGCGTAGAC
Kz1	GAACAGACCACCATG
C κ -f/F	GCACTCTCCCCCTGTTGAAGCTCTTGTGACGGCGAGCTCAGGCC CTGATGGGTGACTTCGCAGGCGTAGACTTTG
ScFv-NcoI/B	GGAACAGACCACCATGGAAGTGCAGCTGG

Underlined are the overlapping sequences for PCR assembly. B = G + T, Y = C + T, N = A + C + T + G, R = A + G, S = G + C, W = A + T. XhoI and NcoI are indicated by *italic*.

^aThis primer is designed to anneal at 5' end of human C κ region when antibody format scFv-C κ is used as the template (1).

13. XhoI and NcoI restriction enzymes (Fermentas).
14. T4 DNA ligase (Roche).
15. Plasmid pET22b(+) (Novagen).
16. Overnight ExpressTM Autoinduction System 1 (EMD Millipore).
17. Ampicillin; stock 100 mg/ml (Sigma).
18. 96 Deep-well plate (Nunc).
19. Maxisorp 96-well plate (Nunc).

2.2. Solutions

1. Phosphate-buffered saline pH 7.4 (PBS).
2. Antigen solution (1–50 μ g/ml) in PBS.
3. Blocking buffer: 1% BSA and 0.05% Tween 20 in PBS.
4. Washing buffer: PBS containing 0.05% Tween 20 and 5 mM MgAc.
5. 2 \times dilution buffer: 4 mM glutathione (GSSG: GSH = 1:1) and 10 mM MgAc.
6. Sucrose extraction buffer: 20% sucrose and 1 mM EDTA in 50 mM Tris-HCl pH 8.0.

3. Methods

Figure 1 show the ribosome display cycle. It has the following key steps:

1. Generation of the PCR construct/libraries (see Subheading 3.1 for the format of the construct).
2. Generation of PRM complexes by cell-free transcription and translation.
3. Panning on immobilized antigen.
4. In situ RT-PCR recovery.
5. Regeneration of the full-length construct for a subsequent ribosome display cycle.
6. Cloning, expression, and screening for specific antibody binders.

3.1. Generation of the PCR Construct(s)/Libraries

Linear PCR DNA constructs designed for eukaryotic ribosome display require the following essential elements (Fig. 2):

- T7 promoter and a kozak sequence upstream of the gene of interest.
- A spacer domain located at the 3' region of the construct (1–3).
- Absence of stop codon using a 3' primer lacking a stop codon.

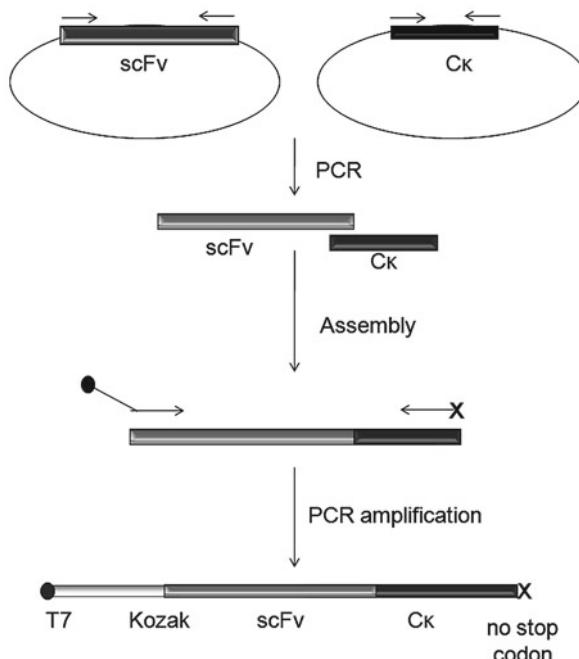


Fig. 2. Construction of scFv-C κ fragments for ribosome display. The primers used for PCR are listed in Table 1. T7: T7 promoter, Kozak: kozak sequence, scFv-single-chain antibody fragment, C κ : the constant region of κ chain X: stop codon removed.

For display of scFv fragments, we use a 3-domain construct scFv-C κ , where C κ is used as a spacer. The scFv-C κ construct is generated by the following protocols (Fig. 2).

1. Generate scFv fragment(s) by setting up the PCR mixture as follows:

DreamTaq™ Green PCR Master Mix (2 \times)	12.5 μ l (see Note 1)
Primer N-Ab/B (10 μ M)	1 μ l
Primer scFv-link/F (10 μ M)	1 μ l
scFv DNA template(s) (10 ng/ μ l)	1 μ l (see Note 2)
H ₂ O to	25 μ l

Carry out 30 thermal cycles: 94°C 30 s; 54°C 1 min; 72°C 1 min. Then, 72°C for 7 min. Finally, hold at 10°C.

2. Generate the C κ domain using primers Link-C κ /B and C κ /F (Table 1) from a plasmid encoding the κ light chain (1), using the PCR protocol as above.
3. Analyze the products by agarose gel electrophoresis. Purify amplified scFv fragment(s) (~700 bp) and the C κ domain (~300 bp) from the agarose gel using GenElute™ Gel Extraction Kit.
4. Assemble equal amounts of scFv and C κ DNA fragments, 1–10 ng DNA in total, followed by amplifying the assembled product as follows:

DreamTaq™ Green PCR Master Mix (2 \times)	12.5 μ l
scFv PCR fragment	x μ l
C κ fragment	y μ l
Primer T7Ab/B (10 μ M)	1 μ l
Primer C κ /F (10 μ M)	1 μ l
H ₂ O to	25 μ l

Carry out 30 thermal cycles: 94°C 30 s; 54°C 1 min, 72°C 1 min. Then, 72°C for 7 min. Finally, hold at 10°C.

5. Analyze the PCR product (scFv-C κ , size 1 kb) by agarose gel electrophoresis (see Note 3).

3.2. Generation of PRM Complexes

To generate PRM complexes, the PCR DNA constructs (in the format of scFv-C κ , see Subheading 3.1) are directly added (with or without purification) into a coupled rabbit reticulocyte lysate (TNT) system. Typically, for display of an scFv recombinant library, 0.5–1 μ g of PCR product (0.5–1 \times 10¹² molecules) is used in a standard 50 μ l reaction.

- Set up the TNT rabbit reticulocyte cell-free system as follows:

TNT T7 Quick for PCR	40 µl (see Note 4)
Linear PCR DNA fragment(s)	0.5–1 µg
Methionine (1 mM)	1 µl
MgAc (100 mM)	1 µl
H ₂ O to	50 µl

Incubate at 30°C for 60 min.

- Remove the original PCR DNA construct by adding 120 U RNase-free DNase I, 7 µl 10× DNase I digestion buffer and H₂O to final volume of 70 µl. Incubate at 30°C for a further 20 min (see Note 5).
- Dilute with 70 µl of ice-cold 2× dilution buffer and the mixture is ready for panning (see below).

3.3. Panning on Immobilized Antigen

Immobilized antigens are prepared by direct coating onto a TopYield wells as follows (see Note 6):

- Add 20 µl of protein (1–50 µg/ml in PBS) to each well of TopYield Strips and incubate at 4°C overnight.
- Remove the solution and block the wells with 200 µl 1% BSA in PBS for 1 h at RT.
- Wash wells three times with 300 µl PBS and either use immediately for panning or stored at 4°C for 2 weeks.
- Pan the PRM complexes by adding 50–140 µl of the translation mixture from Subheading 3.2, step 3 to each of the antigen-coated wells and incubate at 4°C for 2 h with gentle shaking.
- Wash the wells five times with 200 µl ice-cold washing buffer, followed by two quick washes with 100 µl ice-cold H₂O (see Note 7). The wells carrying specific PRM complexes can be directly used for *in situ* single-primer RT-PCR recovery or stored at –20°C.

3.4. In Situ RT-PCR

In situ single-primer RT-PCR recovery is performed using a protocol described by He and Taussig (1). In this method, an internal primer RTKz1 is designed to contain both a sequence for hybridizing to the upstream region of 3' mRNA (to avoid the stalling ribosome) and a sequence identical to the 5' region of the mRNA (Table 1). cDNA synthesis using RTKz1 leads to the generation of single-stranded cDNAs with a complementary flanking sequence at both 5' and 3' ends, which can then be effectively amplified by PCR using a single primer Kz1 (1).

- Set up the reverse transcription reaction by adding the following to each well containing PRM complexes from step 5 in Subheading 3.3.

H ₂ O	9 µl
dNTPs (10 mM)	2 µl
Primer RTKz1 (10 µM)	1 µl

Incubate at 65°C for 5 min and then quickly chill on ice.

- Add the following components to each well:

5× first-strand buffer	4 µl
100 mM DTT	1 µl
RNase inhibitor (20 U)	1 µl
SuperScript III (200 U)	1 µl
H ₂ O	1 µl

Continue incubation at 50°C for 45 min followed by 5 min at 85°C.

- Transfer the cDNA synthesis mixture from the step 2 to a fresh tube. Set up single-primer PCR mixture as follows (see Note 8):

DreamTaq™ Green PCR Master Mix (2×)	12.5 µl
Primer Kz1 (10 µM)	1 µl
cDNA from previous step	0.2–1 µl
H ₂ O to	25 µl

Carry out 20–35 cycles of thermal cycling: 94°C 30 s, 48°C 1 min, 72°C 1 min; then 72°C for 8 min. Finally, hold at 10°C (see Note 9).

- Analyze the PCR product by agarose gel electrophoresis. If multiple bands are produced, extract the band of expected size. The eluted PCR is used as a template either for regeneration of the full length construct for subsequent ribosome display cycles (Subheading 3.5 below) (see Note 10) or cloning and expression of the scFv fragments in *E. coli* (Subheading 3.6 below and Fig. 1).

3.5. Generation of Full-Length PCR for Subsequent Ribosome Display

The use of an internal primer RTKz1 in the in situ RT-PCR recovery leads to shortening of the DNA fragment compared to the original fragment, therefore, a further PCR step is required to regenerate the full-length construct using a long 3' primer Ck-f/F in combination with the 5' primer T7Ab/B (Table 1).

- Set up PCR mixture as follows:

DreamTaq™ Green PCR Master Mix (2×)	12.5 µl
Primer T7Ab/B (10 µM)	1.5 µl
Primer Ck-f/F (10 µM)	1.5 µl
PCR template from Subheading 3.4	1–10 ng
H ₂ O to	50 µl

Carry out 30 thermal cycles: 94°C 30 s, 54°C 1 min, 72°C 1 min; followed by 72°C for 7 min. Finally, hold at 10°C.

- Analyze the PCR by loading 5 µl of the sample onto an agarose gel. The full-length PCR product (~1 kb) can be used directly for subsequent ribosome display cycles.

3.6. Cloning and Expression for Screening Binders

Ribosome display-selected populations can be cloned and expressed in *E. coli* in 96-deep well plates for high-throughput screening of individual binders. We have used the plasmid pET22b(+) system for expression of His-tagged scFv into the periplasm of *E. coli* BL21 (DE3). Protein expression is induced by growing clones in individual wells overnight in an Overnight Express™ Autoinduction System 1.

- Generate scFv fragments with the cloning sites NcoI and XhoI by PCR:

DreamTaq™ Green tag master mix	12.5 µl
Primer ScFv-NcoI/B (10 µM)	1 µl
Primer scFv-link/F (10 µM)	1 µl
PCR template from Subheading 3.4	1–10 ng
H ₂ O to	25 µl

Carry out 30 thermal cycles: 94°C 30 s, 54°C 1 min, 72°C 1 min; then, 72°C for 8 min. Finally, hold at 10°C.

- Analyze the PCR by agarose gel electrophoresis. If multiple bands are present, extract the band of correct size (~700 bp).
- Digest the PCR fragments using NcoI and XhoI.

Tango buffer (10×)	5 µl
NcoI (10 U/µl)	2.5 µl
XhoI (10 U/µl)	2.5 µl
Purified PCR fragment	50–100 ng
H ₂ O to	50 µl

Microwave the solution in the open 1.5 ml tube for 30 s at 800 W. At the same time, digest the pET 22b (+) using the same condition.

4. Analyze the digested fragments on agarose gel and extract the required band of correct size.
5. Set up DNA ligation as follows.

T4 Ligation buffer (10×)	1 µl
NcoI and XhoI digested pET 22b (+) (50 ng)	$x\mu\text{l}$ (see Note 11)
NcoI and XhoI digested fragments	$y\mu\text{l}$
T4 DNA ligase (1 U/µl)	1 µl
H ₂ O to	10 µl

Incubate at 16°C overnight, followed by transformation of BL21(DE3) according to the manufacturer's protocol using 2–4 µl of the ligation mixture.

6. Expression of scFvs in 96-deep well plates:
 - (a) Grow individual colonies in 96-deep well plates filled with 1 ml/well of auto inducible media containing ampicillin (100 µg/ml) for 16 h at 30°C with shaking at 250 rpm.
 - (b) Spin down at 2,500 rpm ($1,329 \times g$) for 20 min and remove the supernatant.
 - (c) Add 100 µl ice cold sucrose buffer into each well. Mix and incubate on ice for 30 min.
 - (d) Spin down at 2,500 rpm ($1,329 \times g$) for 20 min and collect the supernatants for direct use in ELISA or analysis by western blotting.

4. Notes

1. Any PCR polymerase, including high fidelity enzymes, can be used to construct libraries.
2. scFv DNA template(s) for the PCR can originate from either a single construct or various libraries.
3. A clean PCR fragment of the expected size indicates the successful construction. To confirm the construct, direct DNA sequencing is performed.
4. In this chapter, a coupled system was used for direct translation from linear PCR products. Alternatively, uncoupled rabbit reticulocyte lysate system (Promega) can be used by adding mRNA template(s), which are generated by a separate in vitro transcription step.
5. This step should avoid any input DNA amplification in the RT-PCR recovery, although the single primer PCR procedure

described has been designed not to amplify the input DNA construct (1).

6. Direct or indirect antigen immobilization can be used for panning of PRM complexes. If using direct approach, antigen is directly coated on a plastic surface, while in second approach biotinylated antigen is immobilized on streptavidin-coated wells. A negative control surface with the blocking reagent but lacking the immobilized antigen should also be used.
7. The length and number of washing steps can be increased when required. The quick washes with water produce cleaner RT-PCR without affecting the efficiency of DNA recovery.
8. A real-time PCR can be performed at this step to analyze the amount of cDNA recovered from ligand-selected PRM complexes. By comparing the DNA recovery between ligand and control wells, it is possible to validate the selected population and thus choose whether or not to repeat the cycles prior to DNA cloning and expression in *E. coli*.
9. The number of PCR cycles required depends on the amount of recovered cDNA. Adjust the PCR cycles to generate a visible band on agarose gel.
10. The number of ribosome display cycles required to enrich the ligand-binding molecules depends on the nature of the ligand, as well as the quality and diversity of the library used. Generally, three to five display cycles are usually sufficient to enrich the required protein from a library.
11. $x:y = 1:1$ in molar ratios.

Acknowledgments

Research at the Babraham Institute is supported by Biotechnology and Biological Sciences Research Council (BBSRC), UK. BBT is the subdivision of Babraham Institute. We thank Ms Hong Liu for the technical support.

References

1. He M, Taussig M (2007) Eukaryotic ribosome display with *in situ* DNA recovery. *Nat Methods* 4:281–288
2. He M, Khan F (2005) (Review) Ribosome display: next-generation display technologies for production of antibodies *in vitro*. *Expert Rev Proteomics* 2:421–430
3. He M, Taussig M (1997) Antibody-ribosome-mRNA (ARM) complexes as efficient selection particles for *in vitro* display and evolution of antibody combining sites. *Nucleic Acids Res* 25:5132–5134
4. Luginbühl B, Kanyo Z, Jones RM, Fletterick RJ, Prusiner SB, Cohen FE, Williamson RA, Burton DR, Plückthun A (2006) Directed evolution of an anti-prion protein scFv fragment to an affinity of 1 pM and its structural interpretation. *J Mol Biol* 363:75–97

5. Zahnd C, Amstutz P, Plückthun A (2007) Ribosome display: selecting and evolving proteins *in vitro* that specifically binds to a target. Nat Methods 4:269–279
6. Takahashi F, Ebihara T, Mie M, Yanagida Y, Endo Y, Kobatake E, Aizawa M (2002) Ribosome display for selection of active dihydrololate reductase mutants using immobilised methotrexate on agarose beads. FEBS Lett 514:106–110
7. Ohashi H, Shimizu Y, Ying BW, Ueda T (2007) Efficient protein selection based on ribosome display system with purified components. Biochem Biophys Res Commun 352:270–276

Chapter 5

Yeast Display of Engineered Antibody Domains

Qi Zhao, Zhongyu Zhu, and Dimiter S. Dimitrov

Abstract

Yeast display is an efficient technology for selection of antibodies and other proteins with high affinity and thermal stability. Here, we describe a method for affinity maturation of engineered antibody domains (eAds) using yeast display. EAd yeast libraries of relatively large size ($\sim 10^9$) were generated and subjected to alternating rounds of magnetic-activated cell sorting (MACS), fluorescent-activated cell sorting (FACS), and random mutagenesis. The highest affinity clones from the final round of maturation were identified and analyzed. We discuss extensively each key step, and provide detailed protocols and helpful notes.

Key words: Yeast display, Antibody domain, MACS, FACS

1. Introduction

Yeast display is emerging as an effective technology for isolating and engineering antibodies or proteins for therapeutics development and a variety of biomedical applications. In the yeast display system, the antibody or protein is displayed on the yeast surface by fusing to the yeast agglutinin protein Aga2p, which attaches to Agalp through two disulfide bonds. Expression of the antibody/protein-Aga2p and Agalp are under the control of galactose-inducible GAL1 promoter (1). One of the main advantages this technology offers is its eukaryotic system providing very sophisticated protein folding and chaperones machinery, which allows efficient and consistent display of variety of proteins. Recently, yeast surface display has been successfully used to engineer not only antibodies (2–5), but also a wide variety of proteins like fibronectin (6, 7), T cell receptors (TCRs) (8), natural killer cell receptors (9), and proteins of the major histocompatibility complex (MHC) (10) with dramatic improvements in stability and affinity. More importantly, a high-efficacy yeast electroporation protocol has been

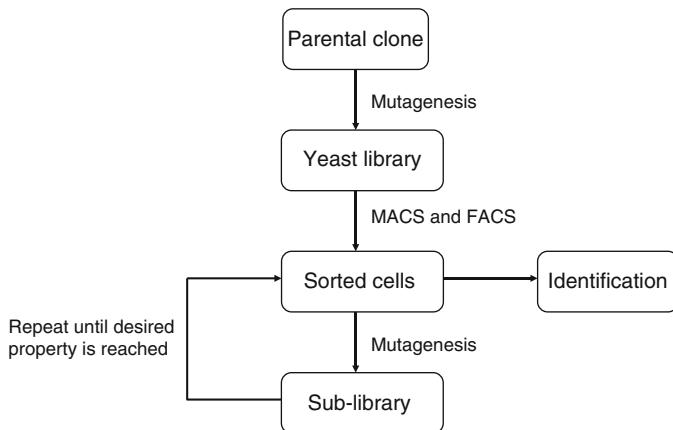


Fig. 1. Sketch of the protocol for selection and maturation of binders.

described (11) recently, which enables construction of yeast display library with large size up to 10^{10} . This makes yeast display comparable to phage display system in terms of library size and thus further simplifies the initial antibody or protein isolation process.

Isolated or engineered human immunoglobulin constant $\gamma 1$ CH2 domains were proposed as novel scaffolds for construction of libraries containing diverse binders of small size but still conferring some effector functions (12–15). Here, we describe a protocol for displaying, engineering, and characterization of CH2 domains against antigens of interest using yeast display (Fig. 1). The yeast-displayed CH2 mutant library with size at 10^9 was generated by transformation of yeast with linearized vector and mutant DNA inserts derived from CH2 domain through error-prone PCR. One round of magnetic-activated cell sorting (MACS) against the biotinylated antigen was performed first to downsize the initial library. The library was then sorted several times by fluorescent-activated cell sorting (FACS) to enrich for specific binders. The enriched CH2 domain mutants were further mutated by error-prone PCR of the entire gene pool to yield a new sub-library. The process of sorting and mutagenesis was then cyclically repeated until the cell population reached the desired property. Clones with the highest affinity from the final round of maturation were identified and their sequences were analyzed.

2. Materials

2.1. Error-Prone PCR

1. Bio-Rad MyCycler™ thermal cycler (Bio-Rad).
2. GeneMorph® II Random Mutagenesis Kit (Agilent).
3. 8-Oxo-deoxyguanosine triphosphate and 2'-deoxy-p-nucleoside-5'-triphosphate.

Table 1
PCR primers used

ERRORF	5' CTTCAGTTGGCCCAGGCAGGCC 3'
ERRORR	5' ACCACTAGTGGCCGGCCTG 3'
YDRDF	5' CTCGCTGTTCAATATTCTGTTATTGCTTCAGTTGGCC CAGGCGGCC 3'
YDRDR	5' GAGCCGCCACCCCTCAGAACGCCACCCTCAGAGCCACCACTAGT TGGGCCGGCCTG 3'

4. Gel extraction kit (Qiagen).
5. Primers shown in Table 1.
- 2.2. Preparation of DNA Inserts and Vector for Library Construction**
1. Bio-Rad MyCycler™ thermal cycler (Bio-Rad).
 2. AccuPrime™ Pfx DNA Polymerase (Invitrogen).
 3. Plasmid isolation kit (Qiagen).
 4. Gel extraction kit (Qiagen).
 5. Primers shown in Table 1.
 6. SfiI restriction enzyme (New England BioLabs).
 7. Yeast display vector pYD7 (see Note 1).
- 2.3. Preparation of Electroporation Competent Cells and Transformation of the Yeast Library**
1. 0.2-cm Gene Pulser/MicroPulser Cuvettes (Bio-Rad).
 2. Gene Pulser (Bio-Rad).
 3. Yeast strain EBY100.
 4. YPD medium: dissolve 20 g dextrose, 20 g peptone, and 10 g yeast extract in deionized H₂O to a volume of 1 L and sterilize by filtration. This medium can be stored for 2 months at 4°C.
 5. Lithium acetate (LiAc) and dithiothreitol (DTT) solution: dissolve 6.6 g LiAc and 1.54 g DTT in 1 L H₂O.
 6. 1 M sorbitol solution: dissolve 182 g sorbitol in 1 L H₂O.
 7. 1 M CaCl₂ solution: dissolve 111 g in 1 L H₂O.
 8. Electroporation buffer: add 1 ml of 1 M CaCl₂ solution into 1 L of 1 M sorbital solution.
 9. SDCAA medium: dissolve 20 g dextrose, 6.7 g Difco yeast nitrogen base w/o amino acid, 5 g Bacto casamino acids, 5.4 g Na₂HPO₄, and 8.56 g NaH₂PO₄·H₂O in H₂O to a volume of 1 L and sterilize by filtration.
 10. SDCAA plate: dissolve 5.4 g Na₂HPO₄ and 8.56 g NaH₂PO₄·H₂O, 182 g sorbitol and 15 g agar in H₂O to 0.9 L and autoclave. Dissolve 20 g dextrose, 6.7 g Difco yeast nitrogen base, and 5 g Bacto casamino acids in 100 ml H₂O and sterilize by filtration. Cool until 50°C, add filter-sterilized solution, and pour plates.

2.4. Growth and Induction of Yeast Cells

- SG/RCAA medium: dissolve 20 g galactose, 20 g raffinose, 1 g dextrose, 6.7 g Difco yeast nitrogen base w/o amino acid, 5 g Bacto casamino acids, 5.4 g Na₂HPO₄, and 8.56 g NaH₂PO₄·H₂O in H₂O to a volume of 1 L and sterilize by filtration (see Note 2).

2.5. MACS Selection

- AutoMACS equipment (Miltenyi).
- Streptavidin-conjugated microbeads (Miltenyi).
- PBSA buffer: dissolve 8 g NaCl, 0.2 g KCl, 1.44 g Na₂HPO₄, 0.24 g KH₂PO₄, 1 g bovine serum albumin in 1 L of H₂O, adjust the pH to 7.4 with HCl, and sterilize by filtration.

2.6. FACS Selection

- FACSAria sorter (BD Biosciences).
- Mouse anti-c-myc IgG (AbD Serotec).
- Alexa Fluor 488 goat anti-mouse IgG (Invitrogen).
- R-phycoerythrin-conjugated streptavidin (Invitrogen).

2.7. Affinity Determination

- FACScalibur Flow cytometer (BD Biosciences).

3. Methods

3.1. Error-Prone PCR

- Randomly mutate the entire CH2 DNA fragment using Stratagene GeneMorph® II Random Mutagenesis Kit. Oligonucleotides used for error-prone PCR are described in Table 1.
- Perform PCR in a 50-μl reaction containing 1× Mutazyme II reaction buffer, 0.5 μM each of primers ERRORF and ERRORR, 0.2 mM (each) dNTPs, 1 ng of plasmid template, 2 μM 8-oxo-deoxyguanosine triphosphate, 2 μM 2'-deoxy-p-nucleoside-5'-triphosphate, and 2.5 U of Mutazyme II DNA polymerase (see Note 3).
- Error-prone PCR was performed using the following conditions: denature at 95°C for 2 min, 35 cycles of 95°C for 1 min, 60°C for 1 min, and 72°C for 1 min, followed by extension at 72°C for 10 min.
- Purify the amplified products from agarose gel with the QIAquick Gel Extraction Kit and determine the DNA concentration through spectrophotometer measurement.

3.2. Preparation of DNA Inserts and Vector for Library Construction (see Note 4)

- Oligonucleotides used for PCR amplification are described in Table 1.
- Perform PCR in four 100-μl PCR reactions containing 1× Accuprime PCR reaction mix, 1 μM of primers YDRDF and

YDRDR, 120 ng of error-prone PCR product, and 2.5 U of Accuprime pfx DNA polymerase.

3. Denature at 95°C for 2 min, cycled 30 times at 95°C for 30 s, 60°C for 30 s, and 72°C for 30 s, and finally extended at 72°C for 2 min.
4. Purify all amplified products from agarose gel with the QIAquick Gel Extraction Kit and determine the DNA concentration through spectrophotometer measurement.
5. Digest the pYD7 vector with SfiI restriction enzyme and purify it with gel extraction kit following the instructions from manufacturer.
6. Combine 12 µg of mutagenized insert DNA and 4 µg of digested vector and concentrate the volume to 20 µl using a vacuum concentrator.

3.3. Preparation of Electroporation

Competent Cells and Transformation of the Yeast Library

1. Inoculate EBY100 yeast colony into 5 ml YPD medium and grow overnight at 30°C.
2. Inoculate the overnight culture into 50 ml fresh YPD medium and grow cells at 30°C until OD at 600 nm reach 1.6 (about 4–5 h).
3. Once cells have reached an absorbance of about 1.6 at 600 nm, pellet cells at $2,500 \times g$ for 3 min at 4°C and wash with 25 ml cold water twice. Wash cell once with 25 ml cold electroporation buffer.
4. Resuspend cells in 10 ml of LiAc/DTT solution and incubate at 30°C with shaking at 250 rpm for 30 min.
5. Pellet cells at $2,500 \times g$ for 3 min at 4°C and washed once with 25 ml of electroporation buffer.
6. Resuspend cells in electroporation buffer to reach 0.4 ml volume. Mix concentrated DNA with resuspended cells and keep cells on ice.
7. Aliquot 0.4 ml of resuspended cell-DNA mixture per prechilled electroporation cuvette. Keep electroporation cuvettes on ice until pulsed.
8. Load cuvette into gene pulser and electroporate at 2.5 kV and 25 µF. Immediately add 1 ml of the warm (30°C) mixture of 1 M sorbitol:YPD medium (1:1) to the cuvette. Typical time constants for electroporation range from about 3 ms to 4 ms without greatly affecting transformation efficiency.
9. Transfer cells from pulsed cuvettes to a 50-ml Falcon tube. Wash each cuvette with an additional 1 ml of the mixture of 1 M sorbitol:YPD medium to recover the remaining cells from the cuvettes.
10. Make up the volume of cells to 10 ml by the mixture of 1 M sorbitol:YPD medium and incubate it at 30°C with shaking at 250 rpm for 1 h.

11. Pellet cells at $2,500 \times g$ for 3 min and remove supernatant. Resuspend cells in 500 ml SDCAA medium. Incubate at 30°C with 250 rpm shaking for 24–48 h. Plate serial dilutions of the transformed cells on SDCAA plates to determine transformation efficiency (see Note 5).
12. Passage the library at least once before use to reduce the number of untransformed cells. Store the library at 4°C or –80°C (see Note 6).

3.4. Growth and Induction of Yeast Cells

3.4.1. Libraries

1. Thaw frozen aliquots of yeast library at room temperature and grow overnight (about 20 h) in 0.5–1 L of SDCAA medium in an incubator shaker set to 30°C with 250 rpm (see Note 7).
2. Pellet at least 5×10^9 cells from a freshly passaged library culture at $2,500 \times g$ for 3 min and resuspend cells in SG/RCAA medium to an absorbance of about 0.5–1 at 600 nm.
3. Induce the library in SG/RCAA medium at 20°C with shaking at 250 rpm for 16–18 h.

3.4.2. Individual Clones

1. Inoculate single colony into 1–2 ml of SDCAA medium and grow at 30°C with shaking at 250 rpm overnight.
2. Resuspend cell pellet in S/GRCAA to an absorbance of about 0.5–1 at 600 nm and incubate at 20°C with 250 rpm for 16–18 h.
3. Wash cells once in wash buffer before staining for flow cytometry, or store the cells by placing the tubes at 4°C for up to 1 month.

3.5. MACS Selection (see Note 8)

1. Pellet 5×10^9 freshly induced yeast cells at $2,500 \times g$ for 5 min and aspirate the supernatant. To wash, resuspend cells in 50 ml PBSA buffer, repellet cells, and discard supernatant.
2. To label yeast, resuspend cells in 10 ml PBSA buffer. Add biotinylated antigen to a final concentration of 100 nM and mix by gentle inversion (see Note 9).
3. Incubate cell suspension at room temperature with gentle rotation on a tube rotator for 1 h, followed by 10 min on ice.
4. Pellet cells at $2,500 \times g$ for 3 min and wash once with 25 ml PBSA buffer.
5. Resuspend pellet in 5 ml PBSA buffer, add 100 µl streptavidin-conjugated microbeads to the suspension and mix by gentle inversion.
6. Incubate on ice for 10 min with gentle mixing by inversion every 2 min.
7. Add 20 ml PBSA buffer to the suspension and gently mix by inversion.

8. Turn on autoMACS and run the “Clean” program. For detailed instructions on using the autoMACS system, refer to the user’s manual.
9. Under the “pos2” port of the instrument, place a 15-ml tube filled with 2 ml SDCAA medium to collect eluted cells. Under the “neg1” port, place an empty 50-ml tube to collect the flow-through. Place the cell suspension under the intake port and choose separation protocol “Possel_s” to begin the separation.
10. Pellet eluted cells at $2,500 \times g$ for 3 min to get rid of the sodium azide in the mixture and resuspend cells in fresh 10 ml SDCAA medium (see Note 10).
11. Propagate eluted cells overnight at 30°C in a shaker with speed set at 250 rpm, followed by induction in SG/RCAA medium at 20°C before subsequent sorting by FACS.

3.6. FACS Selection

1. Induce 5×10^7 cells from MACS sorted pool at 20°C with shaking speed set at 250 rpm for 16–18 h.
2. Pellet 5×10^7 induced cells at $2,500 \times g$ for 3 min.
3. Label yeast cells with mouse anti-c-myc IgG (2 µg/ml) and an appropriate concentration of biotinylated antigen in an appropriate final volume of PBSA buffer. Antigen concentrations are chosen based on the expected dissociation constant (K_D) of the parental binder. The antigen incubation volume must be large enough to allow at least tenfold excess of antigen over binders displayed on yeasts (see Note 11).
4. Incubate at room temperature for 3 h or 4°C overnight with rotation (see Note 12).
5. After incubation, wash cells three times with PBSA buffer and then resuspend cells in 1 ml PBSA buffer.
6. Stain cells with both 1:100 dilution of R-phycoerythrin-conjugated streptavidin and Alexa Fluor 488 conjugated goat anti-mouse IgG antibody at 4°C for 30 min.
7. Wash the cells three times with 5 ml PBSA buffer again and then resuspend in PBSA buffer for flow cytometric sorting. Sorting gates are determined to select only the population with higher antigen-binding signals (see Note 13).
8. Grow collected cells overnight in SDCAA medium at 30°C and induce in 10 ml SG/RCAA at 20°C for 18 h for the next round of sorting.
9. Repeat steps 2–8 for additional rounds. Approximately 1×10^7 yeast cells were used for staining at a concentration of $0.1 \times$ to $0.01 \times$ the concentration at K_D . The addition of anti-c-myc and the staining with secondary reagents is as described above (see Note 14).

10. Isolate yeast plasmids using Zymoprep yeast plasmid kit according to the manufacturer's instructions and use them as templates of sub-library construction (see Note 15).
11. Spread cells from the final round on SDCAA plates to identify single clones; affinities of monoclonal yeast display antibody can be measured and compared by FACS-based analysis.

3.7. Affinity Determination

1. Inoculate a 2-ml SDCAA culture with a clone of interest and grow overnight at 30°C.
2. Inoculate a 2-ml SG/RCAA culture with 5×10^6 cells and induce at 20°C for at least 18 h.
3. Set up nine tubes containing 1×10^5 induced cells. Pellet at $2,500 \times g$ for 3 min and wash once using 200 µl PBSA buffer.
4. Choose nine different concentrations of biotinylated antigens around the equilibrium dissociation constant. Incubation volumes and number of yeast stained are chosen to keep the number of antigen molecules in tenfold excess above the number of binder (see Note 16). An example experimental setup is in Table 2.
5. To each tube, add the appropriate volumes of buffer, cells, and antigen.
6. Place tubes for 3 h at room temperature. This allows ample time for the binding reaction to reach equilibrium (see Note 17).

Table 2
Example setup for K_d determination

Tube	Concentration of Ag	Molecules of Ag	Molecules of Ab ^a	Volume
1	2 µM	6×10^{13}	5×10^9	50 µl
2	1 µM	3×10^{13}	5×10^9	50 µl
3	500 nM	3×10^{13}	5×10^9	100 µl
4	250 nM	1.5×10^{13}	5×10^9	100 µl
5	100 nM	6×10^{12}	5×10^9	100 µl
6	50 nM	3×10^{12}	5×10^9	100 µl
7	25 nM	1.5×10^{12}	5×10^9	100 µl
8	12.5 nM	7.5×10^{11}	5×10^9	500 µl
9	0	0	5×10^9	100 µl

Ag antigen, Ab antibody

^a 1×10^5 cells are used. Assuming 5×10^4 Abs per cell

7. Pellet cells at $2,500 \times g$ for 3 min, aspirate supernatant, and wash cells with 500 μ l PBSA buffer.
8. Add 50 μ l PBSA buffer with streptavidin-phycoerythrin (1:100 dilution) to each tube. Resuspend the cells and mix by pipetting.
9. Incubate on ice for 30 min, shielding from light.
10. Pellet cells at $2,500 \times g$ for 3 min, aspirate supernatant, and wash cells with 500 μ l PBSA buffer.
11. Analyze cells of each tube using a flow cytometry.
12. Plot the total mean fluorescence intensity (MFI) from the phycoerythrin channel against the concentration of antigen and use a nonlinear least squares to fit the curve (such as found in GraphPad) and determine the K_D using the following equation: $Y = B_{\max} \times X / (K_D + X) + B_{\min}$, where Y =MFI at given antigen concentration, X =antigen concentration, B_{\max} =MFI at saturation, and B_{\min} =MFI of no antigen control.

4. Notes

1. pYD7 was modified from vector pCTCON2 (6). Two SfiI restriction sites before and after the insert match the cloning sites of phage display vector pComb3X were introduced into pYD7, which allows identified antibody genes shuttling among different vector for display on yeast and soluble expression in bacteria. In addition, protein of interest for display was put at the N-terminus of Aga2p instead of the C-terminus as it is in pCTCON2.
2. Raffinose and low concentration of dextrose in the SGCAA medium increase the efficiency of protein display on yeast cell surface.
3. Two nucleotide analogues (8-oxo-deoxyguanosine triphosphate and 2'-deoxy-p-nucleoside-5'-triphosphate) were used in the error-prone PCR reaction mixture to further broaden the mutation diversity during the error-prone PCR process. The amount of the template and the number of cycles used in the PCR can be adjusted according to desired mutation rate, decreasing the template concentration or increasing the PCR cycle number will increase the mutation rate. For determining the mutation frequency, the mutagenized genes can be ligated into sequencing vector, e.g., TA cloning vector, for sequencing.
4. In order to obtain large amounts of DNA insert and increase the efficacy of gap-repairing process *in vivo*, mutated DNA repertoire derived from error-prone PCR must be re-amplified using primers YDRDF and YDRDR under normal PCR conditions.

5. Typically, $1\text{--}4 \times 10^8$ transformants could be obtained from each transformation. The number of electroporation reaction is determined based on the desired size of the yeast library.
6. Grown library culture can be stored at 4°C for about 1 month. Frozen aliquots with 10% glycerol can be made for long-term storage at -80°C freezer or liquid nitrogen. Cells should be freshly passaged before induction.
7. The number of cells thawed from the frozen stock should be at least tenfold of the initial library size to reduce the probability of losing the diversity of the library.
8. The sorting capacity of MACS can reach to 5×10^9 yeast cells per sorting, which allows elimination of yeast cells that do not express antibodies, and quickly downsizes the initial library size and makes it compatible for subsequent FACS-based sorting. All of these steps except antigen incubation should use ice-cold buffers.
9. Antigen concentrations for initial sorting with MACS vary with the antigen of interest. 100 nM is generally sufficient to enrich antibody binders.
10. Remove buffer from sorted cells due to the presence of sodium azide in MACS running buffer before amplifying the cells in SDCAA medium. To avoid bacterial contamination after magnetic sorting, grow yeast cultures in SDCAA medium with pen-strep antibiotics.
11. Selections are generally performed after allowing the reaction to reach equilibrium. The volume for incubation of yeast is chosen to keep antigens at least a tenfold excess of the number of antibodies. For example, 5×10^7 cells are incubated in 1 ml labeling volume. Assuming 5×10^4 Ab per yeast cell, the Ab concentration in the sample is calculated as following: $(5 \times 10^4 \text{ Ab per cell}) \times (5 \times 10^7 \text{ cells in 1 ml}) = 4.15 \text{ nM}$. Therefore, the lowest antigen concentration is 41.5 nM.
12. Incubation times are chosen to approach equilibrium of the reaction. The time constant is defined as: $t = -\ln(1 - \theta) / (k_{\text{on}} \times C + k_{\text{off}})$, where t is time, θ is equilibrium, k_{on} is association rate constant, k_{off} is dissociation rate constant, and C is antigen concentration (2). In general, 3 h at room temperature is sufficient for antigen concentrations that are in nanomolar range. Alternatively, incubate the antigen with cells overnight at 4°C .
13. Typically, sorting gate should be set to sort out the brightest 0.1–0.3% of antigen-binding and c-myc positive clones.
14. The sorting at $1/10$ antigen concentration at K_D of the parental clone can differentiate affinity improved mutants from the parental clone and avoid the presence of the parental clone in the sort gate window.

15. Copies of plasmids are very low in yeast cells. Amount of the yeast cell culture used for the miniprep need to be optimized to ensure good-quality plasmid when used as template for subsequent error-prone PCR, using too many cells may result in even lower yield and quality.
16. If the K_D is unknown, use several concentrations across a wide range and then focus the concentration range around K_D in a subsequent experiment.
17. For the K_D in nanomolar range, 3 h is usually sufficient. For the higher affinity binders, longer incubation times are required, see Note 12.

Acknowledgments

We thank Professor Dane Wittrup for providing the yeast display vector pCTCON2 and yeast strain EBY100 and members of our group for helpful discussions. This project was supported by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research and by the Gates Foundation (DSD).

References

1. Chao G, Lau WL, Hackel BJ, Sazinsky SL, Lippow SM, Wittrup KD (2006) Isolating and engineering human antibodies using yeast surface display. *Nat Protoc* 1:755–768
2. Garcia-Rodriguez C, Levy R, Arndt JW, Forsyth CM, Razai A, Lou J, Geren I, Stevens RC, Marks JD (2007) Molecular evolution of antibody cross-reactivity for two subtypes of type A botulinum neurotoxin. *Nat Biotechnol* 25:107–116
3. Walker LM, Bowley DR, Burton DR (2009) Efficient recovery of high-affinity antibodies from a single-chain Fab yeast display library. *J Mol Biol* 389:365–375
4. Wang Z, Kim GB, Woo JH, Liu YY, Mathias A, Stavrou S, Neville DM Jr (2007) Improvement of a recombinant anti-monkey anti-CD3 diphtheria toxin based immunotoxin by yeast display affinity maturation of the scFv. *Bioconjug Chem* 18:947–955
5. Zhao Q, Feng Y, Zhu Z, Dimitrov DS (2011) Human monoclonal antibody fragments binding to insulin-like growth factors 1 and 2 with picomolar affinity. *Mol Cancer Ther* 10:1677–1685
6. Hackel BJ, Kapila A, Wittrup KD (2008) Picomolar affinity fibronectin domains engineered utilizing loop length diversity, recursive mutagenesis, and loop shuffling. *J Mol Biol* 381:1238–1252
7. Lipovsek D (2011) Adnectins: engineered target-binding protein therapeutics. *Protein Eng Des Sel* 24:3–9
8. Richman SA, Healan SJ, Weber KS, Donermeyer DL, Dossett ML, Greenberg PD, Allen PM, Kranz DM (2006) Development of a novel strategy for engineering high-affinity proteins by yeast display. *Protein Eng Des Sel* 19:255–264
9. Dam J, Guan R, Natarajan K, Dimasi N, Chlewicki LK, Kranz DM, Schuck P, Margulies DH, Mariuzza RA (2003) Variable MHC class I engagement by Ly49 natural killer cell receptors demonstrated by the crystal structure of Ly49C bound to H-2K(b). *Nat Immunol* 4:1213–1222
10. Esteban O, Zhao H (2004) Directed evolution of soluble single-chain human class II MHC molecules. *J Mol Biol* 340:81–95

11. Benatuil L, Perez JM, Belk J, Hsieh CM (2010) An improved yeast transformation method for the generation of very large human antibody libraries. *Protein Eng Des Sel* 23:155–159
12. Dimitrov DS (2009) Engineered CH2 domains (nanoantibodies). *MAbs* 1:26–28
13. Xiao X, Feng Y, Vu BK, Ishima R, Dimitrov DS (2009) A large library based on a novel (CH2) scaffold: identification of HIV-1 inhibitors. *Biochem Biophys Res Commun* 387:387–392
14. Gong R, Vu BK, Feng Y, Prieto DA, Dyba MA, Walsh JD, Prabakaran P, Veenstra TD, Tarasov SG, Ishima R, Dimitrov DS (2009) Engineered human antibody constant domains with increased stability. *J Biol Chem* 284:14203–14210
15. Gong R, Wang Y, Feng Y, Zhao Q, Dimitrov DS (2011) Shortened engineered human antibody CH2 domains: increased stability and binding to the human neonatal receptor. *J Biol Chem* 286:27288–27293

Chapter 6

Expression, Purification, and Characterization of Engineered Antibody CH2 and VH Domains

Rui Gong, Weizao Chen, and Dimiter S. Dimitrov

Abstract

Most of the FDA-approved therapeutic monoclonal antibodies are full-size IgG molecules with a molecular weight of about 150 kDa. A major problem for such large molecules is their poor penetration into tissues (e.g., solid tumors) and poor or absent binding to regions on the surface of some molecules (e.g., on the HIV envelope glycoprotein) which are fully accessible only by molecules of smaller size. Therefore, much work especially during the last decade has been aimed at developing novel scaffolds of much smaller size and high stability. Immunoglobulin-based scaffolds including Fab (~50 kD), ScFv (~30 kD), and VH domain (termed domain antibody, dAb) (~15 kD) have been well established. Recently, a new scaffold based on human IgG1 CH2 domain (~15 kD) was also proposed (termed nanoantibody, nAb). Binders based on a CH2 scaffold could also confer some effector functions. Here, we describe the design, expression, purification, and characterization of engineered CH2 and VH domains.

Key words: Domain antibody, VH, Nanoantibody, CH2, Library construction, Phage display, Stability, Engineered antibody domains

1. Introduction

Monoclonal antibodies (mAbs) with high affinity and specificity are now well-established therapeutics and invaluable tools for biological research. It appears that their use will continue to expand in both targets and disease indications. However, because of the fundamental problem for full-size mAbs, a large amount of work has been aimed at developing small-size binders with scaffolds based on various highly stable human and nonhuman molecules during the last decade (1–8). A promising direction is the development of binders based on the heavy or light chain variable region of an antibody; these fragments with size ranging from 11 to 15 kDa were called “domain antibodies” or “dAbs” (7, 9). A unique kind of antibodies composed only of heavy chains (designated HCabs) (10)

are naturally formed in camels, dromedaries and llamas, and their variable regions (referred to as $V_H H$) can also recognize antigens as single domain fragments (11). Not only is the overall size of the dAbs much smaller than that of full-size antibodies but also their paratopes are concentrated over a smaller area so that the dAbs provide the capability of interacting with novel epitopes that are inaccessible to conventional antibodies or antibody fragments with paired light and heavy chain variable domains.

The structure of the antibody constant domains is similar to that of the variable domains consisting of β strands connected mostly with loops or short helices. The second domain of the α , δ , and γ heavy chain constant regions, CH2, is unique in that it exhibits very weak carbohydrate-mediated interchain protein–protein interactions in contrast to the extensive interchain interactions that occur between other domains. The expression of murine and human CH2 in bacteria which does not support glycosylation results in a monomeric protein (12, 13). It has been hypothesized that the CH2 domain (CH2 of IgG, IgA and IgD, and CH3 of IgE and IgM) could be used as a scaffold and could offer additional advantages compared to those of the dAbs because it contains binding sites or portions of binding sites conferring effector and stability functions (termed nanoantibodies, Nabs) (14). It was found previously that an isolated murine CH2 was relatively unstable at physiological temperature with a temperature of 50% unfolding (T_m) slightly higher than 37°C, while the T_m value of human CH2 was 54.1°C which was still significantly lower than that of other small scaffolds such as the 10th type III domain of human fibronectin (FN3) (12, 13, 15). The relatively low stability of human CH2 increases the probability of protein aggregation or degradation when it is engineered for binding to antigens and mediating effector functions. Therefore, further improvement of stability of the CH2 scaffold is important.

Here we describe the design, expression, purification, and characterization of engineered CH2 domains, construction of a CH2-based phage display library, a methodology to increase stability of CH2 and comparison with VH-based engineered domains. We use the term engineered antibody domains (eAds) to denote both variable and constant domains that are engineered to confer functions additional to their native ones.

2. Materials

2.1. Cloning of CH2 into Phagemid pComb3X

- Primers are designed based on sequence of human IgG1 constant domain.

Sfi I

CH2 upstream: 5'- TTC GCT ACC GTG GCC CAG GCG
GCC GCA CCT GAA CTC CTG GGG GGA CC -3'

Sfi I

CH2 downstream: 5'- GTG ATG GTG CTG GCC GGC CTG
GCC TTT GGC TTT GGA GAT GGT TTT CTC -3'

2. High Fidelity PCR Master (Roche, Indianapolis, IN), or other high-fidelity PCR systems may be used.
3. QIAquick Gel Extraction Kit (Qiagen, Valencia, CA).
4. Restriction enzyme Sfi I (New England Biolabs, Ipswich, MA).
5. Vector phagemid pComb3X ([16](#)).
6. Ligation kit: Rapid DNA Ligation Kit (Roche, Indianapolis, IN).
7. Plasmid Mini Kit (Qiagen, Valencia, CA).
8. 2× YT medium (1 L): Tryptone, 16 g; yeast extract, 10 g and NaCl, 5 g.
9. Ampicillin (Sigma-Aldrich, St. Louis, MO): stock 100 mg/ml store -20°C, working at 100 µg/ml.
10. *E. coli* K12ΔD(*lac-proAB*) supE thi hsdD5/F' traD36 proA+B lacIq lacZΔM15.

2.2. Expression and Purification of CH2 Domain

1. SB medium (1 L): Tryptone, 30 g; yeast extract, 20 g; MOPS, 10 g; adjust pH value to 7.0 with 1 M NaOH.
2. IPTG (BioGolden, MO): stock 1 M, working at 1 mM as inducer on the lacZ suppressor for HB2151 cell expression.
3. Buffer A: 50 mM Tris-HCl, 450 mM NaCl, pH 8.0.
4. Buffer B: Buffer A + 200 mM Imidazole.
5. Polymyxin B sulfate: 0.5 µg/ml (Sigma-Aldrich, St. Louis, MO).
6. Nickel column: 1 ml HiTrap Chelating HP Ni-NTA column (GE Healthcare, NJ).
7. FPLC (GE Healthcare, NJ).
8. Protein loading buffer (6×): 0.35 M Tris-HCl pH 6.8, 10.28% SDS, 0.6 M dithiothreitol (DTT), 36% glycerol (V/V), and 0.06% bromophenol blue, store at -20°C.
9. *E. coli* HB2151: K12 ara Δ(*lac-proAB*) thi/F' proA+B lacIq lacZΔM15.

2.3. Primers for Library Construction

A published strategy for library construction was used here as an example ([17](#)).

Sfi I

N terminus primer: 5' ACGT GGCCCAGGCGGCC GCA CCT
 GAA CTC CTG 3'

Sfi I

C terminus primer: 5' ACGT GGCCGGCCT GGCC TTT GGC
 TTT GGA GAT GGT TTT CTC GAT G 3'

Loop BC primer Fw: 5' AAG TTC AAC TGG TAC GTG 3'

Loop BC primer Rv: 5' CAC GTA CCA GTT GAA CTT GCC
 AKM
 CAC CAC CAC GCA TGT GAC 3'

Loop FG primer Rv: 5' GAT GGT TTT CTC GAT GGG GCC
 AKM AKM AKM AKM AKM AKM GTT GGA GAC CTT
 GCA CTT G 3'

2.4. Ligation of CH2 Fragments Containing Mutations with Phagemids

1. Restriction enzymes SfiI: see above.
2. T4 DNA Ligase, 400,000 units/ml (New England Biolabs, Ipswich, MA).

2.5. Purification, Concentration, and Desalting of Ligations

1. QIAquick PCR Purification Kit (Qiagen, Valencia, CA).
2. Centrifugal filter: Amicon Ultra-4 with a cutoff of 3,000 MW (Millipore, Billerica, MA).

2.6. Electroporations

1. TG1 electroporation-competent cells (Stratagene, La Jolla, CA).
2. Gene Pulser/MicroPulser Cuvettes (Bio-Rad, Hercules, CA).
3. Gene Pulser (Bio-Rad, Hercules, CA).
4. S.O.C. Medium: 2% tryptone, 0.5% yeast extract, 10 mM sodium chloride, 2.5 mM KCl, 10 mM MgCl₂, 10 mM MgSO₄, 20 mM glucose.

2.7. Preparation of CH2-Based Library

1. 2× YT medium: see above.
2. Glucose (Sigma-Aldrich, St. Louis, MO): Stock 20% keep at room temperature, sterilized by filtration.
3. M13KO7 helper phage (Invitrogen, Carlsbad, CA).
4. Antibiotics: 100 mg/ml ampicillin and 50 mg/ml kanamycin.
5. HiSpeed Plasmid Maxi Kit.

2.8. Primers for Construction of CH2 Mutant (m01) with Increased Stability

Omp: 5' AAG ACA GCT ATC GCG ATT GCA G 3'

gIIIF: 5' ATC ACC GGA ACC AGA GCC ACC AC 3'

L/C primer Fw: 5' TCA GTC TTC TGC TTC CCC CCA AAA
 CCC AAG GAC 3'

L/C primer Rv: 5' TGG GGG GAA GCA GAA GAC TGA CGG
 TCC CCC CAG 3'

K/C primer Fw: 5' CCC ATC GAG TGC ACC ATC TCC AAA
 GCC AAA GGC 3'

K/C primer Rv: 5' GGA GAT GGT GCA CTC GAT GGG GGC
 TGG GAG GGC 3'

2.9. Measurement of the Antibody Domain Stability

1. Circular dichroism (CD): AVIV Model 202 CD Spectrometer (Aviv Biomedical, NJ).
2. Differential scanning calorimetry (DSC): VP-DSC MicroCalorimeter (MicroCal, MA).
3. Spectrofluorometry: Fluorometer Fluoromax-3 (HORIBA Jobin Yvon, NJ).

3. Methods

Cloning, expression, and purification of a CH2 domain are performed based on modification of methods used for dAbs.

3.1. RNA Isolation and cDNA Synthesis

1. Lymphocyte isolation and cDNA synthesis are performed according to previous protocol ([16](#)).

3.2. Cloning of CH2 into Phagemid pComb3X

1. PCR for CH2 domain amplification:

Mix (for one reaction)

Reagent	Volume (μl)	Final concentration
Sterile double-dist. water	21	
PCR Master Mix (2×)	25	1×
CH2 upstream primer (10 μM)	1.5	300 nM
CH2 downstream primer (10 μM)	1.5	300 nM
cDNA (1 μg/μl)	1	1 μg
Final volume	50	

Thermal cycling.

	Temperature	Time	Cycles
Initial denaturation	94°C	2 min	1×
Denaturation	94°C	15 s	10×
Annealing	55°C	30 s	
Elongation	72°C	25 s	
Denaturation	94°C	30 min + 5 s	25×
Annealing	55°C	Cycle elongation	
Elongation	72°C	for each successive cycle	
Final elongation	72°C	7 min	1×
Cooling	4°C	Forever	

2. Digestion of insert DNA and vector.

Components	CH2 fragment (μ l)	pComb3X (μ l)	Final concentration
Insert DNA	x (2 μ g)	—	
Vector pComb3x	—	y (10 μ g)	
10× NEBuffer 2	5	5	1×
BSA (10 mg/ml)	0.5	0.5	100 μ g/ml
SfiI (20 u/ μ l)	4	3	
ddH ₂ O	40.5 - x	41.5 - y	
Final volume	50		

3. Purification of digested insert and vector.

The digested products are run on agarose gel and purified according to the manual of the Kit.

4. Ligation.

Dissolve insert DNA and vector DNA in Dilution Buffer (5×) to a final volume of 5 μ l in a sterile reaction vial as follows:

Components	Volume (μ l)	Final concentration
Insert DNA (digest CH2 fragment) (see Note 1)	3	
Vector DNA (digested pComb3X)	1	
DNA dilution buffer (5×)	1	1×
ddH ₂ O	—	
Total	5	

Add 5 μ l T4 DNA Ligation Buffer (2×) to each reaction vial;

Add 0.5 μ l T4 DNA Ligase (5 U/ μ l);

Mix thoroughly;

Incubate for 5 min at 15–25°C.

5. Transformation of TG1 competent cells with the ligation product.

6. Plasmid extraction.

The positive clone is verified by direct sequencing and used for transformation of HB2151 for expression.

3.3. Expression and Purification of CH2

- Pick up single colony from fresh transformants and grow it in 5 ml SB medium containing 100 μ g/ml Amp with shaking at 250 rpm and 37°C.

2. After incubation for about 3 h, transfer the 5 ml culture to a 2-L flask containing 500 ml SB medium with 100 µg/ml Amp. Shake the flask at 250 rpm and 37°C.
3. When the OD600 reaches 0.7–1, add 500 µl 1 mM IPTG to induce protein expression with shaking at 250 rpm and 37°C overnight (see Note 2).
4. Harvest the cells by centrifugation at 6,000 rpm (~ 4,500×*g*, rotor: JLA-10.500, Beckman Coulter, CA) for 15 min at 4°C. Remove the supernatant and resuspend the pellet in 50 ml PBS (pH 7.4), and add 500 µl polymyxin B sulfate.
5. Rotate the tube at room temperature for 30 min.
6. Centrifuge the sample at 15,000 rpm (~18,000×*g*, rotor: JA-20, Beckman Coulter, CA) at 4°C for 45 min. The supernatant is filtered by using a 0.45-µM filter (Nalgene, NY). Then FPLC is used for protein purification.
7. Wash the pump A and B with corresponding buffer A and B.
8. Before loading the sample from the Superloop to the HiTrap Chelating HP, allow 5 ml buffer A to flow through the route.
9. Load the sample from the Superloop to the HiTrap Chelating HP. Collect flow through for analysis on SDS-PAGE to examine the efficiency of the HiTrap Chelating HP.
10. Wash the column by 15 ml buffer A.
11. Wash the column by 10 ml buffer B with increasing concentration from 0 to 20% (see Note 3).
12. Increase the concentration of buffer B to 100%.
13. Collect the fractions eluted by 8 ml buffer B with 0.8-ml aliquot per tube.
14. Stop the pump after elution. Wash the system by ddH₂O including Superloop, pump, injector, etc.
15. Analyze the collections on SDS-PAGE.

3.4. Construction of a CH2-Based Library of Mutants

The strategies for library construction are various. Here, one published strategy is used as an example (Fig. 1) (17). The CH2 structure is represented by VMD 1.8.7 (18). Limited mutagenesis (A, S, Y, D) is introduced to loop BC (loop 1) and loop FG (loop 3). The procedure is shown (Fig. 2)

1. First round of PCR to get fragment 1 with mutations on loop BC (loop 1) and fragment 2 with mutations on loop FG (loop 3).

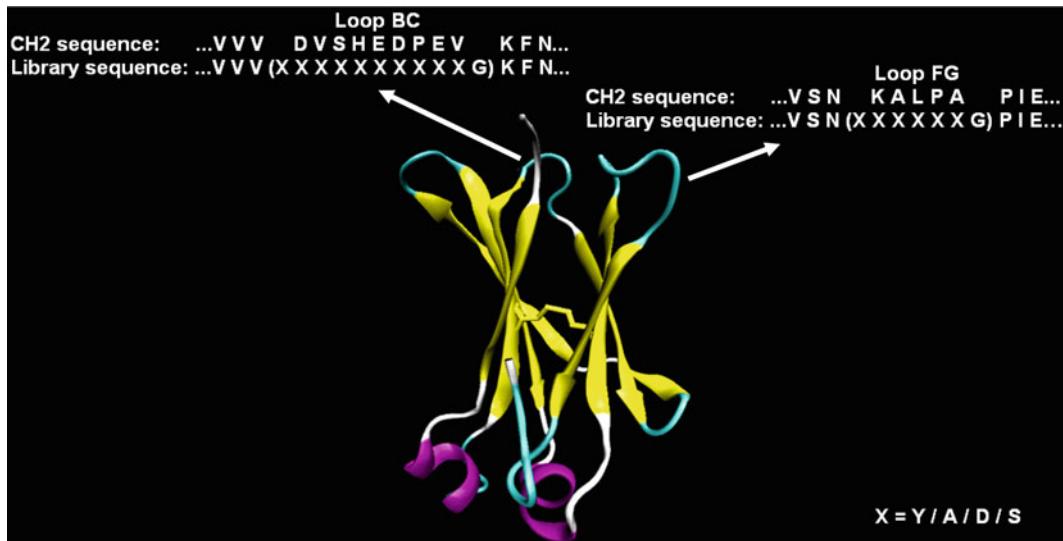


Fig. 1. Design of a CH2-based library (17). Mutations are introduced in loop BC and loop FG with two additional amino acids.

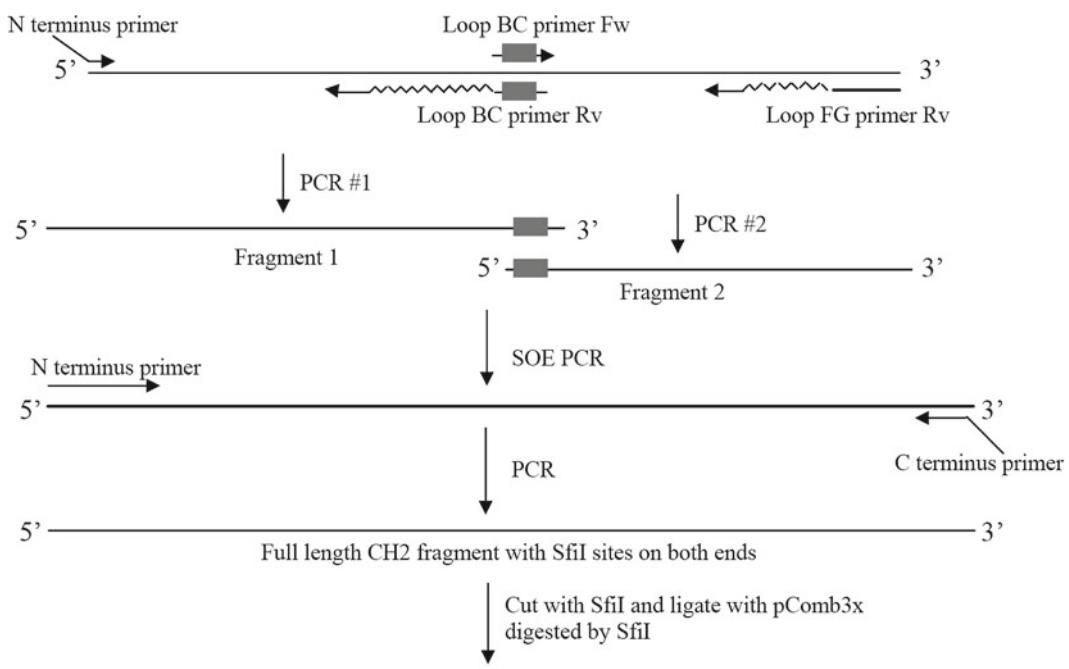


Fig. 2. PCR strategy for amplification of CH2 fragments with mutations for library construction.

For amplification of fragment 1.

Reagent	Volume (μl)	Final concentration
Sterile double-dist. water	43	
PCR Master Mix (2 \times)	50	1 \times
N terminus primer (10 μM)	3	300 nM
Loop BC primer Rv (10 μM)	3	300 nM
pComb3X-CH2 (2 ng/ μl)	1	0.02 ng/ μl
Final volume	100	

For amplification of fragment 2.

Reagent	Volume (μl)	Final concentration
Sterile double-dist. water	43	
PCR Master Mix (2 \times)	50	1 \times
Loop BC primer Fw (10 μM)	3	300 nM
Loop FG primer Rv (10 μM)	3	300 nM
pComb3X-CH2 (2 ng/ μl)	1	0.02 ng/ μl
Final volume	100	

- After purification of these two fragments, SOE (Splicing by Overlapping Extension) PCR is performed to get intact DNA fragment with mutations on both loop BC and loop FG. The volume of fragment 1 ($X \mu\text{l}$) and fragment 2 ($y \mu\text{l}$) is determined by the 1:1 molar ratio of fragment 1 to 2

Reagent	Fragment	Volume (μl)
Sterile double-dist. Water	–	
Primers and Template DNA	1	x
	2	y

Pipet the mixture with $x + y \mu\text{l}$ High Fidelity PCR Master in a thin-walled PCR tube on ice and mix well.

Thermal cycling.

	Temperature	Time	Cycles
Initial denaturation	94°C	2 min	1 \times
Denaturation	94°C	15 s	10 \times
Annealing	55°C	30 s	
Elongation	72°C	30 min	
Final elongation	72°C	10 min	1 \times
Cooling	4°C	Forever	

3. Then, PCR is performed again to amplify the intact DNA fragment.

For one reaction.

Reagent	Volume (μ l)	Final concentration
Sterile double-dist. water	43	
PCR Master Mix (2 \times)	50	1 \times
N terminus primer (10 μ M)	3	300 nM
C terminus primer (10 μ M)	3	300 nM
SOE PCR product (see Note 4)	1	
Final volume	100	

Prepare about 100 μ g intact DNA after purification by QIAquick Gel Extraction Kit (see Note 5) for digestion to get 30 μ g digested insert DNA.

4. Digest intact DNA and phagemid vector pComb3X, and ligate them.

Digestion.

Components	Volume (μ l)	Final concentration
Intact DNA	x (up to 100 μ g)	–
pComb3X	–	y (up to 300 μ g)
10 \times NEBuffer 2	200	1 \times
BSA (10 mg/ml)	20	100 μ g/ml
SfiI (20 u/ μ l)	200	–
ddH ₂ O	1,580 – x	800 – y
Final volume	2,000	1,000

Incubate at 50°C overnight. Run the digested products on 1% agarose gels, purify the DNA with gel extraction kit (elute the DNA with ultra pure water), and quantify it (see Note 6).

Ligation (see Note 7).

Components	Volume (μ l)	Final concentration
Digested insert DNA	x (up to 30 μ g)	
Digested pComb3X	y (up to 100 μ g)	
Molar ratio of mole of insert DNA/mole of vector	3:1	
10× buffer for T4 DNA ligase buffer	100	1×
T4 DNA ligase (400 u/ μ l)	100	
ddH ₂ O	$800 - x - y$	
Final volume	1,000	

Incubate at 16°C for 72 h.

3.5. Concentration and Desalting of Ligated Products

1. The ligated products are purified by QIAquick PCR Purification Kit according to the manufacturer's instructions.
2. Concentrate the purified ligated products by Microcon Ultracel YM-3 (Millipore) to 50–100 μ l (see Note 8).

3.6. Electroporations and Library Preparation

1. Pre-warm 850 ml 2× YT medium containing 2% glucose (w/v) and 150 ml S.O.C. medium at 37°C. Chill 50 gene pulser cuvettes on ice. At the same time thaw, on ice, 2 ml of TG1 electroporation-competent cells.
2. Divide 2 ml of TG1 competent cells into five prechilled 1.5-ml Eppendorf tubes with 400 μ l each. Add 10–20- μ l ligations to each tube and pipet gently to mix. Transfer 41–42 μ l mixtures to each cuvette. Gently tap the cuvette on the bench to make the mixture fill out the bottom of the cuvette.
3. Electroporate at 1.8 kV, 25 μ F, and 200 Ω . Flush the cuvette immediately with 960 ml and then twice with 2 ml of prewarmed S.O.C. medium and combine the 3 ml in a 2-L flask. After all electroporations are completed, about 150 ml electroporation product is obtained.
4. Shake at 250 rpm for 45–60 min at 37°C. Spread the electroporation product with serial dilution on 2× YT agar plates containing 2% glucose (w/v) and 100 μ g/ml of ampicillin. Incubate the plates overnight at 37°C for calculation of the electroporation efficiency and size of the library.

5. Transfer 150 ml culture to 850 ml 2×YT medium to make 1 L culture containing 100 µg/ml ampicillin and 2% glucose. Shake for additional 2 h at 37°C.
6. Take 1 ml of the culture and measure the cell density by reading OD600. Calculate the total number of cells by multiplying the OD600 value by 5×10^8 (estimated number of cells in 1 ml culture when OD600 reaches 1) and the culture volume (1,000 ml in this case). Add 10 MOI (multiplicity of infection) of M13KO7 helper phage to the culture. Incubate at 37°C for 30 min, shaking for homogenization every 10 min.
7. Spin down the cells at 6,000 rpm (~4,500 × g, rotor: JLA-10.500, Beckman Coulter, CA) for 15 min. Resuspend in 2 L 2×YT medium containing 100 µg/ml of ampicillin and 50 µg/ml of kanamycin. Incubate at 250 rpm overnight at 30°C.
8. Spin at 6,000 rpm (~4,500 × g, rotor: JLA-10.500, Beckman Coulter, CA) for 15 min at 4°C. Save the bacterial pellet for phagemid preparation using, for example, the Qiagen HiSpeed Plasmid Maxi Kit. For phage precipitation, transfer the supernatant to a clean 2L flask and add ¼ volume of 20% (w/v) PEG8000 and 2.5 M NaCl solution. Mix well and incubate on ice for at least 1 h.
9. Spin at 12,000 rpm (~14,000 × g, rotor: JA-14, Beckman Coulter, CA) for 20 min at 4°C. Discard the supernatant. Resuspend the phage pellet in 50 ml PBS, pH 7.4 by pipetting up and down along the side of the centrifuge bottle by using a 10-ml pipet.
10. Spin at 6,000 rpm (~4,500 × g, rotor: JLA-10.500, Beckman Coulter, CA) for 10 min at 4°C. Transfer the supernatant to a clean 200-ml flask and add ¼ volume of 20% (w/v) PEG8000 and 2.5 M NaCl solution. Mix well and incubate on ice for 1 h.
11. Spin at 12,000 rpm (~14,000 × g, rotor: JA-14, Beckman Coulter, CA) for 20 min. Discard the supernatant. Resuspend the phage pellet in 50 ml PBS, pH 7.4.
12. Spin at 6,000 rpm (~4,500 × g, rotor: JLA-10.500, Beckman Coulter, CA) for 10 min at 4°C. Transfer the supernatant to a clean 200-ml flask.
13. Measure the concentration of phage by reading OD280 (1 OD280 = 2.33×10^{12} /ml). Add the same volume of sterilized glycerol and mix well. Aliquot the phage to make sure that each contains phage particles at least 100 times the total number of transformants (calculated in step 4). Store the phage at -80°C. The CH2 phage library is now ready for panning.

3.7. Construction of the CH2 Mutant m01

An additional disulfide bond is designed to be introduced between the N-terminal strand A and the C-terminal one G (Fig. 3) (13). The CH2 structure is represented by VMD 1.8.7 (18).

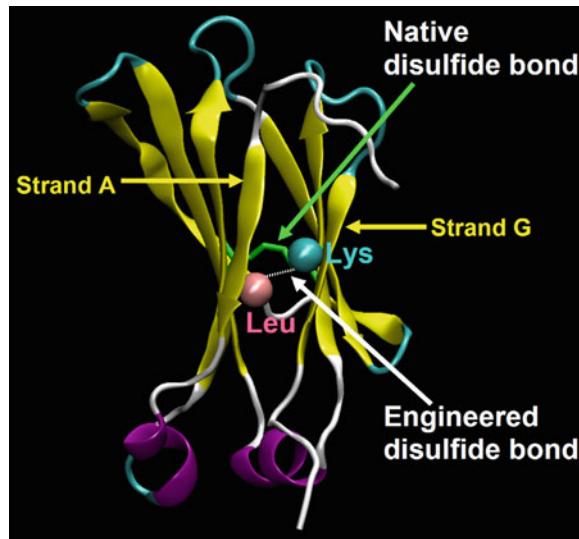


Fig. 3. Design of m01 based on the CH2 structure (13). The distance between two $\text{C}\alpha$ s forming the native disulfide bond (indicated by green arrow) is 6.53 Å. An engineered disulfide bond is introduced between Leu12 and Lys104 which were replaced by cysteines (indicated by white arrow).

1. PCR for amplification of three fragments with mutations.

For amplification of fragment A.

Reagent	Volume (μl)	Final concentration
Sterile double-dist. water	21	
PCR Master Mix (2 \times)	25	1 \times
Omp (10 μM)	1.5	300 nM
L/C primer Rv (10 μM)	1.5	300 nM
pComb3X-CH2 (2 ng/ μl)	1	0.02 ng/ μl
Final volume	50	

For amplification of fragment B.

Reagent	Volume (μl)	Final concentration
Sterile double-dist. water	21	
PCR Master Mix (2 \times)	25	1 \times
L/C primer Fw (10 μM)	1.5	300 nM
K/C primer Rv (10 μM)	1.5	300 nM
pComb3X-CH2 (2 ng/ μl)	1	0.02 ng/ μl
Final volume	50	

For amplification of fragment C.

Reagent	Volume (μ l)	Final concentration
Sterile double-dist. water	21	
PCR Master Mix (2 \times)	25	1 \times
K/C primer Fw (10 μ M)	1.5	300 nM
gIIIF (10 μ M)	1.5	300 nM
pComb3X-CH2 (2 ng/ μ l)	1	0.02ng/ μ l
Final volume	50	

2. After purification of these three fragments, SOE PCR is performed to get intact DNA fragment with replacement of L and K by two Cs. The volume of fragment A (x μ l), fragment B (y μ l) and fragment C (z μ l) is determined by the 1:1:1 molar ratio of three fragments

Reagent	Fragment	Volume
Sterile double-dist. water	—	
Primers and template DNA	A	y
	B	x
	C	z

Pipet the mixture with $x+y+z$ μ l High Fidelity PCR Master in a thin-walled PCR tube on ice and mix well.

Thermal cycling.

	Temperature	Time	Cycles
Initial denaturation	94°C	2 min	1 \times
Denaturation	94°C	15 s	10 \times
Annealing	55°C	30 s	
Elongation	72°C	30 min	
Final elongation	72°C	10 min	1 \times
Cooling	4°C	forever	

3. Amplification of SOE PCR product.

Reagent	Volume (μ l)	Final concentration
Sterile double-dist. water	43	
PCR Master Mix (2 \times)	50	1 \times
Omp (10 μ M)	3	300 nM
gIIIF (10 μ M)	3	300 nM
SOE PCR product	1	
Final volume	100	

4. Digestion, ligation, and transformation.

See the method in Subheading 3.2

3.8. Stability Measurements of CH2, m01, and m36

CH2, m01, and m36, a domain antibody against HIV (19), are expressed and purified by the method in Subheading 3.3. The native disulfide bond in CH2 and the introduced disulfide bond are verified by using mass spectrometry.

1. Circular dichroism (CD).

- (a) Dissolve the purified proteins in PBS (see Note 9) at the final concentration of 0.49 mg/ml.
- (b) Record the wavelength spectra at 25°C using a 0.1-cm path-length cuvette for native structure measurements.
- (c) Measure the thermal stability at 216 nm by recording the CD signal in the temperature range of 25–90°C with heating rate 1°C/min.

2. Differential scanning calorimetry (DSC).

- (a) Concentrate the three proteins to 1.5 mg/ml (see Note 10) in PBS (pH 7.4).
- (b) Use 1°C /min as heating rate and scan the samples from 25 to 100°C.

3. Spectrofluorometry.

- (a) Dilute all the proteins in buffer A to final concentration of 10 µg/ml in the presence of urea from 0 to 8 M.
- (b) Record the emission spectra from 320 to 370 nm at 25°C with excitation wavelength at 280 nm.
- (c) Correct the fluorescence spectra by the background fluorescence (buffer + denaturant).
- (d) Use fluorescence intensity at 340 nm to evaluate the unfolding.

The stabilities of CH2, m01, and m36 are summarized in Table 1. The T_m value of m01 is higher than that of CH2 and

Table 1
Comparison of stabilities of CH2, m01, and m36 by different methods

Midpoint			
	CD	DSC	Spectrofluorometry
Protein	(°C)		Urea concentration (M)
CH2	54.1	55.4	4.2
m01	73.8	73.4	6.8
m36	63.7	62.1	4.4

m36. Interestingly, thermostability of an eAd is typically increased by 10°C after introduction of an additional disulfide bond (20). The strategy is similar to that used in the design of m01.

4. Notes

1. Normally, 1–2- μ g DNA fragment is more than enough for subsequent ligation and transformation.
2. Sufficient air exchange is very important for high yield expression. If the flask is capped too tightly, the yield will be very poor. It is an option to use flasks with vented caps.
3. The sample eluted by imidazole at low concentrations (0–20 mM) could be collected and analyzed on SDS-PAGE to minimize the loss of the proteins.
4. The amount of SOE product could be optimized for high amplification efficiency. For example, 1:1, 1:10, and 1:100 dilutions of the template (SOE product) could be tested.
5. It is very important to remove the trace of the buffer. Otherwise, the residual salts will decrease the efficiency of subsequent digestion, ligation, and electroporation.
6. In some cases the digestion of phagemid vectors may not be complete due to bad quality of DNA. To address this problem, additional treatment is needed to further purify the phagemids before digestion, or use more SfiI to digest for longer time, for example, overnight.
7. Before large-scale ligation can be performed, it is highly recommended to take two ligation tests. One is to assess the suitability of the vector and inserts for high-efficiency ligation and transformation. This can be accomplished through assembling small reactions either with vector only (test for vector self-ligation) or with both vector and insert, and transforming chemically competent cells like DH5 α . The other is to determine the optimal ratio between insert and vector for the highest efficiency of ligation. This can be accomplished through assembling small reactions with insert and vector in different molar ratios, such as 3:1, 2:1, and 1:1, and transforming chemically competent cells.
8. The desalting of DNA samples is a key step to the success of electroporations. High concentration of ions in the DNA solution will result in a long and intense pulse in electroporations, which causes cell damage or rupture. We found that at least 1,000-time dilution of DNA solution was needed to generate time constants of 4.6–5.0 ms in electroporations that generally gave the highest efficiency.

9. The use of the buffer could be optimized to get low background noise and obtain more reliable data.
10. The concentration of the protein could be optimized. If the signal is low, then concentration could be increased.

Acknowledgments

This work was supported by the Intramural AIDS Targeted Antiviral Program (IATAP), National Institutes of Health (NIH), the NIAID (NIH) Intramural Biodefense Program, and by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research.

References

1. Kolmar H, Skerra A (2008) Alternative binding proteins get mature: rivalling antibodies. *FEBS J* 275:2667
2. Skerra A (2007) Alternative non-antibody scaffolds for molecular recognition. *Curr Opin Biotechnol* 18:295–304
3. Nygren PA, Skerra A (2004) Binding proteins from alternative scaffolds. *J Immunol Methods* 290:3–28
4. Binz HK, Amstutz P, Pluckthun A (2005) Engineering novel binding proteins from non-immunoglobulin domains. *Nat Biotechnol* 23:1257–1268
5. Hey T, Fiedler E, Rudolph R, Fiedler M (2005) Artificial, non-antibody binding proteins for pharmaceutical and industrial applications. *Trends Biotechnol* 23:514–522
6. Holliger P, Hudson PJ (2005) Engineered antibody fragments and the rise of single domains. *Nat Biotechnol* 23:1126–1136
7. Holt LJ, Herring C, Jespers LS, Woolven BP, Tomlinson IM (2003) Domain antibodies: proteins for therapy. *Trends Biotechnol* 21:484–490
8. Saerens D, Ghassabeh GH, Muylldermans S (2008) Single-domain antibodies as building blocks for novel therapeutics. *Curr Opin Pharmacol* 8:600–608
9. Ward ES, Gussow D, Griffiths AD, Jones PT, Winter G (1989) Binding activities of a repertoire of single immunoglobulin variable domains secreted from *Escherichia coli*. *Nature* 341:544–546
10. Hamers-Casterman C, Atarhouch T, Muylldermans S, Robinson G, Hamers C, Songa EB, Bendahman N, Hamers R (1993) Naturally occurring antibodies devoid of light chains. *Nature* 363:446–448
11. Muylldermans S, Cambillau C, Wyns L (2001) Recognition of antigens by single-domain antibody fragments: the superfluous luxury of paired domains. *Trends Biochem Sci* 26:230–235
12. Feige MJ, Walter S, Buchner J (2004) Folding mechanism of the CH2 antibody domain. *J Mol Biol* 344:107–118
13. Gong R, Vu BK, Feng Y, Prieto DA, Dyba MA, Walsh JD, Prabakaran P, Veenstra TD, Tarasov SG, Ishima R, Dimitrov DS (2009) Engineered human antibody constant domains with increased stability. *J Biol Chem* 284: 14203–14210
14. Dimitrov DS (2009) Engineered CH2 domains (nanoantibodies). *MAbs* 1:26–28
15. Hackel BJ, Kapila A, Wittrup KD (2008) Picomolar affinity fibronectin domains engineered utilizing loop length diversity, recursive mutagenesis, and loop shuffling. *J Mol Biol* 381:1238–1252
16. Chen W, Zhu Z, Xiao X, Dimitrov DS (2009) Construction of a human antibody domain (VH) library. *Methods Mol Biol* 525:81–99, xiii
17. Xiao X, Feng Y, Vu BK, Ishima R, Dimitrov DS (2009) A large library based on a novel (CH2) scaffold: identification of HIV-1 inhibitors. *Biochem Biophys Res Commun* 387:387–392
18. Humphrey W, Dalke A, Schulter K (1996) VMD: visual molecular dynamics. *J Mol Graph* 14(33–38):27–38

19. Chen W, Zhu Z, Feng Y, Dimitrov DS (2008) Human domain antibodies to conserved sterically restricted regions on gp120 as exceptionally potent cross-reactive HIV-1 neutralizers. *Proc Natl Acad Sci USA* 105:17121–17126
20. Hagiwara Y, Mine S, Uegaki K (2007) Stabilization of an immunoglobulin fold domain by an engineered disulfide bond at the buried hydrophobic region. *J Biol Chem* 282:36489–36495

Chapter 7

Engineering of Affibody Molecules for Therapy and Diagnostics

Joachim Feldwisch and Vladimir Tolmachev

Abstract

Affibody molecules are small and robust non-immunoglobulin affinity ligands capable of binding to a wide range of protein targets. They are selected from combinatorial libraries based on a 58 amino acid, three-alpha-helical Z-domain scaffold. They share no sequence or structural homologies to antibodies and in contrast to antibodies they can be functionally produced both by peptide synthesis and by recombinant expression in *Escherichia coli*. Protein engineering is used to adapt Affibody molecules binding to a target of interest to the specific demands imposed by the intended application. Obviously, the optimal molecule for molecular imaging will be different from the optimal molecule for therapy. Here, we describe general strategies to optimize Affibody molecules for diagnostic imaging and therapy applications.

Key words: Affibody molecules, Affinity ligand, Scaffold protein, Albumin-binding domain, ABD, Molecular Imaging, Targeted therapy, Biodistribution, Radionuclides, Chelator, Site-specific modification

1. Introduction

Affinity proteins, i.e., large or small proteins with high affinity and specificity for a defined target, are a rich source for the development of novel therapeutics and diagnostics. Antibodies, large multidomain proteins (approximately 150 kDa) are currently the most commonly used class of affinity proteins in life science and medical applications (1). However, antibodies have a number of intrinsic disadvantages related to their molecular structure. Since they are large, multidomain proteins with 12 intra-chain and four inter-chain disulfide bridges and complex glycosylation patterns, their production is more difficult and more expensive than production of small proteins (2). In addition, the large size implicates high

protein doses (up to several 100 mg per dose) and thus the need to produce considerable larger amounts than those required for smaller proteins. High antibody doses in small volumes often demand extensive formulation development to reach the desired protein concentrations especially for subcutaneous injection (max 1–2 mL, e.g., 162 mg are administered s.c. in clinical trials with the anti-IL-6 receptor antibody tocilizumab). Furthermore, the large size of antibodies leads to slow biodistribution and elimination and limits their use in applications where rapid distribution and elimination is desired such as molecular imaging using single-photon emission computed tomography (SPECT) or positron emission tomography (PET). Antibodies modified with prosthetic groups or payloads, e.g., chelators or toxins, are usually nonhomogenous drug preparations consisting of complex mixtures with 1–8 or even more modifications at different sites. Site-specific modification of antibodies using the intein-fusion technology has only recently been described (3).

An alternative to antibodies are novel and usually smaller affinity proteins based on various different protein scaffolds. Each class of scaffold proteins has a defined frame of amino acids determining the overall fold or tertiary structure and an exposed surface with variable amino acid compositions creating the binding site and determining binding specificity and affinity for different targets (4). Scaffold proteins can be selected from combinatorial libraries by, e.g., phage display and are easy to produce with high expression levels and short development times, easy to optimize by protein engineering, and are amenable to site-specific modifications. They have therefore become a rich source for the development of next-generation protein therapeutics and diagnostics (5, 6). The first candidate drugs derived from protein scaffolds have already entered clinical trials (7, 8).

Affibody molecules are one of the more advanced members of these scaffold proteins. They are small (58 amino acids, approximately 6.5 kDa) and robust non-immunoglobulin affinity ligands with no sequence or structural homologies to antibodies. The origin of the Affibody scaffold is one of the domains of *Staphylococcus aureus* protein A (SPA). The extracellular part of SPA consists of the five homologous immunoglobulin (Ig)-binding domains E, D, A, B, and C, all capable of binding to the Fc-fragment and certain Fab-fragments of immunoglobulins containing a heavy chain derived from the V_H3 family of gene segments (9, 10). The Z-domain scaffold, the basis for all Affibody molecules, was derived from the B-domain by the mutation of Gly²⁹ to Ala²⁹, which removed a hydroxylamine cleavage site (N²⁸G²⁹), and by replacing Ala¹ with Val¹, which created a suitable cloning site (11). The Z-domain has excellent biophysical properties, including high melting temperature and reversible and rapid refolding. The folding time for the three helical bundle structure of the Z-domain is 3 μs,

i.e., the shortest folding time yet reported (12). Novel binding surfaces are created on the Z-domain by randomizing the 13 surface-exposed amino acids on helix 1 and helix 2 originally involved in binding of immunoglobulins. The novel proteins generated in this way are typically displayed on phages making up libraries of more than 10^{10} different Affibody molecules. These libraries are then used to select Affibody molecules for a desired target protein. The outcomes of these selections are Affibody molecules binding specifically and with high affinity to their corresponding targets (13–16).

This review summarizes recent protein engineering efforts to optimize Affibody molecules for molecular imaging and therapeutic applications. Obviously, the optimal molecule for molecular imaging will be different from the optimal molecule for therapy. For molecular imaging small molecules with rapid biodistribution and rapid elimination, i.e., short plasma residence times are desired, whereas molecules with long plasma half-life are desired for therapeutic applications. Thus, engineering of an Affibody molecule specific for the target of interest is necessary to adapt the protein for the specific demands imposed by these widely different applications. Engineering is focused to regions outside the target-binding site on helix 1 and 2 with the goal to find general modification strategies for all Affibody molecules binding different targets.

In the first part of this review, factors influencing the design of Affibody molecules for therapy and diagnostic imaging are summarized. In the second part, strategies to optimize Affibody molecules for labeling with the most common clinically used radionuclides for SPECT or PET are described, followed by strategies to engineer an optimized scaffold for Affibody molecules. In the third part, we summarize strategies for engineering of Affibody molecules for therapy.

2. Factors Influencing the Design of Affibody Molecules for Therapy and Diagnostic Imaging

Novel imaging techniques play an important role in cancer diagnosis. In particular, molecular imaging, i.e., the visualization, characterization, and quantification of a disease target, such as a cell surface receptor, can provide a global view of all metastatic lesions in the body and has evolved as a valuable tool for stratifying patients for targeted therapy. For both diagnostic imaging and targeted radionuclide therapy, it is essential that the accumulation of radioactivity in tumor is much higher than in normal healthy tissues. Differences in radioactivity accumulation determine image contrast and sensitivity of a radionuclide imaging agent and the tumor dose must exceed the dose in normal tissues for radionuclide therapy to be effective. A tissue where radioactivity concentration is of special

importance is blood, since blood radioactivity content can contribute to overall radioactivity in each organ and cause contrast reduction. For this reason, the tumor-to-blood radioactivity concentration ratio is used to follow the effects of engineering efforts in the design of new imaging agents. The success of novel imaging or therapy tracers for efficient radionuclide tumor targeting depends not only on the properties of the targeting protein, but also on the properties of a radionuclide. The use of an inappropriate labeling strategy can appreciably decrease the clinical value of an imaging or a therapeutic agent. Thus, the protein may have to be engineered in different ways depending on the radionuclide and the labeling chemistry used.

2.1. Role of Catabolism, Cellular Retention, and Excretory Organs

Although there are over 2,000 known radionuclides, only a few meet the requirements for a potential label for radionuclide therapy or imaging (17). For the choice of the optimal radionuclide, it is important to follow the fate of the radionuclide and not only the distribution of the radiolabeled protein conjugate. Radiometals may be transchelated to blood proteins or the protein conjugate may be catabolized after injection in blood by peptidases. Furthermore, catabolism may occur inside tumor cells and excretory organs and the resulting radiocatabolites may reenter the circulation and be redistributed. This nonprotein-associated, free and therefore non-targeting radioactivity can appreciably contribute to image background. Thus, the fate of radiocatabolites should be taken into account during selection of the radiochemistry used for a particular imaging or therapy agent.

Cellular processing and retention of the radiolabeled conjugate by tumor cells and in excretory organs is essential for the optimal design of imaging probes. Binding of a ligand to a transmembrane receptor often causes rapid receptor-mediated clathrin-dependent endocytosis, i.e., internalization of a ligand–receptor complex. This effect has been demonstrated for many ligands to G protein-coupled peptide receptors, receptor tyrosine kinases and also for the bivalent binding of antibodies to a cell-surface antigen. Binding of antagonists do not trigger this clathrin-dependent process, but they can be slowly internalized by clathrin-independent mechanisms as a part of membrane renewal. Approximately 30% of the cellular membranes are processed in this way per day. Internalization is typically followed by a transfer through endosomal and finally lysosomal compartments and results in proteolysis of targeting proteins. In vitro studies using radiolabeled antibodies and peptide ligands demonstrated that the fate of radionuclides after lysosomal degradation depends on the lipophilicity of radiocatabolites (18–21). Lipophilic radiocatabolites can penetrate through phospholipid membranes and leave the targeted malignant cells. This problem is encountered with the vast majority of radiohalogen labels (22, 23). Charged or bulky polar radiocatabolites cannot penetrate cellular membranes and are retained inside the cells

before excretion by relatively slow externalization. This is typical for radiometals (24). Radionuclides, which are trapped inside the cells after internalization and degradation of the protein conjugate, are called “residualizing” (25). Internalization of ligands labeled using residualizing radiometals is considered as an important and useful factor for accumulation of radiotracer in the cell (24).

However, internalization may also occur in healthy tissues, such as the excretory organs. Small proteins and peptides passing through the glomerular membrane in kidneys are reabsorbed in the proximal tubuli by several endocytotic mechanisms (26, 27). If a residualizing label was used, long and stable retention of radioactivity in the kidneys is observed. Hepatic uptake of drugs and xenobiotics is mediated by several mechanisms (28), and some of them may cross-react with targeting proteins and peptides. Depending on their chemical nature, radiocatabolites may leak from liver back to blood circulation, be trapped inside hepatocytes or excreted by efflux pumps into bile. This hepatobiliary excretion of tracers or catabolites gives rise to an elevated background in the lower abdomen. Thus, the behavior of a radiolabeled tracer in the excretory organs is essential for the clinical outcome and the desired properties are different for therapy and imaging applications. For radionuclide therapy, any elevated uptake in excretory organs is undesirable as it contributes to an elevated dose burden for healthy tissues. Since renal clearance is typical for peptides and proteins with the size below glomerular cut-off of 60 kDa (26), the goal for engineering an optimal therapy tracer is mainly focused on the reduction of renal uptake.

For molecular imaging, elevated hepatic uptake has to be avoided, since the liver is a major metastatic site for many malignancies. High liver uptake and a diffuse background due to hepatobiliary excretion can reduce or even prevent detection of liver metastases or extrahepatic abdominal metastases. Minimizing high renal uptake and retention of radioactivity is less critical for radionuclide imaging. Dechristoforo and Mather (29) wrote: “For optimal imaging sensitivity in the detection of receptor-positive tumors a high target uptake must be combined with a low background activity and, for tumors in the pelvis and abdomen, the main sources of diffuse background activity are the gastrointestinal tract and vasculature. Whereas high renal activity can also obscure tumor uptake in the immediate area, the defined shape of the kidneys and the fact that they are seldom a site of metastasis means that renal excretion is the desired route of elimination for these tracers.”

In summary, residualizing radionuclides are preferred for rapidly internalizing imaging or therapeutic agents, since they increase the radioactivity retention by malignant cells, however at the price of elevated radioactivity in excretory organs. For tracers with a slower internalization rate, also non-residualizing radionuclides may be used. Even though this may reduce tumor uptake, the lower uptake in excretory organs could be advantageous.

2.2. Labeling Chemistry

Finally, the labeling chemistry chosen for different radionuclides is of utmost importance for the successful design of new therapeutic or diagnostic proteins. The harsh conditions often necessary for efficient radiolabeling of a protein may diminish or destroy its functional activity. For example, proteins and peptides can be labeled directly with radiohalogens by *in situ* oxidation of radiohalide and subsequent electrophilic attack of tyrosine residues of protein (30, 31). This method is straightforward and robust and often provides efficient labeling. However, tyrosines that are critical for binding of the imaging agent to its molecular target may be modified (32). In fact, the direct radioiodination of anti-HER2 Affibody molecules resulted in loss of binding capacity, presumably due to modification of a tyrosine in the binding site (33). The inactivation of the HER2-binding Affibody molecule could be prevented using indirect radiohalogenation methods which avoid the direct exposure of the protein to oxidizing and reducing agents and directs the modification to ϵ -amino group of lysines or the thiol group of cysteines (14, 34).

In contrast to radiohalogens, radiometals cannot form stable covalent bond with organic compounds. Therefore, proteins need to be modified with chelators, i.e., polydentate complexing agents to allow labeling with radiometals. Several bifunctional chelating prosthetic groups are available with different functional groups for their coupling to proteins, e.g., isothiocyanato, succinimidyl, or maleimido groups and different chelating functional groups, e.g., diethylene triamine pentaacetic acid (DTPA) or 1,4,7,10-tetraaza-cyclododecane-1,4,7,10-tetraacetic acid (DOTA). The very different chemical properties of different metals (ionic radii, coordination numbers) require careful selection of the chelator-radiometal pair. For example, acyclic DTPA derivatives are suitable for labeling with ^{111}In , ^{90}Y , and ^{177}Lu (35–37), but cannot provide sufficient *in vivo* stability for ^{68}Ga . Derivatives of DOTA are versatile, enabling labeling with ^{111}In , ^{68}Ga , ^{177}Lu , and ^{90}Y (38), but are suboptimal for copper isotopes (39).

3. Engineering of Affibody Molecules for Diagnostic Imaging

Both non-residualizing radiohalogens (e.g., ^{18}F , ^{123}I , ^{211}At) and residualizing radiometals (e.g., ^{111}In , ^{68}Ga , $^{99\text{m}}\text{Tc}$, ^{186}Re , ^{177}Lu) have been used for labeling of Affibody molecules for diagnostic imaging or therapy (Tables 1 and 2). *In vitro* studies demonstrated that internalization of Affibody molecules is relatively slow, with only 12% of total radioactivity internalized after 4 h for the HER2-targeting Affibody molecules labeled with the residualizing radiometal ^{111}In (36, 40–42). A similar internalization rate was also found for EGFR-targeting Affibody molecules (43). Most likely, the

Table 1
Nuclides for imaging used with Affibody molecules

Nuclide	Half-life	Reference
<i>Nuclides for SPECT</i>		
^{123}I (^{125}I as a surrogate)	13.3 h	(14, 34, 43, 44)
$^{99\text{m}}\text{Tc}$	6 h	(42, 44, 52, 55–60, 78)
^{111}In	2.8 d	(36, 41, 50, 51, 64, 79)
<i>Nuclides for PET</i>		
^{18}F	110 min	(46, 47, 80)
^{76}Br	16.2 h	(67)
^{124}I	4.18 d	(45)
^{55}Co (^{57}Co as a surrogate)	17.5 h	(81)
^{64}Cu	137 h	(82, 83)
^{68}Ga	68 min	(7, 84)

Table 2
Nuclides for therapy used with Affibody molecules

Nuclide	Half-life	Reference
<i>High energy beta</i>		
^{90}Y	64 h	(85)
^{186}Re	3.7 days	(68)
$^{114\text{m}}\text{In}/^{114}\text{In}$	49.5 days/72 s	(86)
<i>Low energy beta</i>		
^{131}I	8 days	(34)
^{177}Lu	6.7 days	(66, 85)
<i>Alpha</i>		
^{211}At	7.2 h	(87)

internalization is not receptor-mediated but caused by turnover of cellular membranes. Thus, the use of residualizing labels is not absolutely critical for successful tumor imaging with Affibody molecules. Accordingly, Affibody molecules could be labeled with radiohalogens, such as $^{124}\text{I}/^{125}\text{I}$ (14, 44, 45) or ^{18}F (46, 47), and successfully used for imaging of HER2-expressing xenografts in mice.

Figure 1 illustrates the difference between Affibody molecules labeled with residualizing and non-residualizing radionuclides. Tumor uptake of the non-residualizing radioiodine label is lower than the uptake of residualizing ^{111}In . There is no large difference

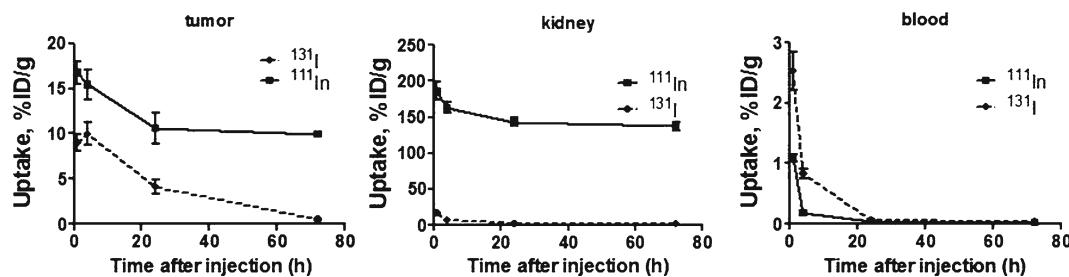


Fig. 1. Distribution of radioactivity after injection of Affibody molecules labeled with residualizing (^{111}In) and non-residualizing (^{131}I) nuclides into nude mice bearing HER2-expressing SKOV-3 xenografts. The data are summarized from refs. 34, 64.

in tumor uptake early after injection due to the slow internalization of Affibody molecules by the cancer cells; however, over time only the residualizing radionuclide is detectable in the tumor. In kidneys, however, where internalization and degradation is rapid, the difference between ^{131}I and ^{111}In uptake is much bigger. The non-residualizing radioiodine is excreted much more rapidly than residualizing indium. Reentry of radiocatabolites from kidneys into blood circulation causes a transient increase of blood-borne radioiodine activity in comparison with indium-111 during first hours after injection. Taken together, the reduced tumor uptake and increased blood radioactivity decreases the tumor-to-blood ratio and overall contrast for non-residualizing halogen labels.

In the first studies, Affibody molecules were radiolabeled by non-site-specific iodination or conjugation of chelators to the $\epsilon\text{-NH}_2$ -group of lysines using amine-directed chemistry and the results have been summarized in several reviews (48, 49). Although these studies have demonstrated the significant potential of Affibody molecules for imaging, the tracers were not homogenous. For example, even after careful optimization of the conjugation conditions for benzyl-DTPA, LC-MS analysis revealed that the final preparation was a mixture of differently modified Affibody molecules. The majority (66%) of $Z_{\text{HER2}:342}$ Affibody molecules bore a single benzyl-DTPA group, 14% were conjugated with two benzyl-DTPA molecules and 20% remained unconjugated (50). Due to the difference in overall charge and lipophilicity, nonhomogeneously-modified Affibody molecules will behave differently *in vivo* and, obviously this complicates the design of an optimized radiotracer.

To solve this problem, chemically and structurally homogeneous imaging tracers with well-defined biodistribution properties are required. Therefore, an appreciable effort has been applied to find general strategies for site-specific labeling of Affibody molecules and to allow the production of well-defined homogenous products for clinical use.

3.1. N-Terminal Modifications

In the first experiment, site-specific coupling of a chelator to the N-terminal amine in the last step of peptide synthesis was used. The resulting synthetic HER2-binding Affibody molecule ABY-002 was labeled with ^{111}In and revealed efficient tumor uptake, high tumor-to-blood and tumor-to-organ ratios allowing imaging of HER2 overexpressing xenografts as early as 30 min after injection (51). In patients, ^{111}In - or ^{68}Ga -labeled ABY-002 was rapidly cleared from blood with a first half-life of 4–11 min for [^{111}In]ABY-002 and 10–14 min for [^{68}Ga]ABY-002. This rapid kinetics allowed high contrast SPECT and PET images to be obtained as early as 2–3 h after injection (7). The successful use of peptide synthesis for designing the site-specifically labeled Affibody molecule ABY-002 opened up for a whole series of experiments using mercaptoacetyl-containing peptide-based chelators for labeling with the radiometal technetium ($^{99\text{m}}\text{Tc}$). In these peptide-based chelators, amide nitrogen atoms and thiols are used for the chelation of Tc (Fig. 2). The use of different amino acids in the chelator allowed using their side-chains for modulation of local charge and lipophilicity and thereby modulation of the pharmacological properties of the Affibody molecule. Thus, the effect of single amino acid substitutions in the chelator part of the peptide could be studied in detail and used for fine tuning the biodistribution of $^{99\text{m}}\text{Tc}$ -labeled Affibody molecules. Figure 3 gives an overview of the protein engineering efforts, i.e., the amino acid changes made to optimize Affibody molecules for molecular imaging.

Initially, the $^{99\text{m}}\text{Tc}$ -labeling of Affibody molecules was performed using the mercaptoacetyl-glycyl-glycyl-glycyl (maGGG) chelator coupled to the N-terminus of the anti-HER2 Affibody molecule Z_{HER2:342} (52). The chelator was selected because it provided stable labeling of macromolecules with $^{99\text{m}}\text{Tc}$ and $^{188/186}\text{Re}$ (53). Biodistribution studies in mice showed that the $^{99\text{m}}\text{Tc}$ -labeled

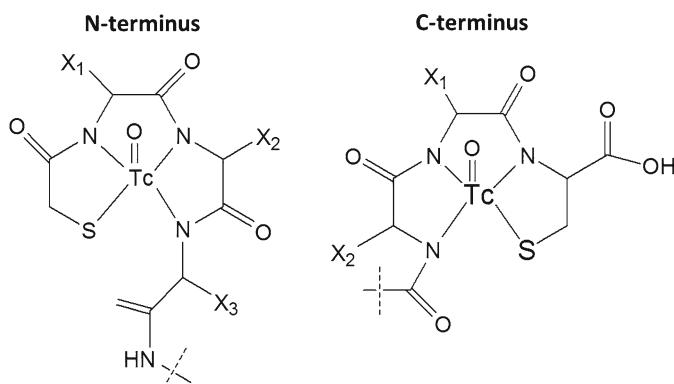


Fig. 2. Thiol-containing peptide-based chelators for $^{99\text{m}}\text{Tc}$ and ^{186}Re . X1, X2, X3 designate side-chains of amino acids.

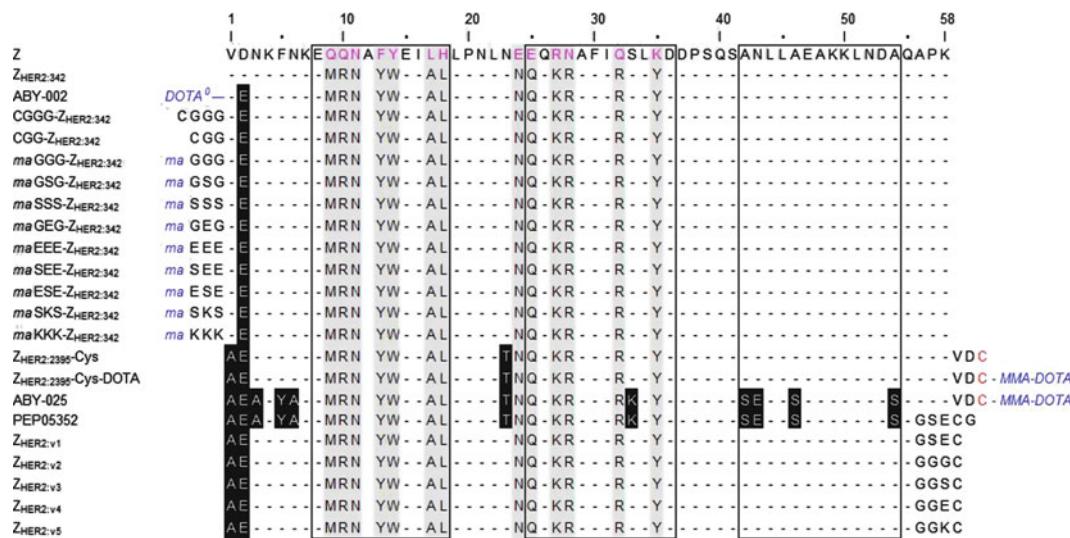


Fig. 3. Primary structure of HER2-binding Affibody molecules used for molecular imaging. The sequence of the Z-domain scaffold from which all Affibody molecules are derived is shown for comparison. The 13 amino acids in the binding site of Affibody molecules are marked gray for HER2-binding Affibody molecules. Amino acids substituted in the scaffold are marked with white letters and black background. The three boxes indicate the position of helix 1, 2, and 3.

Affibody molecules were stable *in vivo*. However, approximately 30% of the radioactivity was recovered from the intestine content due to hepatobiliary excretion. Thus, imaging of abdominal lesions could not be performed until 1 day after injection. Elevated hepatobiliary excretion is usually associated with higher lipophilicity (54). Thus, a decrease of the hepatobiliary excretion may be obtained by increasing the hydrophilicity of the chelator. Experiments in mice demonstrated that replacement of one glycine in the chelator sequence with the more hydrophilic amino acid serine resulted in a shift from the hepatobiliary to the renal excretion pathway (Fig. 4a). Substitution of all three glycines in the chelator with serines reduced the radioactivity of the intestine content in mice to one third (55). However, gamma camera imaging of tumor-bearing mice was still hampered by elevated background radioactivity in the abdomen. Thus, further reduction of the hepatobiliary excretion was necessary. This was achieved using the charged, hydrophilic amino acid glutamic acid instead of serine (Fig. 4b). The substitution of all three glycines by glutamic acids reduced the total radioactivity in the gastrointestinal tract of mice to 3% of the injected radioactivity, and allowed high-contrast imaging in the abdomen (56). The substantial increase of the renal uptake to $95 \pm 23\%$ ID/g for [99m Tc]maEEE-Z_{HER2:342} at 4 h after injection indicate the shift from hepatobiliary to predominantly renal excretion. Further experiments (Fig. 4c) showed that combination of serine and glutamic acids in mercaptoacetyl-containing chelators both reduced the hepatobiliary excretion and the renal

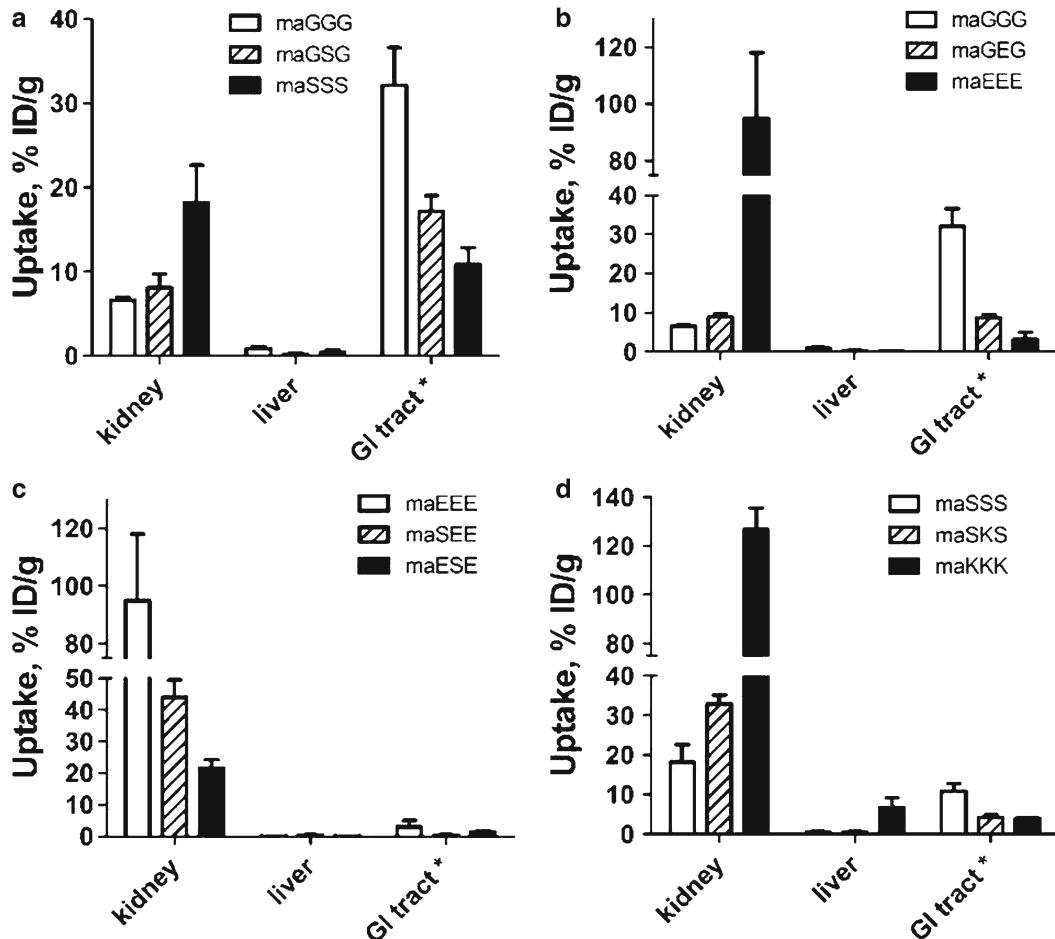


Fig. 4. Uptake of Affibody molecules labeled with ^{99m}Tc using mercaptoacyl-containing peptide-based chelators in excretory organs of NMRI mice at 4 h after injection. (a) effect of glycine and serine in the chelator sequence, (b) effect of glutamic acid, (c) effect of combining glutamic acid and serine, (d) effect of combining serine and lysine. Data are presented as an average percent of injected dose per gram (% ID/g) for four mice \pm standard deviation. *Data for gastrointestinal tract with its content are a measure of hepatobiliary excretion. They are presented as %ID per whole sample.

retention of radioactivity (57). Interestingly, the order of amino acid in the chelator was important. The mercaptoacetyl-glutamyl-seryl-glutamyl (maESE) chelator provided twice lower renal retention than mercaptoacetyl-seryl-glutamyl-glutamyl (maSEE) and mercaptoacetyl-glutamyl-glutamyl-seryl (maEES) chelators. Urine analysis demonstrated that the reduction of radioactivity was due to more rapid excretion of radiocatabolites from the maESE chelator, indicating that the residualizing properties of the ^{99m}Tc -label were diminished. As internalization of HER2-targeting Affibody molecules by HER2-expressing cells is slow, the residualizing properties are not as crucial for successful tumor targeting. In fact, there was no significant difference in radioactivity uptake in tumors when HER2-expressing xenografts in mice were targeted by Affibody

molecules labeled with ^{99m}Tc using maEEE, maESE, maEES, or maSEE chelators (57). Interestingly, the effect of substituting serine with positively charged lysine was different from the substitutions with negatively charged glutamic acid (58). Experiments in normal NMRI mice demonstrated (Fig. 4d) that substitution of a single serine by a lysine led to slightly lower levels of hepatobiliary excretion but a moderate increase of kidney uptake (radioactivity levels were slightly higher for ^{99m}Tc -maSKS-Z_{HER2:342} than for ^{99m}Tc -maESE-Z_{HER2:342}). However, the use of triple-substituted maKKK chelator increased not only renal, but also hepatic uptake. The exact molecular mechanism underlying the elevated radioactivity hepatic uptake of ^{99m}Tc -labeled maKKK chelator-modified Affibody molecules is currently unknown. However, these data indicate that not only the lipophilicity of the chelator, but also their charge is essential for uptake and retention of radiolabeled conjugates in excretory organs.

3.2. C-Terminal Modifications

The use of integrated mercaptoacetyl-based chelators is possible only for Affibody molecules produced by peptide synthesis. It would be desirable, however, to obtain a solution for site-specific ^{99m}Tc -labeling of bacterially produced Affibody molecules as well. The thiophilic nature of technetium requires the presence of a thiol group in the chelator to provide labeling site-specificity. The use of cysteine as a thiol-bearing moiety permits design of conjugates containing only natural amino acids and suitable for recombinant production. The first experiments with ^{99m}Tc -labeled Affibody molecules containing cysteinyl-glycyl-glycyl-glycyl-(CGGG-) and cysteinyl-glycyl-glycyl-(CGG-) at the N-terminus confirmed the feasibility of this approach (59). However, these conjugates suffered from a substantial degree of hepatobiliary excretion ($16 \pm 4\%$ of injected activity was found in intestines content at 4 h after injection). In addition, the elevated radioactivity uptake in salivary glands and stomach suggested release of free technetium during catabolism.

A new chelator with the common N₃S format was obtained by introducing a unique cysteine at the C-terminus (Fig. 2). The resulting C-terminal chelator with the sequence valyl-aspartyl-cysteinyl (-VDC) allowed site-specific and stable labeling of Affibody molecules (Fig. 3) (60). Animal studies with [^{99m}Tc]Z_{HER2:2395}-Cys showed low accumulation of the radioactivity in liver and the intestinal content but high uptake in the kidney indicating predominantly renal excretion (Fig. 5). Excellent tumor targeting (uptake of $15 \pm 3\%$ ID/g and a tumor-to-blood ratio of 121 ± 24) in combination with high in vivo stability of the label (indicated by the low radioactivity uptake in stomach and salivary gland) made [^{99m}Tc]Z_{HER2:2395}-Cys a good imaging tracer. However, due to the residualizing properties of the label leading to high kidney retention, the tracer was still not optimal. To reduce the renal uptake, a GSEC chelator (which is a “mirror” homologue of maESE) was designed (42). In addition, a glycine was added to C-terminus to

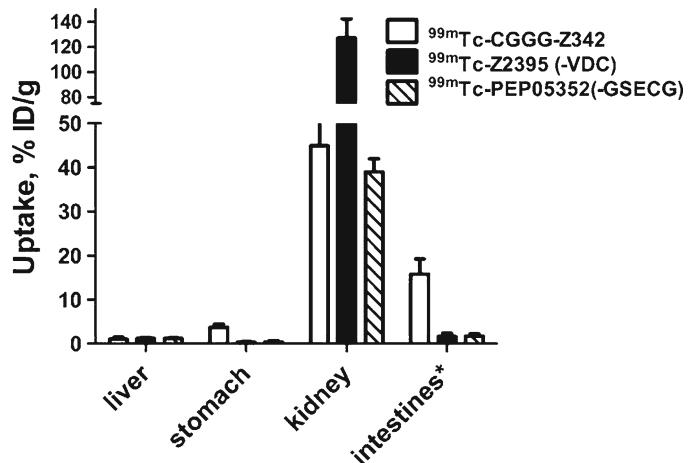


Fig. 5. Uptake of Affibody molecules labeled with ^{99m}Tc using cysteine-containing peptide-based chelators in excretory organs of NMRI mice at 4 h after injection. Data are presented as an average percent of injected dose per gram (% ID/g) for four mice \pm standard deviation. *Data for gastrointestinal tract with its content are a measure of hepatobiliary excretion. They are presented as %ID per whole sample.

prevent interaction of cysteine with resins. This design allowed the production of the Affibody molecule (designated as PEP05352, Fig. 3) both by peptide synthesis and recombinantly in *Escherichia coli*. Biodistribution studies with [^{99m}Tc]PEP05352 showed nearly threefold reduced renal retention of the radioactivity as compared to [^{99m}Tc]Z_{HER2:2395}-Cys, low levels of hepatic uptake and hepatobiliary excretion and good in vivo stability (low uptake in stomach) (Fig. 5). This study has demonstrated that engineering of the amino acid composition of cysteine-containing peptide-based chelators at the C-terminus provides a valuable tool for improving biodistribution properties of Affibody molecules.

Based on the results of the studies described above, an optimization of cysteine-based chelators has been performed (61). A series of Z_{HER2:342} derivatives containing GSEC, GGGC, GGSC, GGEC, or GGKC chelators at the C-terminus (Fig. 3) were labeled with ^{99m}Tc and evaluated in vitro and in vivo. The in vitro evaluation suggested that increasing hydrophilicity of amino acids in the chelators increased the residualizing effect. In vivo data (Fig. 6) showed that the GGGC chelator provided kidney uptake of $6.4 \pm 0.6\%$ ID/g at 4 h after injection while uptake of the GGKC-containing conjugate was $120 \pm 9\%$ ID/g at this time point. Thus, substitution of a single amino acid in the peptide-based chelator resulted in 19-fold increase of the renal uptake. In addition, the lysine-containing chelator caused hepatic retention of radioactivity that was more than threefold higher in comparison with other chelators. The best variant containing the GGGC chelator provided tumor uptake of $22.6 \pm 4.0\%$ ID/g and tumor-to-blood ratio of 186 ± 29 in mice bearing SKOV-3 xenografts at 4 h after injection.

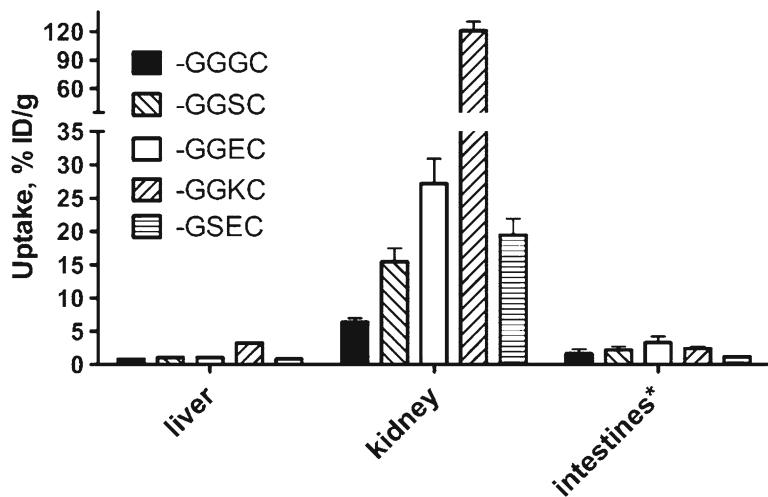


Fig. 6. Influence of amino acid composition of cysteine-containing peptide-based chelators at the C-terminus on uptake of ^{99m}Tc -labeled Affibody molecules in excretory organs of NMRI mice at 4 h after injection. Data are presented as an average percent of injected dose per gram (% ID/g) for four mice \pm standard deviation. *Data for gastrointestinal tract with its content are a measure of hepatobiliary excretion. They are presented as %ID per whole sample.

3.3. Engineering of an Optimized Scaffold for Affibody Molecules

In the preceding sections, we have described strategies to optimize Affibody molecules for in vivo imaging applications with various radionuclides by changing and/or adding a limited number of amino acids at the N- or C-terminus. In this section, we will go further and discuss strategies for improvement of the Affibody scaffold without interfering with the already good properties for the original scaffold.

The goal was to identify amino acids in the nontarget-binding surface that might be replaced without changing the overall 3-D structure of the original scaffold. It was hypothesized that careful replacement of the identified amino acids would lead to a designed scaffold with further reduced similarity to the Z-domain, reduced residual interactions with immunoglobulins, increased overall hydrophilicity, improved thermal and storage stability, enhanced amenability for peptide synthesis, and retained target-binding capacity. The amino acid substitutions were first analyzed in the context of the HER2-specific Affibody molecule $Z_{\text{HER2:342}}$ and later in five Affibody molecules targeting other proteins than HER2 (62).

The first step in this process was a thorough analysis of the primary sequence of the Z-domain scaffold and comparison to all other protein A domains (E, D, A, B, C) as well as analysis of available NMR and crystal structures of these domains. This analysis identified structurally important amino acids, i.e., amino acids buried in the hydrophobic core and surface exposed amino acids potentially involved in stabilizing the overall three-helical fold of

the Affibody scaffold molecules, and therefore these positions were not considered for amino acid substitutions.

The second step was generation of a structural model for $Z_{\text{HER2:342}}$ based on the refined NMR structure of the Z-domain (PDB ID: 1Q2N) (63) followed by extensive analysis and modeling to identify amino acids amenable for replacement.

The third step was an iterative process of replacing one or up to 11 amino acids located at the N-terminus, loop 1, helix 2, and helix 3 and analyzing the effect of the introduced changes on thermal stability and melting temperature (T_m), target-binding activity and interaction with immunoglobulins.

This process resulted in a designed Affibody scaffold with 11 amino acid substitutions in the nonbinding surface. These substitutions resulted in higher hydrophilicity, higher thermal stability, diminished background interactions with immunoglobulins, full production flexibility as well as fully retained *in vitro* and *in vivo* functionality. Together with the 13 randomized positions involved in target binding, 41% (24 of 58) of all amino acids are substituted in the optimized Affibody molecules as compared to the original Z-domain. The HER2-binding Affibody molecule in the optimized scaffold was designated $Z_{\text{HER2:2891}}$. Further coupling of maleimido-DOTA to the C-terminal cysteine led to the new lead candidate drug for molecular imaging (MMA-DOTA-Cys⁶¹)- $Z_{\text{HER2:2891}}$ -Cys (ABY-025) (62). Biodistribution studies in mice bearing HER2 overexpressing SKOV-3 xenografts revealed excellent tumor targeting, high tumor-to-organ ratios, high contrast SPECT images shortly after injections and rapid blood clearance. Preclinical characterization of ABY-025 revealed no toxicity in rats and cynomolgus monkeys and no induction of antibody formation upon repeated administration in rats (64). A recent pilot clinical trial with [¹¹¹In]ABY-025 in breast cancer patients with recurrent disease at the Uppsala University Hospital (Sweden) confirmed the good tumor targeting and safety properties of ABY-025 (88) manuscript in preparation).

4. Engineering of Affibody Molecules for Therapy

One of the goals for targeted therapy is to direct a payload to specific disease areas, e.g., a tumor or an inflammatory patch, but not to normal tissues or organs and, especially not to the excretory organs. Engineering of Affibody molecules for targeted therapy applications aims to improve and adapt the protein for some features which are very different from those demanded by the imaging agents described above. One obvious property is extension of the blood residence time often measured as plasma or serum half-life. Sufficient drug exposure is mandatory to obtain a desired therapeutic

effect and this is often correlated with a blood residence time optimized for the therapeutic regime. Technologies available for half-life extension of small protein drugs range from PEGylation, Fc-fusions, and albumin-fusions to fusions of a small protein domain, i.e., an albumin-binding domain (ABD) mediating the binding of the fusion-protein to the patient's own serum albumin. The third ABD of *Staphylococcus* G148 (G148-GA3, ABD₀₀₁ or ABD) is often used for these applications. Another engineering goal is to reduce or even suppress targeting of Affibody molecules to critical organs like bone marrow and the excretory organs, kidney and liver to prevent severe side effects. Finally, Affibody molecules need to be modified to allow easy and efficient coupling of a payload necessary for therapy. Payloads could be therapeutic radionuclides like ⁹⁰Y, ¹²²Lu, ¹³¹I, or ¹⁸⁶Re, small molecule toxins or fusions of toxic small protein domains.

4.1. Affibody-ABD Fusion Proteins

As highlighted in the previous section, the small size of a "monomeric" Affibody molecule (58 amino acids) is ideal for molecular imaging applications where rapid biodistribution, tumor targeting, and clearance of unbound protein is required for optimal imaging contrast (7). This is, however, suboptimal for most therapy applications where longer blood residence time is required. One exception is radiotherapy where a long residence time of circulating conjugate leads to increased irradiation of the very radiosensitive bone marrow, a main problem in radioimmunotherapy with full length antibodies.

Dosimetry calculations indicated that the radioactive dose to the kidneys was unacceptable high for both non-residualizing iodine- and residualizing indium-labeled first-generation Affibody molecules. Since monomeric Affibody molecules are smaller than the glomerular filtration limit of approximately 60 kDa (65) they are freely filtered through the glomerular membranes and subsequently reabsorbed, leading to accumulation of radioactive molecules in the kidney parenchyma. Thus, a reduction of renal accumulation is mandatory for efficient Affibody radiotherapy. This was achieved by fusing the HER2-binding Affibody molecule Z_{HER2:342} to ABD (66). Two fusion proteins, ABD-Z_{HER2:342} and ABD-(Z_{HER2:342})₂, were engineered with an N-terminal ABD domain and a C-terminal Affibody monomer or dimer (i.e., head-to-tail fusions of two Z_{HER2:342}). Both fusion proteins were modified with the chelator CHX-A''-DTPA to allow subsequent labeling with radiometals. The construct with only one Affibody molecule, CHX-A''-DTPA-ABD-Z_{HER2:342}, showed a 20-fold lower binding affinity to HER2, indicating that N-terminal ABD fusion may sterically interfere with the target-binding site on helix 1 and helix 2 of the Affibody molecule. In later experiments, the C-terminus of Affibody molecules has been proven to be the preferred fusion site for ABD (see below). In vivo studies revealed that [¹⁷⁷Lu]CHX-A''-DTPA-ABD-(Z_{HER2:342})₂ had a 70-fold slower blood clearance, a 25-fold

reduced kidney uptake and, a fivefold increased dose delivered to the tumor as compared to the non-ABD fused Affibody dimer [^{177}Lu]CHX-A''-DTPA-(Z_{HER2:342})₂. Treatment of mice bearing SKOV-3 tumor micro-xenografts with [^{177}Lu]CHX-A''-DTPA-ABD-(Z_{HER2:342})₂ completely prevented tumor formation (66).

Further improvements were obtained by a nearly complete redesign of the HER2-binding Affibody-ABD fusion protein. The new candidate drug, ABY-027, is composed of an N-terminal monomeric Affibody molecule in the optimized scaffold, Z_{HER2:2891}, a new affinity-matured ABD with femtomolar affinity for albumin-designated ABD₀₃₅ (see section below), and a DOTA chelator site-specifically coupled to a unique cysteine at the C-terminus (Z_{HER2:2891}-ABD₀₃₅-Cys-MMA-DOTA) [Orlova (2011), manuscript in preparation]. Biodistribution studies with [^{177}Lu]ABY-027 in SKOV-3 tumor-bearing mice showed efficient and specific tumor targeting. Both renal and hepatic uptake were reduced two- to threefold as compared to [^{177}Lu]CHX-A''-DTPA⁰-(Z_{HER2:342})₂. Dosimetry calculation revealed a 3–5 times higher dose to the tumor than to blood, kidney, and liver. The elimination half-life for ABY-027 increased 80-fold as compared to the non-ABD fused Z_{HER2:2891}.

In conclusion, fusion of Affibody molecules to ABD can be used to transform them from optimal imaging tracers for molecular imaging to potent candidates for targeted therapy.

4.2. Monomeric Affibody Molecules for Radiotherapy

Another approach for reduction of the renal dose of therapeutic radiolabeled Affibody conjugates is based on their slow internalization by tumor cells but rapid internalization in kidneys. Thus, monomeric Affibody molecules labeled with non-residualizing radionuclides may yield a reduced kidney dose without reducing the tumor dose. A potential benefit of this approach is that the reduction of renal radioactivity is not associated with longer residence of radiolabeled substance in blood, thereby minimizing the risk for high dose to the radiosensitive bone marrow.

One possible nuclide for therapy is ^{131}I . The use of radioiodine for labeling of Affibody molecules via coupling of *para*-iodobenzoate (Fig. 7) provided reasonably high tumor uptake, but much lower renal uptake than for radiometals (14). Further analysis showed that the area under curve (AUC) for tumors was only 1.5-fold higher than the AUC for kidneys, which is associated with unacceptably high risk of irreversible renal damage during radionuclide therapy. During the development of site-specific radiobromination, it was found that the use of 3-bromo-((4-hydroxyphenyl)ethyl)maleimide (Fig. 7) for labeling of cysteine-containing Affibody molecules provides sevenfold lower renal radioactivity accumulation at 4 h after injection than the use of *para*-bromobenzoate (67). The chemical similarity between iodine and bromine suggested that the use of ((4-hydroxyphenyl)ethyl)maleimide as a linker for radioiodination may reduce the dose to kidneys more

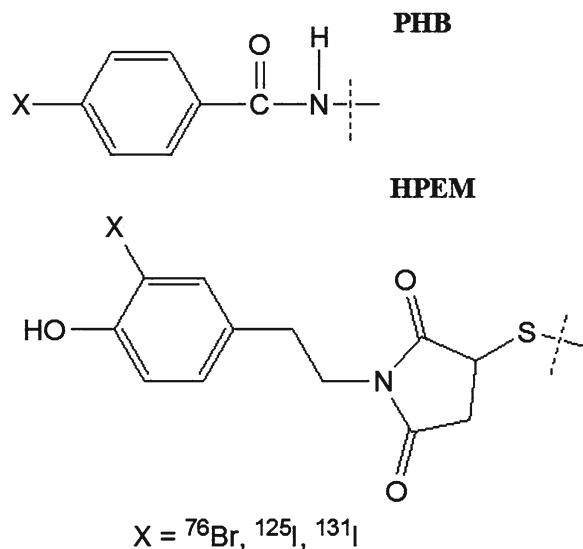


Fig. 7. Prosthetic groups for coupling of radiohalogens to Affibody molecules. *PHB* *para*-halobenzoate coupled to the ϵ -amino group of lysines or the N-terminus; *HPEM* radiohalogenated ((4-hydroxyphenyl)ethyl)maleimide site-specifically coupled to the thiol group of a C-terminal cysteine.

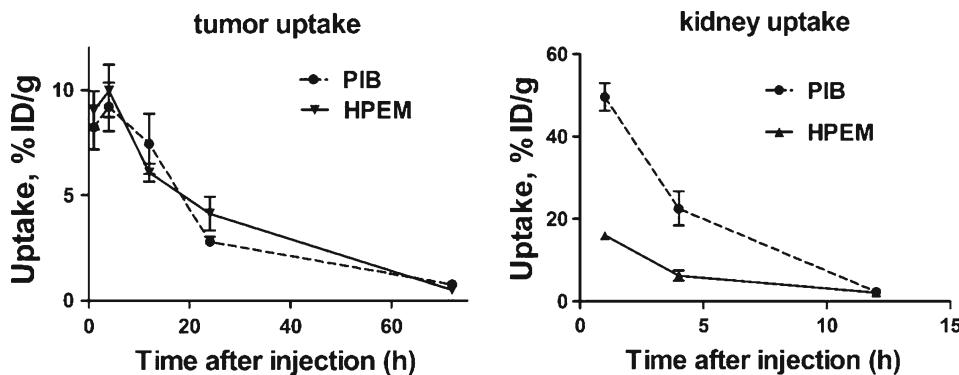


Fig. 8. Comparison of radioiodine accumulation in tumors and kidneys of BALB/c *nu/nu* mice bearing SKOV-3 xenografts after injection of Affibody molecules radioiodinated using *para*-iodobenzoate (PIB) or 3-iodo-((4-hydroxyphenyl)ethyl)maleimido (HPEM).

efficiently as compared to the use of *para*-iodobenzoate. Direct in vivo comparison of both labeling chemistries demonstrated that the AUC for tumors was approximately equal for both labels, but appreciably reduced for kidneys with iodo-HPEM (Fig. 8) (34). These results indicated that the use of monomeric Affibody molecules has a great potential for radionuclide therapy.

Experiments with mercaptoacetyl-containing chelators for $^{99\text{m}}\text{Tc}$ suggested that the use of maGGG and maGSG is associated with low renal retention of radioactivity (52, 55). Due to the nature of the emitted radiation, $^{99\text{m}}\text{Tc}$ is unsuitable for radionuclide therapy.

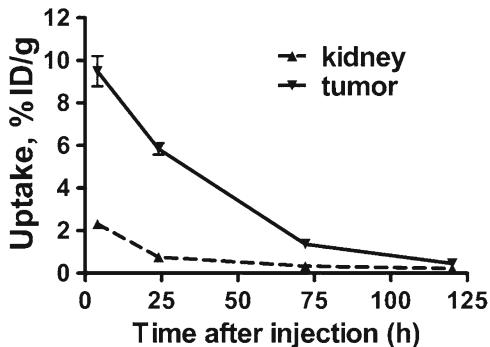


Fig. 9. Radioactivity accumulation in tumor and kidneys after injection of [¹⁸⁶Re] maGSG-Z_{HER2:342} in BALB/c *nu/nu* mice bearing SKOV-3 xenografts.

However, the two high-energy-beta-emitting isotopes of rhenium, ¹⁸⁶Re and ¹⁸⁸Re, a close chemical analogue of technetium, may have a potential for radionuclide therapy of bulky non-operable tumors (17). To evaluate this, maGGG-Z_{HER2:342} and maGSG-Z_{HER2:342} were labeled with ¹⁸⁶Re, and their biodistribution was compared in nude mice bearing HER2-expressing xenografts (68). The comparison suggested that [¹⁸⁶Re]maGSG-Z_{HER2:342} provides the lowest radioactivity accumulation in both kidneys and intestine content. Further evaluation of this conjugate (Fig. 9) showed that the AUC for tumor is more than fivefold higher than AUC for kidneys, making [¹⁸⁶Re]maGSG-Z_{HER2:342} a promising candidate for radionuclide therapy. Further improvement of dosimetry of rhenium-labeled Affibody molecules might be obtained using the C-terminal GGGC chelator described above, as in this case a low renal dose would not be associated with elevated dose to intestines.

4.3. Engineering of an Optimized Albumin-Binding Domain for Therapy

The third ABD of *Staphylococcus* G148 (G148-GA3, ABD₀₀₁ or ABD) was used in the first experiment to extend the half-life of Affibody molecules. ABD binds human, monkey, rat, and mouse serum albumin with low nanomolar affinity (69). ABD₀₀₁ is a small protein (46 amino acids, approximately 5 kDa) with a robust three helical bundle structure. The binding site is located on helix 2 and 3 of ABD and interacts with domain II of HSA (70). Upon binding to albumin, ABD seems to have the same kinetics and distribution properties as albumin itself (71). Studies using the Fab-fragment of trastuzumab fused to different albumin-binding peptides showed a close correlation between the affinity of these peptides for albumin and the serum half-life (72). Thus, engineering of ABD variants with high affinity might be useful to improve the serum half-life of biopharmaceuticals. This goal was achieved by using a combination of combinatorial protein engineering, in vitro phage display selection and rational design (73). In the first step, 15 of the 46

amino acids of ABD were randomized to different degrees and displayed on bacteriophages. Selected binders showed improved binding to albumin with affinities in the low picomolar range. Combination of different sequence features form the first set of binders and rational design lead to the identification of ABD₀₃₅, with a very high affinity for HSA (in the femtomolar range), relative high melting temperature (Tm), and excellent refolding properties. Pharmacokinetic experiments in mice using the high affinity variant ABD₀₃₅ (K_D 1.8 nM for MSA), the low affinity ABD variant (ABD_{Y20A}; K_D 634 nM for MSA), and the original ABD (K_D 21.4 nM for MSA) as fusion partner for a single-chain diabody, showed that the ABD₀₃₅-fusion protein had a considerable longer terminal half-life (47.5 h) than the two low affinity ABD-fusions (28.4 and 36.4 h, respectively) (74).

Like all other biopharmaceuticals, ABD could in principle stimulate an immune response. At least one T-cell epitope was identified in ABD₀₀₁ in earlier studies but this epitope was not removed during the affinity maturation process leading to ABD₀₃₅ (75). Therefore, ABD₀₃₅ has been subjected to a deimmunization program in order to remove potential T-cell epitopes, without affecting its high affinity, stability, and solubility (Ekblad C et al. Affibody AB). Potential T-cell epitopes were identified using open source in silico T-cell epitope prediction programs (76, 77) and a large number of ABD variants were created and analyzed using an iterative amino acid substitution process. Two variants with the best properties were analyzed in a T-cell proliferation assay using PBMC cells derived from 52 human donors and compared to the original ABD₀₀₁ (Algonomics/Lonza). T-cell proliferation in response to ABD₀₉₄ was obtained with PBMC cells from two donors only as compared to ABD₀₀₁ eliciting a response in PBMC cells from ten donors. Two individuals responded to rHSA alone and 51 individuals to the positive control (KLH). Thus, the number of T-cell epitopes in the optimized ABD variant were substantially decreased as no significant immunogenicity could be detected with ABD₀₉₄, indicating that this construct has very low immunogenic potential (Ekblad C et al. Affibody AB).

5. Conclusion and Outlook

Taken together, Affibody molecules have favorable properties for the development of novel therapeutic, diagnostic imaging and biotechnological applications. The first clinical trials indicate safety and efficacy. The use of the deimmunized, high affinity albumin-binding domain (ABD₀₉₄) is currently being evaluated for half-life extension of not only Affibody molecules, but also a range of different biopharmaceuticals.

Acknowledgments

The authors would like to thank Anders Wennborg for critical reading of the manuscript.

References

1. An Z (2010) Monoclonal antibodies – a proven and rapidly expanding therapeutic modality for human diseases. *Protein Cell* 1:319–330
2. Kenanova V, Wu AM (2006) Tailoring antibodies for radionuclide delivery. *Expert Opin Drug Deliv* 3:53–70
3. Mohlmann S et al (2011) Site-specific modification of ED-B-targeting antibody using intein-fusion technology. *BMC Biotechnol* 11:76
4. Friedman M, Stahl S (2009) Engineered affinity proteins for tumour-targeting applications. *Biotechnol Appl Biochem* 53:1–29
5. Löfblom J et al (2010) Affibody molecules: engineered proteins for therapeutic, diagnostic and biotechnological applications. *FEBS Lett* 584:2670–2680
6. Zoller F, Haberkorn U, Mier W (2011) Miniproteins as phage display-scaffolds for clinical applications. *Molecules* 16:2467–2485
7. Baum RP et al (2010) Molecular imaging of HER2-expressing malignant tumors in breast cancer patients using synthetic ¹¹¹In- or ⁶⁸Ga-labeled Affibody molecules. *J Nucl Med* 51:892–897
8. Tolcher AW et al (2011) Phase I and pharmacokinetic study of CT-322 (BMS-844203), a targeted Adnectin inhibitor of VEGFR-2 based on a domain of human fibronectin. *Clin Cancer Res* 17:363–371
9. Uhlén M et al (1984) Complete sequence of the staphylococcal gene encoding protein A. A gene evolved through multiple duplications. *J Biol Chem* 259:1695–1702
10. Moks T et al (1986) Staphylococcal protein A consists of five IgG-binding domains. *Eur J Biochem* 156:637–643
11. Nilsson B et al (1987) A synthetic IgG-binding domain based on staphylococcal protein A. *Protein Eng* 1:107–113
12. Arora P, Oas TG, Myers JK (2004) Fast and faster: a designed variant of the B-domain of protein A folds in 3 microsec. *Protein Sci* 13:847–853
13. Nord K et al (1995) A combinatorial library of an alpha-helical bacterial receptor domain. *Protein Eng* 8:601–608
14. Orlova A et al (2006) Tumor imaging using a picomolar affinity HER2 binding Affibody molecule. *Cancer Res* 66:4339–4348
15. Grönwall C et al (2007) Selection and characterization of Affibody ligands binding to Alzheimer amyloid beta peptides. *J Biotechnol* 128:162–183
16. Lindborg M et al (2011) Engineered high-affinity Affibody molecules targeting platelet-derived growth factor receptor beta in vivo. *J Mol Biol* 407:298–315
17. Tolmachev V (2008) Choice of radionuclides and radiolabeling techniques. In: Stigbrand T (ed) Targeted radionuclide tumor therapy – biological aspects. Springer Science + Business Media B.V, Dordrecht, pp 145–174
18. Mattes MJ et al (1994) Processing of antibody-radioisotope conjugates after binding to the surface of tumor cells. *Cancer* 73:787–793
19. Shih LB et al (1994) The processing and fate of antibodies and their radiolabels bound to the surface of tumor cells in vitro: a comparison of nine radiolabels. *J Nucl Med* 35:899–908
20. Press OW et al (1996) Comparative metabolism and retention of iodine-125, yttrium-90, and indium-111 radioimmunoconjugates by cancer cells. *Cancer Res* 56:2123–2129
21. Orlova A et al (2000) Cellular processing of ¹²⁵I- and ¹¹¹In-labeled epidermal growth factor (EGF) bound to cultured A431 tumor cells. *Nucl Med Biol* 27:827–835
22. Tolmachev V, Orlova A, Lundqvist H (2003) Approaches to improve cellular retention of radiohalogen labels delivered by internalising tumour-targeting proteins and peptides. *Curr Med Chem* 10:2447–2460
23. Behr TM et al (2001) Imaging tumors with peptide-based radioligands. *Q J Nucl Med* 45:189–200
24. Reubi JC (2003) Peptide receptors as molecular targets for cancer diagnosis and therapy. *Endocr Rev* 24:389–427
25. Thorpe SR, Baynes JW, Chroneos ZC (1993) The design and application of residualizing labels for studies of protein catabolism. *FASEB J* 7:399–405

26. Behr TM, Goldenberg DM, Becker W (1998) Reducing the renal uptake of radiolabeled antibody fragments and peptides for diagnosis and therapy: present status, future prospects and limitations. *Eur J Nucl Med* 25:201–212
27. Vegt E et al (2010) Renal toxicity of radiolabeled peptides and antibody fragments: mechanisms, impact on radionuclide therapy, and strategies for prevention. *J Nucl Med* 51:1049–1058
28. Hagenbuch B (2010) Drug uptake systems in liver and kidney: a historic perspective. *Clin Pharmacol Ther* 87:39–47
29. Decristoforo C, Mather SJ (1999) 99m-Technetium-labelled peptide-HYNIC conjugates: effects of lipophilicity and stability on biodistribution. *Nucl Med Biol* 26:389–396
30. Wilbur DS (1992) Radiohalogenation of proteins: an overview of radionuclides, labeling methods, and reagents for conjugate labeling. *Bioconjug Chem* 3:433–470
31. Sundin J et al (1999) High yield direct ⁷⁶Br-bromination of monoclonal antibodies using chloramine-T. *Nucl Med Biol* 26: 923–929
32. Nikula TK et al (1995) Impact of the high tyrosine fraction in complementarity determining regions: measured and predicted effects of radioiodination on IgG immunoreactivity. *Mol Immunol* 32:865–872
33. Steffen AC et al (2005) In vitro characterization of a bivalent anti-HER-2 Affibody with potential for radionuclide-based diagnostics. *Cancer Biother Radiopharm* 20:239–248
34. Tolmachev V et al (2009) Influence of valency and labelling chemistry on in vivo targeting using radioiodinated HER2-binding Affibody molecules. *Eur J Nucl Med Mol Imaging* 36: 692–701
35. Camera L et al (1994) Evaluation of the serum stability and in vivo biodistribution of CHX-DTPA and other ligands for yttrium labeling of monoclonal antibodies. *J Nucl Med* 35: 882–889
36. Tolmachev V et al (2008) Evaluation of a maleimido derivative of CHX-A'' DTPA for site-specific labeling of Affibody molecules. *Bioconjug Chem* 19:1579–1587
37. Kelly MP et al (2009) Therapeutic efficacy of ¹⁷⁷Lu-CHX-A''-DTPA-hu3S193 radioimmuno-therapy in prostate cancer is enhanced by EGFR inhibition or docetaxel chemotherapy. *Prostate* 69:92–104
38. De Leon-Rodriguez LM, Kovacs Z (2008) The synthesis and chelation chemistry of DOTA-peptide conjugates. *Bioconjug Chem* 19:391–402
39. Anderson CJ et al (2008) Cross-bridged macrocyclic chelators for stable complexation of copper radionuclides for PET imaging. *Q J Nucl Med Mol Imaging* 52:185–192
40. Wällberg H, Orlova A (2008) Slow internalization of anti-HER2 synthetic Affibody monomer ¹¹¹In-DOTA-Z_{HER2:342-pep2}: implications for development of labeled tracers. *Cancer Biother Radiopharm* 23:435–442
41. Ahlgren S et al (2008) Evaluation of maleimide derivative of DOTA for site-specific labeling of recombinant Affibody molecules. *Bioconjug Chem* 19:235–243
42. Tran TA et al (2009) Design, synthesis and biological evaluation of a multifunctional HER2-specific Affibody molecule for molecular imaging. *Eur J Nucl Med Mol Imaging* 36:1864–1873
43. Tolmachev V et al (2009) Affibody molecules for epidermal growth factor receptor targeting in vivo: aspects of dimerization and labeling chemistry. *J Nucl Med* 50:274–283
44. Orlova A et al (2006) Comparative in vivo evaluation of technetium and iodine labels on an anti-HER2 Affibody for single-photon imaging of HER2 expression in tumors. *J Nucl Med* 47:512–519
45. Orlova A et al (2009) On the selection of a tracer for PET imaging of HER2-expressing tumors: direct comparison of a ¹²⁴I-labeled Affibody molecule and trastuzumab in a murine xenograft model. *J Nucl Med* 50:417–425
46. Kramer-Marek G et al (2008) [¹⁸F]FBEM-Z_{HER2:342}-Affibody molecule-a new molecular tracer for in vivo monitoring of HER2 expression by positron emission tomography. *Eur J Nucl Med Mol Imaging* 35:1008–1018
47. Cheng Z et al (2008) Small-animal PET imaging of human epidermal growth factor receptor type 2 expression with site-specific ¹⁸F-labeled protein scaffold molecules. *J Nucl Med* 49:804–813
48. Tolmachev V et al (2007) Affibody molecules: potential for in vivo imaging of molecular targets for cancer therapy. *Expert Opin Biol Ther* 7:555–568
49. Orlova A et al (2007) Update: Affibody molecules for molecular imaging and therapy for cancer. *Cancer Biother Radiopharm* 22:573–584
50. Tolmachev V et al (2006) ¹¹¹In-benzyl-DTPA-Z_{HER2:342}, an Affibody-based conjugate for in vivo imaging of HER2 expression in malignant tumors. *J Nucl Med* 47:846–853
51. Orlova A et al (2007) Synthetic Affibody molecules: a novel class of affinity ligands for molecular imaging of HER2-expressing malignant tumors. *Cancer Res* 67:2178–2186

52. Engfeldt T et al (2007) Imaging of HER2-expressing tumours using a synthetic Affibody molecule containing the 99m Tc-chelating mercaptoacetyl-glycyl-glycyl-glycyl (MAG3) sequence. *Eur J Nucl Med Mol Imaging* 34:722–733
53. Wang Y, Liu X, Hnatowich DJ (2007) An improved synthesis of NHS-MAG3 for conjugation and radiolabeling of biomolecules with 99m Tc at room temperature. *Nat Protoc* 2:972–978
54. Lister-James J, Moyer BR, Dean RT (1997) Pharmacokinetic considerations in the development of peptide-based imaging agents. *Q J Nucl Med* 41:111–118
55. Engfeldt T et al (2007) 99m Tc-chelator engineering to improve tumour targeting properties of a HER2-specific Affibody molecule. *Eur J Nucl Med Mol Imaging* 34:1843–1853
56. Tran T et al (2007) 99m Tc-maEEE-Z_{HER2:342}, an Affibody molecule-based tracer for the detection of HER2 expression in malignant tumors. *Bioconjug Chem* 18:1956–1964
57. Ekblad T et al (2008) Development and pre-clinical characterisation of 99m Tc-labelled Affibody molecules with reduced renal uptake. *Eur J Nucl Med Mol Imaging* 35:2245–2255
58. Tran TA et al (2008) Effects of lysine-containing mercaptoacetyl-based chelators on the biodistribution of 99m Tc-labeled anti-HER2 Affibody molecules. *Bioconjug Chem* 19:2568–2576
59. Tran T et al (2007) In vivo evaluation of cysteine-based chelators for attachment of 99m Tc to tumor-targeting Affibody molecules. *Bioconjug Chem* 18:549–558
60. Ahlgren S et al (2009) Targeting of HER2-expressing tumors with a site-specifically 99m Tc-labeled recombinant Affibody molecule, Z_{HER2:2395}, with C-terminally engineered cysteine. *J Nucl Med* 50:781–789
61. Wällberg H et al (2011) Molecular design and optimization of 99m Tc-labelled recombinant Affibody molecules improves their biodistribution and imaging properties. *J Nucl Med* 52: 461–469
62. Feldwisch J et al (2010) Design of an optimized scaffold for Affibody molecules. *J Mol Biol* 398:232–247
63. Zheng D, Aramini JM, Montelione GT (2004) Validation of helical tilt angles in the solution NMR structure of the Z domain of Staphylococcal protein A by combined analysis of residual dipolar coupling and NOE data. *Protein Sci* 13:549–554
64. Ahlgren S et al (2010) Targeting of HER2-expressing tumors using 111 In-ABY-025, a second-generation Affibody molecule with a fundamentally reengineered scaffold. *J Nucl Med* 51:1131–1138
65. Christensen EI, Verroust PJ, Nielsen R (2009) Receptor-mediated endocytosis in renal proximal tubule. *Pflugers Arch* 458:1039–1048
66. Tolmachev V et al (2007) Radionuclide therapy of HER2-positive microxenografts using a 177 Lu-labeled HER2-specific Affibody molecule. *Cancer Res* 67:2773–2782
67. Mume E et al (2005) Evaluation of ((4-hydroxyphenyl)ethyl)maleimide for site-specific radio-bromination of anti-HER2 Affibody. *Bioconjug Chem* 16:1547–1555
68. Orlova A et al (2010) 186 Re-maSGS-Z_{HER2:342}, a potential Affibody conjugate for systemic therapy of HER2-expressing tumours. *Eur J Nucl Med Mol Imaging* 37:260–269
69. Johansson MU et al (2002) Structure, specificity, and mode of interaction for bacterial albumin-binding modules. *J Biol Chem* 277:8114–8120
70. Lejon S et al (2004) Crystal structure and biological implications of a bacterial albumin binding module in complex with human serum albumin. *J Biol Chem* 279:42924–42928
71. Andersen JT et al (2011) Extending half-life by indirect targeting of the neonatal Fc receptor (FcRn) using a minimal albumin binding domain. *J Biol Chem* 286:5234–5241
72. Nguyen A et al (2006) The pharmacokinetics of an albumin-binding Fab (AB.Fab) can be modulated as a function of affinity for albumin. *Protein Eng Des Sel* 19:291–297
73. Jonsson A et al (2008) Engineering of a femtomolar affinity binding protein to human serum albumin. *Protein Eng Des Sel* 21:515–527
74. Hopp J et al (2010) The effects of affinity and valency of an albumin-binding domain (ABD) on the half-life of a single-chain diabody-ABD fusion protein. *Protein Eng Des Sel* 23:827–834
75. Goetsch L et al (2003) Identification of B- and T-cell epitopes of BB, a carrier protein derived from the G protein of *Streptococcus* strain G148. *Clin Diagn Lab Immunol* 10:125–132
76. Singh H, Raghava GP (2001) ProPred: prediction of HLA-DR binding sites. *Bioinformatics* 17:1236–1237
77. Nielsen M, Lund O (2009) NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction. *BMC Bioinformatics* 10:296
78. Ekblad T et al (2009) Positioning of 99m Tc-chelators influences radiolabeling, stability and biodistribution of Affibody molecules. *Bioorg Med Chem Lett* 19:3912–3914

79. Tolmachev V et al (2010) Imaging of EGFR expression in murine xenografts using site-specifically labelled anti-EGFR ^{111}In -DOTA-Z_{EGFR:2377} Affibody molecule: aspect of the injected tracer amount. *Eur J Nucl Med Mol Imaging* 37:613–622
80. Namavari M et al (2008) Direct site-specific radiolabeling of an Affibody protein with 4-[^{18}F]fluorobenzaldehyde via oxime chemistry. *Mol Imaging Biol* 10:177–181
81. Wällberg H et al (2010) Evaluation of the radiocobalt-labeled [MMA-DOTA-Cys⁶¹]-Z_{HER2:2395}-Cys Affibody molecule for targeting of HER2-expressing tumors. *Mol Imaging Biol* 12:54–62
82. Cheng Z et al (2010) ^{64}Cu -labeled Affibody molecules for imaging of HER2 expressing tumors. *Mol Imaging Biol* 12:316–324
83. Miao Z et al (2010) Small-animal PET imaging of human epidermal growth factor receptor positive tumor with a ^{64}Cu labeled Affibody protein. *Bioconjug Chem* 21:947–954
84. Tolmachev V et al (2010) A HER2-binding Affibody molecule labelled with ^{68}Ga for PET imaging: direct in vivo comparison with the ^{111}In -labelled analogue. *Eur J Nucl Med Mol Imaging* 37:1356–1367
85. Fortin MA et al (2007) Labelling chemistry and characterization of [$^{90}\text{Y}/^{177}\text{Lu}$]-DOTA-Z_{HER2:342-3} Affibody molecule, a candidate agent for locoregional treatment of urinary bladder carcinoma. *Int J Mol Med* 19:285–291
86. Tolmachev V et al (2009) The influence of Bz-DOTA and CHX-A"-DTPA on the biodistribution of ABD-fused anti-HER2 Affibody molecules: implications for $^{114\text{m}}\text{In}$ -mediated targeting therapy. *Eur J Nucl Med Mol Imaging* 36:1460–1468
87. Steffen AC et al (2007) Biodistribution of ^{211}At labeled HER-2 binding Affibody molecules in mice. *Oncol Rep* 17:1141–1147
88. Sandberg D et al (2011) First-in-human whole-body HER2-receptor mapping using Affibody molecular imaging. *Cancer Res* 71[24Suppl.]: 273s Abstract P2-09-01

Chapter 8

Protein Design for Diversity of Sequences and Conformations Using Dead-End Elimination

Karl J.M. Hanf

Abstract

Proteins, especially antibodies, are widely used as therapeutic and diagnostic agents. Computational protein design is a powerful tool for improving the affinity and stability of these molecules. We describe a protein design method which employs the dead-end elimination (DEE) and A* discrete search algorithms with a few improvements aimed at making the procedure more useful for actual projects to design proteins for better affinity or stability. DEE/A* and related algorithms allow vast search spaces of protein sequences and their alternative side chain conformations (“rotamers”) to be systematically explored, to find those with the best free energy of folding or binding. To maximize a protein design project’s chance of success, it needs to find a diverse set of sequences to experimentally synthesize. It should also find structures that score well, not only on the pairwise-additive energy function which DEE/A* and related search algorithms must use, but also on a post-search energy function with accurate treatment of solvation effects. Straight DEE/A*, however, typically finds vast numbers of very similar low-energy conformations, making it infeasible to find a diverse set of sequences or conformations. Herein, we describe a three-level DEE/A* procedure that uses DEE/A* at the level of sequences, at the level of rotamers, and at an intermediate “fleximer” level, to ensure a wide variety of sequences as well as a diverse set of conformations for each sequence.

A physics-based method is also described herein for calculating the free energy of folding based on a thermodynamic cycle with a model of the unfolded state. The free energies of both folding and binding may be used for the final evaluation of the designed structures. For example, when designing for improved affinity (binding), we can also ensure that stability is not degraded by screening on the free energy of folding.

Key words: Protein design, Dead-end elimination, A*, Rotamers, Fleximers, Electrostatics, Finite-difference Poisson–Boltzmann, Free energy of folding, Protein stability

1. Introduction

Use of protein-based drugs and diagnostic agents has expanded significantly over the past 20 years, and candidate agents can now often be produced by *in vivo* or *in vitro* affinity maturation.

Computational protein design—i.e., *in silico* affinity maturation—can further improve the candidates' target affinity in order to achieve better drug potency or detection sensitivity. Computational design can also improve the candidates' stability in order to improve storage characteristics, drug circulation times, and other properties.

Computational design can be done by stochastic Monte Carlo search methods, which can quickly but incompletely search the possible conformations for those with better affinity or stability, or by discrete deterministic methods, which can also guarantee that the minimum-energy conformation(s) will be found. For discrete search methods, the space of all possible conformations of the system is represented by a discrete set — a library — of side chain conformations, called “rotamers,” for each amino acid type. The excellent and widely used Dunbrack backbone-dependent and backbone-independent rotamer libraries (1) are based on clustering the most common dihedral angles found in the Protein Data Bank (PDB) for each amino acid type. For example, most side chain dihedral angles have local minima in the vicinity of -60° , $+60^\circ$, and $+180^\circ$.

Dead-end elimination (DEE) is a way to reduce the size of a very large search space by repeatedly eliminating rotamers which cannot be in the global minimum-energy conformation (GMEC) (2). For example, the simple Goldstein criterion compares two rotamers at one position and eliminates the first rotamer if it gives worse energy than the second rotamer for all possible sets of rotamers at the other positions (3) The A* search algorithm can be used to search for the GMEC in the reduced search space remaining after DEE. Both DEE and A* are easily extended to find all conformations within a given energy cutoff of the minimum energy, though this reduces the eliminating power of DEE and therefore leaves a much larger search space for A* to traverse (4–6).

We mitigate the inherent discreteness of the rotamer library by expanding each rotamer in the library into a “fleximer,” i.e., a set of up to 9 very similar rotamers. We use three levels of description: the sequence (a list of the amino acid type at each mobile residue), the fleximer state (a list of the fleximer state at each mobile residue), and the rotamer state (a list of the rotamer at each mobile residue, defining a specific conformation). The DEE/A* search algorithms are very efficient, but require an energy function which is pairwise-additive, i.e., which has no three-body terms (2). Because the pairwise-additive energy function will generally favor some sequences over others—notably, its lack of solvation effects gives it a false bias toward charged residues—doing DEE/A* directly on the full rotamer space can find such a vast list of conformations of over-charged sequences that it cannot feasibly find other, less-charged sequences. Similarly, doing DEE/A* on the rotamer subspace of one sequence finds such a vast list of nearly identical conformations that it cannot feasibly find a few more-diverse conformations (7). We solve both of these problems

by doing DEE/A* on three levels: first to rank sequences (by treating each amino acid as a combined rotamer); second, for each sequence, to rank fleximer states; and third, for each of the ten best fleximer states per sequence, to find the one minimum-energy rotamer state in that fleximer state (8).

Lastly we reevaluate the list of conformations found by our three-level DEE/A* procedure with a final energy function which has a more accurate treatment of solvation: finite-difference solution of the Poisson–Boltzmann equation (FDPB) and a hydrophobic term proportional to the solvent-accessible surface area (SASA). FDPB is an accurate and computationally feasible way to estimate the electrostatic free energy, including non-pairwise-additive solvation effects (9).

For both the pairwise-additive and the final energy functions, we represent stability not only by the energy of the final state but also by a proper free energy of folding—the energy difference of the final state from the unfolded state—which we model by treating each side chain as a separate solvated single-amino-acid peptide fragment. Final evaluation of the designs for stability and/or affinity is done using the final folding and/or binding energy.

2. Materials

The following software resources are needed to implement this procedure.

1. Molecular modeling: CHARMM (10) and SHARPE (11, 12) are among the many software packages available to do the molecular modeling tasks required for this procedure: place all atoms of a side chain given the set of its dihedral angles, place polar hydrogens on a heavy-atom structure, calculate interaction energies between two specified atom sets, and do some simple energy minimization of a structure.
2. DEE/A*: We use proprietary software to do DEE and A*, but the algorithms are in the open literature referenced herein and are also available in a few open-source packages such as EGAD (13, 14) and OSPREY (15, 16) (see Note 1).
3. Poisson–Boltzmann electrostatics: We use the open-source APBS package (17, 18) to do FDPB. It uses a few parameters for the continuum solvent model, a van der Waals radius parameter set, and an atomic charge parameter set.
4. Solvent-accessible surface area (SASA): SASA can be calculated (19), in total or for each atom, by a number of available programs, including CHARMM.
5. Most of this procedure can be automated by programs to tie the preceding parts together. Also needed are programs to

manipulate data structures such as: the amino acid types to allow at each position, the number of rotamers in each fleximer, the number of fleximers at each position, the atomic coordinates of each rotamer, a flag for each rotamer if it is to be kept, APBS input and output files, rotamer self and pair energy terms, and self and pair terms of fleximer energy. The programs we used for these purposes were proprietary to MIT and/or Biogen Idec.

3. Methods

3.1. System Setup

1. Prepare the starting structure: If it is from X-ray crystallography, select one crystallographic subunit or do the entire procedure on each crystallographic subunit. Identify missing heavy atoms. Identify disulfide bonds by the proximity of cysteine pairs. Add polar hydrogen atoms to the structure using the HBUILD facility (20) of the CHARMM package (or using the pdb2pqr program available with APBS). Minimize the structure briefly, e.g., CHARMM minimization for 100 steps.
2. Choose the objective function: The goal of your design project determines the objective function which you seek to minimize. To design for improved stability, use the free energy of folding as your objective function. To design for tighter binding of two binding partners, use the free energy of binding (see Fig. 1).
3. Choose evolving residues: Screen for single mutations by doing the entire procedure for each residue (or, if designing for binding, only for each residue within 10 Å of the binding partner). In each run, declare only that one residue to be “evolving,” i.e., allowed to explore conformations of all amino acid types (see Note 2).
4. Choose molten residues: Declare all residues within 8 Å of the evolving residue(s) to be “molten,” i.e., of fixed amino acid type

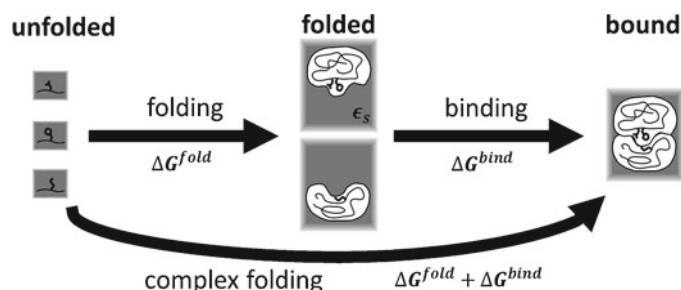


Fig. 1. The unfolded state model, a separate solvated single-amino-acid peptide fragment for each mobile residue, depends only on the amino acid sequence. The folded (but unbound) state model is each separate solvated binding partner.

but allowed to explore new conformations. All residues not evolving or molten, as well as the entire backbone, are declared “fixed,” and their atomic coordinates are held constant.

5. Obtain a rotamer library: Start with the Dunbrack backbone-independent (or backbone-dependent) rotamer library. We allow 19 natural amino acid types (all but proline, because its backbone is different), including three forms of histidine (protonated on N_{α} , N_{β} , or both), for a total of 370 rotamers. We do not allow protonated forms of Asp or Glu, though this could be done if the correct free energy cost of protonation were taken into account (8).
6. Expand the rotamer library into fleximers: Using a discrete rotamer library rather than a continuous search space can cause a minimum-energy conformation to be missed if the rotamer library is too coarse. Mitigate this problem by making the energy function softer (see below, about scaling the van der Waals radii) and by making the rotamer library finer by expanding each of the 370 rotamers of the rotamer library into a “fleximer” (21), i.e., a small family of up to 9 similar rotamers, by permuting each of the first two dihedral angles ϕ_1 and ϕ_2 by -5° , 0° , or $+5^\circ$. So we use 370 fleximers which comprise 3,386 rotamers. Store the coordinates of each side chain atom of each rotamer.
7. Choose a pairwise-additive energy function: The DEE/A* search algorithms require a pairwise-additive energy function, meaning that the interaction of two atoms does not depend on the location or properties of any other atoms. A pairwise-additive energy function can be expressed as the sum of one “fixed” term, a “self” term for each mobile residue’s interactions with the fixed atoms and with itself, and a “pair” term for the interaction of each mobile residue pair. For our pairwise energy function, we use the covalent (bond, angle, dihedral, improper dihedral), van der Waals, and distance-dependent Coulombic energy terms, all calculated by the CHARMM package using the PARAM19 parameters (see ref. 10) (see Note 3) with atom types patched for disulfide bonds. As mentioned above, to mitigate the problem of a rotamer library’s inherent coarseness, we scale the van der Waals radii by 0.9 so as not to penalize slight atomic clashes which could easily be relieved in reality by small movements not allowed by the rotamer library. (22) For electrostatics, we use the “distance-dependent Coulombic” energy, which is Coulomb’s law but with the screening effect of solvent crudely modeled by an effective dielectric constant $\epsilon_{ij} = \epsilon r_{ij} / 1 \text{ \AA}$ that increases linearly with the distance r_{ij} between atoms i and j . The electrostatic and entropic/hydrophobic components of the free energy cannot be captured well by any pairwise-additive energy function.

This distance-dependent Coulombic energy generally favors charged residues too much, because it does not capture the energetic cost to desolvate them on folding or binding.

8. Make unfolded model compounds: We model the unfolded state, which is actually an ill-defined ensemble of conformations, as a set of totally separate solvated model compounds, one per mobile position, with each unfolded model compounds depending only on the amino acid type at the position. The unfolded model compound for each amino acid type X is an N -acetyl- X methylamide $\text{CH}_3 \square (\text{CO}) \square (\text{NH}) \square (\text{C}_{\square} X) \square (\text{CO}) \square (\text{NH}) \square \text{CH}_3$, with the backbone held in an extended conformation and the side chain atoms beyond C_{\square} minimized to completion using the pairwise-additive free energy function but with full van der Waals radii.

3.2. Initial Optimization

1. Calculate rotamer self and pair energy terms: Using your chosen pairwise-additive energy function, calculate and store the self and pair terms of the free energy of folding, and also of the free energy of binding if that is your objective function. Constant terms of the free energy—the interactions among the fixed atoms in the bound, the unbound “folded” state, and in unfolded model compounds representing the fixed atoms’ unfolded state—will not change the relative free energy differences between conformations, so they are ignored. In our rigid binding model, with this pairwise-additive energy function, each binding self term is a rotamer’s interaction with the fixed atoms in the other binding partner; and each binding pair term is the interaction between a pair of rotamers if they are on the opposite binding partners or 0 if they are on the same binding partner. Each folding self term is a rotamer’s interaction with itself and with the fixed atoms in its binding partner minus the interaction of the side chain of the unfolded model compound of its amino acid type with the whole unfolded model compound. Each folding pair term is the interaction between a pair of rotamers if they are on the same binding partner or 0 if they are on the opposite binding partners.
2. Reduce rotamer search space with self term cutoffs: Eliminate every rotamer with a self term (binding or folding) over a cut-off, e.g., 25 kcal/mol. This will save time in the upcoming DEE/A* steps. Eliminating these rotamers entails writing new, shorter files for the energy terms and the rotamer coordinates.
3. Calculate approximate fleximer energy self and pair terms: We use three levels of description: the sequence (a list of the amino acid type at each mobile residue), the fleximer state (a list of the fleximer states at each mobile residue), and the rotamer state (a list of the rotamers at each mobile residue). Only a rotamer state defines an exact conformation, but we define a pairwise-additive approximate “fleximer energy” in preparation

for using DEE/A* once for each sequence to find the lowest-energy fleximer states in their much smaller search space (the fleximers of one sequence). We use the simple form from ref. 21 (see Note 4) for the fleximer energy G^{flex} self and pair terms, based on the energy G self and pair terms of the rotamers $\{i_r\}$ and $\{j_u\}$ which comprise fleximers i_R and j_U at mobile positions i and j , respectively:

$$\begin{aligned} G_{\text{self}}^{\text{flex}}(i_R) &= \min_{r \sqsubseteq R} [G_{\text{self}}(i_r)], \\ G_{\text{pair}}^{\text{flex}}(i_R, j_U) &= \min_{r \sqsubseteq R, u \sqsubseteq U} [G_{\text{self}}(i_r) + G_{\text{self}}(j_u) + G_{\text{pair}}(i_r, j_u)] \\ &\quad \square G_{\text{self}}^{\text{flex}}(i_R) \square G_{\text{self}}^{\text{flex}}(j_U). \end{aligned}$$

The value of G^{flex} for a fleximer state (i_R, j_U, \square) approximates the minimum G of any of its constituent rotamer states (i_r, j_u, \square) , but is often an underestimate, because a fleximer i_R gets credit for its best possible interaction with a neighboring fleximer j_U and for its best possible interaction with another neighboring fleximer k_X , even if no single rotamer state (i_r, j_u, k_X, \square) could make both interactions. Less often, G^{flex} can overestimate G because of the way the $G_{\text{self}}(i_r)$ terms picked out by the min function may be different for $G_{\text{pair}}^{\text{flex}}(i_R, j_U)$ than for $G_{\text{pair}}^{\text{flex}}(i_R, k_X)$. It is a simple matter to calculate the table of fleximer energy self and pair terms from the table of energy self and pair terms.

4. Calculate approximate sequence energy self and pair terms: In an exactly analogous fashion, we can consider each amino acid type at each position as one fleximer comprised of all its rotamers, and so define an approximate “sequence energy” G^{seq} , and easily calculate and store the table of sequence energy self terms $G_{\text{self}}^{\text{seq}}$ and pair terms $G_{\text{pair}}^{\text{seq}}$ (see Note 5).
5. Do DEE on sequence energy to reduce the search space: Do these DEE elimination steps in order: Goldstein singles (3), split singles (23, 24), and Goldstein magic bullet pairs (25, 26). Repeat all three steps until no more rotamers can be eliminated. DEE will greatly reduce the search space but not generally reduce it to only the GMEC.

We will briefly define these types of DEE by giving the criterion to eliminate rotamer i_r , using the rotamer notation i_r here for generality, although the algorithm is precisely the same when considering fleximers i_R or amino acid types:

To do Goldstein singles DEE: Eliminate rotamer i_r if there is some alternative rotamer i_t such that

$$G_{\text{self}}(i_r) \square G_{\text{self}}(i_t) + \bigcup_{j, j \sqsubseteq i} \left\{ \min_u \square G_{\text{pair}}(i_r, j_u) \square G_{\text{pair}}(i_t, j_u) \right\} > 0,$$

i.e., if switching from i_r to i_t gives a lower energy regardless of which rotamers are at the other positions j . To do split singles: Eliminate rotamer i_r if, for each rotamer k_x at some position $k \square i$, there is some alternative rotamer i_t such that

$$\begin{aligned} & G_{\text{self}}(i_r) \square G_{\text{self}}(i_t) \\ & + \bigcup_{j, j \square i, j \square k} \left\{ \min_u \left[G_{\text{pair}}(i_r, j_u) \square G_{\text{pair}}(i_t, j_u) \right] \right\} \\ & + \left[G_{\text{pair}}(i_r, k_x) \square G_{\text{pair}}(i_t, k_x) \right] > 0, \end{aligned}$$

i.e., if switching from i_r to some i_{t_1} gives a lower energy for any rotamer state including k_{x_1} , and switching from i_r to some i_{t_2} gives a lower energy for any rotamer state including k_{x_2} , and so on for each possible k_x .

To do magic bullet doubles: Define energy terms to treat rotamer pairs like single rotamers:

$$\begin{aligned} & G([i_r, j_u]) \square G_{\text{self}}(i_r) + G_{\text{self}}(j_u) + G_{\text{pair}}(i_r, j_u), \\ & G([i_r, j_u], k_x) \square G_{\text{pair}}(i_r, k_x) + G_{\text{pair}}(j_u, k_x). \end{aligned}$$

For position pair $[i, j]$, find one “magic bullet” rotamer pair $[i_*, j_*]$ with lowest maximum energy; i.e., let $[i_*, j_*]$ equal the $[i_r, j_u]$ which minimizes this quantity:

$$G([i_r, j_u]) + \bigcup_{k, k \square i, k \square j} \left\{ \max_x \left[G([i_r, j_u], k_x) \right] \right\}$$

Then, eliminate each rotamer pair $[i_r, j_u]$ if, for the “magic bullet” alternative rotamer pair $[i_*, j_*]$,

$$\begin{aligned} & G([i_r, j_u]) \square G([i_*, j_*]) \\ & + \bigcup_{k, k \square i, k \square j} \left\{ \min_x \left[G([i_r, j_u], k_x) \square G([i_*, j_*], k_x) \right] \right\} > 0. \end{aligned}$$

Check for when rotamer pairs $[i_r, j_u]$ are eliminated for all u of any one j , because then you can eliminate the entire rotamer i_r .

6. Do A* on sequence energy to find the GMEC (unless, of course, DEE alone reduced the search space down to only the GMEC). The A* algorithm treats the search for the GMEC as the traversal of a tree, where each node on level 1 represents having chosen a state for position 1, each node on level 2 represents having also decided on a state for position 2, and so on (see Fig. 2). A* seeks to reach the GMEC, which is one of the nodes on the bottom row, without having needed to explore most nodes of the tree. The algorithm consists of maintaining a list of nodes currently under consideration (they may be on various rows), and repeat-

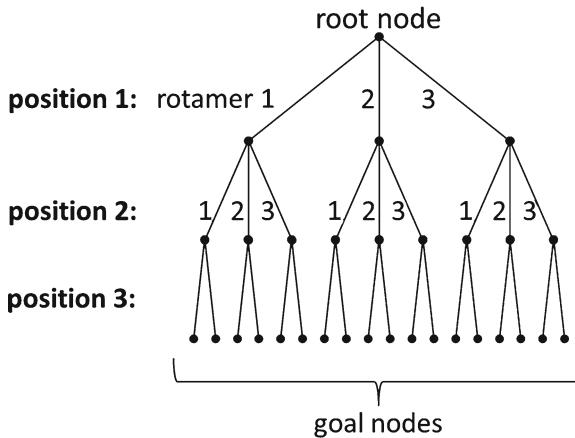


Fig. 2. A* treats the search space for the GMEC as a tree in which going down one level represents deciding on the rotamer state of one mobile position.

edly “expanding” the listed node with minimum estimated energy f^* , replacing it by all of its “daughter” nodes below it. For a node with states placed at p_f of p positions, let $f^* = g^* + h^*$ where g^* , the energy of the states already placed, is just the sum of their self and pair terms, and h^* is a lower bound on the energy for choosing a state at each of the remaining locations,

$$h^* = \min_{j=p_f+1}^{p_f} \left(\min_u \left[\sum_{i=1}^{p_f} G_{\text{pair}}^*(i_r, j_u) \right] + \min_{k=j+1}^p G_{\text{pair}}^*(j_u, k_x) \right)$$

where

$$G_{\text{pair}}^*(i_r, j_u) = \frac{G_{\text{self}}(i_r) + G_{\text{self}}(j_u)}{p-1} + G_{\text{pair}}(i_r, j_u)$$

is defined to split the self terms up among the pair terms (6) (see Note 6). The result will be a list of sequences in order of their sequence energy (see Note 7). You may limit the length of the list, or its span of sequence energies; otherwise, the list will be exhaustive, except possibly for some sequences totally disallowed by the self energy cutoff done above (see Note 8).

7. Repeat DEE/A* to find not only the GMEC but also all conformations with energy within a chosen cutoff G_{cut} of the minimum. Repeat DEE as in step 5 of Subheading 3.2 but with “>0” in the criteria replaced by “ $> G_{\text{cut}}$,” then do A* as in step 6 of Subheading 3.2 but continuing to expand nodes with daughters until their f^* is G_{cut} above the GMEC. At this point, you have a list of sequences with the lowest sequence energy.
8. For each sequence, do DEE/A on fleximer energy to get a list of low-energy fleximer states: For each sequence on the list from the previous step, use the previously stored table of all fleximer energy self and pair terms to write a table of the self and pair terms in the subspace of just that sequence. Do DEE followed by A* on these fleximer energy terms to find a list of up to ten fleximer states for that sequence.

9. For each of the ten lowest-energy fleximer states of each sequence, do DEE/A* on rotamer energy to get one rotamer state: For each fleximer state on the list from the previous step, use the previously stored table of all self and pair terms of the pairwise-additive rotamer energy to write a table of the self and pair terms in the subspace of just that fleximer state. Do DEE followed by A* on these energy terms to find the one minimum-energy rotamer state for that fleximer state. The result will be a list of rotamer states, usually ten per sequence.

3.3. Final Optimization

1. Minimize each structure: Some of the rotamer states on the list from the previous step will have minor or moderate van der Waals clashes which the rotamer library could not avoid but can be relieved by a brief minimization. And some rotamer states may be overpacked, having taken advantage of the fact that we scaled down the van der Waals radii by 0.9. Both of these problems can be addressed by briefly minimizing (for 100 steps, e.g.) the bound state of each rotamer state on the final list (or the folded state, if you are not designing for binding), with only the mobile residues allowed to move, using the pairwise-additive energy function (or preferably, a better energy function) but with full van der Waals radii (see Note 9).
2. Choose a final energy function: The last step will be re-evaluating each design’s minimized structure with a better, non-pairwise-additive “final” energy function. We use only the electrostatic term from FDPB (see Note 10), although the covalent terms and van der Waals term with radii scaled down by 0.9 may also be recalculated (see Note 11). A hydrophobic term of the form $G_{\text{hyd}} = \square \cdot \text{SASA}$ may also be calculated using a total or per-atom SASA calculated by the CHARMM program, for example, and a coefficient of $\square = 5.0 \text{ cal / mol } \text{\AA}^2$ (27). The FDPB electrostatic term is based on a continuum solvent model, with parameters $\square_b = 4$ and $\square_s = 80$ for the dielectric constants inside and outside the macromolecular solvent-accessible surface (the region accessible to the surface of a probe of radius 1.4 Å outside the macromolecule), and positive and negative monovalent ions of radius 2 Å at a concentration of 0.145 M each in the solvent region accessible to the center of a probe of radius 2 Å. The atomic charges have magnitudes given by the PARSE charge set (see ref. 27), but the atomic charge can not generally be precisely at the atomic center when doing FDPB on a grid. Instead, each atomic charge is split up into eight partial charges on the surrounding eight grid points, with the magnitude of each partial charge scaled linearly by its grid point’s proximity to the atomic center in each of the three dimensions. The result of FDPB is the electric potential at each grid point, from which the total electrostatic free energy is

easily calculated. We typically use a grid spacing of about 1 Å (see Note 12), which usually requires $65 \times 65 \times 65$ to $97 \times 97 \times 97$ grid points (not necessarily cubic) to cover a region 20 Å wider in each dimension than the macromolecules. The FDPB calculation is preceded by a similar FDPB run on the same number of grid points but covering a larger region 1.7 times wider in each dimension than the macromolecular complex to determine correct electrostatic potentials at the boundary of the final grid. We repeat each FDPB calculation ten times at a series of translations smaller than the grid spacing, in order to quantify and to average out the error caused by how the grid lines happen to land with respect to the charges.

3. For each rotamer state kept, calculate the final free energies of binding and folding: We calculate these non-pairwise-additive FDPB-based energies *after* our DEE/A* search procedure to get the final binding and/or folding free energies by which each searched structure will be judged. For all but the FDPB term, the free energy of folding term is simply $\square G^{\text{fold}} = G^{\text{folded}} - G^{\text{unfolded}}$, and the free energy of binding (if applicable) is simply $\square G^{\text{bind}} = G^{\text{bound}} - G^{\text{folded}}$. For the FDPB terms, we must take care to cancel out the “grid energy”. The output of a FDPB calculation is the sum $G_{\text{PB}} = G_{\text{ES}} + G_{\text{grid}}$ of the desired electrostatic energy and the grid energy, a non-physical artifact of how the charges and the electric potential are represented only as values at grid points rather than as continuous functions over the space. The grid energy is a function of the grid, the charges and their locations, and the dielectric constant at the charge locations. Only for certain energy differences does the grid energy cancel out: $\square G_{\text{ES}}^{\text{bind}} = G_{\text{PB}}^{\text{bound}} - G_{\text{PB}}^{\text{folded}}$ is valid if, for each binding partner, we take care that its unbound “folded” state lies on the same grid in the same way that it did in the bound state, though the other binding partner is missing (see Fig. 3). Any energy of solvation $\square G_{\text{ES}}^{\text{solvation}} = G_{\text{PB}}^{\text{solvated}} - G_{\text{PB}}^{\text{desolvated}}$ is valid if we take care that the desolvated version of the state lies on the grid in the same way that it did in the solvated state, but with the dielectric constant in the solvent (where there are no explicit charges) set to \square_p . The electrostatic free energy of folding can be constructed by following this thermodynamic cycle (see Fig. 4): desolvate the unfolded model state, then simply use Coulomb’s law (see Note 13) for the electrostatic energy of rearranging all the charges to form the folded “unbound” state, then solvate that folded state.

$$\begin{aligned} \square G_{\text{ES}}^{\text{fold}} &= \square \square G_{\text{ES}}^{\text{solvation}}(\text{unfolded}) - \square G_{\text{coulombic}}^{\text{desolvated}}(\text{unfolded}) \\ &\quad + G_{\text{coulombic}}^{\text{desolvated}}(\text{folded}) + \square G_{\text{ES}}^{\text{solvation}}(\text{folded}) + C, \end{aligned}$$

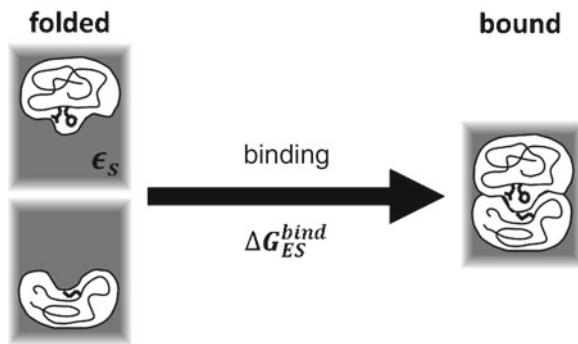


Fig. 3. For the electrostatic energy of binding $\square G_{ES}^{bind}$, the artifactual FDPB “grid energy” cancels out if each unbound binding partner is calculated on the same grid as the bound complex.

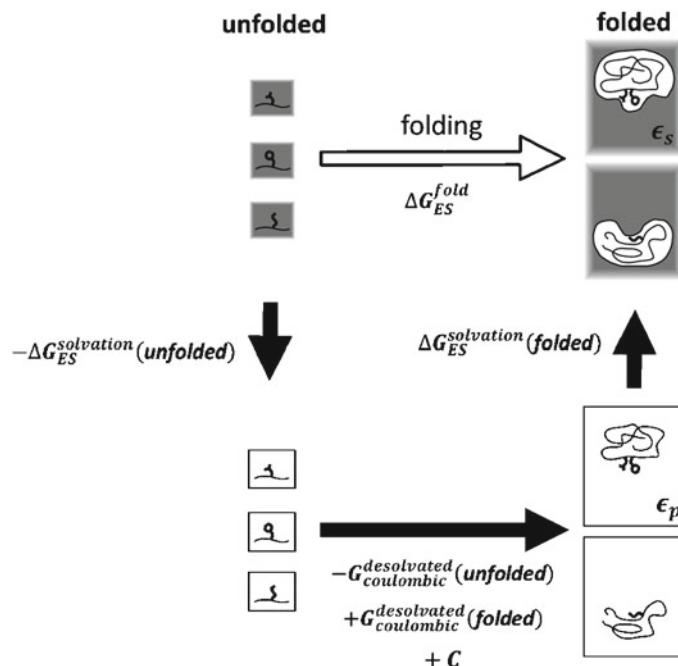


Fig. 4. The electrostatic energy of folding $\square G_{ES}^{fold}$ is calculated by first desolvating the unfolded state, then rearranging atoms to form the folded state, and finally solvating the folded state. The artifactual FDPB “grid energy” cancels out for each energy of solvation. The fixed atoms are not shown in the unfolded state here; they need no unfolded state model because their contribution to $\square G_{ES}^{fold}$ is constant for all conformations in a design run.

where C is a constant for all conformations from this design run, which depends only on the fixed atoms, and therefore on the original structure and the set of mobile position numbers. In practice, we can ignore C because we only care about $\square \square G_{ES}^{fold}$, comparisons between conformations (see Note 14).

4. Choose a representative rotamer state with the wild-type sequence: If your objective function was the free energy of

folding, then simply choose the one with minimum final free energy of folding. If your objective function was the free energy of binding, choose the one with the minimum sum of final (folding + binding) free energy. This sum is the “complex folding” free energy; it is the stability of the bound complex relative to its unbound, unfolded state.

5. Reject all structures with a free energy of folding more than 1 kcal/mol worse than the wild-type representative (see Note 15).
6. Keep only one representative rotamer state for each sequence: Choose the one with the minimum final free energy of folding (or binding), if your objective function was folding (or binding). The final result of the procedure is a list of sequences, each with the final free energy of folding, final free energy of binding, and the coordinates of its representative structure. Sort the list by final free energy of folding (or binding), then start viewing the structures and considering which of the top-ranked sequences to synthesize (see Note 16).

4. Notes

1. The A* algorithm is simple and published. We have found, however, that clever programming and compiling can speed up the A* algorithm by a factor of 600 and save memory by a factor of 300, making some larger search spaces feasible. One big speed up is to expand the positions in order of their energy gap, i.e., the difference between the global minimum energy and the minimum energy of any rotamer state with a different rotamer at that position.
2. Your decision on which residues should be evolving, molten, or fixed will determine the size of your search space, and therefore the computational tractability of the design run. Although the increasing speed of computer processors now allows for design runs with several or many evolving residues, almost all of the favorable mutations found in practice are fairly independent. So each of these favorable single mutations can be found quickly by running a single-mutant design run for each residue. Some very rough estimates of the computational expense of an affinity design run on a 3.2 GHz processor, with 1 evolving position and 15 molten positions on an antibody Fab: 30 min to calculate self terms and 5.5×10^6 pair terms; 14 min to do three-level DEE/A*, mostly spent doing fleximer-level DEE/A* for each of the 21 sequences; 46 min to do FDPB on the model compounds (15 s per binding/folding calculation, for 10 grid offsets of 21 amino acid types), 6,000 min to do FDPB (3 min per binding/folding calculation on the large

antibody Fab system, times 10 grid offsets, times 1 rotamer state per fleximer state, times 10 fleximer states per sequence, times 21 sequences). With multiple evolving positions, DEE/A* will take significantly longer, but the longest part will still be the FDPB.

3. For the pairwise-additive energy function, we omit the van der Waals and electrostatic terms between atom pairs which are bonded (a “1–2” pair), or connected via two bonds (“1–3”); and between atom pairs connected via three bonds (a “1–4” pair), we scale down the electrostatic term by 0.4, as intended for the CHARMM PARAM19 parameter set. In CHARMM, these are parameters “ELEC NBXMOD 5 E14FAC 0.4”
4. In ref. 21, a temperature T is defined, each rotamer i_r in the fleximer is considered occupied according to the Boltzmann factor $e^{\square G_{\text{self}}(i_r)/k_B T}$, and the fleximer energy terms are Boltzmann-weighted averages. The equations shown here are simply the result of setting the temperature to zero, i.e., choosing the minimum energy rotamer of each fleximer.
5. For a single-mutant design run, DEE/A* is overkill when applied to ranking only approximately 21 sequences. You could instead replace steps 4–7 of Subheading 3.2 with simply iterating over those sequences.
6. A* is quite fast when finding only the minimum energy, but it can slow down considerably when finding all states within an energy difference $\square G_{\text{cut}} > 0$ of the minimum, and this is needed to get several design candidates. If $\square G_{\text{cut}}$ is chosen too high, A* will waste time or run out of memory. In order to get multiple design candidates, our goal when running A* is to get a list of lowest-energy states (all, or thousands, of the lowest-energy sequences; and later, for conformational diversity, the ten lowest-energy fleximer states for a sequence). Therefore, we must run A* with $\square G_{\text{cut}} > 0$, but it is not known a priori what minimum $\square G_{\text{cut}}$ value will yield the desired number of low-energy states. Running A* with a $\square G_{\text{cut}}$ too high will likely take a long time and then run out of memory, whereas using a $\square G_{\text{cut}}$ too low will return fewer states than desired, or only the GMEC. So, regardless of how efficiently the A* algorithm is programmed, we have found it necessary to make heuristic guesses each time A* is called for what $\square G_{\text{cut}}$ to use and what maximum memory to allow A* to use before giving up. The very first guess for $\square G_{\text{cut}}$ is found by doing a preliminary DEE/A* run with $\square G_{\text{cut}} = 0$ used for DEE but $\square G_{\text{cut}} = 25 \text{ kcal/mol}$ used for A*; rarely this DEE with $\square G_{\text{cut}} = 0$ narrows the search space down to only the GMEC, but usually it does not, so then the A* with $\square G_{\text{cut}} = 25 \text{ kcal/mol}$ finds a few states other than the GMEC in this (incomplete) search space. This preliminary DEE/A* run is done only to

find the energy gap between the GMEC and the next few states. Setting our first guess for $\square G_{\text{cut}}$ to this energy gap is likely to give some but not too many low-energy states. Then DEE/A* is called repeatedly with its $\square G_{\text{cut}}$ (the same for DEE and A* now) as well as its maximum allowed memory both carefully controlled to ensure that the procedure does not get “stuck” or waste time for searching a too-large search space: After each of these DEE/A* runs, if fewer states are found than desired, $\square G_{\text{cut}}$ is raised by a factor “states desired/states found”. If the maximum allowed memory is exceeded, it is raised by 40%, but $\square G_{\text{cut}}$ is lowered halfway back to the most recent $\square G_{\text{cut}}$ for which states were found, or is multiplied by 1/3 or less. These DEE/A* runs are repeated until one finds the desired number of states or finds all states within $\square G_{\text{cut}}^{\max} = 25 \text{ kcal / mol}$ of the minimum energy.

7. If you have more than two or three evolving positions, and you want to find a potentially very long list of states, then A* could run out of memory. In such a case, depth-first A* is a simple alternative to A* that finds every state within $\square G_{\text{cut}}$ of the GMEC, not in order, but using little memory. To get a complete sorted list of the lowest-energy states, you must allow depth-first A* to run to completion, and then sort its output by energy. The downside is that, if you used too high an $\square G_{\text{cut}}$, then depth-first A* may run longer and output more states than are feasible. No stage of our three-level DEE/A* procedure (sequences, then fleximers, then rotamers) needs to find vast numbers of states, so in this procedure, A* always has the advantage of running more quickly than depth-first A*.
8. If you have more than two or three evolving positions, it will not be feasible to continue the procedure on the entire list of sequences, and if you do continue the procedure starting with the beginning of the list, you may waste time on many of the over-charged sequences that the pairwise-additive energy function favors. So, first break the sequence list up by the number of charged evolving residues, then interleave these lists, and continue the procedure by stepping up this interleaved list of sequences as far as is practical.
9. This minimization procedure has some disadvantages: (a) If it uses a pairwise-additive energy function, then it could make the final energy worse. (b) By minimizing the bound state energy, it could make the binding or folding energy worse. (c) It may cause several or all of the ten conformations for a sequence to minimize to nearly identical conformations, reducing the diversity of the structures passed to the final energy function for the final evaluation. Passing the ten unminimized structures as well as one or all of the ten minimized structures on to the final energy function would address all of these disadvantages.

10. We have found the nonlinear PB equation to give results fairly similar to the linearized PB equation for our purposes, but to take 5× as long.
11. We have limited data for a set of designs (with or without the final minimization of step 1 of Subheading 3.3) for which the FDPB-based electrostatic energy term $\square G_{\text{ES}}^{\text{fold}}$ alone correlates well with experimental stability, while $\square G^{\text{fold}}$ correlates poorly, presumably because the van der Waals and covalent terms are overly sensitive to imperfectly modeled structural details.
12. Determining a sufficient grid resolution: We have found that the $\square G_{\text{ES}}^{\text{bind}}$ and/or $\square G_{\text{ES}}^{\text{fold}}$ between alternative designs converges with respect to increasing grid resolution at a grid spacing of about 1 Å.
13. If you use CHARMM to calculate these Coulombic energies, be sure to include all atom pairs, by using “ELEC NBXMOD 0 E14FAC 1” as opposed to the “ELEC NBXMOD 5 E14FAC 0 .4” setting described above for the van der Waals and distance-dependent Coulombic terms.
14. In ref. 8, we used a different but equivalent expression for $\square G_{\text{ES}}^{\text{fold}}$ which replaced one large FDPB calculation, on the desolvated folded state, with many quicker FDPB calculations, each on one desolvated rotamer for a mobile position:

$$\begin{aligned} \square G_{\text{ES}}^{\text{fold}} = & \square G_{\text{PB}}^{\text{solvation}}(\text{unfolded}) \square G_{\text{coulombic}}^{\text{desolvated}}(\text{unfolded}) \\ & + G_{\text{coulombic}}^{\text{desolvated}}(\text{separated desolvated rotamers}) \\ & \square G_{\text{PB}}^{\text{desolvated}}(\text{separated desolvated rotamers}) \\ & + G_{\text{PB}}(\text{folded}) + C_2, \end{aligned}$$

where C_2 is an ignored function of the grid, the fixed atoms' charges and positions, and the set of mobile positions.

15. We screen on the free energy of folding to avoid unstable designs, of course; but even if you could afford to pay some stability to get better affinity, designs predicted to sacrifice folding energy for better binding energy may not have better binding in reality. Our method assumes rigid binding, but in reality the actual unbound structure of such a design may be different and lower in free energy than our design. It will have to pay a free energy cost to reconfigure from this unmodeled unbound structure to the bound structure we designed, thus worsening its actual, *non-rigid* free energy of binding. This is why, to design for tighter binding, we use the free energy of binding as the objective function for DEE/A*, but as a final screen we keep only sequence designs with FDPB free energy of folding no more than 1 kcal/mol worse than the wild-type sequence. Another perspective on this is that we choose to only seek designs that will bind well *rigidly*.

16. It is permissible but generally unfair to compare free energy of folding differences from different design runs D_1, D_2, \dots which have the same initial structure and the same reference structure R but different sets of mobile positions:

$$\begin{aligned} & \square\square G_{\text{ES}}^{\text{fold}}(\text{structure } S \text{ of design run } D_i) \\ &= \square G_{\text{ES}}^{\text{fold}}(\text{structure } S \text{ of design run } D_i) \\ &\quad \square\square G_{\text{ES}}^{\text{fold}}(\text{reference structure } R \text{ of design run } D_i). \end{aligned}$$

Such a comparison is permissible in that the neglected constant terms cancel out, but unfair in practice because a position which is mobile in one design run may find a lower-energy conformation not available to the other design run where the position is held fixed. This can cause a spurious energy difference between the two design runs, even for the same sequence.

References

1. Dunbrack RL Jr, Karplus M (1993) Backbone-dependent rotamer library for proteins: application to side-chain prediction. *J Mol Biol* 230:543–574
2. Desmet J et al (1992) The dead-end elimination theorem and its use in protein side-chain positioning. *Nature* 356:539–542
3. Goldstein RF (1994) Efficient rotamer elimination applied to protein side-chains and related spin glasses. *Biophys J* 66:1335–1340
4. Winston PH (1992) Artificial intelligence. Addison-Wesley, Reading, Massachusetts
5. Leach AR, Lemon AP (1998) Exploring the conformational space of protein side chains using dead-end elimination and the A* algorithm. *Protein Struct Funct Genet* 33:227–239
6. Gordon DB, Mayo SL (1999) Branch-and-terminate: a combinatorial optimization algorithm for protein design. *Structure* 7:1089–1098
7. Caravella JA (2002) Electrostatics and packing in biomolecules: accounting for conformational change in protein folding and binding. Thesis, Massachusetts Institute of Technology, p. 112. <http://hdl.handle.net/1721.1/16823>
8. Hanf KJM (2002) Protein design with hierarchical treatment of solvation and electrostatics. Thesis, Massachusetts Institute of Technology, p. 143–165. <http://hdl.handle.net/1721.1/29223>
9. Sharp KA, Honig BH (1990) Electrostatic interactions in macromolecules: theory and applications. *Annu Rev Biophys Chem* 19:301–332
10. Brooks BR et al (1983) CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J Comput Chem* 4: 187–217
11. Snow C et al, SHARPEN website. <http://sharpen.engr.colostate.edu/>
12. Loksha IV et al (2008) SHARPEN—systematic hierarchical algorithms for rotamers and proteins on an extended network. *J Comput Chem* 30:999–1005
13. Handel TM et al, EGAD website. <http://egad.berkeley.edu>
14. Pokala N, Handel TM (2005) Energy functions for protein design: adjustment with protein-protein complex affinities, models for the unfolded state, and negative design of solubility & specificity. *J Mol Biol* 347(1): 203–227
15. Donald BR et al, OSPREY website. <http://www.cs.duke.edu/donaldlab/osprey.php>
16. Chen C et al (2009) Computational structure-based redesign of enzyme activity. *Proc Natl Acad Sci USA* 106(10):3764–3769
17. Baker NA et al, APBS website. <http://www.poissonboltzmann.org>
18. Baker NA et al (2001) Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc Natl Acad Sci USA* 98: 10037–10041
19. Lee B, Richards FM (1971) The interpretation of protein structures: estimation of static accessibility. *J Mol Biol* 55:379–400
20. Brünger AT, Karplus M (1988) Polar hydrogen positions in proteins: empirical energy placement and neutron diffraction comparison. *Protein Struct Funct Genet* 4:148–156

21. Mendes J et al (1999) Improved modeling of side-chains in proteins with rotamer-based methods: a flexible rotamer model. *Protein Struct Funct Genet* 37:530–543
22. Dahiyat BI, Mayo SL (1997) Probing the role of packing specificity in protein design. *Proc Natl Acad Sci USA* 94:10172–10177
23. Pierce NA et al (2000) Conformational splitting: a more powerful criterion for dead-end elimination. *J Comput Chem* 21(11): 999–1009
24. Gordon DB et al (2002) Exact rotamer optimization for protein design. *J Comput Chem* 24:232–243
25. Lasters I, 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
26. Gordon DB, Mayo SL (1998) Radical performance enhancements for combinatorial optimization algorithms based on the dead-end elimination theorem. *J Comput Chem* 19(13):1505–1514
27. Sitkoff D, Sharp KA, Honig B (1994) Accurate calculation of hydration free energies using macroscopic solvent methods. *J Phys Chem* 98:1978–1988

Chapter 9

Design and Generation of DVD-IgTM Molecules for Dual-Specific Targeting

Enrico DiGiammarino, Tariq Ghayur, and Junjian Liu

Abstract

The dual variable domain immunoglobulin (DVD-IgTM) protein is a new type of dual-specific IgG. As a novel therapeutic class, the great potential of the DVD-Ig protein is to simultaneously target two mediators of disease by a single pharmaceutical entity. The molecule contains an Fc region and constant regions in a configuration similar to a conventional IgG; however, the DVD-Ig protein is unique in that each arm of the molecule contains two variable domains (VDs). The VDs within an arm are linked in tandem and can possess different binding specificities. Here, we discuss critical design features of the DVD-Ig protein and describe a methodology for cloning, expressing, and purifying the molecules.

Key words: Dual variable domain immunoglobulin, Dual-specific targeting, DVD-Ig, Immunotherapy, Linker, Monoclonal antibody, Variable domain, Antibody engineering

1. Introduction

Immunotherapy has clearly become an important and rapidly growing therapeutic category over the last two decades, it is a category dominated by monoclonal antibodies (mAbs) that have proven effective in the treatment of cancer, autoimmune and inflammatory diseases. There are currently over two dozen therapeutic mAbs on the market and many more in clinical trials and development (1). In addition to mAbs, antibody-like molecules (for example, Fabs, receptor-Fc fusions and a host of other formats—both mono-specific and multi-specific) also fall into the immunotherapy category; several are on the market and many are being researched or are in development (2). The various formats offer specific advantages and disadvantages related to effector functions, serum stability, immunogenicity, tissue penetration, and manufacturability (among others).

MAbs and other antibody-like formats that retain the Fc portion of the immunoglobulin (Ig) G have the advantage of being able to engage the immune system to initiate cytotoxic effector functions at the site of the target. They can also benefit from endosomal recycling *via* the FcRn receptor system which can prolong the serum half-life of these molecules. However, mono-therapy with mAbs can sometimes only benefit a subset of patients because of multiple and possibly redundant disease mechanisms. As such, the increased disease fighting potential of combination therapy with monoclonal antibodies, to simultaneously target multiple mediators of disease, has been recognized as a route to greater therapeutic efficacy (3). Multi-specific immunotherapy approaches seek to achieve the advantage of combination therapy with a single therapeutic entity while reducing the cost and effort required for pre-clinical and clinical development (and manufacturing) for two independent pharmaceuticals. The dual variable domain immunoglobulin (DVD-IgTM) protein is a novel dual-specific therapeutic modality.

First described by Wu et al. in 2007, the DVD-Ig protein is a dual-specific, tetravalent IgG-like molecule (4). The four-chain molecule is composed of two heavy chains and two light chains as depicted in Fig. 1. The molecule has a functional Fc region and constant domains in a similar configuration as a conventional IgG;

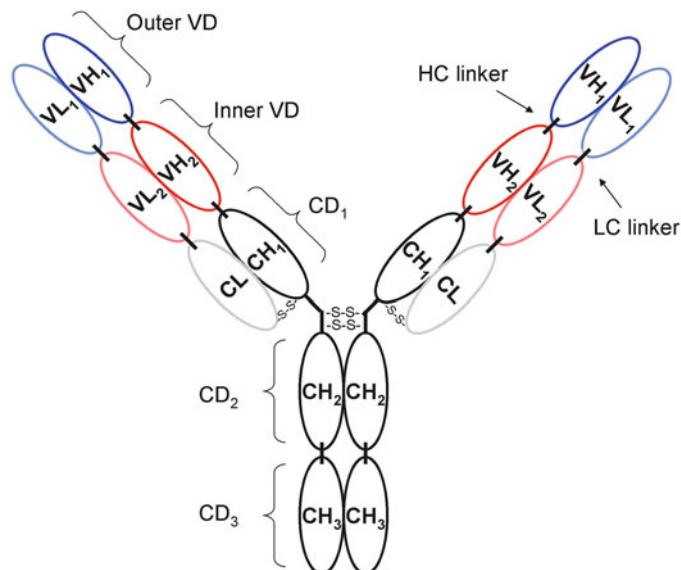


Fig. 1. DVD-Ig protein schematic. The figure depicts the general structure of a DVD-Ig protein where CH and VH refer to constant and variable heavy chain regions, respectively, and CL and VL refer to constant and variable light chain regions, respectively, and disulfide linkages are depicted. CD and VD refer to constant domain and variable domain, respectively. The variable regions form tandem inner and outer variable domains, connected by both light and heavy chain linkers, on each arm of the DVD.

however, each arm possesses two variable domains (VDs). The VDs are connected in tandem by separate heavy and light chain linkers; together they form an inner VD and an outer VD capable of different ligand-binding specificities. The inner and outer VDs of a DVD-Ig protein can essentially be engineered by linking variable region sequences from any pair of monoclonal antibodies. The DVD-Ig protein can be engineered from fully human sequences, thus reducing potential immunogenicity, and can possess effector functions and serum stability similar to a conventional IgG. In addition to unique structural features, the DVD-Ig protein can also be efficiently produced by standard mammalian expression systems and they exhibit physiochemical properties comparable to conventional IgGs, despite their increased size and complexity.

Anti-IL12/IL18 and anti-IL1 α /IL1 β DVD-Ig proteins have been described as potential therapeutics for autoimmune indications and inflammatory arthritis, respectively (4–6). In each case, the DVD-Ig protein was shown to be capable of simultaneous high affinity binding to both ligands. It was also clear from these studies that to achieve desirable binding properties required a carefully designed DVD-Ig protein. For example, in 2009, Wu et al. (5) described instances where a given VD shows reduced affinity as an inner VD, compared to when engineered as an outer VD or to the reference antibody. This effect was attributed, at least partially, to steric hindrance due to the proximity of the outer VD to the ligand-binding site of the inner VD; longer linkers were shown to reduce the affinity loss. Here, we describe critical design considerations of the DVD-Ig protein, in general, and a methodology for cloning, expressing, and purifying these molecules.

The key design elements for the DVD-Ig protein are (1) selection of the VD pairing from parental mAbs, (2) the inner/outer orientation of the selected VDs, and (3) the amino acid sequences used to link the VDs. In practice, the optimal parental VDs, VD orientation, and linker sequences for a given DVD-Ig protein are difficult to predict and may need to be determined empirically by constructing a series of exploratory molecules. The series should be screened for desirable physiochemical properties, *in vivo* pharmacokinetics, and efficacy (see Note 11). Details and special consideration regarding each of the design elements are discussed in more detail below.

1.1. Variable Domain Selection

Variable domain sequences required to construct the DVD-Ig protein are derived from the so-called parental mAbs. The intrinsic physiochemical and performance properties of the parental mAb should be considered carefully when choosing VDs—properties such as expression yield, solubility and stability, binding kinetics, PK, efficacy, etc. These properties are thought to be derived largely from the sequence and structure of the VDs themselves. It is advisable to avoid VDs from poorly behaved mAbs when constructing

DVD-Ig proteins. It is important to note however that the properties of the parental mAb may not be predictive of the final DVD-Ig protein performance with respect to these parameters. It is therefore good practice to select multiple parental mAbs for a given target and empirically determine which combinations of VDs lead to DVD-Ig proteins with desirable properties.

1.2. Orientation of Variable Domains

The orientation of the two VDs in a DVD-Ig protein (i.e., whether a given VD should be engineered as an inner VD or outer VD) must also be optimized. The optimal domain orientation may rely on factors such as the molecular sizes of the two targets and the general nature of the two targets (for example, whether they are soluble molecules in circulation or found on the cell surface). A given VD may show reduced affinity when engineered as the inner VD (compared to either the parental mAb or when engineered as the outer VD) (4, 5). This effect may be attributed to potential steric hindrance arising from the proximity of the outer variable domain to the ligand-binding site of the inner VD—an effect which may, in turn, be impacted by the size, structure, and location of the target molecule. It may logically follow that VDs which bind to an antigen of larger molecular size or more restricted accessibility be engineered as outer VD. Linker design also plays a role in the binding performance of the inner variable domain, as discussed below.

1.3. Linker Design Approaches

The two VDs within an arm of the DVD-Ig protein are connected by peptide sequences called “linkers” which span from the C-termini of the outer VD heavy and light chain to the N-termini of the inner VD heavy and light chain, respectively. In a conventional IgG, the sequences between the constant domain 1 and the variable domain are known to impart flexibility between domains (7) and this flexibility is thought to be critical to both the physical properties of the molecule as well as antigen binding. Similarly, linkers within the DVD-Ig protein should be designed to impart structural flexibility while also reducing potential immunogenicity. For any given DVD-Ig protein, linker design may affect expression yields, physiochemical properties, activity, pharmacokinetics, and efficacy in ways that may be difficult to predict and therefore choice of linker sequence may need to be empirically determined. Three categories of example linkers, and their corresponding sequences, that have been found to work well in DVD-Ig proteins are listed in Table 1. Elbow linkers are naturally occurring sequences derived from the linkage between the constant domain 1 and the variable domain of human IgG1; long and short versions of these linkers, as well as kappa and lambda-specific light chain versions are described. Alternative naturally occurring linker sequences are the hinge-based linkers which are derived from sequences found between the CH₁ and CH₂ of human IgG1. Glycine/serine linkers are another notable category; these sequences are thought to impart flexibility, pose low immunogenicity risks and have been

Table 1
Example linkers for DVD-Ig proteins

Linker type	Heavy chain		Light chain	
	Name	Sequence	Name	Sequence
Elbow	HC- γ S	ASTKG	LC- κ S	TVAAP
	HC- γ L	ASTKGPSVFPLAP	LC- κ L	TVAAPSVFIFPP
			LC- λ S	QPKAAP
			LS- λ L	QPKAAAPSVTLFPP
Hinge	HC-hS	PNLLGGP	LC-hS	PAPNLLGGP
	HC-hL	PAPNLLGGP	LC-hL	PTISPAPNLLGGP
Glycine/serine	HC-GSS	GGGGSG	LC-GSS	GGSGG
	HC-GSM	GGGGSGGGGS	LC-GSM	GGSGGGGSG
	HC-GSL	GGGGSGGGGSGGGG	LC-GSL	GGSGGGGSGGGGS

widely used in the construction of antibody-like molecules (8). Linker sequences may not necessarily need to be paired according to Table 1 and unique combinations may be required to achieve desired performance criteria. In some specialized cases, it may also be desirable to introduce specific protease cleavage sites within a given linker.

2. Materials

2.1. Reagents

- High fidelity PCR kit, Invitrogen Inc, Carlsbad, CA, USA.
- QIAquick Gel Extraction Kit, Qiagen, German.
- pHybE vectors, Abbott Laboratories, Abbott Park, IL, USA.
- Restriction endonuclease AfeI, BsiWI, NruI and SalI, New England Biolabs, MA, USA.
- LB agar, Sigma-Aldrich, USA.
- Pluronic F-68, Invitrogen Inc, Carlsbad, CA, USA.
- G418, Invitrogen Inc, Carlsbad, CA, USA.
- Freestyle Expression Medium, Invitrogen Inc, Carlsbad, CA, USA.
- Polyethylenimine (PEI), Polysciences, warrington, WI, USA.
- Protein A binding, washing, and elution buffers, Thermo Scientific, Rockford, IL, USA.
- Protein A resin, GE Healthcare, Piscataway, NJ, USA.
- Tryptone N1, Organotechnie, France.
- 1.0 M Tris-hydrochloride, Hampton Research, Aliso Viejo, CA.
- Coomassie Plus Protein Assay Reagent, Thermo Scientific, Rockford, IL, USA.

2.2. *E. coli* Strain and Cell Line

MAX Efficiency® DH5 α ™ Competent Cells kit, Invitrogen Inc, Carlsbad, CA, USA.

HEK 293-6E cells, American Type Culture Collection, Manassas, VA, USA.

3. Methods

Generation of DVD-Ig™ molecules begins with molecular cloning of separate pHybE DVD-Ig VH and pHybE DVD-Ig VL vectors (one vector for the heavy chain and one vector for the light chain, respectively). A schematic describing basic features of DVD-Ig vector construction is illustrated in Fig. 2. Generally, there are three steps involved in construction of a DVD-Ig vector: (1) primer design, (2) overlapping PCR to generate VH1-linker-VH2 and VL1-linker-VL2, and (3) homologous recombination to incorporate

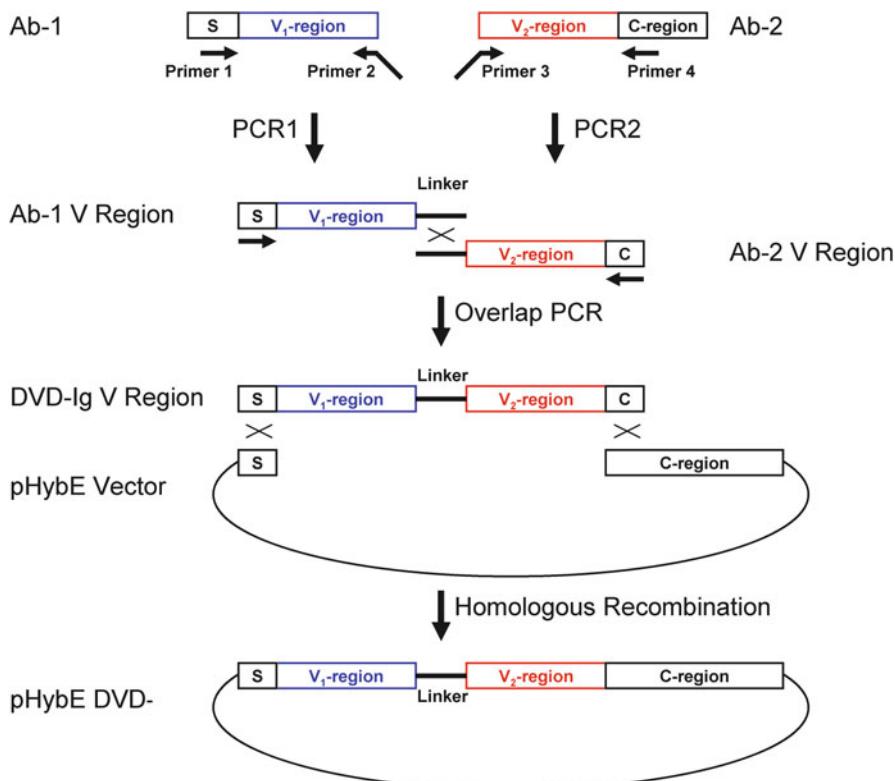


Fig. 2. Summary of DVD-Ig molecular construction. The variable domains of two parental mAbs (Ab-1 and Ab-2) are first amplified by PCR (PCR1 and PCR2) to introduce the linker sequence (S) and a portion of the constant region (C). The two variable regions (Ab-1 V Region and Ab-2 V Region) are combined, in frame, by a second round of PCR (Overlap PCR) to generate a DVD-Ig VH or VL (DVD-Ig V Region). For transient expression, the DVD-Ig VL and VH were cloned into separate pHybE expression vectors by homologous recombination.

PCR generated variable domain fragments (VH and VL) of DVD-Ig into respective mammalian expression vectors containing HC and LC constant regions. Vector construction is followed by transient expression and DVD-Ig protein purification.

3.1. Primer Design

- Primer 1: 100% anneal to signal peptide could be used for similar DVD-Ig construction.
- Primer 2: Include 20 bp complementary region of V₁, linker region, and possibly some of V₂ sequence (see Note 1).
- Primer 3: Include 20 bp complementary region of V₂, linker region, and possibly some of V₁ sequence (see Note 1).
- Primer 4: 100% anneal to constant region could be used for similar DVD-Ig construction.

3.2. Overlapping PCR Amplification of two V Domains

DVD-Ig molecule construction includes two rounds of PCR followed by homologous recombination as depicted in Fig. 2. In the first round, two separated PCR are performed: PCR1 amplifies partial signal peptide sequence, the N-termini V-region, and linker sequence using Primer 1 and Primer 2, while PCR2 amplifies linker sequence, C-termini V-region and partial constant region sequence using Primer 3 and Primer 4. Since the two PCR products overlap at linker region, they can be combined in the second round of overlapping PCR. The resulting combined PCR product includes partial signal peptide sequence at the 5' end, followed by the two V-regions joined by the linker sequence, then partial constant region sequence at the 3' end. This PCR product is ready for homologous recombination in next step (see Note 1). A typical PCR reaction and program are listed below. The annealing temperature may need to be optimized depending on the template/primers.

PCR reaction:

10× High Fidelity PCR buffer	5 µl
10 mM dNTP mixture	1 µl
50 mM MgSO ₄	2 µl
Forward primer (10 µM)	1 µl
Reverse primer (10 µM)	1 µl
DNA template	20 ng (first round PCR)

Or

Gel purified PCRI and PCR2 20 ng/each (second round overlapping PCR)	
Platinum Taq High Fidelity	0.2 µl
H ₂ O	to 50 µl

PCR program:

Initial denaturation: 94°C for 2 min.
 25 cycles of: Denature at 94°C for 30 s.
 Anneal at 55°C for 30 s.
 Extend at 68°C for 1 min/kb of PCR product.
 Incubation at 68°C for 5 min.
 Store at 4°C.

After PCR reaction, purify PCR products using QIAquick Gel Extraction Kit (Qiagen, German) according to the protocol supplied by manufacturer (see Note 2).

3.3. Homologous Recombination for Construction of Mammalian Cell Expression Vectors

The mammalian cell expression vectors used for DVD-Ig expression are pHybE huIgG1 and pHybE huC κ or pHybE huC λ as described previously (9) (see Note 3).

Table 2 addresses restriction endonucleases used for linearization of each vector. Perform enzyme digestion according to the protocol supplied by manufacturer.

The PCR products from Subheading 3.2 can be introduced into linearized pHybE vectors using 5' and 3' overlapping sequences for homologous recombination. The heavy and light chain plasmids are constructed separately as described below:

1. Aliquot 50 μ l of chilled DH5 α competent cells.
2. Add 30–50 ng linearized vector and 100–150 ng of PCR product into DH5 α .
3. Incubate on ice for 30 min (see Note 4).
4. Heat shock at 42°C for 1 min.
5. Cool down on ice for 2 min.
6. Add 80ul of SOC (included with MAX Efficiency® DH5 α ™ Competent Cells kit, see materials) to each transformation tube.
7. Incubate at 37°C for 60 min (see Note 5).
8. Plate all on LB agar plate with 50 μ g/ml ampicillin.

Table 2
Restriction endonucleases for linearization of pHybE vectors

	5' End	3' End
pHybE huIgG1	NruI	SalI
pHybE huC κ	NruI	BsiWI
pHybE huC λ	NruI	AfeI

The expression vectors with correct inserts can be identified with following steps:

1. Run eight colony PCRs for each construct using primer 1 and primer 2 and save the cultures, respectively (see Note 6).
2. Send colony PCR product for sequencing confirmation.
3. Make Maxi preps for the positive clones and sequence confirm again.

3.4. Transient Expression of DVD-Ig Molecules in HEK293 Cells

Constructed heavy chain expression vector and light chain expression vector are co-transfected in 293-6E cells as the cells containing stably transfected functional forms of EBNA1. This cell line allows episomal persistence of pHybE vectors containing the EBV origin of replication oriP for high yield expression (10).

Transfection protocol:

1. On the day of transfection, seed 293-6E cells to a density of 1.2×10^6 cells/ml into 80% of desired transfection volume in Freestyle Expression Medium supplemented with 0.1% (w/v) Pluronic F-68. The cells are cultured for 3–5 h before transfection.
2. Mix heavy and light chain DNA at 2–3 ratio (see Note 7) in 5% final transfection volume of Freestyle Expression Medium (Invitrogen), and then add PEI diluted in 5% final transfection volume of Freestyle Expression Medium (Invitrogen), followed by incubation at room temperature for 15–20 min at room temperature.
3. Add the mixture to cell suspension, with a final DNA concentration at 0.5 µg DNA/ml and final PEI concentration at 1 µg/ml.
4. Feed transfected 293-6E cells with prewarmed Freestyle Expression Medium supplemented with 5% Tryptone N1 on next day of transfection and monitor cell viability daily.
5. Harvest transfection medium when cell viability drops to 50–60%, usually on day 6 or day 7 after transfection (see Note 8).

3.5. Purification of DVD-Ig Molecules Using Protein A Chromatography

Since DVD-Ig molecules have a regular Fc region, purification of DVD-Ig proteins is essentially the same as purification of monoclonal antibodies:

1. Prepare protein A column as instructed (GE Healthcare).
2. Gently apply cell culture medium (diluted 1:1 with binding buffer) to the column by layering onto the top of the resin. Be careful not to disturb the bed surface (see Note 9).
3. Wash column with 10 volumes of the 1× wash/binding buffer, or until the absorbance of eluate at 280 nm approaches the background level (see Note 9).

4. Add 100 μ l 1 M Tris-HCl buffer (pH 8.0) to each collection tube so the eluent could be immediately neutralized.
5. To elute the antibody, gently add 1 \times elution buffer to the top of the resin, collecting the eluate in a prepared collection tube (0.9 ml/tube).
6. Repeat until the entire volume has been collected, up to eight tubes.
7. Identify positive fractions by adding 10–20 μ l of eluted fractions to 300 μ l of Coomassie Plus Protein Assay Reagent (Thermo Scientific) (in a microtiter plate). Positive fractions show a blue reaction.
8. Combine positive fractions and dialyze against 1,000-fold of sample volume of PBS overnight.
9. Determine DVD-Ig protein concentration by measuring OD₂₈₀.
10. Check purity of the sample by SDS-PAGE. Single band of >200 kDa should be observed for DVD-Ig under nonreducing condition, and two bands of 37.5 kDa (LC) and 62.5 kDa (HC) should be seen under reducing condition. Figure 3 shows an example SDS-PAGE gel (see Note 10).
11. Store purified protein at –20°C (for comments on the characterization of DVD-Ig molecules, see Note 11).

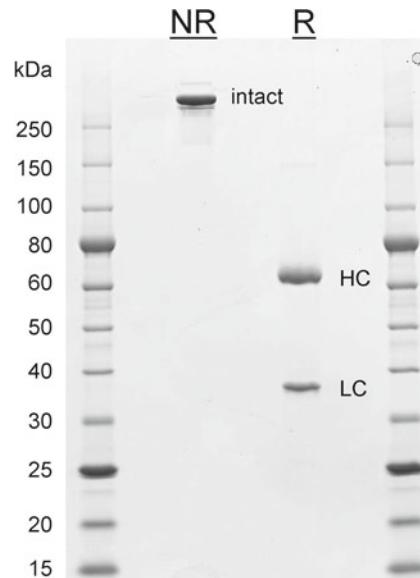


Fig. 3. SDS-PAGE analysis of a representative DVD-Ig protein. This 4–20% gradient, Tris-glycine SDS-PAGE shows that under nonreduced (NR) conditions the DVD-Ig protein migrates as an intact molecule of >200 kDa (see Note 10); when reduced (R), the heavy chain (HC) and light chain (LC) are apparent at ~63 and ~38 kDa, respectively. The molecular weight standards are from New England BioLabs (P7703S).

4. Notes

1. A recombinant/overlapping PCR primer contains two parts: a 20–30 bp region at 3' end for specific PCR amplification of fragment of interest, and a 20–30 bp region at 5' end for recombination. We routinely used high fidelity PCR kit (Invitrogen Life Science, Carlsbad, California) for PCR amplification, but other kits may also work for this application.
2. This purification step can be omitted if PCR amplification is efficient and specific. We generally get expected result from next recombination step by directly applying 1–2 μ l of PCR product to the recombination reaction.
3. For light chain vector construction, use pHyBE huC κ when V₂ is a V κ , use pHyBE huC λ when V₂ is a V λ .
4. This incubation time can be reduced to as short as 5 min because the efficiency of recombinant *E. coli* transformation is generally very high.
5. This incubation time can be reduced to as short as 15 min.
6. The number of colony PCR reactions varies according to the background of *E. coli* transformation. It can be as many as 24–48 if necessary.
7. Sometimes the ratio needs further optimization for better transient expression.
8. As with conventional IgGs, if cell viability drops to as low as 50–60%, DVD-Ig protein aggregation/fragmentation could be an issue. If confirmed, medium harvest should be performed slightly earlier when cell viability is around 70–80%.
9. Save the flow-through culture medium and washing buffer in case of inefficient binding.
10. The intact DVD-Ig protein can show anomalous migration on SDS-PAGE (as also observed for conventional IgGs) and often observed at higher than expected molecular weight. Mass spectrometry was used to verify the molecular weight of this example DVD-Ig protein is actually ~200 kDa.
11. The methods used to characterize conventional IgGs should also be used for characterization of DVD-Ig molecules. DVD-Ig proteins should be assessed for composition and purity, molecular weight, fraction of aggregate and glycosylation profile as determined by SDS-PAGE, size exclusion chromatography, and mass spectrometry. ELISA can be used for initial screening of antigen binding to verify general functionality; however, the antigen-binding kinetics of each variable domain should also be determined by SPR. The binding kinetic performance is a key measure closely coupled to the design

elements of the DVD-Ig protein. Further functional aspects of the DVD-Ig protein should be probed with specifically designed *in vitro* cellular assays. Finally, the pharmacokinetic properties and efficacy of the molecules should be evaluated in appropriate model systems.

References

1. Nelson AL, Dhimolea E, Reichert JM (2010) Development trends for human monoclonal antibody therapeutics. *Nat Rev Drug Discov* 9:767–774
2. Enever C, Batuwangala T, Plummer C, Sepp A (2009) Next generation immunotherapeutics - honing the magic bullet. *Curr Opin Biotechnol* 20:405–411
3. Uno T, Takeda K, Kojima Y, Yoshizawa H, Akiba H, Mittler RS et al (2006) Eradication of established tumors in mice by a combination antibody-based therapy. *Nat Med* 12:693–698
4. Wu C, Ying H, Grinnell C, Bryant S, Miller R, Clabbers A et al (2007) Simultaneous targeting of multiple disease mediators by a dual-variable-domain immunoglobulin. *Nat Biotechnol* 25:1290–1297
5. Wu C, Ying H, Bose S, Miller R, Medina L, Santora L, Ghayur T (2009) Molecular construction and optimization of anti-human IL-1alpha/beta dual variable domain immunoglobulin (DVD-Ig) molecules. *MAbs* 1(4):339–347
6. Wu C, Ghayur T, Salfeld J (2010) Generation and characterization of a dual variable domain immunoglobulin (DVD-IgTM) molecule, antibody engineering, vol 2, Chap 19, Roland Kontermann and Stefan Dübel, pp 239–250
7. Sandin S, Ofverstedt LG, Wikström AC, Wrangle O, Skoglund U (2004) Structure and flexibility of individual immunoglobulin G molecules in solution. *Structure* 12: 409–415
8. Robinson CR, Sauer RT (1998) Optimizing the stability of single-chain proteins by linker length and composition mutagenesis. *Proc Natl Acad Sci U S A* 95:5929–5934
9. Hsieh C-M (2009) Improved mammalian expression vectors and uses thereof. US PCT/US2009/031136
10. Sun X, Hia HC, Goh PE et al (2008) High-density transient gene expression in suspension-adapted 293 EBNA1 cells. *Biotechnol Bioeng* 99:108–116

Chapter 10

Engineering and Expression of Bibody and Tribody Constructs in Mammalian Cells and in the Yeast *Pichia pastoris*

Steve Schoonooghe

Abstract

Bibodies and tribodies are therapeutic antibody derivatives with sizes of approximately 75 and 100 kDa, respectively. This makes them smaller than full-size monoclonal antibodies, leading to better tissue penetration. Compared to the smaller scFv and Fab fragments, the bi- and tribody formats have the additional advantage of a slower renal clearance. However, the cost-effective and efficient production of these and other antibody derivatives is crucial for their further success as therapeutics. Here, we describe the construction and initial transient production in mammalian cells of bibodies and tribodies, followed by their stable production in *Pichia pastoris*. The purification of the antibody derivatives from the yeast supernatant is also explained.

Key words: Bibodies, Tribodies, Antibody derivatives, Bivalent, Multivalent, Yeast expression, *Pichia pastoris*

1. Introduction

A number of physiological and practical factors should be taken into account when designing therapeutic antibody derivatives. The molecules should be small enough to allow for efficient tissue distribution, but should also be large enough to avoid fast renal clearance. Also, avoiding immunogenic heterodimerization domains helps to prevent unwanted reactions to the construct. An equally important and often overlooked factor in the development of (recombinant) protein therapies, such as monoclonal antibody therapy, is the need for high treatment doses (>1 g/patient/year). In addition, many proteins can only be generated in relatively expensive mammalian cell fermentors (1, 2). So any

new recombinant antibody derivative must be efficiently produced and purified; otherwise, their clinical applicability is severely hampered. In the past few years, yeasts, and especially *Pichia pastoris*, have gained a significant interest for the production of recombinant antibody fragments. *P. pastoris* grows on cheap mineral defined media and requires a shorter process time as compared to mammalian cell culture (3, 4). Yeasts can be grown to high cell densities of up to 100 g/l dry biomass and the availability of strong, inducible promoters, such as the alcohol oxidase gene (AOX1) promoter, are further advantages of heterologous expression in *P. pastoris* (5).

Using the disulfide stabilized Fab fragment as a scaffold for heterodimerization, we engineered multivalent antibody derivatives of intermediate size (75–100 kDa). By recombinantly adding a flexible linker and scFv to the C-terminus of the Fab Fd or L chain, a “bibody” is created. When molecules are expressed with scFvs attached to both chains, “tribodies” are produced. Furthermore, these proteins are almost exclusively expressed as correctly heterodimerized products in both mammalian and *P. pastoris* expression systems (6). Here, we describe the construction and initial expression in mammalian cells of these bibodies and tribodies, followed by the stable expression in *P. pastoris* and purification from the yeast culture supernatant.

2. Materials

2.1. DNA Construction and Manipulation

1. Plasmids or PCR fragments containing Fab Fd chain, Fab L chain, and scFv(s).
2. pES33Hneo and pES33Ezeo for mammalian expression (see Note 1).
3. pKai61 and pKai51.2 for yeast expression (see Note 2).
4. Primers: These oligos should be designed specifically for the antibody genes used in the project. The locations on the genes and any necessary extensions are listed in Table 1.
5. Restriction enzymes, restriction buffers, BSA (buffer and BSA come included with enzyme), T4 DNA kinase, riboATP (rATP) (New England Biolabs, Promega).
6. Ready-to-ligation tubes (GE Healthcare), DNA-PCR cleanup kit (Promega), Qiaex gel extraction kit (Qiagen), plasmid DNA purification kit (Qiagen).
7. Vent DNA polymerase, thermopol buffer, Mg₂SO₄ (100 mM, included in Vent kit) (New England Biolabs).
8. dXTPs (10 mM) (Promega), DMSO, 5 M NaCl.

Table 1
Overview of primers needed for a bi- or tribody construction project

Name Annealing site 5' extension to be added

FdF	5' start of V _H	
LF	5' start of V _L	
FdR	3' end of C _H 1	AGGCCTTA
LR	3' end of C _L	AGGCCTTA
CH1 F	5' start of CH1	
CH1 R	3' end of CH1	AGATCCACCTCCGCCACTACCGCCTCC GCCGGGCC
scFv F	5' end of scFv	GTTAGTGGCGGGAGGTGGATCTGGAGGC GGCGGTAGTCCCAGG
scFv R	3' end of scFv	AGGCCTTA
HY R	3' end of Fd-scFv	TAACTAGTTA
LY R	3' end of L-scFv	TAACTAGTTA

9. TE buffer: 10 mM Tris–HCl pH 8.0, 1 mM EDTA.
 10. 0.22-μm syringe filters for DNA sterilization and plastic syringes.
 11. 1.5-ml tube centrifuge capable of at least 14,000×*g*, PCR thermocycler.
 12. Agarose gel-electrophoresis material.
 13. Heater block or warm/cold water bath capable of 16–95°C.
 14. Bacteria for DNA preparation and selection: *E. coli* MC1061: F⁻ araD139Δ(ara-leu)7696 galE15 galK16 Δ(lac)X74 rpsL- (*Srr*) *hsdR2* (r_k m_k⁺) *mcrA* *mcrB1L*.
 15. LB: 10 g/l Bacto-Tryptone, 5 g/l Bacto-Yeast extract, 5 g/l NaCl.
 16. 100 μg/ml ampicillin and 25 μg/ml zeocin LB + 1.2% agar plates.
- 2.2. Transient Expression in Mammalian Cells**
1. HEK293T human embryonic kidney cell line with SV40 large T-Ag (SV40T^{tsA1609}).
 2. DMEM medium with standard supplements: 0.03% L-glutamine, 20 U/ml penicillin, 20 μg/ml streptomycin, 0.4 mM sodium pyruvate.

3. Foetal Bovine Serum (FBS) and insulin-transferrin-selenium (ITS) (Gibco).
4. Hepes/BS: 5.96 g/l Hepes (25 mM)–NaOH, 16 g/l NaCl, 0.74 g/l KCl, 0.5 g/l $\text{Na}_2\text{HPO}_4 \cdot 12\text{H}_2\text{O}$, 2 g/l dextrose, pH 7.05.
5. CaCl_2 in Hepes: 12.5 M CaCl_2 , 125 mM Hepes–NaOH, pH 7.05.
6. Centrifugal protein concentration device with 10 kDa < cut-off < 50 kDa (depending on protein size) (Millipore).
7. Centrifuge for 15–50-ml tubes capable of $4,000 \times g$.

2.3. Expression in *P. Pastoris*

1. Competent methylotrophic yeast *P. pastoris* GS115(his4) cells.
2. 1 M sterile sorbitol, methanol, PBS, Tween-20.
3. 20% sterile glucose: Dissolve 20 g glucose in 100 ml H_2O , filter sterilize through 0.22- μm filter.
4. 1 M potassium phosphate pH 6.0: Dissolve 136.09 g KH_2PO_4 in 800 ml H_2O , adjust to pH 6.0 with KOH, add H_2O to 1,000 ml.
5. YP: 10 g Yeast extract, 20 g Peptone, fill up to 900 ml with H_2O , autoclave.
6. YPD: Take 900 ml of YP and add 100 ml of 20% sterile glucose.
7. YPD-agar: Add 20 g of agar to YP before autoclaving. After autoclaving, leave until temperature is below 50°C, then add 100 ml of 20% sterile glucose and 100 $\mu\text{g}/\text{ml}$ zeocin. Immediately pour plates, before mixture solidifies.
8. 13.4% yeast nitrogen base (YNB): 134 g YNB with ammonium sulfate in 1 l H_2O , sterilize over a 0.22- μm filter.
9. YPNG: Like YP, but dissolved in 700 ml H_2O . After autoclaving, add 100 ml of 10 % glycerol, 100 ml of 13.4 % YNB, 100 ml of 1 M potassium phosphate pH 6.0, and 2 ml of 0.02 % (w/v) biotine.
10. YPNM: Like YP, but dissolved in 700 ml H_2O . After autoclaving, add 100 ml of 10 % methanol, 100 ml of 13.4 % YNB, 100 ml of 1 M potassium phosphate pH 6.0, and 2 ml of 0.02 % (w/v) biotine.
11. Electroporation instrument capable of electric pulse parameters: 1,500 V, 40 μF , 200 Ω , and 8-ms duration; 0.2 cm gap cuvettes (Bio-Rad).
12. Shakers with temperature control.
13. 24 and 96 deep-well plates, shake flasks.

2.4. Purification

1. Centrifuge capable of 13,000×*g* and 0.22-μm bottle filters.
2. Chromatography equipment with UV and conductivity measurement and fraction collector.
3. Sephadex G-25 column (XK16/31), Chelating Sepharose Fast Flow column, Superdex 200 C26 column (GE Healthcare).
4. 3 mM NiSO₄.
5. 20 mM NaH₂PO₄-NaOH, pH 7.5.
6. (NH₄)₂SO₄, NaCl, imidazole, trehalose, glycerol.
7. Pierce Micro BCA™ Protein Assay Kit with IgG standard protein.

2.5. Quality Control

1. Standard SDS-PAGE and blotting materials, Coomassie dye, NBT/BCIP substrate.
2. Immunodetection buffer: 50 mM Tris-HCl pH 8.0, 80 mM NaCl, 5% nonfat milk powder, 0.2% Nonidet P40, 0.02 % NaN₃.
3. Anti-Fab antibodies and enzyme-conjugated secondary antibody (alkaline phosphatase, peroxidase).
4. S-tag FRETWorks assay kit, 96-well plates (Novagen).
5. Fluorescence 96-well plate reader with 485/20-nm excitation, 530/25-nm emission filter.

3. Methods**3.1. Construction of Fab-scFv Bibodies and Tribodies and Initial Expression in Mammalian Cells**

Although it is possible to construct yeast expression vectors straightaway, it is recommended to perform bi- and tribody construct optimization in a transient mammalian expression system due to the faster generation of test-samples and the more native expression environment (see Fig. 1). Once the bi- and tribodies with the desired characteristics are isolated, their expression cassettes can be transferred to *P. pastoris* vectors in two straightforward steps.

3.1.1. Cloning of Fd and L Chains in the pES33 Mammalian Expression Vector

Before making more advanced constructs, the basic elements of the Fab-based molecules are introduced into the pCAGGS-based pES33 mammalian expression vectors. Should no further modification of your molecule(s) be needed, please continue with Subheading 3.1.3.

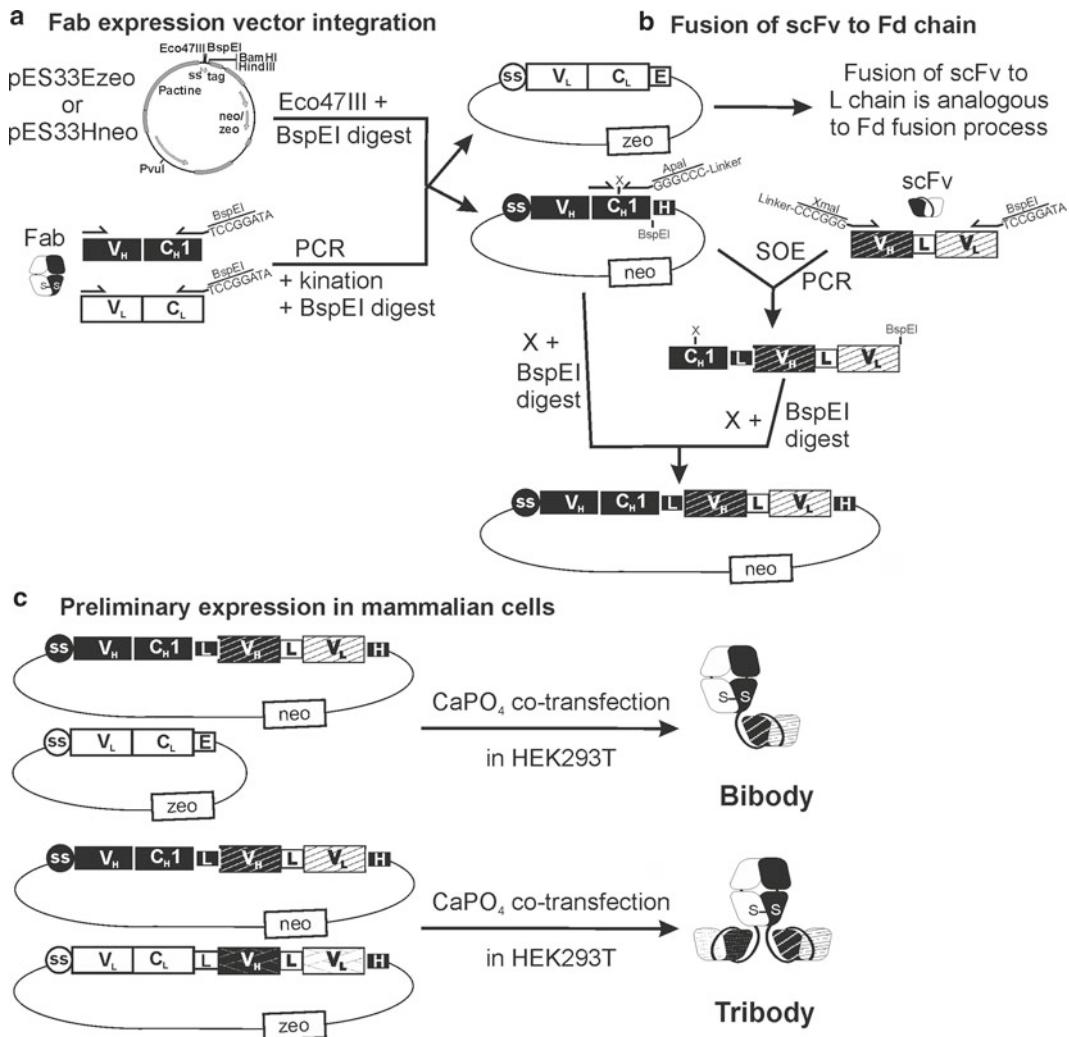


Fig. 1. Schematic outline of bibody and tribody construction and expression. (a) Both Fab chains are integrated in the mammalian expression system. (b) An scFv is fused to a Fab chain through SOE PCR, and the linker between Fab and scFv is created during this reaction. (c) Co-expression of heavy and light chains in mammalian cells results in bi- or tribody production. ss: signal sequence, V_H, V_L, C_H1, C_L: antibody genes, L: linker sequence, H: His₆-tag, E: E-tag, neo: neomycin resistance gene, zeo: zeocin resistance gene, X: single cutter restriction enzyme chosen in C_H1 domain, SOE: splice overlap extension.

- First, the heavy and light chains of the chosen Fab are introduced into the pES33 mammalian expression vector. In order to achieve this, four primers should be designed. Two forward primers (FdF and LF) for the 5' start of the V_H and V_L genes and two reverse primers (FdR and LR) for the 3' end of both C_H1 and C_L genes. The reverse primers should have a (AGGCCTTA) 5' extension, containing a BspEI site.

2. For PCR amplification of the Fd and L chains, the following mix is used:

PCR mix for Fd amplification	PCR mix for L amplification
0.2 µl pDNA with Fd gene	0.2 µl pDNA with L gene
10 µl Termopol buffer	10 µl Termopol buffer
1 µl Mg ₂ SO ₄	1 µl Mg ₂ SO ₄
4 µl dXTPs (10 mM stock)	4 µl dXTPs (10 mM stock)
2 µl FdF primer	2 µl LF primer
2 µl FdR primer	2 µl LR primer
1 µl Vent	1 µl Vent
80 µl H ₂ O	80 µl H ₂ O

3. The default PCR cycling program used is 2 min 95°C, followed by 30 cycles of 1 min 95°C, 1 min 55°C, and 1 min 72°C. The program is concluded by a final step of 10 min 72°C (see Note 3).
4. Gel-purify the PCR fragments (\pm 700 bp) and perform the following restriction digests:

40 µl Fd PCR DNA	40 µl L PCR DNA	10 µl pES33Hneo	10 µl pES33Ezeo
9 µl NEB3 buffer			
1 µl BSA (100 \times)			
5 µl (50 U) BspEI	5 µl BspEI	5 µl BspEI	5 µl BspEI
35 µl H ₂ O	35 µl H ₂ O	65 µl H ₂ O	65 µl H ₂ O
1.5 h at 37°C			
+1.5 µl 100 mM rATP	+1.5 µl 100 mM rATP	+1 µl NEB3	+1 µl NEB3
+1 µl T4 kinase	+1 µl T4 kinase	+1 µl 5 M NaCl	+1 µl 5 M NaCl
30 min at 37°C	30 min at 37°C	+5 µl (50 U) Eco47III	+5 µl Eco47III
		+3 µl H ₂ O	+3 µl H ₂ O
		1.5 h at 37°C	1.5 h at 37°C

5. Gel-purify all digests and prepare the ligation mix for Fd and L chain plasmids:
- Fd chain: Mix equimolar quantities of Fd PCR fragment and pES33Hneo fragments in a ready-to-go ligation tube to a maximum of 20 µl.

- L chain: Mix equimolar quantities of L PCR fragment and pES33Ezeo fragments in a ready-to-go ligation tube to a maximum of 20 µl.

Incubate both tubes at 16°C for 1 h.

6. Transform *E. coli* MC1061 bacteria with the ligation mixes and plate out on ampicillin containing LB + agar plates.
7. Pick a number of colonies and prepare DNA, and screen through restriction digest or DNA sequencing. Correct plasmids are designated pES33 <name Fab> HHneo for the Fd chain and pES <name Fab> LEzeo for the L chain, further referred to here as pES33FdHneo and pES33LEzeo.

3.1.2. Construction of Fd-scFv and L-scFv Chains

1. In order to fuse a scFv to the Fab, primers should be designed for the scFv and the Fab chain. The linker between both is incorporated in these primers. The process is described for fusion of a VH-VL-formatted scFv to the Fd chain, but the process is identical for the L-chain or fusion of VL-VH scFvs. The CH1 F primer is designed at the 5' start of the CH1 domain of the Fab. The CH1 R primer should anneal to the 3' of the CH1 domain. This primer carries the first part of the (Gly₄Ser)₃ linker and ApaI site in a 5' extension (AGATCCACCTCCGCCACTACCGCCTCCGCCGG GCCC). The partially complementary primer scFv F primer anneals to the 5' end of the VH should contain be extended with the second part of the linker and an XmaI site (GTAGTGGCGGAGGTGGATCTGGAGGCGGCGGT AGTCCCAGGG) (see Note 4). Finally, the scFv R primer anneals at the 3' end of the scFv and must have an (AGGCCTTA) 5' extension, containing a BspEI site (see Note 5).
2. For PCR amplification of the CH1 and scFv, the following mix is used:

PCR mix for CH1 amplification	PCR mix for scFv amplification
0.2 µl pES33FdHneo DNA	0.2 µl DNA with scFv gene
10 µl Termopol buffer	10 µl Termopol buffer
1 µl Mg ₂ SO ₄	1 µl Mg ₂ SO ₄
4 µl dXTPs (10 mM stock)	4 µl dXTPs (10 mM stock)
2 µl CH1 F primer	2 µl scFv F primer
2 µl CH1 R primer	2 µl scFv R primer
1 µl Vent	1 µl Vent
80 µl H ₂ O	80 µl H ₂ O

3. The default PCR cycling program used is 2 min 95°C, followed by 30 cycles of 1 min 95°C, 1 min 55°C, and 1 min

72°C. The program is concluded by a final step of 10 min 72°C (see Note 4).

- Gel-purify the PCR fragments (CH1 \pm 350 bp, scFv \pm 700 bp) and perform the following splice overlap extension (SOE) reaction:

SOE mix 1	SOE temperature program
5 μ l CH1 PCR fragment ($>=$ 50 ng)	
5 μ l scFv PCR fragment (equimolar to CH1)	Repeat 7 \times
5 μ l Termopol buffer	1 min 95°C
1 μ l Mg ₂ SO ₄	4 min 63°C
4 μ l dXTPs (10 mM stock)	
2 μ l DMSO	
1 μ l Vent	
25 μ l H ₂ O (adjust if more PCR DNA is used)	

- Starting from the previous reaction, perform the following rescue PCR reaction:

SOE mix 2	res.PCR temperature program
50 μ l SOE mix 1	2 min 95°C
5 μ l Termopol buffer	Repeat 30 \times
2 μ l dXTPs (10 mM stock)	1 min 95°C
2 μ l CH1 F primer	1 min 55°C
2 μ l scFv B primer	2 min 72°C
1 μ l Vent	10 min 72°C
38 μ l H ₂ O	

- Gel-purify the PCR fragment (\pm 1,150 bp) and perform the following restriction digests:

40 μ l CH1-scFv PCR DNA	10 μ l pES33FdHneo
9 μ l NEB3 buffer	9 μ l NEB3 buffer
1 μ l BSA (100 \times)	1 μ l BSA (100 \times)
5 μ l BspEI	5 μ l BspEI
5 μ l Enzyme X	5 μ l Enzyme X
30 μ l H ₂ O	30 μ l H ₂ O
2 h at 37°C	2 h at 37°C

(see Note 6).

7. Gel-purify all digests and prepare the ligation mix with equimolar quantities of the CH1-scFv PCR fragment and pES33FdHneo fragment in a ready-to-go ligation tube to a maximum of 20 µl. Incubate at 16°C for 1 h.
8. Transform *E. coli* MC1061 bacteria with the ligation mixes and plate out on ampicillin containing LB + agar plates.
9. Pick a number of colonies, prepare DNA, and screen through restriction digest or sequencing. Correct plasmids are designated pES33<name Fab>H<linker>sc<name scFv>Hneo, further generically referred to here as pES33FdHLscFvHneo. When an L chain-scFv fusion is made, the plasmids are designated pES33<name Fab>L<linker>sc<name scFv>Ezeo and further referred to here as pES33LLscFvEzeo.

3.1.3. Transient Expression in HEK293T Cells

For transient bi- or tribody expression, HEK293T cells were transfected according to the $\text{Ca}_3(\text{PO}_4)_2$ precipitation method (7). Protocol is described for one culture flask of 175 cm².

1. Seed 4.3×10^6 HEK293T cells in 32 ml DMEM with FBS.
2. The next day, mix and filter sterilize 14 µg of heavy- and 14 µg of light-chain pDNA.
3. Suspend the pDNA in a final volume of 1,400 µl 0.1 TE and add 350 µl $\text{CaCl}_2/\text{Hepes}$ buffer.
4. While shaking, add 1,750 µl Hepes/BS drop by drop to guarantee a fine precipitation, shake vigorously for 1 additional minute, and add immediately to the HEKT293T cells.
5. Remove the medium the next day and replace with 35 ml DMEM with ITS.
6. Harvest every 48 h. After three harvests, the production level generally drops dramatically.
7. Concentrate the supernatant and check for Fab-scFv expression by loading samples representing 0.5, 1, and 2 ml of culture supernatant on SDS-PAGE gels.

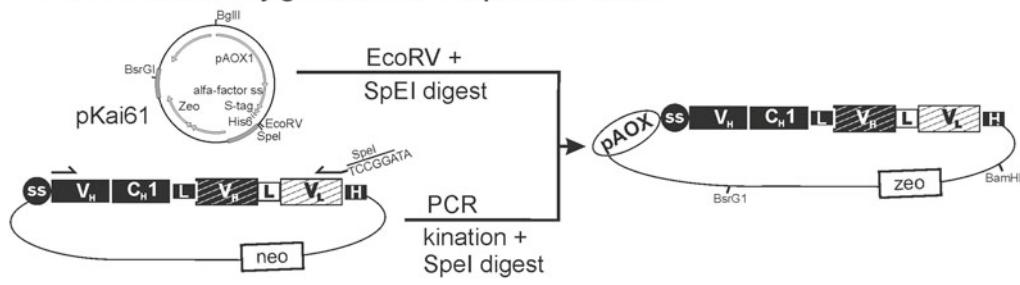
3.2. Construction of Fab-scFv Bibodies and Tribodies for Yeast Expression

3.2.1. Cloning of Bi- and Tribody Heavy and Light Chains in the pKai51 Yeast Expression Vector

After confirming the presence of all desired characteristics of the antibody construct in small-scale transient HEK293T expression, the genes coding for both heavy and light chains can be transferred to a yeast vector for stable expression in *P. pastoris* (see Fig. 2).

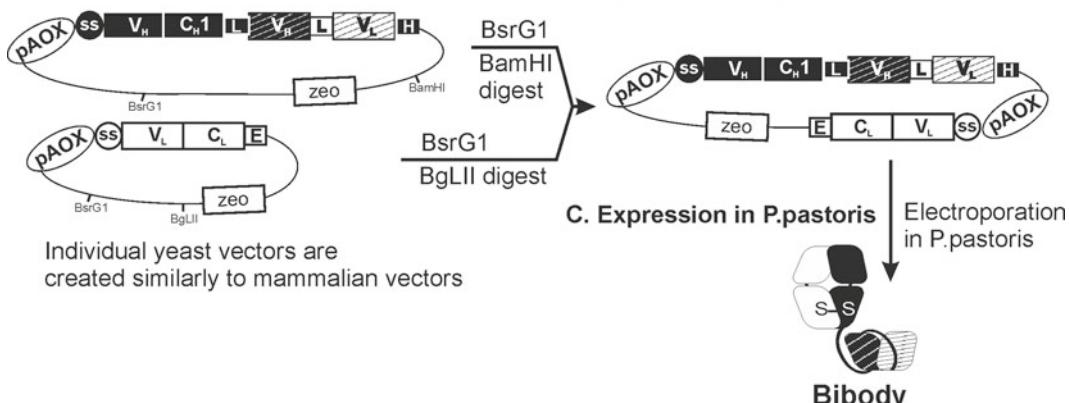
1. Only two reverse primers need to be designed specific for the 3' end of the Fd-scFv and L-scFv expression cassette. These HY R and LY R primers should contain a 5' TAACTAGTTA extension which codes for a stop-codon and SpeI restriction site.

a Transfer of antibody genes to Yeast expression vectors



Transfer of L-chain to yeast vector is completely analogous

b Construction of bi-cistronic vector for stable expression in *P. pastoris*



Individual yeast vectors are created similarly to mammalian vectors

Fig. 2. Schematic outline for bibody and tribody yeast expression vector construction. (a) For yeast expression, the bibody or tribody genes are transferred to pKai yeast expression plasmids by using PCR to create appropriate restriction enzyme sites at the end of the antibody genes. After digestion of both the PCR fragment and the yeast vector, the antibody genes are ligated into the vector. (b) In order to create a bicistronic vector, the plasmids for the antibody heavy and light chains are digested and ligated together. (c) This vector is subsequently introduced in *P. pastoris* through electroporation. ss: signal sequence, V_H, V_L, C_H1, C_L: antibody genes, L: linker sequence, H: His₆-tag, E: E-tag, neo: neomycin resistance gene, zeo: zeocin resistance gene, X: single cutter restriction enzyme chosen in C_H1 domain, SOE: splice overlap extension, pAOX: AOX promoter for *P. pastoris*.

2. For PCR amplification of the heavy and light chains, the following mix is used:

PCR mix for heavy chain amplification	PCR mix for light chain amplification
0.2 µl pDNA with Fd gene	0.2 µl pDNA with L gene
10 µl Termopol buffer	10 µl Termopol buffer
1 µl Mg ₂ SO ₄	1 µl Mg ₂ SO ₄
4 µl dXTPs (10 mM stock)	4 µl dXTPs (10 mM stock)
2 µl FdF primer (see Table 1)	2 µl LF primer (see Table 1)
2 µl HYR primer	2 µl LYR primer
1 µl Vent	1 µl Vent
80 µl H ₂ O	80 µl H ₂ O

3. The default PCR cycling program used is 2 min 95°C, followed by 30 cycles of 1 min 95°C, 1 min 55°C, and 1 min 72°C. The program is concluded by a final step of 10 min 72°C. For Fd or L-scFv fusion genes, an extension time of 2 min is required.
4. Gel-purify the PCR fragments (700 bp without, 1,500 bp with scFv fusion) and perform the following restriction digests:

40 µl heavy chain PCR DNA	40 µl light chain PCR DNA	10 µl pKai61	10 µl pKai51.2
9 µl NEB4 buffer	9 µl NEB4 buffer	9 µl NEB2 buffer	9 µl NEB2 buffer
1 µl BSA (100×)	1 µl BSA (100×)	1 µl BSA (100×)	1 µl BSA (100×)
5 µl SpeI	5 µl SpeI	5 µl SpeI	5 µl SpeI
35 µl H ₂ O	35 µl H ₂ O	7.5 µl EcoRV	7.5 µl EcoRV
		57.5 µl H ₂ O	57.5 µl H ₂ O
1.5 h at 37°C	1.5 h at 37°C	2 h at 37°C	2 h at 37°C
+1,5 µl 100 mM rATP	+1,5 µl 100 mM rATP		
+1 µl T4 kinase	+1 µl T4 kinase		
30 min at 37°C	30 min at 37°C		

5. Gel-purify all digests and prepare the ligation mix for heavy and light chain plasmids:
 - Fd chain: Mix equimolar quantities of heavy-chain PCR fragment and the large fragment from pKai61 in a ready-to-go ligation tube to a maximum of 20 µl.
 - L chain: Mix equimolar quantities of light-chain PCR fragment and pKai51.2 large fragment in a ready-to-go ligation tube to a maximum of 20 µl.

Incubate both tubes at 16°C for 1 h.
6. Transform *E. coli* MC1061 bacteria with the ligation mixes and plate out on zeocin containing LB + agar plates.
7. Pick a number of colonies and prepare DNA, and screen through restriction digest or sequencing. Correct plasmids are designated pKai61<name Fab-(scFv)> for the Fd chain and pKai51.2<name Fab(-scFv)> for the L chain, further referred to here as pKai61Fd and pKai51.2 L.

3.2.2. Cloning of the Heavy + Light Chain Bicistronic Expression Cassette in the pKai51 Yeast Expression Vector

Before transformation of *P. pastoris*, the expression vectors for the bibody or tribody heavy and light chains are combined in one bicistronic vector.

1. Perform the following restriction digests:

10 µl heavy-chain pDNA	10 µl light-chain pDNA
9 µl NEB2 buffer	9 µl NEB2 buffer
1 µl BSA (100×)	1 µl BSA (100×)
5 µl BamHI	7.5 µl BgIII
5 µl BsrGI	5 µl BsrGI
60 µl H ₂ O	57.5 µl H ₂ O
2 h at 37°C	2 h at 37°C

Gel-purify both digests and prepare the ligation reaction by mixing equimolar quantities of heavy- and light-chain fragment (both ±3,300 bp without scFv and ±3,900 bp with scFv fusion). Incubate at 16°C for 1 h.

2. Transform *E. coli* MC1061 bacteria with the ligation mixes and plate out on zeocin containing LB + agar plates.
3. Pick a number of colonies, prepare DNA, and screen through restriction digest or sequencing. Correct plasmids are designated pKai61<name bibody or tribody>, further referred to here as pKai61Fab-scFv.

3.3. Expression of Bi- and Tribodies in *P. pastoris*

P. pastoris is transformed using the newly created bicistronic yeast expression vector.

1. Linearize 10 µg of pKai61Fab-scFv with a PmeI restriction digest.
2. Gel-purify and sterilize the linear DNA expression cassettes.
3. Mix 2.5 µg of the pure expression DNA with 100 µl competent *P. pastoris* GS115 cells.
4. Transfer the mixture to an ice cooled 0.2-cm-gap cuvette, insert in an electroporation instrument, and pulse with the following electric pulse parameters: 1,500 V, 40 µF, 200 Ω, and 8-ms duration.
5. Add 1 ml of 1 M ice-cold sorbitol immediately after the pulse and transfer the mixture to tubes containing 2 ml YPD medium.
6. Incubate at 30°C for 1–1.5 h without shaking.
7. Plate 50 µl of the culture on YPD-agar plates containing 100 µg/ml zeocin and incubate at 28°C for 3–4 days.
8. Select 22 colonies for bi- or tribody expression by transferring them to a new YPD-zeocin plate with a 24-cell numbered grid. Also inoculate one positive and one negative control.
9. Transfer the clones to a 24 deep-well plate containing 2 ml YPNG medium with 100 µg zeocin/well and incubate at 28°C, 250 rpm shaking, for 24 h.

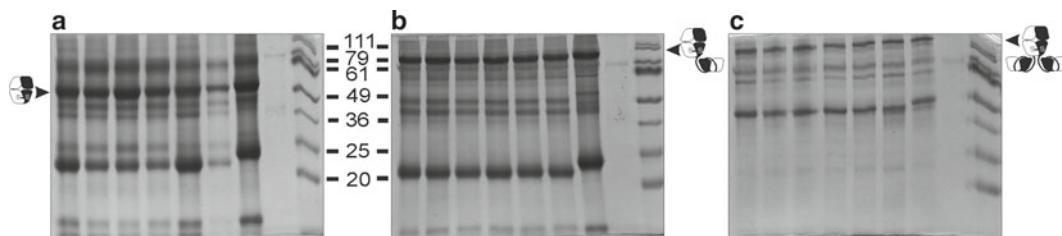


Fig. 3. Characterization of Fab, bi- and tribody producing *P. pastoris* clones. Coomassie-stained 15% SDS-PAGE gels loaded with concentrated protein samples equivalent to 200 µl culture supernatant from anti-MUC1 Fab (a), bibody (b) or tribody (c) expressing *P. pastoris* clones. Arrows and *antibody chain symbols* indicate the running height for anti-MUC1-derived proteins. The *uttermost right lane* of each gel contains a relative molecular mass marker. Based on these results, production clones were selected for all anti-MUC1 derivatives.

10. Centrifuge the plate for 8 min at $500 \times g$, replace the medium with 2 ml YPNM (without zeocin), and incubate for a further 18 h at 28°C and 250 rpm shaking.
11. Induce antibody expression by adding 50 µl of 50% methanol to each well, continue incubation at 28°C and 250 rpm shaking, and repeat this step after 8 and 24 h.
12. Centrifuge the culture 8 h after the last methanol addition for 8 min at $500 \times g$ and harvest the supernatant.
13. The productivity of the clones is examined using SDS-PAGE (see Fig. 3) and S-tag quantification. Highly productive clones are saved by making a 1/1,000 dilution in PBS-0.2%Tween-20 in 96-well deep polypropylene plates and freezing at -20°C.
14. Choose the best producing clone and scale-up production to shake flasks, scaling up all volumes accordingly.

3.4. Purification of His₆-Tagged Bi- and Tribodies

After the bibodies and tribodies have been produced in *P. pastoris*, the products need to be cleared of impurities. The products constructed according to the previously described protocols all carry a His₆-tag that can be used for purification by immobilized metal affinity chromatography (IMAC). Alternatively, bi- and tribodies without this tag can also be purified through protein-L or hydrophobic charge induction chromatography.

Three phases are distinguished during the purification protocol: getting the product in the right buffer and volume for purification, the purification step, and removal of unwanted buffer components and possible dimers and degradation products.

3.4.1. Buffer Changes and Capture Steps

Before the product can be efficiently purified, it has to be present in an appropriate volume and interfering medium components should be removed. Yeast supernatant contains peptides and other substances that interfere with IMAC purification. Therefore, the bi- or tribodies were precipitated by ammonium sulfate precipitation (see Note 7).

1. Centrifuge harvested cell supernatant for 30 min at $13,000 \times g$ and filter over a $0.22\text{-}\mu\text{m}$ filter.
2. Add 70% $(\text{NH}_4)_2\text{SO}_4$ to the harvested yeast medium to precipitate the antibody derivatives.
3. Centrifuge for 30 min at $13,000 \times g$ and dissolve the pellet in 10 ml 20 mM NaH_2PO_4 , 300 mM NaCl, and 20 mM imidazole pH 7.5.
4. Inject this solution on a 62 ml desalting sephadex G-25 column (XK16/31) equilibrated with the same buffer and collect the desalted protein fraction.

3.4.2. IMAC Purification

During IMAC purification, the His tag interacts with the immobilized Ni^{2+} on the column, allowing the bi- and tribodies to be efficiently purified (see Note 8).

1. Adjust the buffer of the protein sample to contain 20 mM imidazole and adjust the pH to 7.5.
2. Equilibrate a Ni^{2+} loaded Chelating Sepharose Fast Flow column run at 2 ml/min with 20 mM phosphate running buffer containing 0.5 M NaCl.
3. Load the sample on the column and wash with 10 column volumes of running buffer afterwards (see Note 9).
4. Elute the protein with running buffer containing 400 mM imidazole.

3.4.3. Product Polishing

More than 90% pure bi- or tribody can be obtained by size exclusion chromatography (SEC). For these proteins, SEC columns should be used with a cutoff >150 kDa.

1. Equilibrate a 400 ml Superdex 200 C26 column to PBS at a constant speed of 1–2 ml/min.
2. Inject the sample ($\leq 2\%$ of column volume) on the column and start collecting samples of 2–5 ml. Record the UV chromatogram.
3. When the sample has passed the column, inject a molecular weight standard and record the UV chromatogram (see Fig. 4).
4. Determine the peak (usually the largest at this stage) which contains the bi- or tribody molecule and pool the fractions from this peak.
5. Protein concentrations of pure protein are measured with the Pierce Micro BCA™ Protein Assay Kit with IgG standard protein according to the manufacturer's instructions.
6. Pure bibody and tribody proteins can be safely frozen after adding 10% trehalose or 50% glycerol.

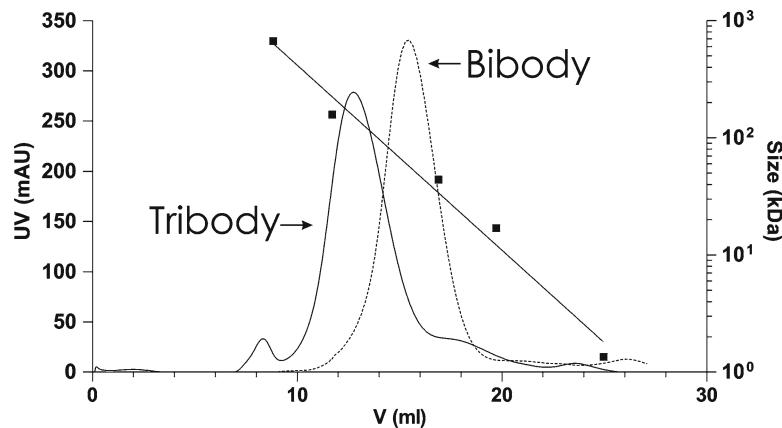


Fig. 4. Characterization and final purification step of yeast produced anti-MUC1 bi- and tribody. UV 280 nm size exclusion chromatogram of the *P. pastoris* produced PH1 bibody (solid line) and tribody (dotted line) purified on a superdex 200 HR10/30 column. Arrows depict PH1 bibody and tribody peaks. Also indicated is the Bio-rad molecular mass standard run under similar conditions on the same column (filled square).

3.5. Quality Control

During the selection and production of bibodies and tribodies, a number of quality control methods are used. These protocols are generally well known and only the details specific to this chapter are discussed.

3.5.1. SDS-PAGE

1. Protein fractions are separated on two 10% SDS-PAGE gels.
2. One is visualized using Coomassie Brilliant Blue dye, and the second is blotted to a nitrocellulose membrane.
3. Immunodetection of blots is performed by incubating for 1 h consecutively with anti-Fab antibodies and an alkaline phosphatase-conjugated secondary antibody. Anti-tag antibodies can also be used as primary detection antibodies.
4. Subsequent visualization is performed with NBT/BCIP substrate.
5. Protein recovery and purity are determined on scanned Coomassie gels using Quantity One software from Bio-Rad.

3.5.2. 2S-Tag Assay

The *P. pastoris*-expressed bi- and tribodies carry an N-terminal S-tag that can be used for quantification in complex media. In this protocol, the FRETWorks S-tag assay kit was used.

1. Make a calibration curve in a 96-well plate by 1/2 serial dilution starting at 5 fmol/ μ L.
2. Make several dilutions of yeast supernatant in PBS-Tween-20, starting at least at 1/1,000, and add 20 μ L of this sample to 180 μ L working solution present in the 96-well plate.

Table 2
Peptide sequence of Fab-scFv linkers (6,9 & unpublished data)

Name	Sequence									
H1	E	P	S	G	P					S
H2	E	P	S	G	P	G	G	S	G	G
L2			G	P	G	G	G	S	G	S P G
H3	E	P	T	S	G	S	G	K	P	G
L4	D	V	P	S	P	G				
L5	D	V	D	G	G	S	R	G	D	G
L6				G	P	P	S	P	P	G
H6	E	P	S	G	P	P	S	P	G	
L7				G	P	Q	P	Q	P	Q
H7	E	P	S	G	P	Q	P	Q	P	Q

Stop the reaction after 30 min and read in a fluorescence plate reader using a 485/20-nm excitation and a 530/25-nm emission filter.

4. Notes

1. The pES33Ezeo and pES33Hneo expression plasmids were constructed using pCDNA3, pCDNA3.1zeo- (Invitrogen), and pCAGGS (8). The vector contains the following features: a β -actin- β globulin promoter from pCAGGS vector, a kozak sequence (GCCACCATGG), a consensus excretion signal sequence (ss) from a mouse antibody (MGWSCIIFLVATA TGVHS), an Eco47III site 3' in frame of the ss for easy insertion of genes, a BspEI restriction site 5', and a BamHI site 3' from the tag sequence for tag management and zeocin or neomycin resistance genes from pCDNA3.1zeo- and pCDNA3, respectively. The construction strategies described here will work for other expression vectors, but the appropriate insertion sites should be chosen for the system in question.
2. The pKai51.2 yeast expression vector originated from pGAPZalphaA (Invitrogen). An N-terminal His₆ tag was introduced followed by a Caspase 3 protease site to allow tag removal. The GAP promoter was replaced with AOX1 from pPICZA (Invitrogen) and the PmeI site in the middle of the AOX1 promoter was removed. pKai61 carries an additional S-tag in front of the His6-tag.
3. The precise annealing temperature is dependent on the primer design, so optimization of this parameter may be required. When several bands are detected on gel, increasing the temperature is recommended to lower false priming. In case no bands are found of the expected size, the temperature should be lowered to increase the primer annealing.
4. The linker described in the text is based on the “classic” (Gly₄Ser)₃ linker. However bi- and tribodies have been successfully made with a host of other linkers (Table 2).
5. The sequence covered by the CH1 F and CH1 R primer should contain at least one unique restriction site that generates ends different from those generated by BspEI. If this is not the case, the primer should be shifted further upstream into the VH sequence. If no site is present there, the VH F primer from the previous construction can be used, but it is advised to keep the PCR fragments as small as possible for the efficiency of the SOE procedure. Also, the ApaI and XmaI sites in the primers are nonessential for this construction but allow the easy exchange of Fab and scFv fragments in secondary constructs.

6. Enzyme X is the restriction enzyme for a unique site, chosen as described in Note 5. Preferentially this enzyme is compatible with the BspEI restriction buffer; otherwise, supplementary cleanup and restriction digest steps are necessary. Examples for Enzyme X from our sequences are EcoNI and Van9II for the mouse IgG2b CH1 and AgeI or BstXI for the human IgG1 CH1. If no enzyme is available and the cloning takes place with primer VH F, a blunt/BspEI fragment can be created as described in the very first cloning step. In this case the resulting Fd-scFv gene needs to be inserted in pES33Hneo similar to the preceding steps in Subheading 3.1.
7. Other capture/purification column-based protocols are also prone to interference from the yeast medium. Some components accumulate on the matrix of the column and are very difficult to clean afterwards.
8. The volume of the column should be proportional to the amount of expected bi- or tribody product in order to avoid overloading the column or overly diluting the eluted protein. For HEK293T productions a 1 ml column is usually sufficient. For NS0 and yeast production a 20 ml column is routinely used.
9. At this time the sample should be free of chelating substances; otherwise, the Ni²⁺ will leech from the column and the purification will fail.

References

1. Holliger P, Winter G (1993) Engineering bispecific antibodies. *Curr Opin Biotechnol* 4: 446–449
2. Werner RG (2004) Economic aspects of commercial manufacture of biopharmaceuticals. *J Biotechnol* 113:171–182
3. Freyre FM, Vazquez JE, Ayala M, Canaan-Haden L, Bell H, Rodriguez I, Gonzalez A, Cintado A, Gavilondo JV (2000) Very high expression of an anti-carcinoembryonic antigen single chain Fv antibody fragment in the yeast *Pichia pastoris*. *J Biotechnol* 76:157–163
4. Ning D, Junjian X, Qing Z, Sheng X, Wenyin C, Guirong R, Xunzhang W (2005) Production of recombinant humanized anti-HBsAg Fab fragment from *Pichia pastoris* by fermentation. *J Biochem Mol Biol* 38:294–299
5. Cregg JM, Vedvick TS, Raschke WC (1993) Recent advances in the expression of foreign genes in *Pichia pastoris*. *Biotechnology (N Y)* 11:905–910
6. Schoonooghe S, Kaigorodov V, Zawisza M, Dumolyn C, Hastrae J, Grooten J, Mertens N (2009) Efficient production of human bivalent and trivalent anti-MUC1 Fab-scFv antibodies in *Pichia pastoris*. *BMC Biotechnol* 9:70
7. O'Mahoney JV, Adams TE (1994) Optimization of experimental variables influencing reporter gene expression in hepatoma cells following calcium phosphate transfection. *DNA Cell Biol* 13:1227–1232
8. Niwa H, Yamamura K, Miyazaki J (1991) Efficient selection for high-expression transfectants with a novel eukaryotic vector. *Gene* 108:193–199
9. Schoonjans R, Willem A, Schoonooghe S, Fiers W, Grooten J, Mertens N (2000) Fab chains as an efficient heterodimerization scaffold for the production of recombinant bispecific and trispecific antibody derivatives. *J Immunol* 165(12):7050–7057

Chapter 11

Use of *E. coli* for the Production of a Single Protein

Lili Mao and Masayori Inouye

Abstract

E. coli has been widely used for recombinant protein production. Here, we introduce a novel expression method in *E. coli*, the Single Protein Production (SPP) system, in which *E. coli* is converted into a bioreactor producing only the target protein. In the SPP system, all *E. coli* cellular mRNAs are eliminated by the induction of MazF, an ACA-specific mRNA interferase, which results in complete cell growth arrest. However, the mRNA for a target protein was engineered to be devoid of ACA sequences, thus resistant to MazF cleavage. Therefore, the SPP system is unique and ideal for expression of toxic proteins and incorporation of toxic amino acid analogues. We have also demonstrated that the SPP system is a cost-effective protein production method for NMR structural studies because cell culture can be highly condensed without affecting protein yields.

Key words: *E. coli*, MazF, The SPP system, NMR protein structure

1. Introduction

The Single Protein Production (SPP) system was developed on the basis of an *E. coli* mRNA interferase, MazF, which cleaves mRNA at the ACA sequences (1). Almost all the *E. coli* cellular mRNAs are subjected to MazF cleavage and thereafter degraded rapidly, leading to cell growth arrest (Fig. 1a). However, cells in the SPP system are still metabolically active in that they are capable of producing nucleic acids, amino acids, ATP, RNA, and proteins. Therefore, if a gene of interest is designed to have no ACA sequences, it will be expressed as the only protein produced in the living *E. coli* cells (Fig. 1b).

The SPP system has a few advantages over conventional expression methods. First of all, since only the target protein is produced, purification will be much easier and less time consuming. Secondly, since cell growth is completely arrested, cells in the SPP system are

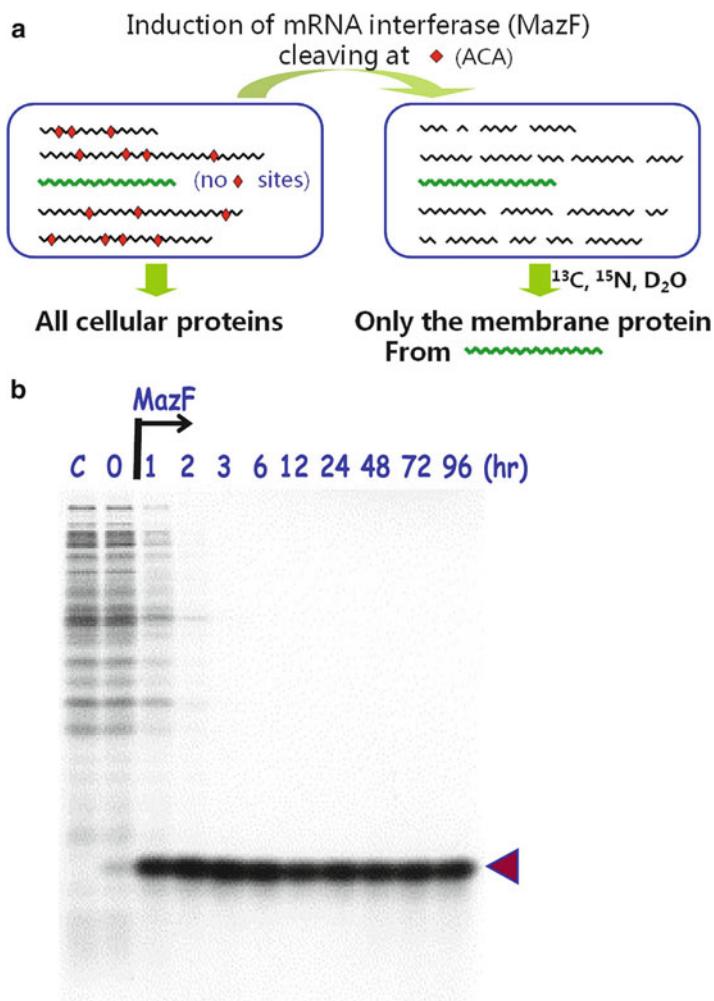


Fig. 1. The production of a single protein in *E. coli*. (a) Single Protein Production (SPP) system in *E. coli*. (b) The gene for eotaxin, consisting of 74 amino acid residues, was synthesized without ACA-sequences and inserted in pColdI vector. After MazF and the target gene was induced by IPTG, cells were pulse-labeled with (^{35}S)-methionine (reproduced from (4) with permission from Springer).

resistant to most toxic chemicals for protein synthesis, such as toxic amino acid analogues, D_2O , etc., that affect cell division or cell growth-related functions (2, 3). Thirdly, an *E. coli* culture in the SPP system can be condensed at least 20-fold without affecting protein yield per cell. If expensive chemicals are used, such as isotopes for NMR sample preparation, their cost in the SPP system can be reduced to less than 5% of that in the conventional methods (4). Lastly, if the target protein is a membrane protein and expressed in the isotope-containing medium, the protein will be the only protein in the *E. coli* membrane labeled with isotopes. NMR measurements of the target protein can be performed directly on the *E. coli*

native membrane or with simple detergent extraction, which avoids any reconstitution steps (5, 6).

Significant optimization of the SPP system has also been achieved since it was first established in 2005. In this protocol, we describe how to carry out expression and isotope labeling of a target protein with the most recently developed dual-inducible SPP system to achieve nearly 100% incorporation of isotope-labeled amino acids or amino acid analogues.

2. Materials

2.1. *E. coli* Strains and Plasmids

1. *E. coli* strains: *E. coli* BL21 (DE3) strain was used for expression in the original SPP system (1, 2, 7). In order to develop the dual-inducible SPP system, the *E. coli* histidine auxotroph, BL21 (DE3) ($\Delta hisB$), was constructed by deleting *hisB* gene and employed as the host strain for experiments in this protocol (see Note 1) (8). A tryptophan auxotroph may also be used in a similar manner as the histidine auxotroph, since MazF (ΔW) is also available (8). In this MazF (ΔW), Trp14 and Trp83 were replaced with Phe and Leu, respectively, without affecting the enzymatic activity of the cleavage specificity.
2. pACYCmazF(ΔH): For the construction of the dual-inducible SPP system, MazF(ΔH) was expressed from pACYCmazF(ΔH), which is chloramphenicol resistant. The His28 residue is removed from MazF by replacing it with Arg and Gly27 with Lys without compromising its mRNA interferase activity (see Note 2) (8).
3. pCold vectors: Expression of target proteins is carried out by using pCold (sp-4) vectors, which are ampicillin resistant. Four pCold vectors are currently available, which are pColdI, II, III, and IV (see Note 3). They differ at the sequences between 5'-UTR and the multiple cloning sites (MCSs). pColdI contains translation enhancing element (TEE), a hexa-His tag, and a factor Xa cleavage site. pColdII contains TEE and the hexa-His tag. pColdIII contains TEE only, while pColdIV does not have any extra sequences (7).

2.2. Buffer and Medium

1. Inoue transformation buffer: pH 6.7. 10 mM PIPES, 15 mM CaCl₂, 250 mM KCl, and 55 mM MnCl₂. Weigh 0.302 g of PIPES, 0.22 g of CaCl₂·2H₂O, and 1.86 g of KCl. Dissolve them in about 90 ml of distilled water. Adjust pH to 6.7 with 1 N KOH. Add 1.09 g of MnCl₂·4H₂O and fill the solution up to 100 ml with distilled water. Filter the solution to sterilize with a 0.2-μm cellulose acetate cartridge filter. Keep the buffer at 4°C.

2. M9 phosphate buffer: 50 mM Na_2HPO_4 , 22 mM KH_2PO_4 , and 8.5 mM NaCl.
3. 10× M9: 132 g/l $\text{Na}_2\text{HPO}_4 \cdot 7\text{H}_2\text{O}$ (or 70 g/l Na_2HPO_4), 30 g/l KH_2PO_4 , 5 g/l NaCl, and 10 g/l NH_4Cl .
4. M9 minimal medium: To make 1 l of M9 minimal medium, dilute 100 ml of 10× M9 with 880 ml of distilled water and autoclave it before adding other components (see Note 4). After temperature cools down, add 1.26 ml of 0.81 M MgSO_4 , 10 ml of 40% glucose, 4 ml of 0.5 mg/ml thiamine (VB1), 1 ml 100 $\mu\text{g}/\text{ml}$ of Ampicillin, and 1 ml of 25 $\mu\text{g}/\text{ml}$ chloramphenicol. For M9 (CAA) medium, add additional 10 ml of 20 % casamino acid. For M9 (His) medium, add additional 5 ml of 10 mg/ml histidine.
5. M9 (CAA) plate: For 1-l scale, add 15 g of agar into 880 ml of distilled water, which is then autoclaved. After temperature cools down, add 100 ml of 10× M9, 1.26 ml of 0.81 M MgSO_4 , 10 ml of 40% glucose, 4 ml of 0.5 mg/ml thiamine (VB1), and 10 ml of 20% casamino acid. Ampicillin(100 $\mu\text{g}/\text{ml}$), chloramphenicol (25 $\mu\text{g}/\text{ml}$), and histidine (50 $\mu\text{g}/\text{ml}$) are also added for the dual-inducible SPP system, in which the host strain is histidine auxotroph. Pour about 25 ml to each Petri dish (100×15 mm).
6. LB plate: For 1-l scale, weigh 10 g of tryptone, 5 g of yeast extract, and 10 g of NaCl. Dissolve in 900 ml of distilled water and fill up to 1,000 ml. Add 15 g of agar and autoclave the mixture. For transformation with pACYCmazF(ΔH), chloramphenicol (25 $\mu\text{g}/\text{ml}$) is added after the medium is cooled down. Pour about 25 ml to each Petri dish (100×15 mm).

2.3. ACA-Less Genes

All the ACA sequences have to be removed from the DNA sequence of a target gene without altering its amino acid sequence (7). Codon-optimized, ACA-less DNA sequences are generated using the program (SPP-ACA). All NdeI and BamHI restriction enzyme sites are also removed without changing the amino acid sequence (see Note 5). The ACA-less gene is commercially synthesized.

3. Methods

All the equipment and solutions should be sterilized.

3.1. Preparation of Competent Cells for the SPP Methods

All the competent cells are prepared following the Inoue Method from the Book of *Molecular Cloning*, with a few modifications. The transformation buffer is filtered to be sterilized and has to be ice-cold before use. Overnight incubation is about 16–21 h.

(a) Preparation of competent cells of BL21 (DE3) ($\Delta hisB$)

1. Streak the *E. coli* strain BL21 (DE3) ($\Delta hisB$) on an LB plate and incubate the plate at 37°C overnight.
2. Inoculate one colony into 50 ml of LB medium in a 250-ml flask and incubate the flask on a shaker at 18°C for about 40 h, until OD₆₀₀ reaches 0.5 (see Note 6).
3. Keep the flask on ice for 10 min and then pour the culture into a 50-ml centrifuge tube. Centrifuge at 1,000 $\times g$ for 10 min at 4°C.
4. Discard the supernatant and resuspend the cell pellet in 16 ml of inoue transformation buffer by gently rotating the tube on ice (see Note 7).
5. Leave the cell suspension on ice for 10 min, followed by centrifugation at 3,000 rpm for 10 min at 4°C.
6. Discard the supernatant and resuspend the cell pellet in 4 ml of transformation buffer by gently rotating the tube on ice (see Note 8). Add 300 μ l of dimethyl sulfoxide (DMSO) slowly to the cell suspension (see Note 8). Leave the tube on ice for 10 min.
7. Aliquot the cell suspension to 200 μ l each into 500- μ l microtubes and freeze them quickly in liquid nitrogen. Keep the tubes in -80°C freezer.

(b) Preparation of competent cells of BL21 (DE3) ($\Delta hisB$)/*pACYCmazF* (ΔH)

1. Thaw one tube of BL21 (DE3) ($\Delta hisB$) competent cells on ice and transfer 80 μ l to an ice-cold 1-ml microtube which contains 1 μ l of *pACYCmazF* (ΔH). Mix it well by pipetting up and down a few times.
2. Heat shock at 42°C for 90 s, followed by chilling on ice for 1 min.
3. Incubate the mixture at 37°C for 1 h and spread it on a pre-warmed LB plate (25 μ g/ml chloramphenicol). Incubate the plate at 37°C overnight.
4. Prepare competent cells of BL21 (DE3) ($\Delta hisB$)/*pACYCmazF*(ΔH) as described in (a) (see Note 9).

3.2. Transformation of BL21 ($\Delta hisB$)/*pACYCmazF* (ΔH) with *pCold* Vectors

1. Thaw one tube of BL21 (DE3) ($\Delta hisB$)/*pACYCmazF* (ΔH) competent cells on ice and transfer 80 μ l to an ice-cold 1-ml microtube that contains 1 μ l of the *pCold* vector in which the target gene was cloned (see Note 10). Mix it well by pipetting up and down a few times.
2. Heat shock at 42°C for 90 s, followed by chilling on ice for 1 min.
3. Spread it on a pre-warmed M9 (CAA) plate, which was prepared as described in step 5 in Subheading 2.2. Incubate the plate at 37°C for overnight.

3.3. Pre-culture

1. Prepare 50 ml of M9 (CAA) medium as described in step 4 Subheading 2.2.
2. Inoculate one colony (from Subheading 3.2) into the medium and incubate at 37°C with shaking overnight.

3.4. Scaled-Up Culture

1× M9 was made by diluting 10× M9 with distilled water. M9 minimal medium was prepared as described in step 4 Subheading 2.2. Both 15°C and 37°C shakers are set at 180 rpm. All the experiments after cold-shock (see step 3 Subheading 3.4) are carried out on ice or at 4°C.

1. Collect the pre-culture in a 50-ml tube and centrifuge at $3,000 \times g$ for 15 min at room temperature. Discard the supernatant and resuspend the cell pellet in 5 ml of 1× M9 by gentle vortexing.
2. Inoculate the cell suspension into 1 l of M9 (His) minimal medium in a 4-l flask and grow at 37°C with shaking. Measure OD₆₀₀ of the culture every 1 h, until it reaches 0.5–0.6 (see Note 11).
3. Incubate the culture in an ice-water bath for 5 min and then shift it to a 15°C shaker (see Note 12). This step is called cold-shock and it takes about 45–60 min (see Note 12).
4. Collect the culture in a 1-l centrifuge bottle. Centrifuge at 3,000 rpm for 30 min at 4°C.
5. Wash the cell pellet with 50 ml of 1× M9 (ice-cold). Centrifuge at 5,000 rpm, for 15 min at 4°C.
6. Resuspend the cell pellet with 100 ml of M9 minimal medium (ice-cold) in a 2-l flask (see Note 13). 1 mM IPTG was added subsequently to the culture for induction of MazF. Incubate the culture on the 15°C shaker overnight.
7. If the target protein has to be labeled with chemicals like isotopes or amino acid analogues, go to step 8. Otherwise, add 50 µg/ml Histidine to the condensed culture and continue incubating the culture on the 15°C shaker overnight, for induction of the target protein (Fig. 2).
8. Collect the cell culture by centrifugation (5,000 rpm, 15 min, 4°C,) and wash the cell pellet with 50 ml of M9 phosphate buffer (ice-cold, step 2 Subheading 2.2). Resuspend the cell pellet with 50 ml of M9 (His) minimal medium (ice-cold) in a 2-l flask (see Notes 13 and 14), in which 1 mM of IPTG was added for induction of the target protein. Incubate the culture on the 15°C shaker overnight
9. Collect the cell culture and keep the cell pellets at -80°C for future use.

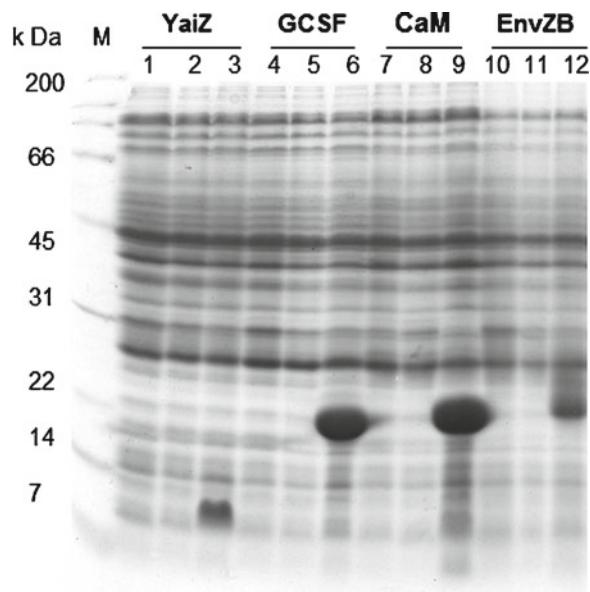


Fig. 2. Expression profiles of different target proteins using His-inducible SPP system. *YaiZ* an *E. coli* inner membrane protein; *GCSF* granulocyte colony-stimulating factor; *CaM* calmodulin, a calcium-binding protein expressed in all eukaryotic cells; *EnvZB* the ATP-binding domain of *E. coli* EnvZ histidine kinase. Lanes 1, 4, 7 and 10, before addition of 1 mM IPTG; Lanes 2, 5, 8, and 11, overnight with IPTG but no histidine; Lanes 3, 6, 9, and 12, overnight in the presence of both IPTG and histidine; and Lane M, molecular weight markers (reproduced from (8) with permission from *Applied and Environmental Microbiology*).

4. Notes

1. *E. coli* BL21(DE3) (Δ hisB) strain is not capable of synthesizing histidine because it is a histidine auxotroph. For production of proteins that contain His residue, M9 minimal medium has to be supplemented with histidine.
2. In the dual-inducible SPP system, both MazF(Δ H) and the target gene are IPTG inducible. However, only MazF(Δ H) can be induced by IPTG in the M9 minimal medium which is not supplemented with histidine. Expression of the target protein, which contains a His residue(s), requires both IPTG and histidine in M9 minimal medium.
3. pCold vectors are under control of the Lac operator and the cspA promoter. Therefore, if a target gene is cloned into pCold vectors, it is IPTG inducible at low temperature (15°C).
4. Do not autoclave M9 minimal medium after adding MgSO₄, glucose, and VB1. Make separate stock solution for each component (0.81 M MgSO₄, 40% glucose, and 0.5 mg/ml VB1), which is autoclaved separately.

5. NdeI and BamHI are the most commonly used restriction enzymes in cloning for the SPP system as NdeI and BamHI are the 5'- and 3'-end cloning sites, respectively. If there is His-Met sequence in the amino acid sequence, there is an NdeI site on the DNA. The NdeI site cannot be removed because it will generate an ACA sequence. In this case, other restriction enzyme sites in the MCS, such as SacI, may be chosen.
6. The culture with OD_{600} between 0.5 and 0.6 is the best for preparation of competent cells. However, it is also acceptable if OD_{600} is above 0.3 until early stationary phase.
7. Do not use a vortex for suspending cell pellets. Finger vortexing or rotating the tube on ice is recommended. It takes a few minutes to get homogeneous suspension.
8. Suspend the cell pellet carefully as described in Note 7. The cell pellet at this step is very easily suspended. DMSO has to be added to the cell suspension drop by drop while slowly hand rotating the tube at the same time.
9. LB medium to prepare competent cells of *BL21 (DE3) (ΔhisB)/pACYCmazF (ΔH)* contains 25 µg/ml chloramphenicol.
10. Expression level of a target protein may be different on each pCold vector. The gene of interest is cloned to all the four pCold vectors. Test experiments are needed to find in which pCold vector yield of the target protein is the highest, before performing large-scale production.
11. The culture with OD_{600} between 0.5 and 0.6 is the best for the SPP system. If OD_{600} is lower than 0.4, the yield of the target protein will be low. If OD_{600} is higher than 0.7, it is hard to achieve cell growth arrest by MazF and therefore, the SPP condition cannot be established.
12. It is important to cool the culture quickly from 37°C to low temperature. A large amount of cold-shock proteins, including cspA which binds to the cspA promoter on pCold vectors, are produced during the rapid adaptation process to low temperature.
13. Always use a larger flask for the condensed culture than the regular culture, to provide enough aeration, for example a 5-ml culture of 20-fold condensation in a 250-ml flask, instead of a 30-ml test tube.
14. If the target has to be labeled with isotopes or amino acid analogues, add them in the M9 minimal medium of this step, for example replacing $^{14}\text{N-NH}_4\text{Cl}$ with $^{15}\text{N-NH}_4\text{Cl}$ and/or $^{12}\text{C-glucose}$ with $^{13}\text{C-glucose}$.

Acknowledgments

This work was supported by National Institutes of Health Grants, 5R01GM085449 (to M.I.).

References

1. Suzuki M, Zhang J, Liu M, Woychik NA, Inouye M (2005) Single protein production in living cells facilitated by an mRNA interferase. *Mol Cell* 18:253–261
2. Suzuki M, Roy R, Zheng H, Woychik N, Inouye M (2006) Bacterial bioreactors for high yield production of recombinant protein. *J Biol Chem* 281:37559–37565
3. Schneider WM, Tang Y, Vaiphei ST, Mao L, Maglaqui M, Inouye M, Roth MJ, Montelione GT (2010) Efficient condensed-phase production of perdeuterated soluble and membrane proteins. *J Struct Funct Genomics* 11:143–154
4. Mao L, Vaiphei ST, Shimazu T, Schneider WM, Tang Y, Mani R, Roth MJ, Montelione GT, Inouye M (2010) The *E. coli* single protein production system for production and structural analysis of membrane proteins. *J Struct Funct Genomics* 11:81–84
5. Mao L, Inouye K, Tao Y, Montelione GT, McDermott AE, Inouye M (2011) Suppression of phospholipid biosynthesis by cerulenin in the condensed Single-Protein-Production (cSPP) system. *J Biomol NMR* 49:131–137
6. Mao L, Tang Y, Vaiphei ST, Shimazu T, Kim SG, Mani R, Fakhouri E, White E, Montelione GT, Inouye M (2009) Production of membrane proteins for NMR studies using the condensed single protein (cSPP) production system. *J Struct Funct Genomics* 10:281–289
7. Suzuki M, Mao L, Inouye M (2007) Single protein production (SPP) system in *Escherichia coli*. *Nat Protoc* 2:1802–1810
8. Vaiphei ST, Mao L, Shimazu T, Park JH, Inouye M (2010) Use of amino acids as inducers for high-level protein expression in the single-protein production system. *Appl Environ Microbiol* 76:6063–6068

Chapter 12

Folding Engineering Strategies for Efficient Membrane Protein Production in *E. coli*

Brent L. Nannenga and François Baneyx

Abstract

Membrane proteins are notoriously difficult to produce at the high levels required for structural and biochemical characterization. Among the various expression systems used to date, the enteric bacterium *Escherichia coli* remains one of the best characterized and most versatile. However, membrane protein overexpression in *E. coli* is often accompanied by toxicity and low yields of functional product. Here, we briefly review the involvement of signal recognition particle, trigger factor, and YidC in α -helical membrane protein biogenesis and describe a set of strains, vectors, and chaperone co-expression plasmids that can lead to significant gains in the production of recombinant membrane proteins in *E. coli*. Methods to quantify membrane proteins by sodium dodecyl sulfate polyacrylamide gel electrophoresis are also provided.

Key words: Molecular chaperone, Insertase, Trigger factor, Signal recognition particle, YidC

1. Introduction

1.1. Membrane Proteins and Their Production in *E. coli*

Integral membrane proteins account for 20–30% of sequenced open reading frames and play essential roles in cell physiology and survival (1), including—but not limited to—signal transduction (2, 3), energy conversion (4, 5), and selective ion transport (6, 7). This class of proteins is important in the etiology of human diseases (8, 9) and it has been estimated that they are the target of nearly 60% of all pharmaceutical drugs, a yearly market of nearly \$50 billion (9). Membrane proteins also play an important role in the emerging field of nanobiotechnology (10), where they have been used as molecular motors (11), nanomachines for energy conversion and storage (12, 13), components of optical and electronic devices (14, 15), and templates for self-assembled nanostructures (16), sensors (17–19), and nanopores (20, 21).

With a few exceptions (e.g., bacteriorhodopsin in the purple membrane of *Halobacterium*), membrane proteins are not abundant and must be overproduced in bacteria, yeasts, or insect cells for biochemical or structural characterization. Among these expression systems, the gram-negative bacterium *Escherichia coli* is attractive due to its rapid growth, extensive characterization, and availability of a large number of genetic tools (22, 23). Unfortunately, membrane protein overexpression in *E. coli* is often accompanied by misfolding and cellular toxicity, leading to low yields of functional material. Here, we describe recently developed folding engineering strategies for enhanced expression of functional membrane proteins in the plasma membrane of *E. coli*.

1.2. Inner Membrane Protein Biogenesis in *E. coli*: The Players

1.2.1. The Signal Recognition Particle-Dependent Pathway

In *E. coli*, protein synthesis only takes place in the cytoplasm, but about 40% of all polypeptides are inserted into the inner and outer membranes, targeted to the periplasm or excreted into the growth medium (24). Most inner membrane proteins (IMPs) are delivered to the plasma membrane through the co-translational, signal recognition particle (SRP)-dependent pathway (Fig. 1; (25, 26)), which is also responsible for the export of secretory proteins containing highly hydrophobic signal sequences (27, 28). Bacterial SRP consists of Ffh, a 48-kDa GTPase encoded by *ffh* (for fifty-four homolog of the mammalian SRP54 subunit), and a stable,

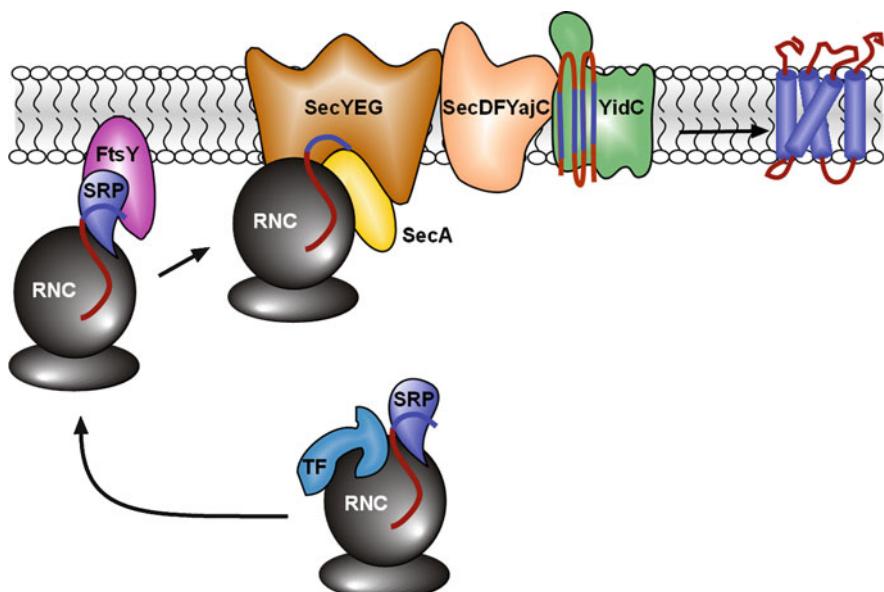


Fig. 1. SRP-dependent pathway for IMP biogenesis. As they emerge from the ribosome, the highly hydrophobic transmembrane segments (TMSs) of integral inner membrane proteins are preferentially bound by the signal recognition particle (SRP) rather than by trigger factor (TF). The ribosome nascent chain (RNC) complex is then targeted to the membrane-bound FtsY receptor and shuttled to the SecYEG translocon. YidC collaborates with the SecYEG/SecDFYajC system to integrate TMS within the inner membrane and also helps IMPs reach a properly folded conformation within the lipid bilayer.

114-nucleotides-long 4.5S RNA encoded by *ffs*. Both of these genes are essential in *E. coli* (29, 30). Ffh is a three domain protein that consists of an amino-terminal four helix bundle (the N domain), a Ras-like GTPase domain (the G domain), and a C-terminal methionine-rich M domain that binds to both its cognate 4.5S RNA and hydrophobic stretches of amino acids emerging from the ribosome (31).

The function of SRP is to identify nascent IMPs or SRP-dependent secretory proteins, bind these chains as they exit the ribosomal tunnel, and shuttle the resulting ribosome/nascent chain (RNC) complexes to the SecYEG translocon, which is responsible for IMP insertion into the inner membrane and translocation of secretory proteins to the periplasm with the aid of the SecA molecular motor (Fig. 1). To this end, the N domain of SRP interacts with the L23 ribosomal subunit, positioning the particle near the peptide exit tunnel on the ribosome (32–34). Once docking is achieved, the M-domain binds to emerging transmembrane segments (TMSs) of IMPs, or to the highly hydrophobic signal sequences of SRP-dependent secretory proteins. The SRP-RNC is next targeted to the inner membrane through interaction with FtsY, the membrane-associated SRP receptor (35, 36). Following GTP hydrolysis, FtsY and SRP dissociate from the RNC which becomes free to engage the SecYEG translocon for co-translational insertion of the growing chain into (or across) the inner membrane (37).

1.2.2. *YidC*

E. coli YidC is an inner membrane protein that associates with the SecYEG and SecDFYajC components of the translocon (Fig. 1). It is homologous to Alb3 from chloroplasts and Oxa1 from mitochondria (38), which can both substitute for it (39, 40). YidC spans the membrane six times and features a relatively large periplasmic domain between TMS 1 and 2 (41). It can act as a chaperone that promotes the proper folding of IMPs within the inner membrane (42), but also as an insertase that facilitates IMP transfer into the lipid bilayer (43). For the latter function, YidC may either interact sequentially with successive TMS associated with the lateral gate of SecYEG (41, 44), or act as a hydrophobic assembly site for groups of TMS, promoting membrane insertion only after core pre-assembly has taken place (45). Interestingly, YidC also acts as a SecYEG-independent insertase for the F_1F_0 ATP synthase subunit c (46), the viral M13 procoat protein (47), and a subset of other IMPs (41, 43).

1.2.3. Trigger Factor

Trigger factor (TF) is a molecular chaperone whose primary role is to shield the hydrophobic regions of certain newly synthesized polypeptides exiting from the ribosome, thus preventing their aggregation in the cytoplasm (48, 49). TF binds with moderate affinity to the L23 ribosomal protein near the SRP binding site, a position that allows it to sample nascent polypeptides emerging

from the ribosome (34, 50). Although it does not play a direct role in SRP-dependent trafficking, the close proximity of TF and SRP at the peptide exit site leads to competition for substrates. The outcome of this contest is based on the degree of hydrophobicity of emerging segments: SRP captures those nascent proteins exposing the most nonpolar segments (e.g., the TMS of IMPs), while TF captures less hydrophobic stretches exposed by slow folding or multidomain proteins (51). Because TF binds to 90% of the cell's ribosomes (52) and because TF and SRP can both bind to the L23 region (33, 53), TF continues to affect SRP-dependent trafficking until docking of the RNC to FtsY and initiation of co-translational translocation lead to its ejection (54).

In *E. coli*, the TF gene (*tig*) is not essential and can be inactivated without obvious phenotypical effects due to the fact that its cytoplasmic substrate pool overlaps with that of DnaK (55), an abundant and highly efficient molecular chaperone (56, 57).

1.3. Folding Engineering Strategies for Membrane Protein Expression

1.3.1. Expression Vectors

Toxicity and protein misfolding often accompany membrane protein overexpression in *E. coli*, and the use of strong and/or leaky promoters (e.g., P_{T7}) can intensify these negative side effects (58). Using a tightly regulated promoter abrogates preinduction toxicity, while using a weaker promoter leads to a decrease in the flux of overexpressed membrane proteins that will need to be inserted into the inner membrane through the SRP-dependent pathway, often mitigating post-induction toxicity (59). One such promoter is P_{BAD} from the arabinose operon (60) which has been used to express a variety of membrane proteins in *E. coli* (61–64).

1.3.2. Expression in Δtig Strains

We have shown elsewhere that *E. coli* cells lacking functional TF accumulate overexpressed IMPs (64) or SRP-dependent secretory proteins (65) at levels that are much higher than those present in the isogenic wild type. When improved cell fitness is taken into accounts, gains in productivity can be as high as sevenfold, leading to yield of tens to hundreds of milligram of protein per liter of culture in shake flasks (64, 65). This applies not only to endogenous IMPs as we show here with YidC (Fig. 2a), and previously demonstrated with the bitopic histidine kinase ZraS (64), but also to topologically complex heterologous heptahelical proteins such as *Haloterrigena turkmenica* deltarhodopsin (HtdR; Fig. 2b) and *Natronobacterium pharaonis* sensory rhodopsin II (64).

The precise mechanisms responsible for improved IMP expression in TF-deficient cells remain to be determined. However, because SRP and TF compete for nascent chains emerging from the ribosome, the flux of overexpressed membrane proteins that are properly directed to the SRP-dependent pathway is likely to increase in Δtig cells. Additionally, because the SecYEG docking site on the RNC overlaps with the ribosomal TF binding site (66–68), engagement of the translocon should be more efficient in Δtig strains.

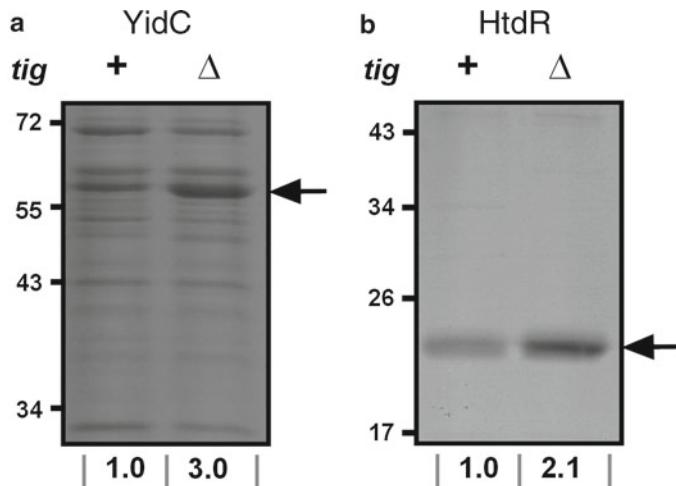


Fig. 2. Effect of trigger factor (TF) inactivation on the accumulation of autologous and heterologous IMPs. Isogenic wild-type (*tig*⁺) and TF-deficient (Δ *tig*) cells harboring pBLN200 derivatives encoding YidC (a) or HtdR (b) were grown in LB at 37°C and harvested 3 h post induction. In the case of YidC, membrane fractions normalized to identical amounts of cells are shown. In the case of HtdR, purified proteins obtained by Ni-NTA chromatography of membrane protein samples corresponding to identical amounts of cells are shown (see ref. 64 for a description of the purification protocol). Arrows indicate the migration position of YidC and HtdR. Numbers below the gels correspond to a videodensitometric quantification of the intensity of membrane protein bands in Δ *tig* cells relative to *tig*⁺ cells. Migration positions of molecular mass standards (kDa) are shown on the left of the gels.

1.3.3. Co-expression of YidC and SRP

The membrane-associated chaperone, YidC (Subheading 1.2.2), has been shown to play a crucial role in the proper folding of IMPs (42). Increasing the amount of YidC available to accept and process IMPs that are laterally transferred to the lipid bilayer by SecEYG should, thus, improve expression levels by promoting proper insertion and folding in the inner membrane. We have indeed observed that co-expression of YidC can double the yields of certain polytopic membrane proteins (Fig. 3; (64)). For unknown reasons, the beneficial effects of TF inactivation and YidC co-expression on IMP accumulation are not additive (e.g., Fig. 3). Therefore, YidC overproduction might be most appropriate when one seeks to express polytopic membrane proteins in wild-type (*tig*⁺) cells.

We have previously shown that co-expression of SRP can increase the translocation and yields of leech carboxypeptidase inhibitor (LCI) fused to an SRP-dependent signal sequence, a phenomenon that has been attributed to improved targeting of LCI RNCs to the SRP-dependent pathway (65). To our surprise, SRP co-expression had a deleterious effect on the accumulation of three model membrane proteins, irrespective of the genetic background (Fig. 3; (64)). We nevertheless believe that SRP co-expression might prove useful under different expression conditions and/or with other membrane proteins.

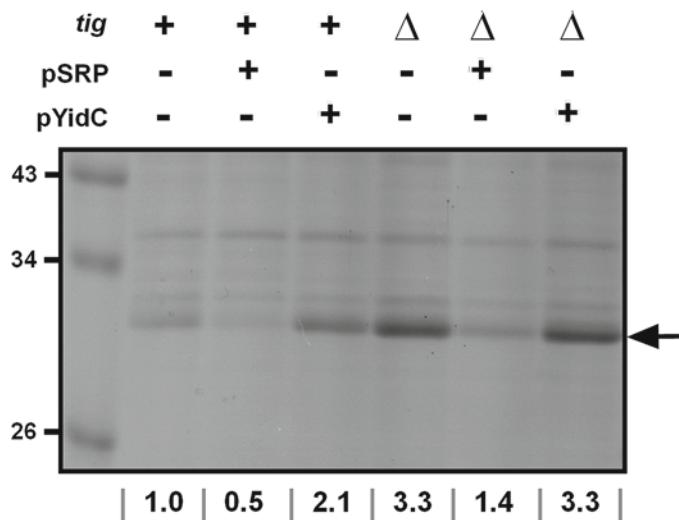


Fig. 3. Effect of SRP and YidC co-expression on HtdR accumulation. Isogenic wild-type (*tig*⁺) and TF-deficient (Δtig) cells harboring a pBLN200 derivative encoding HtdR and the indicated chaperone co-expression plasmids were grown in LB at 37°C and harvested 3 h post induction. Membrane fractions normalized to identical amounts of cells were analyzed by SDS-PAGE. The migration position of HtdR is indicated by an arrow. Numbers below the gels correspond to a videodensitometric quantification of the intensity of the HtdR protein bands relative to *tig*⁺/pMM102 control cells. Migration positions of molecular mass standards (kDa) are shown on the left of the gels.

2. Materials

2.1. Cell Growth

2.1.1. Growth Media

1. LB broth: mix 10 g of BD Bacto tryptone, 5 g of BD yeast extract, and 10 g of NaCl in 950 mL of ddH₂O. Stir to dissolve all solids and bring the final volume to 1 L with ddH₂O before autoclaving.
2. LB plates: add 15 g of agar (Sigma) per liter of LB broth prepared as above before autoclaving.

2.1.2. Antibiotics and Inducer

1. Carbenicillin: prepare a 50 mg/mL stock solution by dissolving 0.5 g of carbenicillin disodium (Sigma) in 10 mL of ddH₂O. Filter sterilize through a 0.2-μm membrane and store at -20°C in 1-mL aliquots. For a 50 μg/mL working solution, dilute 1:1,000.
2. Kanamycin: prepare a 50 mg/mL stock solution by dissolving 0.5 g of kanamycin sulfate (Sigma) in 10 mL of ddH₂O. Filter sterilize through a 0.2-μm filter and store at -20°C in 1-mL aliquots. For a 50 μg/mL working solution, dilute 1:1,000.
3. Arabinose: prepare a 20% (w/v) stock solution by dissolving 2 g of L(+)-arabinose (Sigma) in 8 mL of ddH₂O. Mix until all solid has dissolved, adjust the volume to 10 mL, and filter sterilize. To induce at a concentration of 0.2% (w/v), use a 1:100 dilution of the stock.

2.2. Expression Vector and Chaperone Co-expression Plasmids

2.3. Expression and Fractionation

2.3.1. Preparation of Membrane Fractions

Plasmids pBLN200, pMM102, pYidC, and pSRP are available upon request from Prof. François Baneyx, Department of Chemical Engineering, University of Washington, Box 351750, Seattle, WA 98195-1750 (E-mail: baneyx@uw.edu).

1. Potassium phosphate monobasic (50 mM): dissolve 6.8 g of KH_2PO_4 in 950 mL of ddH₂O, adjust the pH to 7.4 using HCl and the volume to 1 L. Store at 4°C.
 2. 4× Upper Tris–HCl buffer (500 mM): dissolve 30 g of Tris base in 400 mL of ddH₂O, adjust the pH to 6.8 with HCl and the volume to 500 mL. Store at 4°C.
 3. 1× Sodium dodecyl sulfate (SDS) loading buffer: dissolve 182 mg of dithiothreitol (DTT; Sigma) in 5.8 mL ddH₂O. Add 3 mL of 20% (w/v) SDS, 1.8 mL of 4× upper Tris buffer, 1 mL of 100% glycerol, and 0.4 mL 0.05% (w/v) Bromophenol blue. Store at room temperature.
 4. French pressure cell and press.
 5. Ultracentrifuge.
1. Buffer A (50 mM MES, pH 6.5, 300 mM NaCl, 1.0% *n*-dodecyl β-D-maltoside (DDM)): dissolve 0.976 g MES in 80 mL ddH₂O and adjust the pH to 6.5 with HCl. Dissolve 1.752 g NaCl and 1.0 g of DDM (Sigma) in this solution and adjust the volume to 100 mL. Store at 4°C.

3. Methods

3.1. Placing Membrane Protein Genes Under P_{BAD} Transcriptional Control

The vector pBLN200 (Fig. 4; (64)) is a pET24a(+) derivative (kanamycin-resistant, ColE1 *ori*), where the T7 promoter and a majority of the *lacI* gene have been replaced with a DNA segment from pBAD33 (60) encoding the arabinose-inducible P_{BAD} promoter and the *araC* regulatory gene. The ribosome binding site, multiple cloning site (MCS, see Note 1), and C-terminal hexahistidine tag from pET24a(+) are retained in pBLN200.

1. Design primers to amplify the target membrane protein by PCR. The forward primer should introduce an *Nde*I restriction site that overlaps with the ATG start codon at the 5' end of the amplified fragment and the reverse primer should introduce any of the other restriction sites present in the MCS of pBLN200 (see Notes 2 and 3).
2. Verify amplification was successful by performing agarose gel electrophoresis. Excise the PCR product from the gel using a razor blade and recover the DNA using a gel extraction kit.
3. Digest the PCR product and pBLN200 with *Nde*I and the restriction enzyme introduced at the 3' end of the membrane

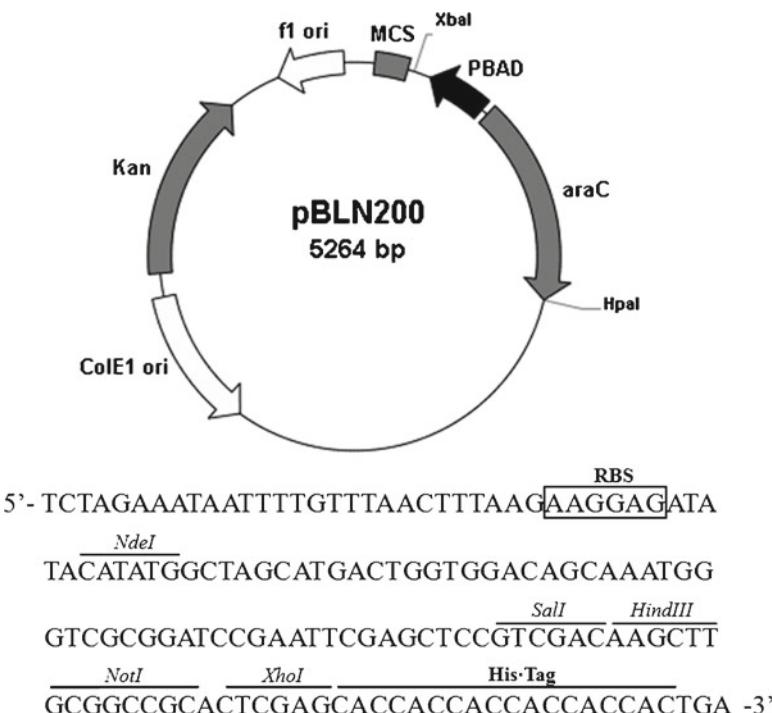


Fig. 4. Structure of expression vector pBLN200. The plasmid features the P_{BAD} promoter, a ColE1 origin of replication, kanamycin resistance, and the *araC* gene whose product regulates the promoter. The sequence of the ribosome binding site (RBS), multiple cloning site (MCS), and the C-terminal hexahistidine tag is shown below the map. All restriction sites shown are unique.

protein gene. Purify the digested vector and insert by agarose electrophoresis and extraction as in step 2.

- Join the two fragments using T4 DNA ligase and an insert-to-vector molar ratio of 4:1 (see Note 4) overnight at 18°C.
- Transform 1–5 µL of the ligation mixture into electrocompetent Top10 cells and incubate at 37°C for 1 h. Plate 100–200 µL of transformed cells on LB agar plates supplemented with 50 µg/mL kanamycin and incubate overnight at 37°C.
- Pick single colonies, grow in 5 mL LB-kanamycin medium, miniprep the plasmid DNA, and verify the presence of the correct insert using restriction digestion followed by DNA sequencing.

3.2. Construction of Δtig Strains

Marked and unmarked Δtig strains (see Note 5) can be constructed using PCR SOEing and λ Red-mediated recombination as described below and illustrated in Fig. 5.

- Using purified genomic DNA as a template, amplify a 452-bp-long 5' homology fragment with primers TigF1 (5'-GCCTCT CCTCCCTGTCGTGGAG-3') which hybridizes 36 bp

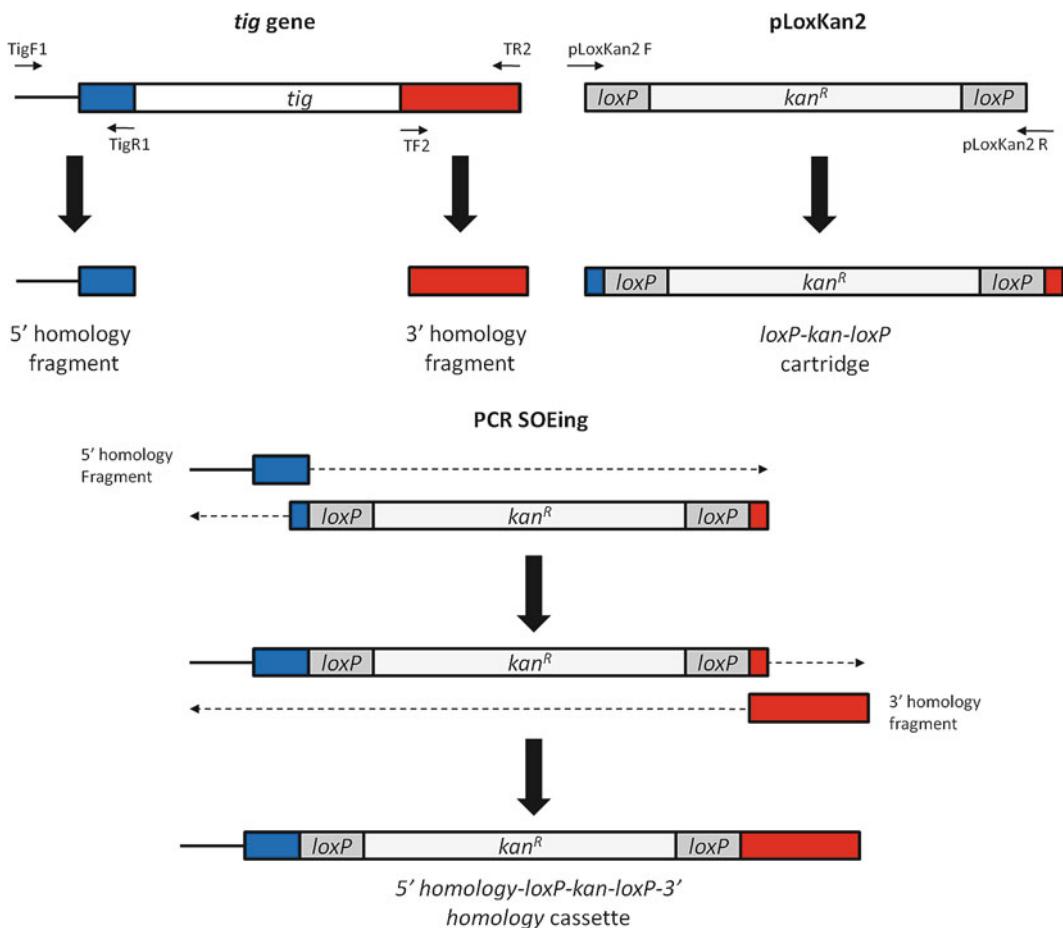


Fig. 5. Schematic illustration of the procedure used for the creation of Δtig strains. The figure corresponds to steps 1–6 of Subheading 3.2.

upstream of the *bola* stop codon, and TigR1 (5'-CTGTC AGCAGCGATAGTAATCG-3') which hybridizes 65 bp downstream of the *tig* start codon.

2. Using purified genomic DNA as a template, amplify a 487 bp-long 3' homology fragment with primers TF2 (5'-GAAGTGCATAAAACATGGAGCG-3') which hybridizes 804 bp downstream of the *tig* start codon, and primer TR2 (5'-TGGTTCATCAGCTCGTTGAAAG-3') which hybridizes 7 bp upstream of the *tig* stop codon.
3. Using plasmid pLoxKan2 as a template (69), primer pLoxKan2 R, and pLoxKan2 F amplify a 1,345-bp-long *loxP-kan^R-loxP* cartridge. Primer pLoxKan2 F (5'-CGATTACTATCGCTGCT GACAGATGCATATGGCGGCCGATAACTT-3') hybridizes 7 bp into the 5' *loxP* site of pLoxKan2 and contains a tail homologous to the last 22 bp of the 5' homology fragment,

and primer pLoxKan2 R (5'-CGCTCCATGTTTTACGCA CTTCGTATCGATAAGCTGGATCCATAAC-3') hybridizes 5 bp downstream of the 3' *loxP* site of pLoxKan2 and contains a tail homologous to the first 23 bp of the 3' homology fragment.

4. Run the three DNA fragments obtained in steps 1–3 (5' homology tail, 3' homology tail, and *loxP-kan-loxP* cartridge) on a 0.8% agarose gel followed by gel extraction and purification of the fragments.
5. Perform PCR using the 5' homology fragment and the *loxP-kan-loxP* fragment to generate DNA with the 5' homology fragment and the *loxP-kan-loxP* cartridge spliced together. Purify the resulting DNA using agarose gel electrophoresis and gel extraction as in step 4.
6. Using the 3' homology fragment along with the DNA obtained in step 5, perform PCR to generate a cassette containing the 5' homology region, *loxP-kan-loxP*, and the 3' homology region.
7. Purify the DNA cassette from step 6 by performing electrophoresis on a 0.8% agarose gel followed by gel extraction.
8. Introduce plasmid pKD46, which encodes the λRed recombinase system (70), into the strain from which the *tig* gene is to be removed. Make electrocompetent cells from the resulting transformant.
9. Transform the competent cells of step 8 with the purified DNA cassette from step 7 by electroporation. Incubate the cells for 3 h at 30°C, plate 50 μL on LB agar plates supplemented with 50 μg/mL neomycin, and transfer to a 37°C incubator for overnight growth.
10. Screen multiple colonies for the loss of pKD46 by duplicate streaking on LB agar plates supplemented with 50 μg/mL neomycin and LB agar plates supplemented with 50 μg/mL carbenicillin. Select carbenicillin-sensitive cells (which have lost pKD46) and perform colony PCR with primers TigF1 and TR2 to verify deletion of the *tig* gene. Select a colony containing the resulting $\Delta tig100::kan$ allele (65) for storage and subsequent steps.
11. In order to create an unm_lred *tig* strain, prepare competent cells from the $\Delta tig100::kan$ strain of step 10 and transform them with pJW168 (see Note 6). This plasmid encodes the *cre* recombinase which will remove genetic material located between consecutive *loxP* sites and contains a temperature-sensitive origin of replication (69). After overnight growth on LB agar plates (see Note 7) at 30°C, streak individual colonies in duplicate on LB agar and LB agar supplemented with 50 μg/mL neomycin; incubate overnight at 37°C. Choose colonies that

are kanamycin sensitive and grow them overnight in LB at 42°C to cure pJW168. Excision of the kanamycin resistance cartridge can be verified by PCR while elimination of pJW168 can be verified by sensitivity to carbenicillin.

3.3. YidC and SRP Co-expression

The genes encoding YidC and SRP have been cloned along with their native promoters in the Cole1-compatible plasmid pMM102 to yield pYidC (64) and pSRP (65) (see Note 8). These plasmids can be co-transformed with the pBLN200 expression vector (described in Subheading 3.1) to test the effect of YidC or SRP co-expression on target membrane protein production.

3.4. Expression and Fractionation Procedures

Before scaling up, optimal conditions for membrane protein expression (including, temperature, induction time, co-expression of chaperones, etc.) should be tested with small culture volumes. This section details how to conduct such tests with an emphasis on fractionation procedures.

3.4.1. Growth and Expression Conditions

1. Grow overnight cultures of *tig*⁺ or Δtig cells harboring the pBLN200 derivative encoding the target membrane protein, and chaperone co-expression plasmids pSRP, pYidC or control vector pMM102 (if desired) in 5 mL of LB supplemented with the appropriate antibiotics at 37°C.
2. Using 125-mL shake flasks, inoculate 25 mL of antibiotic-supplemented LB with overnight cultures so that the OD₆₀₀ is approximately 0.05 (see Note 9).
3. Grow the cells to an OD₆₀₀ of approximately 0.5 at 37°C (see Note 10) and induce membrane protein synthesis by adding L-arabinose to a final concentration of 0.2% (w/v). Record the OD₆₀₀ 3 h post-induction and harvest the cells (see Note 11). Collect a 1 mL sample for the preparation of whole cell fractions, and a 5 mL sample for the preparation of soluble, insoluble, and membrane fractions.

3.4.2. Preparation of Whole Cell Fractions

1. Centrifuge the 1 mL culture sample at 8,000 $\times g$ for 2 min in a microfuge.
2. Discard the supernatant and resuspend the cell pellet into (165 \times OD₆₀₀) μ L of 1× SDS loading buffer. Freeze at -20°C.
3. To visualize whole cell fractions, warm the tubes at 42°C for 5 min (see Note 12), vortex briefly, and load 15 μ L of sample on a 1-mm-thick, ten lanes SDS-PAGE minigel (or 10 μ L of sample on 1-mm-thick, 15 lanes gel).

3.4.3. Preparation of Membrane Fractions

1. Centrifuge the 5 mL culture sample at 5,000 $\times g$ for 10 min at 4°C. Discard the supernatant and resuspend the cell pellet in 3 mL potassium phosphate buffer by tapping.

2. Disrupt the resuspended cells by two cycles of lysis in a French press operated at 10,000 psi and 4°C.
3. Centrifuge the lysate from step 2 at $10,000 \times g$ and 4°C for 10 min to sediment aggregated (inclusion body) species. If the insoluble fraction is to be analyzed, save the pellet for further processing (see Subheading 3.4.4).
4. Transfer the supernatant from step 3 into 3.5-mL thick-wall, polycarbonate ultracentrifuge tubes (Beckman Coulter) and sediment the cell membranes by ultracentrifugation at $150,000 \times g$ and 4°C for 1 h (Beckman Coulter rotor TLA-100.3 or equivalent). If soluble fractions are to be analyzed, recover the supernatant and perform a methanol–chloroform extraction to isolate soluble proteins (71).
5. Resuspend the centrifuged membranes from step 4 in $(495 \times OD_{600}) \mu L$ of 1× SDS loading buffer, and freeze at -20°C. To visualize membrane fractions, warm the samples at 42°C for 5 min (see Note 12), vortex briefly, and load 15 μL of sample on a 1-mm-thick, ten lane SDS-PAGE minigel (or 10 μL of sample on 1-mm-thick, 15 lane gel).

3.4.4. Insoluble Fractions

1. Wash the pellet corresponding to insoluble material (see Subheading 3.4.3, step 3) by resuspending in Buffer A, pipetting up and down until the pellet is redissolved, and vigorous vortexing (see Note 13). Centrifuge at $10,000 \times g$ and 4°C for 10 min and discard the supernatant.
2. Repeat the above step twice.
3. Resuspend the insoluble material in $(495 \times OD_{600}) \mu L$ of 1× SDS loading buffer, and store at -20°C. To visualize membrane fractions, warm the samples at 42°C for 5 min, vortex briefly, and load 15 μL of sample on a 1-mm-thick, ten lane SDS-PAGE minigel (or 10 μL of sample on 1-mm-thick, 15 lane gel).

4. Notes

1. Not all unique restriction sites in the pET24a(+) MCS are unique in pBLN200. Sites that are no longer unique in pBLN200 are *Nhe*I, *Bam*HI, *Eco*RI, and *Sac*I. Sites that remain unique in pBLN200 and should be used for the cloning of target genes are *Nde*I, *Sal*II, *Hind*III, *Not*I, and *Xho*I.
2. We typically use *Xho*I as a restriction site at the 3' end of genes encoding target membrane proteins. In order to make use of the C-terminal hexahistidine tag, the reverse primer should be designed in such a way that it adds an *Xho*I site in-frame and

just before the stop codon. If a hexahistidine tag is not desired, the reverse primer should encode the gene's authentic stop codon before the 3' restriction site.

3. When designing primers, flank the restriction sites with a minimum of six extra base pairs to guarantee efficient digestion.
4. Although a ratio of insert to vector of 4:1 typically works well, it may be varied from 3:1 to 8:1 if the ligation is not initially successful.
5. Marked and unmarked *tig* variants of BW25113, MC4100, and BL21(DE3) can be requested from Prof. François Baneyx, Department of Chemical Engineering, University of Washington, Box 351750, Seattle, WA 98195-1750 (E-mail: baneyx@uw.edu). Marked strains are adequate donors to transfer the $\Delta tig100::kan$ allele to other genetic backgrounds via P1 transduction.
6. When transforming cells with pJW168, do not subject cultures to heat shock and perform all steps at 30°C since the plasmid encodes the temperature-sensitive *rep10*ts replication protein.
7. Leaky *cre* transcription in the absence of IPTG produces sufficient protein to remove genetic material located between *loxP* sites.
8. For expression vectors containing the p15A *ori*, pSRP2 ([65](#)) can be used for SRP co-expression. This plasmid can be requested from Prof. François Baneyx, Department of Chemical Engineering, University of Washington, Box 351750, Seattle, WA 98195-1750 (E-mail: baneyx@uw.edu).
9. This corresponds approximately to a 1:50 dilution of overnight cultures.
10. Cells are typically cultivated at 37°C. Transferring cultures to lower temperatures (15–25°C) before induction, however, may improve the accumulation of certain membrane proteins.
11. The length of the post-induction phase should be varied to determine the harvest time that maximizes target membrane protein accumulation. Three hours at 37°C is generally adequate. Nevertheless, larger amounts of material may be present in the membranes at longer incubation times, especially if the cultures are grown at low temperature.
12. Samples should only be heated at 45°C before loading onto SDS-PAGE gels since heating at 95°C affects the integrity and migration of certain membrane proteins.
13. Washing inclusion bodies helps remove contaminating proteins that adsorb to the aggregates but may not be themselves insoluble. The mild detergent in the wash buffer will not solubilize aggregated material, but will help remove contaminants and provide a cleaner insoluble fraction.

Acknowledgments

BLN gratefully acknowledges NSF-IGERT fellowship support from the University of Washington Center for Nanotechnology. This work was supported by NSF award BBE-0854511.

References

- Wallin E, von Heijne G (1998) Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci* 7:1029–1038
- Warne T, Serrano-Vega MJ, Baker JG, Moukhametzianov R, Edwards PC, Henderson R, Leslie AG, Tate CG, Schertler GF (2008) Structure of a betal-adrenergic G-protein-coupled receptor. *Nature* 454:486–491
- Bilwes AM, Alex LA, Crane BR, Simon MI (1999) Structure of CheA, a signal-transducing histidine kinase. *Cell* 96:131–141
- Neutze R, Pebay-Peyroula E, Edman K, Royant A, Navarro J, Landau EM (2002) Bacteriorhodopsin: a high-resolution structural view of vectorial proton transport. *Biochim Biophys Acta* 1565:144–167
- Capaldi RA, Aggeler R (2002) Mechanism of the F(1)F(0)-type ATP synthase, a biological rotary motor. *Trends Biochem Sci* 27:154–160
- Olesen C, Picard M, Winther AM, Gyrup C, Morth JP, Osvig C, Moller JV, Nissen P (2007) The structural basis of calcium transport by the calcium pump. *Nature* 450:1036–1042
- Gonen T, Walz T (2006) The structure of aquaporins. *Q Rev Biophys* 39:361–396
- Hopkins AL, Groom CR (2002) The druggable genome. *Nat Rev Drug Discov* 1:727–730
- McCusker EC, Bane SE, O’Malley MA, Robinson AS (2007) Heterologous GPCR expression: a bottleneck to obtaining crystal structures. *Biotechnol Prog* 23:540–547
- Curnow P (2009) Membrane proteins in nanotechnology. *Biochem Soc Trans* 37:643–652
- Soong RK, Bachand GD, Neves HP, Olkhovets AG, Craighead HG, Montemagno CD (2000) Powering an inorganic nanodevice with a biomolecular motor. *Science* (New York, NY) 290:1555–1558
- Choi HJ, Montemagno CD (2005) Artificial organelle: ATP synthesis from cellular mimetic polymersomes. *Nano Lett* 5:2538–2542
- Luo TJ, Soong R, Lan E, Dunn B, Montemagno C (2005) Photo-induced proton gradients and ATP biosynthesis produced by vesicles encapsulated in a silica matrix. *Nat Mater* 4:220–224
- Nakamura C, Hasegawa M, Yasuda Y, Miyake J (2000) Self-assembling photosynthetic reaction centers on electrodes for current generation. *Appl Biochem Biotechnol* 84–86:401–408
- Zhang L, Zeng T, Cooper K, Claus RO (2003) High-performance photovoltaic behavior of oriented purple membrane polymer composite films. *Biophys J* 84:2502–2507
- Mo X, Krebs MP, Yu SM (2006) Directed synthesis and assembly of nanoparticles using purple membrane. *Small* 2:526–529
- Gu LQ, Braha O, Conlan S, Cheley S, Bayley H (1999) Stochastic sensing of organic analytes by a pore-forming protein containing a molecular adapter. *Nature* 398:686–690
- Kang XF, Cheley S, Guan X, Bayley H (2006) Stochastic detection of enantiomers. *J Am Chem Soc* 128:10684–10685
- Cheley S, Gu LQ, Bayley H (2002) Stochastic sensing of nanomolar inositol 1,4,5-trisphosphate with an engineered pore. *Chem Biol* 9:829–838
- Kasianowicz JJ, Brandin E, Branton D, Deamer DW (1996) Characterization of individual polynucleotide molecules using a membrane channel. *Proc Natl Acad Sci USA* 93:13770–13773
- Kang XF, Cheley S, Rice-Ficht AC, Bayley H (2007) A storable encapsulated bilayer chip containing a single protein nanopore. *J Am Chem Soc* 129:4701–4705
- Baneyx F (1999) Recombinant protein expression in *Escherichia coli*. *Curr Opin Biotechnol* 10:411–421
- Wagner S, Bader ML, Drew D, de Gier JW (2006) Rationalizing membrane protein over-expression. *Trends Biotechnol* 24:364–371
- Weiner JH, Li L (2008) Proteome of the *Escherichia coli* envelope and technological challenges in membrane proteome analysis. *Biochim Biophys Acta* 1778:1698–1713
- Ulbrandt ND, Newitt JA, Bernstein HD (1997) The *E. coli* signal recognition particle is required for the insertion of a subset of inner membrane proteins. *Cell* 88:187–196
- Yuan J, Zweers JC, van Dijken JM, Dalbey RE (2009) Protein transport across and into cell

- membranes in bacteria and archaea. *Cell Mol Life Sci* 67:179–199
27. Schierle CF, Berkmen M, Huber D, Kumamoto C, Boyd D, Beckwith J (2003) The DsbA signal sequence directs efficient, cotranslational export of passenger proteins to the *Escherichia coli* periplasm via the signal recognition particle pathway. *J Bacteriol* 185:5706–5713
 28. Bowers CW, Lau F, Silhavy TJ (2003) Secretion of LamB-LacZ by the signal recognition particle pathway of *Escherichia coli*. *J Bacteriol* 185:5697–5705
 29. Brown S, Fournier MJ (1984) The 4.5S RNA gene of *Escherichia coli* is essential for cell growth. *J Mol Biol* 178:533–550
 30. Phillips GJ, Silhavy TJ (1992) The *E. coli* ffh gene is necessary for viability and efficient protein export. *Nature* 359:744–746
 31. Luirink J, Sinning I (2004) SRP-mediated protein targeting: structure and function revisited. *Biochim Biophys Acta* 1694:17–35
 32. Pool MR, Stumm J, Fulga TA, Sinning I, Dobberstein B (2002) Distinct modes of signal recognition particle interaction with the ribosome. *Science (New York, NY)* 297:1345–1348
 33. Ullers RS, Houben EN, Raine A, ten Hagen-Jongman CM, Ehrenberg M, Brunner J, Oudega B, Harms N, Luirink J (2003) Interplay of signal recognition particle and trigger factor at L23 near the nascent chain exit site on the *Escherichia coli* ribosome. *J Cell Biol* 161:679–684
 34. Gu SQ, Peske F, Wieden HJ, Rodnina MV, Wintermeyer W (2003) The signal recognition particle binds to protein L23 at the peptide exit of the *Escherichia coli* ribosome. *RNA* 9:566–573
 35. Luirink J, ten Hagen-Jongman CM, van der Weijden CC, Oudega B, High S, Dobberstein B, Kusters R (1994) An alternative protein targeting pathway in *Escherichia coli*: studies on the role of FtsY. *EMBO J* 13:2289–2296
 36. de Leeuw E, Poland D, Mol O, Sinding I, ten Hagen-Jongman CM, Oudega B, Luirink J (1997) Membrane association of FtsY, the *E. coli* SRP receptor. *FEBS Lett* 416:225–229
 37. Bibi E (2011) Early targeting events during membrane protein biogenesis in *Escherichia coli*. *Biochim Biophys Acta* 1801:841–850
 38. Luirink J, Samuelsson T, de Gier JW (2001) YidC/Oxa1p/Alb3: evolutionarily conserved mediators of membrane protein assembly. *FEBS Lett* 501:1–5
 39. Jiang F, Yi L, Moore M, Chen M, Rohl T, Van Wijk KJ, De Gier JW, Henry R, Dalbey RE (2002) Chloroplast YidC homolog Albino3 can functionally complement the bacterial YidC depletion strain and promote membrane insertion of both bacterial and chloroplast thylakoid proteins. *J Biol Chem* 277:19281–19288
 40. van Bloois E, Nagamori S, Koningstein G, Ullers RS, Preuss M, Oudega B, Harms N, Kaback HR, Herrmann JM, Luirink J (2005) The Sec-independent function of *Escherichia coli* YidC is evolutionary-conserved and essential. *J Biol Chem* 280:12996–13003
 41. Xie K, Dalbey RE (2008) Inserting proteins into the bacterial cytoplasmic membrane using the Sec and YidC translocases. *Nat Rev Microbiol* 6:234–244
 42. Nagamori S, Smirnova IN, Kaback HR (2004) Role of YidC in folding of polytopic membrane proteins. *J Cell Biol* 165:53–62
 43. Kol S, Nouwen N, Driessens AJ (2008) Mechanisms of YidC-mediated insertion and assembly of multimeric membrane protein complexes. *J Biol Chem* 283:31269–31273
 44. Houben EN, ten Hagen-Jongman CM, Brunner J, Oudega B, Luirink J (2004) The two membrane segments of leader peptidase partition one by one into the lipid bilayer via a Sec/YidC interface. *EMBO Rep* 5:970–975
 45. Beck K, Eisner G, Trescher D, Dalbey RE, Brunner J, Muller M (2001) YidC, an assembly site for polytopic *Escherichia coli* membrane proteins located in immediate proximity to the SecYE translocon and lipids. *EMBO Rep* 2:709–714
 46. van der Laan M, Bechtluft P, Kol S, Nouwen N, Driessens AJ (2004) F1F0 ATP synthase subunit c is a substrate of the novel YidC pathway for membrane protein biogenesis. *J Cell Biol* 165:213–222
 47. Samuelson JC, Jiang F, Yi L, Chen M, de Gier JW, Kuhn A, Dalbey RE (2001) Function of YidC for the insertion of M13 procoat protein in *Escherichia coli*: translocation of mutants that show differences in their membrane potential dependence and Sec requirement. *J Biol Chem* 276:34847–34852
 48. Maier T, Ferbitz L, Deuerling E, Ban N (2005) A cradle for new proteins: trigger factor at the ribosome. *Curr Opin Struct Biol* 15:204–212
 49. Hoffmann A, Bukau B, Kramer G (2010) Structure and function of the molecular chaperone trigger factor. *Biochim Biophys Acta* 1803:650–661
 50. Kramer G, Rauch T, Rist W, Vordewulbecke S, Patzell H, Schulze-Specking A, Ban N, Deuerling E, Bukau B (2002) L23 functions as a chaperone docking site on the ribosome. *Nature* 419:171–174
 51. Hoffmann A, Merz F, Rutkowska A, Zachmann-Brand B, Deuerling E, Bukau B (2006) Trigger

- factor forms a protective shield for nascent polypeptides at the ribosome. *J Biol Chem* 281:6539–6545
52. Patzelt H, Kramer G, Rauch T, Schonfeld HJ, Bukau B, Deuerling E (2002) Three-state equilibrium of *Escherichia coli* trigger factor. *Biol Chem* 383:1611–1619
 53. Buskiewicz I, Deuerling E, Gu SQ, Jockel J, Rodnina MV, Bukau B, Wintermeyer W (2004) Trigger factor binds to ribosome-signal-recognition particle (SRP) complexes and is excluded by binding of the SRP receptor. *Proc Natl Acad Sci USA* 101:7902–7906
 54. Buskiewicz I, Deuerling E, Gu S-Q, Jöckel J, Rodnina MV, Bukau B, Wintermeyer W (2004) Trigger factor binds to ribosome-signal recognition particle (SRP) complexes and is excluded by binding of the SRP receptor. *Proc Natl Acad Sci USA* 101:7902–7906
 55. Deuerling E, Patzelt H, Vorderwulbecke S, Rauch T, Kramer G, Schaffitzel E, Mogk A, Schulze-Specking A, Langen H, Bukau B (2003) Trigger factor and DnaK possess overlapping substrate pools and binding specificities. *Mol Microbiol* 47:1317–1328
 56. Baneyx F, Nannenga BL (2010) Chaperones: a story of thrift unfolds. *Nat Chem Biol* 6: 880–881
 57. Sharma SK, De los Rios P, Christen P, Lustig A, Goloubinoff P (2010) The kinetic parameters and energy cost of the Hsp70 chaperone as a polypeptide unfoldase. *Nat Chem Biol* 6:914–920
 58. Mujacic M, Cooper KW, Baneyx F (1999) Cold-inducible cloning vectors for low-temperature protein expression in *Escherichia coli*: application to the production of a toxic and proteolytically sensitive fusion protein. *Gene* 238:325–332
 59. Wagner S, Klepsch MM, Schlegel S, Appel A, Draheim R, Tarry M, Hogbom M, van Wijk KJ, Slotboom DJ, Persson JO, de Gier JW (2008) Tuning *Escherichia coli* for membrane protein overexpression. *Proc Natl Acad Sci USA* 105:14371–14376
 60. Guzman LM, Belin D, Carson MJ, Beckwith J (1995) Tight regulation, modulation, and high-level expression by vectors containing the arabinose PBAD promoter. *J Bacteriol* 177: 4121–4130
 61. Ren H, Yu D, Ge B, Cook B, Xu Z, Zhang S (2009) High-level production, solubilization and purification of synthetic human GPCR chemokine receptors CCR5, CCR3, CXCR4 and CX3CR1. *PLoS One* 4:e4509
 62. Hassan KA, Xu Z, Watkins RE, Brennan RG, Skurray RA, Brown MH (2009) Optimized production and analysis of the staphylococcal multidrug efflux protein QacA. *Protein Expr Purif* 64:118–124
 63. Romantsov T, Battle AR, Hendel JL, Martinac B, Wood JM (2010) Protein localization in *Escherichia coli* cells: comparison of the cytoplasmic membrane proteins ProP, LacY, ProW, AqpZ, MscS, and MscL. *J Bacteriol* 192: 912–924
 64. Nannenga BL, BaneyxF (2011) Reprogramming chaperone pathways to improve membrane protein expression in *Escherichia coli*. *Protein Sci* 20:1411–1420
 65. Puertas JM, Nannenga BL, Dornfeld KT, Betton JM, Baneyx F (2010) Enhancing the secretory yields of leech carboxypeptidase inhibitor in *Escherichia coli*: influence of trigger factor and signal recognition particle. *Protein Expr Purif* 74:122–128
 66. Kramer G, Rauch T, Rist W, Vorderwulbecke S, Patzelt H, Schulze-Specking A, Ban N, Deuerling E, Bukau B (2002) L23 protein functions as a chaperone docking site on the ribosome. *Nature* 419:171–174
 67. Menetret JF, Schaletzky J, Clemons WM Jr, Osborne AR, Skanland SS, Denison C, Gygi SP, Kirkpatrick DS, Park E, Ludtke SJ, Rapoport TA, Akey CW (2007) Ribosome binding of a single copy of the SecY complex: implications for protein translocation. *Mol Cell* 28:1083–1092
 68. Ataide SF, Schmitz N, Shen K, Ke A, Shan SO, Doudna JA, Ban N (2011) The crystal structure of the signal recognition particle in complex with its receptor. *Science* (New York, NY) 331:881–886
 69. Palmeros B, Wild J, Szybalski W, Le Borgne S, Hernandez-Chavez G, Gosset G, Valle F, Bolivar F (2000) A family of removable cassettes designed to obtain antibiotic-resistance-free genomic modifications of *Escherichia coli* and other bacteria. *Gene* 247:255–264
 70. Datsenko KA, Wanner BL (2000) One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc Natl Acad Sci USA* 97:6640–6645
 71. Baneyx F, Palumbo JL (2003) Improving heterologous protein folding via molecular chaperone and foldase co-expression. *Methods Mol Biol* 205:171–197

Chapter 13

Transient Expression Technologies: Past, Present, and Future

Sabine Geisse and Bernd Voedisch

Abstract

The first protocols describing transient gene expression in mammalian cells for the rapid generation of recombinant proteins emerged more than 10 years ago as an alternative to the establishment of stable, often amplified clonal cell lines, and relieved somewhat the bias against mammalian cell systems as being too complicated, labor intensive, and tedious to serve as a source for tool proteins in industrial research and academia. Over the past decade, these attempts have been refined and optimized, giving rise to expression protocols applicable in every lab in dependence on available tools, equipment, and envisaged outcome. This chapter summarizes the development of transient expression technologies over the past decade up to its current status and provides an outlook into what may be the future of transient technology development.

Key words: Recombinant protein expression, Transient transfection, HEK293 cells, CHO cells, Lipofection, Polyethylenimine

1. Introduction

The wealth of experimental data publicly available demonstrates impressively the success of recombinant protein production by transient transfection in mammalian cells. Driven by increasing insight into the process and incremental improvements, mature protocols, first for HEK293 cell lines, and more recently also for Chinese hamster ovary (CHO) cell-based expression, have been developed. This chapter aims at presenting a rather concise overview of current knowledge and future perspectives of transient transfection approaches with special attention to recently published data; additional information can be found in some recent reviews ([1–4](#)). Detailed descriptions of protocols are documented in refs. [5–7](#) as well as cited in the text.

2. The Past: Establishment of Standard Protocols for Transient Production of Recombinant Proteins

2.1. Cell Lines, Culture Media, and Expression Vectors for HEK293 Cells and Descendants

The most widely used host system in transient protein expression employs derivatives of the HEK293 cell line. Based on the human embryonic kidney cell line HEK293 established in 1977 by Graham et al. (8) via transformation with sheared Adenovirus 5 DNA, a number of engineered sublines were developed in the late 1990s by means of introduction of viral elements derived from SV40 virus or Epstein–Barr virus (EBV). These viruses have been characterized extensively for their ability to retain plasmid DNA in an episomal, nonintegrated state and for their potential to enhance transcription and translation via unique viral properties. As a result, two interacting components were identified: the SV40 large T-antigen binding to the SV40 origin of replication, *SV40ori*, and similarly the EBV-derived EBNA-1 protein and its counterpart *oriP*. Mechanistic details can be found in numerous publications and are summarized in recent papers (9–12).

Based on these scientific findings, the HEK293 T cell line carrying a stably integrated SV40 large T gene and the HEK.EBNA and 293-6E cell lines harboring an integrated EBNA-1 gene were derived, which give rise to high expression levels when transfected with expression plasmids featuring the corresponding origins of replication of the respective viruses.

The HKB-11 cell line, also belonging to the family of HEK293 descendants, is exceptional as being a hybrid originating from a fusion of HEK293 cells with the B-cell lymphoma line 2B8 (13, 14). The lymphoma cells originally harbored the EBV virus and thus the EBNA-1 gene, even though a recent assessment revealed that the HKB-11 cells have retained only very little of the EBV viral genome (S. Geisse (SG)). Nevertheless, their inherent properties to grow as single cells in suspension and their high secretory capacity render them attractive as host in transient transfection trials.

293 Freestyle™ cells (15) are non-engineered 293 cells carefully selected for high transfection efficiency and production, but their advantage is not correlated with any viral elements introduced.

293 Cells, when grown in fetal calf serum containing media in stationary culture, are moderately adherent and can be easily detached by tapping the culture flask. The division time approximates 30 h, dependent on culture conditions, and maximal cell densities under standard, non-optimized conditions reach $2\text{--}5 \times 10^6$ cells/ml. For transient production of proteins at large scale, adaptation to suspension growth under agitated conditions is mandatory, and along with the development of these protocols a large variety of serum-free culture media came to the market. Most of these support good cell growth and reduce cell aggregation in suspension—a natural disadvantage of 293 cells—however, not all

of them are compatible with large-scale transfection approaches, as explained in more detail below.

Unlike the rapid emergence of culture media, the commercial availability of suitable expression plasmids for transient transfection approaches has lagged behind. The ideal carrier of foreign genes should feature, apart from a strong promoter such as the CMV or EF1 α promoter and an intron splice element to enhance transcription, a selectable marker gene (to enable the selection of stable, episomal pools, in case this is desired) and the SV40 *ori* or *oriP* for interaction with SV40 large T or EBNA-1 gene products. For EBNA-based cell lines, the pCEP4 plasmid is being sold by Invitrogen/Life Technologies (Carlsbad, CA), which carries in addition to *oriP* also the EBNA-1 element—a suboptimal combination, as the EBNA-1 protein needs to be expressed first in order to bind to *oriP* (16). Additionally, this renders the plasmid backbone with 10 kB fairly large, a disadvantage in cellular and nuclear plasmid uptake (17). The pEAK series of plasmids (Edge BioSystems, Gaithersburg, MD), which is frequently described in the literature was unfortunately withdrawn from the market. The well-performing pTT series of plasmids for EBV-mediated enhanced protein production (18) is available under license from NRC Montreal, Canada. For SV40-enhanced protein expression, plasmids such as the pcDNA™ series (Invitrogen/Life Technologies) can be used, available in a variety of different flavors, as they display the SV40 *ori* in the context of an SV40 promoter driving the selectable marker gene. As a note of caution: In case the establishment of stable episomal pools is envisaged, the choice of the selectable antibiotic resistance marker gene on the expression plasmid is important, as, e.g., the 293-6E and 293T cells are already resistant to Geneticin™ (G418) (18).

In very few publications, expression of different genes was systematically compared in this family of HEK293 derivatives, but it seems that EBV-engineered lines are superior to SV40-engineered cells and wild-type cells (4, 19, 20).

2.2. CHO Cells and Derivatives

CHO cells were among the first cell lines established for in vitro cultivation, but their popularity as a host for the production and manufacturing of biological products is undiminished and undisputed. Sublines engineered by chemical treatment or radiation to allow for gene amplification procedures via treatment with methotrexate (MTX) or methionine sulfoximine (MSX) (CHO DUK X B11, DG44, CHO GS cell lines) have a long-standing history as well. One should assume that aligning early activities in research with the establishment of stable manufacturing cell lines for biologics should favor the use of CHO cells, but in fact attempts to engineer or apply CHO cell lines for transient protein production have only gained recognition more recently. Three expression systems based on CHO cells have been described to date: an

EBNA-1-engineered CHO cell line (CHO EBNA1c-3E7) patented by Durocher and Loignon (21); a CHO EBNA LT cell line carrying apart from the EBNA-1 gene also the mouse polyoma virus large T antigen (QMCF system, Icosagen SA (22)); and the *EpiCHO* system described by Codamo et al. (Acyte (23)), a mouse polyoma virus large T antigen transformed CHO cell line. The latter is used in conjunction with expression plasmids displaying apart from the mouse polyoma origin of replication also the human EBV elements EBNA-1 and *oriP*, while expression vectors for the QMCF system bear the polyoma virus origin of replication (Py*ori*) and a partial *oriP* only. For CHO EBNA1c-3E7 cells, the pTT plasmid (harboring the *oriP*) described above can be employed.

The CHO Freestyle Max™ system sold by Invitrogen/Life Technologies relies again on a non-engineered CHO subclone selected and optimized for transfection efficiency and expression. A number of publications, e.g., several papers from the group of Wurm et al. also show promising results using classical CHO DG44 cells as host (24–26), but mostly in association with other production-enhancing mechanisms, such as co-transfection of cytokine genes and chaperones, and treatment of the cultures with histone-modifying reagents (see Subheading 2 below for additional details).

The adaptation of naturally very adherent CHO cells to suspension culture, choosing from the large set of commercially available serum-free, protein-free, or chemically defined culture media, has long been achieved and allows cultivation of wild-type and engineered CHO cells under agitated suspension conditions on small and large scale. High cell densities of 5×10^6 – 1×10^7 cells/ml associated with a cell division time of approx. 18 h can be achieved. Yet, one needs to ensure that the culture medium used is compatible with large-scale transfection reagents such as PEI, an endeavor frequently associated with empirical testing, as the media compositions are proprietary to the vendors.

A summary of the most commonly used HEK293 and CHO cell lines and their sources is given in Table 1.

2.3. Transfection Reagents and Transfection Efficiency

2.3.1. Transfection Based on Cationic Lipid Formulations

Especially if multiple plasmid constructs require rapid assessment of suitability for expression and scale-up, e.g., selected antibody candidates or expression of single domains of proteins, a quick and easy expression trial can be performed using lipofection as transfection reagent. The suspension cell culture can be transfected in plate format, Tubespins™ reactors (27) or Erlenmeyer shake flasks, non-agitated or agitated, in multi-milliliter scale with or without subsequent small-scale purification, thus allowing the testing of a variety of different recombinant vectors and conditions simultaneously. Although expensive, the transfection efficiency of lipofection reagents ranges from mostly acceptable to high efficiency, allowing also the detection of proteins expressed at low levels only.

Table 1
Overview on cell lines and expression systems for transient protein expression

	Features of cell line	Described cultivation medium	Used in conjunction with plasmid vectors	Comments	References
<i>HEK293 lines</i>					
HEK EBNA (HE, 293-EBNA)	EBNA-1 transformed (Invitrogen)	Suspension : e.g. ExCell 293 (SAFC Biosciences), Freestyle 293 (Invitrogen)	pCEP4 (Invitrogen) PEAK8, pcDNA 3.1 pTT (NRC Canada)	Most commonly used cell line in literature, off market	Many, citations e.g. in refs. 1, 2
HEK EBNA (HE, 293-EBNA) 293-SFE (293SF-3F6, NRC)	EBNA-1 transformed (ATCC CRL-10852) EBNA-1 transformed	Adherent: DMEM+10% FCS + 400 µg/ml G418 Suspension: Hybridoma SF medium+1% BCS	Same as above pTT plasmid series	Suspension cultivation not described System licensed out from NRC	ATCC/Stanford University (73)
HEK293 T	SV40 T-antigen transformed (ATCCCRL-11268)	Adherent: DME/M+10% FCS, suspension not routinely done	pCMV/myc/ER (Invitrogen)+derivatives	Common in many labs	(46)
293 Freestyle (293-F)	HEK293 wildtype	Suspension: Freestyle medium (Invitrogen)	pcDNA 3.1, pCMV SPOR T	Subselected cell line for growth in Freestyle medium	(15) manual on website of Invitrogen (13)
HKB-11 (hybrid of kidney and B-cell)	Fusion of 293/B-Cell lymphoma cell line (ATCC CRL-12568)	Suspension: Bayer proprietary pTAT/TAR vector medium	(Bayer)	System licensed out from Bayer Healthcare	
<i>CHO cell lines</i>					
CHO EBNA1c Clone 3E7	EBNA-1 transformed	Freestyle CHO (Invitrogen)	pTT plasmid series	System licensed out from NRC, Montreal Canada	US 2011/0039339 patent application (21)
CHO EBNA LT	EBNA-1 /mouse polyoma virus large T antigen transformed	CD CHO:293 SFM II 1:1 (Invitrogen)	pQMCF plasmids (Icosagen)	System licensed out from Icosagen AS, Tartu, Estonia	(22) patent EPI8513139 patent US7,790,446
CHO T (<i>Epi</i> -CHO)	Mouse polyoma virus large T antigen transformed	CHO-S SFM II (Invitrogen) and others	Vector comprises Pyori/orIP/EBNA-1	System available from Acyte Biotech, Brisbane, AU	(23)
CHO Freestyle MAX™	CHO K1 subline (CHO S)	CHO Freestyle Max™	CMV-based	System commercially available	Invitrogen user manual

The large variety of reagents on the market, each of which is claimed to be the best, makes it difficult to give a recommendation of superiority. However, HEK293 cells are very amenable to the uptake of foreign DNA, achieving transfection efficiencies of >80%, if a protocol optimized for cell density and ratio of plasmid DNA to reagent is applied. We have obtained transfection efficiencies well above 90% using, e.g., FuGene HD (Roche) or Lipofectamine 2000 (Invitrogen). Detailed protocols can be found in refs. 6, 7.

In contrast, transfection of CHO cells unexpectedly proves to be much more difficult. There is a striking dependence on the individual CHO subline transfected—possibly related to the origin and age in culture of the particular cell line and associated genotypic, epigenetic, and regulatory aberrations (28). Additionally, the composition of the culture medium, in conjunction with cell line and reagent, appears to be of critical importance. An overview on transfection efficiencies using two different CHO cell lines and different lipofection reagents is given in Table 2. As a recommendation, it is probably wise to test several reagents combined with different culture media in conjunction with the CHO cell line available in the lab to achieve optimal results.

2.3.2. PEI-Mediated Transfection: The Option for Large-Scale Transfection

It is obvious that the cost of goods precludes the use of lipofection reagents on large to very large scale, i.e., above several hundred milliliters, and thus part of the success of transient transfection technology can be attributed to the discovery of polyethylenimine (PEI) as a cheap transfection reagent by Boussif et al. (29), which can be readily obtained in bulk quantities. Originally evaluated as DNA carrier in gene therapy trials, PEI has become the reagent of choice in large-scale transient transfection, as reflected by a huge number of publications. Some key parameters essential for success when using PEI-mediated gene transfer are discussed below.

1. PEI transfection solution: Despite the wealth of PEI formulations with different molecular weights available and some conflicting data on results, the most commonly used PEI in large-scale transient transfection of HEK293 and CHO cells is linear 25 kDa PEI (PolySciences, Warrington, PA) (30–32). Preparation of a stock solution of 1 mg/ml in water, pH adjustment to 7.0, and sterile filtration are done as described by Durocher et al. (18). The stock solution should be stored in aliquots at -80°C prior to use and repeated freeze/thaw cycles should be avoided, as they seem to impact transfection efficiencies and yields (25). A fully deacetylated 40 kDa PEI formulation (sold as PEI MaxTM, also by PolySciences) has been claimed to improve transfection efficiency and productivity in CHO cells, but not in HEK293 cell lines (21).

Table 2
Transfection efficiencies for CHO cells achieved with different cell lines, media and reagents

Transfection reagent	Cell lines: CHO K1 Novartis (left column)/CHO EBNALT85 Icosagen (right column)				
	Culture medium	Novartis Medium 1	Novartis Medium 2	UltraCHO™ (Lonza)	ProCHO4™ (Lonza)
Lipofectamine 2000 (Invitrogen)	35%/94%	<1%/11%	<1%/1%	n.d./<1%	
LTX (Invitrogen)	n.d./71%	n.d./88%	n.d./13%	n.d./19%	
GeneJammer (Stratagene)	2%/4%	<1%/7%	12%/7%	<1%/7%	
Fecturin (Polyplus)	2%/8%	<1%/<1%	<1%/2%	<1%/<1%	
FuGene HD (Roche)	4%/17%	<1%/19%	n.d./54%	n.d./21%	
PEI 25 kDa lin. (Polysciences)	20%/40%	43%/65%	20%/60%	4%/8%	

Cell lines were adapted to growth in the respective media and transfected with an eGFP expression plasmid according to the recommendations of the manufacturer of the lipofection reagent. Where possible, amounts of cells and reagents used were aligned between different reagents. For PEI-mediated transfections the protocol described in ref. 7 was applied. Forty-eight hours post transfection cells were analyzed by FACS analysis. Only viable single cells were used for analysis. The transfection efficiency was calculated as the percentage of cells showing eGFP fluorescence. A mock-transfected control served as negative control. Analyses were done in duplicates. The choice of the commercial media for analysis was based on the publication by Ye et al. (40)

2. Several parameters govern the success of PEI-mediated transfection, most importantly the absolute quantity of DNA transfected, the DNA:PEI ratio, and cell density at transfection. A ratio of 1:2–1:3 ($\mu\text{g}:\mu\text{g}$) of DNA:PEI is frequently applied with success (5, 33), but also higher ratios have been reported (25, 32, 34). The latter may lead to an increase in transfection efficiency, however frequently accompanied by cell toxicity effects. Partial replacement of coding plasmid DNA with stuffer DNA from, e.g., salmon sperm, does not necessarily lead to a reduction in expression efficiency, and reduces the amount of plasmid DNA required to be prepared (35).
3. The concentration of viable cells at transfection represents another variable. While densities of $0.5\text{--}2.0 \times 10^6$ cells/ml are easily achievable for both HEK293 and CHO cells, transfection

at high cell densities ($1\text{--}2 \times 10^7$ cells/ml) (33, 36) inevitably requires a concentration step, either by centrifugation or by sedimentation of cells cultivated on microcarriers, as recently proposed (37). In summary, there is a distinct interplay between all three parameters, plasmid concentration, PEI concentration, and cell density, and individual fine-tuning is required for the individual cell lines to obtain optimal results.

4. The formation of the DNA:PEI polyplexes with respect to maturation time and resulting size was considered quite important for DNA uptake (38, 39) until most recently it was shown that direct addition of DNA followed by PEI to the cells similarly gives rise to good transfection efficiency and productivity, bypassing additional handling steps and waiting times (5, 25, 36).
5. Moreover, the composition of the cultivation medium impacts heavily the success of PEI-mediated transfection. Several authors have convincingly shown that many commercially available cell culture media support good growth and cell viability of HEK293 and CHO cells, but impede PEI-mediated transfection (33, 36, 40). While media components such as dextran sulfate or heparin for reducing cell aggregation were discussed as potential causes of lower transfection efficiency, a recent publication by Eberhardy et al. reported the inhibition of PEI-mediated transfection in CHO cells by iron(III) citrate, which replaces the more expensive transferrin in serum-free medium (41).

HEK293 cell lines were shown to exhibit a transfection efficiency of approx. 60% upon PEI transfection, independently of the scale, and thus transfections done on large scale in WAVE bioreactors and orbital shakers proved to be very successful, giving rise to titers ranging from multi-milligram up to gram per liter of recombinant proteins. Details on procedures and protocols can be found in the following refs. 5–7, 33, 42, 43.

For PEI-mediated transfection of CHO cells, there is currently less information publicly available, but the general notion is that the transfection efficiency and protein quantities retrieved from CHO cells are three- to fivefold lower than those from transiently transfected HEK293 cells (31, 44, 45). Yet, very few side-by-side comparisons have been published, and in contrast a comparison of 58 proteins expressed in both HEK293-6E and CHO-E7 cells gave in 67% of all cases better titers when an optimized CHO process was employed (Y. Durocher, oral communication at MipTech Conference, Basel, September 2010). We have obtained a PEI-mediated transfection efficiency of approx. 40–60%, in dependence on the CHO cell line and the cell culture medium tested, but also observed in some cases higher product titers or improved protein quality of candidates produced in CHO cells rather than in HEK293 cells.

A few attempts have been made using polymers other than PEI, such as DEA, chitosan, and chitosan–PEI blends (46–48), to replace or complement PEI as transfection reagent, but without a major breakthrough. The only feasible alternative to PEI is calcium phosphate-mediated transfection. Despite being experimentally of age, the approach has its merits with respect to successful outcomes (34), but necessitates usually a change of medium prior to transfection. Suspension media commonly feature a low calcium content to reduce cell aggregation, which leads to precipitation of the calcium phosphate complexes. Therefore, the applicability of this protocol on large scale is rather restricted.

One of the true advantages of transient protein production is the flexibility of scale, ranging from small, i.e., microliter to milliliter scale (5) to large scale of 100 l and beyond (6, 30, 42, 45, 49), allowing a tailored approach to protein production in dependence on titers, requested amounts, and available equipment. Using a plate or flask format, this approach is also suitable for automation in case very-high-throughput applications are envisaged (50, 51).

3. Current Status: Improving Yields by Vector Design, Co-expression, and Process Strategies

Even though the basics of transient transfection, i.e., cell lines, media, and transfection protocols, have been well established for several years, recent publications have identified other points of intervention for further enhancing productivity, in particular in CHO cells. As said, the design and availability of optimized expression plasmids represents still a relevant bottleneck. One of the most frequently used promoters for transgene expression, the human Cytomegalovirus major immediate early promoter (hCMV MIE) and its regulatory elements, was recently dissected for the impact of the individual elements on transient and stable expression in HEK293 and CHO K1 cells, and this confirmed previous findings (52) that inclusion of the first exon and intron A results in increased mRNA and protein expression levels in both cell lines upon transient transfection (53). Similarly, the 5' untranslated region (UTR) of Adenovirus 5, the tripartite leader sequence linked to a major late promoter/enhancer gave rise to enhanced expression levels in both cell lines, whereas including the post-transcriptionally active regulatory element, WPRE, derived from the woodchuck hepatitis virus, enhanced expression only in HEK293 cells in a gene- and promoter-dependent fashion (54–57). Combining elements, such as an intron splice element and WPRE, leads to cumulative effects (54, 57).

Preventing epigenetic silencing by maintenance of an open chromatin structure, resulting in high-level, stable expression has been discussed and described for many years in the context of

establishment of manufacturing cell lines (58). In a transient setting, this appears less relevant due to the episomal nature of the plasmid, yet some reports indicate that transgene expression may also be enhanced when, e.g., certain MAR elements are included on a (modified) plasmid backbone (59).

Along these lines, the treatment of transfected HEK293 cells, and to a lesser extent also CHO cells with inhibitors of histone deacetylases, such as valproic acid (VPA) and sodium butyrate (NaBut) as the most effective ones, has shown beneficial effects on expression rates. VPA, a compound with pleiotropic effects on many cell types, in particular tumor cells, enhances gene transcription rates directly by inhibiting histone deacetylation, thus increasing mRNA levels (60), leading to enhanced protein production in HEK293 and CHO cells (21, 24, 61). In contrast to sodium butyrate, a well-known histone deacetylase inhibitor (62, 63), VPA exhibits less pronounced cytotoxic effects. The search for other, novel, small-molecular-weight compounds is still ongoing (64).

Co-transfection of plasmids intended to boost expression rates along with the actual expression vector appears to be particularly relevant for CHO cell-based expression. For instance, expression of protein kinase B (PKB/Akt) delays the onset of programmed cell death upon nutrient depletion of cells (65); in concert with VPA, PKB overexpression compensates for the inhibition of PKB/Akt induced by VPA (66) and enhances its effect (21). Co-expression of fibroblast growth factors (FGFs), a large family of factors implicated in a variety of diseases and with huge potential as biologics in vivo (67), has also been tried. Co-transfection of basic FGF (FGF-2) plasmid leads to reduced expression rates in HEK293 and enhanced expression in CHO cells (21, 68), while co-expression of acidic FGF (FGF-1) increases expression rates in both HEK293 and CHO DG-44 cells (26).

As a stand-alone approach or in combination with these co-expression efforts, CHO cultures can be maintained under mild hypothermic conditions, i.e., at 30–32°C instead of 37°C during the production phase, leading to several-fold increased expression rates, as shown by several groups independently (21, 31, 69–71). Nota bene, hypothermia does not exert any enhancing effects when using HEK293 cells.

In comparison to processes developed for manufacturing cell lines, feeding strategies applied during the production phase by, e.g., addition of peptones (72) or recombinant insulin-like growth factor (LR3-IGF) (31) have not been studied systematically and in depth, but beneficial effects of a fed-batch versus a batch process have been observed (40, 73).

Last but not least, the expression of monoclonal antibodies versus other proteins by transient means should be addressed. As antibodies reflect a class of molecules with intrinsically high secretory capacity, there is no need to include a heterologous signal

peptide in the expression plasmid to improve secretion levels. Secondly, antibodies are in most instances inherently stable proteins, a feature influencing the length of the production phase. Frequently, a production phase of more than 10–14 days post (transient) transfection has been described without any negative effects on protein integrity, while for some antibodies and regular proteins a susceptibility to protein degradation due to cell death and release of endogenous proteases has been observed (74, 75). To successfully express an antibody, two options for vector construction are possible: the integration of both heavy (HC) and light chain (LC) cassettes on one vector only, driven either by two independent promoters or, more recently, one CMV enhancer element controlling two truncated core promoters (76). Alternatively, a co-transfection approach using two individual plasmids can be performed. While this inevitably necessitates the generation of two plasmid preparations at large scale, the ratio of both plasmids can be modulated. In most instances, a 1:1 (w/w) ratio leads to good results, but exceptions to this rule have also been reported (77, 78). Overall, transient approaches have been applied not only to expression of full antibodies, but also single-chain antibodies and Fab fragments (79, 80), as well as single-domain antibodies (81). Here, optimization may also be the key to success (own results, see Fig. 1).

4. Future Directions of Transient Transfection Technologies

Up to today, the main focus of transient expression approaches has been on employing members of the HEK293 cell family (4) and CHO cell lines, but remarkably little has been published on side-by-side comparisons. A recent review by Durocher and Butler (3) as well as a few single publications (82–85) report differences in expression levels and biological activity *in vitro* and *in vivo* linked to the capacity of individual host cell lines for secondary modifications such as glycosylation patterns based on stably transfected cell lines, but similar differences can be observed with proteins derived from transient approaches. Thus, making the appropriate choice of a host cell system for a novel protein target is currently a question of availability of systems and empirical testing. Furthermore, it is unlikely that these two cell systems and their derivatives will be capable of expressing *all* target proteins, and therefore the emergence of newly established cell lines with different cellular origin and shorter cultivation history may be extremely beneficial for production of research proteins. Examples of these are the Per.C6 cell line derived from human embryonic retina cells (86, 87), the AGE1.HN cell line established out of a human brain tissue sample (88), the CAP/CAP-T cells immortalized from

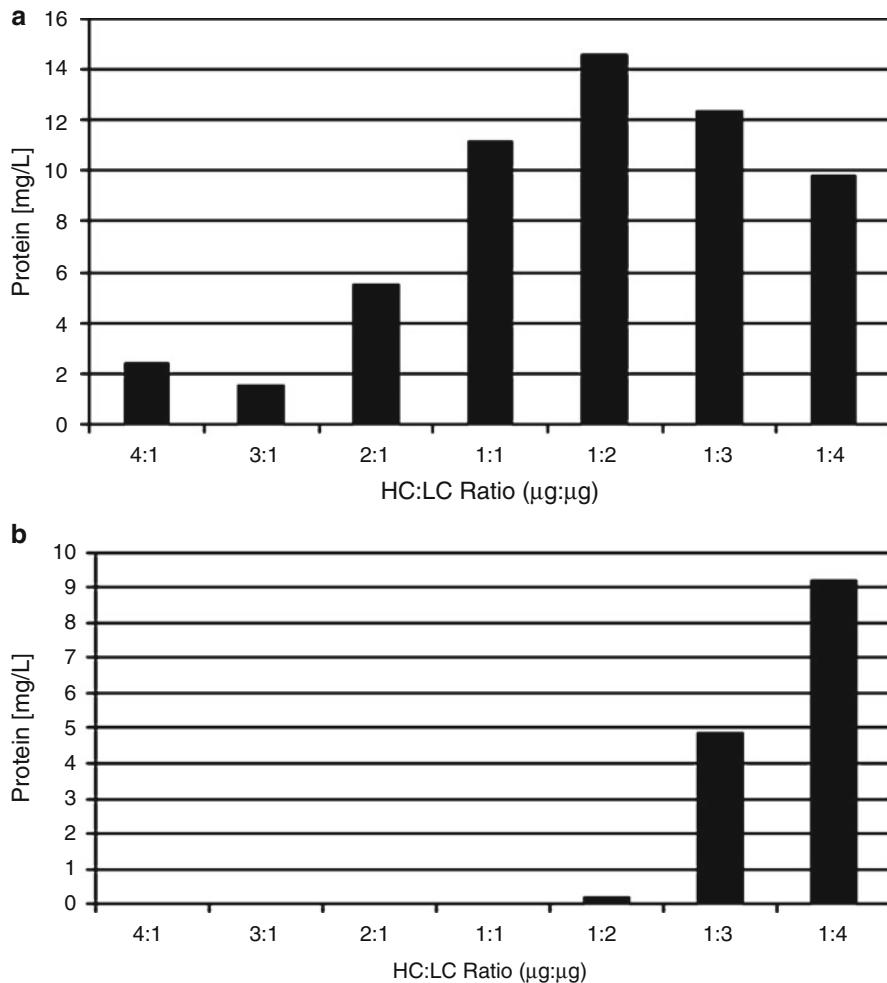


Fig. 1. Two plasmid constructs (panels **(a, b)**) differently designed for expression of a Fab fragment directed against an antigen from the Wnt signaling pathway were transiently expressed by lipofection with FuGene HD in 293-6E cells. The ratio of HC to LC plasmids was varied as shown. Analysis of expression was carried out in duplicates and titers were determined 3 days post transfection by Protein A HPLC. Subsequently, expression was scaled up to a 10 L Wave™ bioreactor applying the optimal conditions found. 228 and 450 mg of Fab proteins could be retrieved after purification.

human amniocytes (89), and the avian cell line EB66 (90). Except for the CAP-T expression system, these cell lines have not been fully exploited for transient expression approaches.

Cellular engineering of CHO cells to improve cell growth, viability, and productivity has been pursued for many years (91), but most recently a new stage has been set with the availability of the CHO miRNA transcriptome (92, 93) and the option to tackle individual genes and pathways specifically for optimized performance (94). Unraveling the genome of CHO cell lines (and the hamster genome as reference) expected to be publicly available soon will again add another layer of information, but also complexity

and challenge to these approaches. As transient expression trials do not differ extensively in this respect, additional information on how to optimize protocols and expression rates will also benefit from these scientific findings.

Acknowledgments

We would like to thank our team in the lab for the generation of the experimental data and our colleague Dr. Peter LeMotte, Novartis Cambridge, MA, for critically reviewing the manuscript.

References

- Pham PL, Kamen A, Durocher Y (2006) Large-scale transfection of mammalian cells for the fast production of recombinant protein. *Mol Biotechnol* 34:225–237
- Baldi L, Hacker DL, Adam M, Wurm FM (2007) Recombinant protein production by large-scale transient gene expression in mammalian cells: state of the art and future perspectives. *Biotechnol Lett* 29:677–684
- Durocher Y, Butler M (2009) Expression systems for therapeutic glycoprotein production. *Curr Opin Biotechnol* 20:700–707
- Geisse S (2009) Reflections on more than 10 years of TGE approaches. *Protein Expr Purif* 64:99–107
- Raymond C, Tom R, Perret S, Moussouami P, L'Abbe D, St-Laurent G, Durocher Y (2011) A simplified polyethylenimine-mediated transfection process for large-scale and high-throughput applications. *Methods* 55(1):44–51
- Geisse S, Jordan M, Wurm F (2005) Large-scale transient expression of therapeutic proteins in mammalian cells. In: Smales C, James D (eds) *Therapeutic proteins*. Humana Press, Totowa NJ, pp 87–98
- Geisse S, Fux C (2009) Recombinant protein production by transient gene transfer into mammalian cells. *Methods Enzymol* 463:223–238
- Graham FL, Smiley J, Russell WC, Nairn R (1977) Characteristics of a human cell line transformed by DNA from human adenovirus type 5. *J Gen Virol* 36:59–72
- Kishida T, Asada H, Kubo K, Sato YT, Shin-Ya M, Imanishi J, Yoshikawa K, Mazda O (2008) Pleiotrophic functions of Epstein–Barr virus nuclear antigen-1 (EBNA-1) and oriP differentially contribute to the efficacy of transfection/expression of exogenous gene in mammalian cells. *J Biotechnol* 133:201–207
- Norseen J, Thomae A, Sridharan V, Aiyar A, Schepers A, Lieberman PM (2008) RNA-dependent recruitment of the origin recognition complex. *EMBO J* 27:3024–3035
- Prasad TK, Rao NM (2005) The role of plasmid constructs containing the SV40 DNA nuclear-targeting sequence in cationic lipid-mediated DNA delivery. *Cell Mol Biol Lett* 10:203–215
- Wang S, Frappier L (2009) Nucleosome assembly proteins bind to Epstein–Barr virus nuclear antigen 1 and affect its functions in DNA replication and transcriptional activation. *J Virol* 83:11704–11714
- Cho MS, Yee H, Chan S (2002) Establishment of a human somatic hybrid cell line for recombinant protein production. *J Biomed Sci* 9:631–638
- Cho M-S, Yee H, Brown C, Jeang K-T, Cahn S (2001) An oriP expression vector containing the HIV Tat/TAR transactivation axis produces high levels of protein expression in mammalian cells. *Cytotechnology* 37:23–30
- Liu C, Dalby B, Chen W, Kilzer JM, Chiou HC (2008) Transient transfection factors for high-level recombinant protein production in suspension cultured mammalian cells. *Mol Biotechnol* 39:141–153
- Laengle-Rouault F, Patzel V, Benavente A, Taillez M, Silvestre N, Bompard A, Sczakiel G, Jacobs E, Rittner K (1998) Up to 100-fold increase of apparent gene expression in the presence of Epstein–Barr virus oriP sequences and EBNA1: implications of the nuclear import of plasmids. *J Virol* 72:6181–6185
- Mairhofer J, Grabherr R (2008) Rational vector design for efficient non-viral gene delivery: challenges facing the use of plasmid DNA. *Mol Biotechnol* 39:97–104

18. Durocher Y, Perret S, Kamen A (2002) High-level and high-throughput recombinant protein production by transient transfection of suspension-growing human 293-EBNA1 cells. *Nucleic Acids Res* 30:E9
19. Parham J, Kost T, Hutchins J (2001) Effects of pCIneo and pCEP4 expression vectors on transient and stable protein production in human and simian cell lines. *Cytotechnology* 35:181–187
20. Berntzen G, Lunde E, Flobakk M, Andersen JT, Lauvrak V, Sandlie I (2005) Prolonged and increased expression of soluble Fc receptors, IgG and a TCR-Ig fusion protein by transiently transfected adherent 293E cells. *J Immunol Methods* 298:93–104
21. Durocher Y, Loignon M (2011) US Patent Application Publication US2011/0039339 A1
22. Silla T, Haal I, Geimanen J, Janikson K, Abroi A, Ustav E, Ustav M (2005) Episomal maintenance of plasmids with hybrid origins in mouse cells. *J Virol* 79:15277–15288
23. Codamo J, Munro TP, Hughes BS, Song M, Gray PP (2011) Enhanced CHO cell-based transient gene expression with the Epi-CHO expression system. *Mol Biotechnol* 48:109–115
24. Backliwal G, Hildinger M, Kuettel I, Delegrange F, Hacker DL, Wurm FM (2008) Valproic acid: a viable alternative to sodium butyrate for enhancing protein expression in mammalian cell cultures. *Biotechnol Bioeng* 101:182–189
25. Rajendra Y, Kiseljak D, Baldi L, Hacker DL, Wurm FM (2011) A simple high-yielding process for transient gene expression in CHO cells. *J Biotechnol* 153:22–26
26. Backliwal G, Hildinger M, Chenuet S, de Jesus M, Wurm FM (2008) Coexpression of acidic fibroblast growth factor enhances specific productivity and antibody titers in transiently transfected HEK293 cells. *N Biotechnol* 25: 162–166
27. Zhang X, Stettler M, De Sanctis D, Perrone M, Parolini N, Discacciati M, De Jesus M, Hacker D, Quarteroni A, Wurm F (2010) Use of orbital shaken disposable bioreactors for mammalian cell cultures from the milliliter-scale to the 1,000-liter scale. *Adv Biochem Eng Biotechnol* 115:33–53
28. Pichler J, Galosy S, Mott J, Borth N (2011) Selection of CHO host cell subclones with increased specific antibody production rates by repeated cycles of transient transfection and cell sorting. *Biotechnol Bioeng* 108:386–394
29. Boussif O, Lezoualc'h F, Zanta MA, Mergny MD, Scherman D, Demeneix B, Behr JP (1995) A versatile vector for gene and oligonucleotide transfer into cells in culture and in vivo: polyethylenimine. *Proc Natl Acad Sci USA* 92:7297–7301
30. Derouazi M, Girard P, Fv T, Iglesias K, Muller N, Bertschinger M, Wurm FM (2004) Serum-free large-scale transient transfection of CHO cells. *Biotechnol Bioeng* 87:537–545
31. Galbraith DJ, Tait AS, Racher AJ, Birch JR, James DC (2006) Control of culture environment for improved polyethylenimine-mediated transient production of recombinant monoclonal antibodies by CHO cells. *Biotechnol Prog* 22:753–762
32. Huh SH, Do HJ, Lim HY, Kim DK, Choi SJ, Song H, Kim NH, Park JK, Chang WK, Chung HM, Kim JH (2007) Optimization of 25 kDa linear polyethylenimine for efficient gene delivery. *Biologicals* 35:165–171
33. Sun X, Hia HC, Goh PE, Yap MG (2008) High-density transient gene expression in suspension-adapted 293 EBNA1 cells. *Biotechnol Bioeng* 99:108–116
34. Chenuet S, Martinet D, Besuchet-Schmutz N, Wicht M, Jaccard N, Bon AC, Derouazi M, Hacker DL, Beckmann JS, Wurm FM (2008) Calcium phosphate transfection generates mammalian recombinant cell lines with higher specific productivity than polyfection. *Biotechnol Bioeng* 101:937–945
35. Kichler A, Leborgne C, Danos O (2005) Dilution of reporter gene with stuffer DNA does not alter the transfection efficiency of polyethylenimines. *J Gene Med* 7:1459–1467
36. Backliwal G, Hildinger M, Hasija V, Wurm FM (2008) High-density transfection with HEK-293 cells allows doubling of transient titers and removes need for a priori DNA complex formation with PEI. *Biotechnol Bioeng* 99:721–727
37. Fliedl L, Kaiser Mayer C (2011) Transient gene expression in HEK293 and vero cells immobilised on microcarriers. *J Biotechnol* 153:15–21
38. Bertschinger M, Backliwal G, Schertenleib A, Jordan M, Hacker DL, Wurm FM (2006) Disassembly of polyethylenimine-DNA particles in vitro: implications for polyethylenimine-mediated DNA delivery. *J Control Release* 116:96–104
39. Bertschinger M, Schertenleib A, Cevey J, Hacker D, Wurm F (2008) The kinetics of polyethylenimine-mediated transfection in suspension culture of Chinese hamster ovary cells. *Mol Biotechnol* 40:136–143
40. Ye J, Kober V, Tellers M, Naji Z, Salmon P, Markusen JF (2009) High-level protein expression in scalable CHO transient transfection. *Biotechnol Bioeng* 103:542–551
41. Eberhardy SR, Radzniak L, Liu Z (2009) Iron (III) citrate inhibits polyethylenimine-mediated transient transfection of Chinese hamster ovary cells in serum-free medium. *Cytotechnology* 60:1–9

42. Tuvesson O, Uhe C, Rozkov A, Lullau E (2008) Development of a generic transient transfection process at 100 L scale. *Cytotechnology* 56:123–136
43. Muller N, Girard P, Hacker DL, Jordan M, Wurm FM (2005) Orbital shaker technology for the cultivation of mammalian cells in suspension. *Biotechnol Bioeng* 89:400–406
44. Bollin F, Dechavanne V, Chevalet L (2011) Design of experiment in CHO and HEK transient transfection condition optimization. *Protein Expr Purif* 78:61–68
45. Haldankar R, Danqing L, Saremi Z, Baikarov C, Deshpande R (2006) Serum-free suspension large-scale transient transfection of CHO cells in wave bioreactors. *Mol Biotechnol* 34:191–199
46. Dang JM, Leong KW (2006) Natural polymers for gene delivery and tissue engineering. *Adv Drug Deliv Rev* 58:487–499
47. Jiang HL, Kim TH, Kim YK, Park IY, Cho MH, Cho CS (2008) Efficient gene delivery using chitosan–polyethylenimine hybrid systems. *Biomed Mater* 3:25013
48. Kusumoto K, Akao T, Mizuki E, Nakamura O (2006) Gene transfer effects on various cationic amphiphiles in CHO cells. *Cytotechnology* 51:57–66
49. Stettler M, Zhang X, Hacker DL, Md J, Wurm FM (2007) Novel orbital shake bioreactors for transient production of CHO derived IgGs. *Biotechnol Prog* 23:1340–1346
50. Chapple SD, Crofts AM, Shadbolt SP, McCafferty J, Dyson MR (2006) Multiplexed expression and screening for recombinant protein production in mammalian cells. *BMC Biotechnol* 6:49
51. Zhao Y, Bishop B, Clay JE, Lu W, Jones M, Daenke S, Siebold C, Stuart DI, Yvonne Jones E, Radu Aricescu A (2011) Automation of large scale transient protein expression in mammalian cells. *J Struct Biol* 175(2):209–215
52. Xia W, Bringmann P, McClary J, Jones PP, Manzana W, Zhu Y, Wang S, Liu Y, Harvey S, Madlansacay MR, McLean K, Rosser MP, MacRobbie J, Olsen CL, Cobb RR (2006) High levels of protein expression using different mammalian CMV promoters in several cell lines. *Protein Expr Purif* 45:115–124
53. Mariati Ng YK, Chao SH, Yap MG, Yang Y (2010) Evaluating regulatory elements of human cytomegalovirus major immediate early gene for enhancing transgene expression levels in CHO K1 and HEK293 cells. *J Biotechnol* 147:160–163
54. Mariati Ho SC, Yap MG, Yang Y (2010) Evaluating post-transcriptional regulatory elements for enhancing transient gene expression levels in CHO K1 and HEK293 cells. *Protein Expr Purif* 69:9–15
55. Klein R, Ruttkowski B, Knapp E, Salmons B, Gunzburg WH, Hohenadl C (2006) WPRE-mediated enhancement of gene expression is promoter and cell line specific. *Gene* 372:153–161
56. Kim K-S, Kim M, Moon J, Jeong M, Kim J, Lee G, Myung P-K, Hong H (2009) Enhancement of recombinant antibody production in HEK293E cells by WPRE. *Biotechnol Bioproc Eng* 14:633–638
57. Backliwal G, Hildinger M, Chenuet S, Wulhfard S, De Jesus M, Wurm FM (2008) Rational vector design and multi-pathway modulation of HEK 293E cells yield recombinant antibody titers exceeding 1 g/l by transient transfection under serum-free conditions. *Nucleic Acids Res* 36:e96
58. Kwaks TH, Otte AP (2006) Employing epigenetics to augment the expression of therapeutic proteins in mammalian cells. *Trends Biotechnol* 24:137–142
59. Harraghy N, Regamey A, Girod PA, Mermod N (2011) Identification of a potent MAR element from the mouse genome and assessment of its activity in stable and transient transfactions. *J Biotechnol* 154:11–20
60. Chateauvieux S, Morceau F, Dicato M, Diederich M (2010) Molecular and therapeutic potential and toxicity of valproic acid. *J Biomed Biotechnol* 2010:479364
61. Wulhfard S, Baldi L, Hacker DL, Wurm F (2010) Valproic acid enhances recombinant mRNA and protein levels in transiently transfected Chinese hamster ovary cells. *J Biotechnol* 148:128–132
62. Jiang Z, Sharfstein ST (2008) Sodium butyrate stimulates monoclonal antibody over-expression in CHO cells by improving gene accessibility. *Biotechnol Bioeng* 100:189–194
63. Yee JC, de Leon Gatti M, Philp RJ, Yap M, Hu WS (2008) Genomic and proteomic exploration of CHO and hybridoma cells under sodium butyrate treatment. *Biotechnol Bioeng* 99:1186–1204
64. Allen MJ, Boyce JP, Trentalange MT, Treiber DL, Rasmussen B, Tillotson B, Davis R, Reddy P (2008) Identification of novel small molecule enhancers of protein production by cultured mammalian cells. *Biotechnol Bioeng* 100:1193–1204
65. Hwang SO, Lee GM (2009) Effect of Akt overexpression on programmed cell death in antibody-producing Chinese hamster ovary cells. *J Biotechnol* 139:89–94
66. Chen J, Ghazawi FM, Bakkar W, Li Q (2006) Valproic acid and butyrate induce apoptosis in

- human cancer cells through inhibition of gene expression of Akt/protein kinase B. *Mol Cancer* 5:71
67. Yun YR, Won JE, Jeon E, Lee S, Kang W, Jo H, Jang JH, Shin US, Kim HW (2010) Fibroblast growth factors: biology, function, and application for tissue regeneration. *J Tissue Eng* 2010: 218142
 68. Sheng Z, Liang Y, Lin CY, Comai L, Chirico WJ (2005) Direct regulation of rRNA transcription by fibroblast growth factor 2. *Mol Cell Biol* 25:9419–9426
 69. Fox SR, Yap MX, Yap MG, Wang DI (2005) Active hypothermic growth: a novel means for increasing total interferon-gamma production by Chinese-hamster ovary cells. *Biotechnol Appl Biochem* 41:265–272
 70. Wulhfard S, Tissot S, Bouchet S, Cevey J, De Jesus M, Hacker DL, Wurm FM (2008) Mild hypothermia improves transient gene expression yields several fold in Chinese hamster ovary cells. *Biotechnol Prog* 24:458–465
 71. Han YK, Koo TY, Lee GM (2009) Enhanced interferon-beta production by CHO cells through elevated osmolality and reduced culture temperature. *Biotechnol Prog* 25:1440–1447
 72. Pham PL, Perret S, Cass B, Carpentier E, St-Laurent G, Bisson L, Kamen A, Durocher Y (2005) Transient gene expression in HEK293 cells: peptone addition posttransfection improves recombinant protein synthesis. *Biotechnol Bioeng* 90:332–344
 73. Sun X, Goh PE, Wong KT, Mori T, Yap MG (2006) Enhancement of transient gene expression by fed-batch culture of HEK 293 EBNA1 cells in suspension. *Biotechnol Lett* 28:843–848
 74. Sandberg H, Lutkemeyer D, Kuprin S, Wrangel M, Almstedt A, Persson P, Ek V, Mikaelsson M (2006) Mapping and partial characterization of proteases expressed by a CHO production cell line. *Biotechnol Bioeng* 95:961–971
 75. Robert F, Bierau H, Rossi M, Agugiaro D, Soranzo T, Broly H, Mitchell-Logean C (2009) Degradation of an Fc-fusion recombinant protein by host cell proteases: identification of a CHO cathepsin D protease. *Biotechnol Bioeng* 104:1132–1141
 76. Andersen CR, Nielsen LS, Baer A, Tolstrup AB, Weilguny D (2011) Efficient expression from one CMV enhancer controlling two core promoters. *Mol Biotechnol* 48:128–137
 77. Schlatter S, Stansfield SH, Dinnis DM, Racher AJ, Birch JR, James DC (2005) On the optimal ratio of heavy to light chain genes for efficient recombinant antibody production by CHO cells. *Biotechnol Prog* 21:122–133
 78. Bentley KJ, Gewert R, Harris WJ (1998) Differential efficiency of expression of humanized antibodies in transient transfected mammalian cells. *Hybridoma* 17:559–567
 79. Nettleship JE, Ren J, Rahman N, Berrow NS, Hatherley D, Barclay AN, Owens RJ (2008) A pipeline for the production of antibody fragments for structural studies using transient expression in HEK 293T cells. *Protein Expr Purif* 62:83–89
 80. Zhao Y, Gutshall L, Jiang H, Baker A, Beil E, Obmolova G, Carton J, Taudte S, Amegadzie B (2009) Two routes for production and purification of Fab fragments in biopharmaceutical discovery research: papain digestion of mAb and transient expression in mammalian cells. *Protein Expr Purif* 67:182–189
 81. Zhang J, Liu X, Bell A, To R, Baral TN, Azizi A, Li J, Cass B, Durocher Y (2009) Transient expression and purification of chimeric heavy chain antibodies. *Protein Expr Purif* 65:77–82
 82. Haack A, Schmitt C, Poller W, Oldenburg J, Hanfland P, Brackmann HH, Schwaab R (1999) Analysis of expression kinetics and activity of a new B-domain truncated and full-length FVIII protein in three different cell lines. *Ann Hematol* 78:111–116
 83. Gaudry JP, Arod C, Sauvage C, Busso S, Dupraz P, Pankiewicz R, Antonsson B (2008) Purification of the extracellular domain of the membrane protein GlialCAM expressed in HEK and CHO cells and comparison of the glycosylation. *Protein Expr Purif* 58:94–102
 84. Van den Nieuwenhof IM, Koistinen H, Easton RL, Koistinen R, Kamarainen M, Morris HR, Van Die I, Seppala M, Dell A, Van den Eijnden DH (2000) Recombinant glycodeolin carrying the same type of glycan structures as contraceptive glycodeolin-A can be produced in human kidney 293 cells but not in Chinese hamster ovary cells. *Eur J Biochem* 267:4753–4762
 85. Suen KF, Turner MS, Gao F, Liu B, Althage A, Slavin A, Ou W, Zuo E, Eckart M, Ogawa T, Yamada M, Tuntland T, Harris JL, Trauger JW (2010) Transient expression of an IL-23R extracellular domain Fc fusion protein in CHO vs. HEK cells results in improved plasma exposure. *Protein Expr Purif* 71:96–102
 86. Tchoudakova A, Hensel F, Murillo A, Eng B, Foley M, Smith L, Schoenen F, Hildebrand A, Kelter AR, Ilag LL, Vollmers HP, Brandlein S, McIninch J, Chon J, Lee G, Caciuttolo M (2009) High level expression of functional human IgMs in human PER.C6 cells. *MAbs* 1:163–171
 87. Jones D, Kroos N, Anema R, van Montfort B, Vooys A, van der Kraats S, van der Helm E, Smits S, Schouten J, Brouwer K, Lagerwerf F, van Berkel P, Opstelten DJ, Logtenberg T, Bout A (2003) High-level expression of recombinant IgG in the human cell line per.c6. *Biotechnol Prog* 19:163–168

88. Niklas J, Schrader E, Sandig V, Noll T, Heinze E (2011) Quantitative characterization of metabolism and metabolic shifts during growth of the new human cell line AGE1.HN using time resolved metabolic flux analysis. *Bioprocess Biosyst Eng* 34: 533–545
89. Schiedner G, Hertel S, Bialek C, Kewes H, Waschutza G, Volpers C (2008) Efficient and reproducible generation of high-expressing, stable human cell lines without need for antibiotic selection. *BMC Biotechnol* 8:13
90. Brown SW, Mehtali M (2010) The avian EB66(R) cell line, application to vaccines, and therapeutic protein production. *PDA J Pharm Sci Technol* 64:419–425
91. Kramer O, Klausung S, Noll T (2010) Methods in mammalian cell line engineering: from random mutagenesis to sequence-specific approaches. *Appl Microbiol Biotechnol* 88:425–436
92. Johnson KC, Jacob NM, Nissom PM, Hackl M, Lee LH, Yap M, Hu WS (2011) Conserved microRNAs in Chinese hamster ovary cell lines. *Biotechnol Bioeng* 108:475–480
93. Hackl M, Jakobi T, Blom J, Doppmeier D, Brinkrolf K, Szczepanowski R, Bernhart SH, Siederdissen CH, Bort JA, Wieser M, Kunert R, Jeffs S, Hofacker IL, Goesmann A, Puhler A, Borth N, Grillari J (2011) Next-generation sequencing of the Chinese hamster ovary microRNA transcriptome: identification, annotation and profiling of microRNAs as targets for cellular engineering. *J Biotechnol* 153:62–75
94. Barron N, Sanchez N, Kelly P, Clynes M (2011) MicroRNAs: tiny targets for engineering CHO cell phenotypes? *Biotechnol Lett* 33:11–21

Chapter 14

Stable Transfection Pools for Large Quantity of Protein Production

Jianxin Ye

Abstract

During the early development phase of therapeutic proteins such as monoclonal antibodies, representative material is often requested before the final production cell line is established. In order to fulfill such requests, technologies capable of delivering large quantity of proteins quickly are essential. This chapter outlines the stable transfection pool technology that generates grams of proteins within 2 months post transfection. This technology shortens the overall developmental time frame for therapeutic proteins.

Key words: CHO cells, Monoclonal antibody production, Stable transfection pool, Glycosylation

1. Introduction

Monoclonal antibodies and other therapeutic proteins have been the fastest-growing new therapies in the past decade (1, 2). The general process for protein production in mammalian cells involves stable cell line generation. The overall time frame for stable cell line generation can take 6–12 months or more (3). To generate materials quickly, transient transfection and stable transfection pool technologies are commonly used (4–7). This chapter focuses on transfection pool technology for large-scale production.

Stable transfection pool refers to the heterogeneous population of transfectants which have been transfected with DNA of gene of interest. Unlike clonal stable cell lines, transfection pools do not need to undergo time-consuming screening procedures; therefore, they can be expanded for production in a relatively short time. Large-scale production from stable transfection pools can be achieved simply by expansion of cell culture volume. For generating stable transfection pools, the same host cell line and expression

vector for the final production cell line can be utilized, which is valuable for generation of representative product for preclinical studies (5). Within 2 months post transfection, the production can be scaled up to 100 L or more. Since stable transfection pools contain a mixed population of cells which have various expression levels, the expression levels of pools are usually lower than those of the final selected clonal stable cell lines. For the same reason, the productivities of transfection pools are not very stable, and subject to decline with increasing number of passaging or time. Therefore, minimizing the time from transfection to production is critical to achieve high yield (5).

2. Materials

1. Linearized plasmid containing the gene of interest.
2. Electroporation equipment or other transfection reagents.
3. Suspension-adapted Chinese Hamster Ovary (CHO) cell line or equivalent, for example CHOK1 from ATCC, adapted in serum-free chemical-defined medium.
4. CD-CHO medium (Invitrogen) or other appropriate cell culture medium and cell culture supplements, such as glutamine.
5. Production medium and feeds for fed-batch process.
6. Appropriate selection agent. For example, if glutamine synthetase is used as the selective marker, methionine sulfoximine (MSX) is used as the selection agent. Appropriate MSX concentration has to be determined by titration.
7. Appropriate tissue culture vessels, such as T-75 flasks, 125–500-mL vented shake flasks.
8. 20–200 L Wave Bioreactor® (GE Healthcare) or other appropriate bioreactors.
9. Standard cell culture equipment, such as CO₂ incubator, orbital shaker, centrifuge, cell counter.
10. Equipment for metabolic measurement, such as BioProfile Analyzers (Nova Biomedical).

3. Methods

3.1. Transfection

Mammalian cell transfection can be accomplished through various approaches, such as electroporation, calcium phosphate coprecipitation, lipofectamine, etc. (3). In our experiments, we use electroporation. The expression vector containing the gene of interest is typically

linearized to enhance the integration events. The linearization site can be somewhere in the plasmid backbone to avoid impact on the expression of the selectable marker or gene of interest.

1. CHO cells are maintained in 125-mL shake flasks at 37°C in a humidified 5% CO₂ orbital shaker. Shaker speed is 100 rpm. The culture medium is CD-CHO supplemented with 2 mM glutamine. Subculture the cells every 3–4 days, seeding at concentration of 2×10^5 cells/mL.
2. Passage the cells 2 days prior to transfection at a cell density of 2×10^5 cells/mL.
3. Prior to transfection, measure cell density and viability. The cell culture viability should be above 90%.
4. For each transfection by electroporation, 40 µg of linearized DNA and 1×10^7 viable cells are used.
5. Pellet 1×10^7 viable cells by centrifugation (5 min at 200 × g).
6. Wash the pellet with 20 mL of CD-CHO medium. Pellet the cells by centrifugation.
7. Resuspend the cell pellet in appropriate amount of CD-CHO medium so that the total volume of cell suspension and DNA is equal to 800 µL.
8. Mix the cell suspension and DNA, and transfer the mixture into a sterile 4-mm-gap cuvette.
9. Load cuvette into electroporation carriage and pulse. The optimal electroporation parameters could be cell line dependent (see Note 1).
10. Immediately following electroporation, transfer the entire contents from the cuvette into 7 mL of culture medium. Transfer this cell suspension into a T-75 flask. Incubate at 37°C in a humidified 5% CO₂ incubator overnight.

3.2. Selection and Expansion

The selection phase will depend on the expression plasmid and cell line. Killing curves can be used to determine the appropriate selection pressure. During expansion, the selection pressure is maintained to avoid the significant loss of productivity. Since the productivities from the transfection pools are not very stable, it is critical to shorten the expansion phase (5). If a large production scale is desired, scale up expansion of the pool as early as possible.

1. Approximately 24 h post transfection, add 42 mL of culture medium with selection drug to the T-75 flask. The total culture volume in the T-75 flask will be about 50 mL. The appropriate concentration of selectable drug will be dependent on the drug and host cell line used. For example, we use 10 µg/mL of Puromycin for CHO cells (see Note 2).

2. Seven to ten days post transfection, transfer the cells from T-75 flask into a 250-mL vented shake flask and incubate at 37°C in a humidified 5% CO₂ orbital shaker.
3. Every 3–4 days, measure the cell viability and cell density; passage the cells in shake flasks at 2 × 10⁵ cells/mL.
4. After 4–5 passages in shake flasks, the cell viability should reach above 90%, and the doubling time should be roughly the same as that of the host cells (see Note 3). At this point, the pool is ready for production. If a large production volume is desired, expand the culture as appropriate.

3.3. Production in Wave Bioreactor

The production phase can be done in shake flasks, Wave Bioreactors, or stirred bioreactors. We used 20–200 L Wave Bioreactor for monoclonal antibody production. The production procedure in a 20 L Wave Bioreactor using a fed-batch process is discussed here. Regular batch culture process can be used for this purpose as well with reduced production duration and productivity. Maintaining the selection pressure during the production phase helps to improve productivity.

1. Seed the cells in the production medium (with selection pressure) at an inoculation density of 2 × 10⁵ cells/mL (see Note 4). For a 20 L Wave Bioreactor, the initial working volume is 8 L.
2. Set the Wave Bioreactor rocking speed at 25 rpm; set the rocking angle at 8°; set the temperature at 36.5°C; apply airflow rate at 0.1 L per minute with a 5% CO₂/air gas overlay.
3. Four days post inoculation, sample the culture for cell viability, cell density, glucose level, lactate level, ammonium level, pH, etc.
4. Based on the glucose measurements, add glucose to the culture to achieve 40 mM glucose on each feed day.
5. Feed the culture with medium feeds for fed-batch process as appropriate.
6. Repeat steps 3–5 on days 6, 8, and 11 post inoculation or as appropriate (see Note 5).
7. Harvest on day 14 or when the cell viability is below 50%.

4. Notes

1. The optimal parameters for electroporation could be cell line dependent. A typical capacitance value is around 1,000 µF. The appropriate voltage ranges from 200 to 350 V.
2. The selective conditions may vary based on cell lines and selectable markers used. For Hygromycin selectable marker, 150–300 µg/mL of hygromycin can be used for CHO cells.

For Puromycin, 7–20 µg/mL can be used for CHO cells. For glutamine synthetase, 25–50 µM of MSX can be used for CHO cells.

3. After the initial 7–10 day selection in T-75 flasks, the cell viability is usually low. Collect all the cells by centrifugation at $200 \times g$ for 5 min. Resuspend the cells in 25 mL of fresh selective medium and transfer the cell suspension in the shake flask. The combination of selection and suspension re-adaption will shorten the overall time frame. After 4–7 days in suspension culture, the culture is ready to be passaged when the cell density reaches 5×10^5 cells/mL. If the cell density is too low, continue the culture for several more days before passaging. The cell viability will increase as passaging continues. If large-scale production is desired, expand the culture to the largest volume possible.
4. Maintaining the selective pressure during the production phase will improve productivity. However, if the selection reagents are hard to remove during the downstream process or trace amounts of the selection reagent are not desirable in the final product, the selection reagent can be kept out of the production medium.
5. To align the principle of quick production, it is recommended to have a platform process established beforehand. However, if it is desired, process optimization can be performed, such as temperature, pH, feed composition, feed schedule, etc.

References

1. Li J, Zhu Z (2010) Research and development of next generation of antibody-based therapeutics. *Acta Pharmacol Sin* 31:1198–1207
2. Reichert JM (2008) Monoclonal antibodies as innovative therapeutics. *Curr Pharm Biotechnol* 9:423–430
3. Birch JR, Racher AJ (2006) Antibody production. *Adv Drug Deliv Rev* 58:671–685
4. Pham PL, Kamen A, Durocher Y (2006) Large-scale transfection of mammalian cells for the fast production of recombinant protein. *Mol Biotechnol* 34:225–237
5. Ye J, Alvin K, Latif H, Hsu A, Parikh V, Whitmer T, Tellers M, de la Cruz Edmonds MC, Ly J, Salmon P, Markusen JF (2010) Rapid protein production using CHO stable transfection pools. *Biotechnol Prog* 26:1431–1437
6. Ye J, Kober V, Tellers M, Naji Z, Salmon P, Markusen JF (2009) High-level protein expression in scalable CHO transient transfection. *Biotechnol Bioeng* 103:542–551
7. Wulhfard S, Tissot S, Bouchet S, Cevey J, De Jesus M, Hacker DL, Wurm FM (2008) Mild hypothermia improves transient gene expression yields several fold in Chinese hamster ovary cells. *Biotechnol Prog* 24:458–465

Chapter 15

Mammalian Stable Expression of Biotherapeutics

Thomas Jostock and Hans-Peter Knopf

Abstract

Many therapeutically relevant proteins, like IgG antibodies, are highly complex, multimeric glycoproteins that are difficult to express in microbial systems and thus usually produced in mammalian host cells. During the past two decades, stable mammalian expression technologies have made huge progress resulting in highly increased speed of cell line development and yield of manufacturing processes. Here, we give an overview of technologies that are applied at different stages of state-of-the-art cell line development processes for biomanufacturing.

Key words: Antibody, Cell line development, Stable expression, CHO, Biomanufacturing

1. Introduction

An increasing proportion of drug candidates entering clinical trials and the market are biotherapeutics, many of which are glycoproteins. The glycan moiety, thereby, often contributes to the biological activity or has major impact on pharmacokinetics of the molecule. For erythropoietin alpha (EPO) for example, it is known that glycosylation as such and the type of glycosylation are critical for in vivo activity (1). The same is true for IgG antibodies, the mode of action of which depends also on Fc-mediated effector functions such as antibody-dependent cell cytotoxicity (ADCC) or complement-dependent cytotoxicity (CDC) (2).

As microbial expression systems are naturally either not able to perform glycosylation or deliver glycan patterns that strongly differ from human-type glycosylation, mammalian cell-line based expression for biomanufacturing was introduced in the late 1980s to produce tissue plasminogen activator (t-PA), one of the first recombinant therapeutic glycoproteins (3). Around that time, increasing genetic engineering capabilities opened also the way for

generating chimeric, humanized, and fully human antibodies directed against almost any target of choice. Thus, the early vision of Paul Ehrlich to use antibodies as “magic bullets” to fight against all kinds of diseases became more and more realistic. Due to the glycosylation and heterotetrameric nature of the molecule, manufacturing of IgG antibodies however turned out to be quite challenging. Initially, the timelines for mammalian cell line development were quite long and the yields were quite low, typically significantly below 1 g/L. In combination with the need of high doses for many antibody treatments, this made development of antibody therapeutics quite time consuming and manufacturing comparably expensive. Introduction of new technologies and optimization of many parameters over the past two to three decades led to a massive reduction of timelines and very substantial increase of the yield of manufacturing processes. Today, the typical cycle time for cell line development from gene to master cell bank (MCB) takes about 20–30 weeks and IgG yields ranging from 3 to 6 g/L, sometimes even up to 10 g/L, can be reached (4). This enables the manufacturing of high quantities of IgG at comparatively low cost of goods, which are today typically in a range of 40–100\$/g. Whereas in earlier days the upstream processes has been the main cost driver, today the down stream process is responsible for the main part of the cost.

In the following paragraphs, we give an overview of technologies that are commonly used in mammalian cell line development for biomanufacturing; however, due to the complexity of the field and the high dynamics of technology development in that area, not every aspect may be covered by our summary.

Figure 1 provides a schematic overview of the individual steps of a typical cell line development process.

2. Host Cells

Host cells for biomanufacturing of therapeutic proteins have to fulfill several requirements. In order to achieve high yields, fast growth to high cell densities in bioreactors is favorable. Culture media of modern cell line development and production processes are chemically defined and protein free. The origin of all raw

Fig. 1. (continued) stringency of the selection system, the result is a more or less diverse pool of surviving cells. (b) Single-cell cloning and screening: Monoclonal cell lines are generated from transfected and selected pools via single-cell cloning, which can be selective or random. Several layers of screening are applied to identify high-performing clone candidates. Screening results from multiwell plates and shake flasks may be confirmed in small-scale bioreactors prior to selecting the final clone and manufacturing of the master cell bank. (Taken from Jostock T (2011) Expression of antibody in mammalian cells. In: AlRubeai M (ed) Cell engineering, vol 7: antibody expression and production. Springer Science and Business Media, pp 1–24).

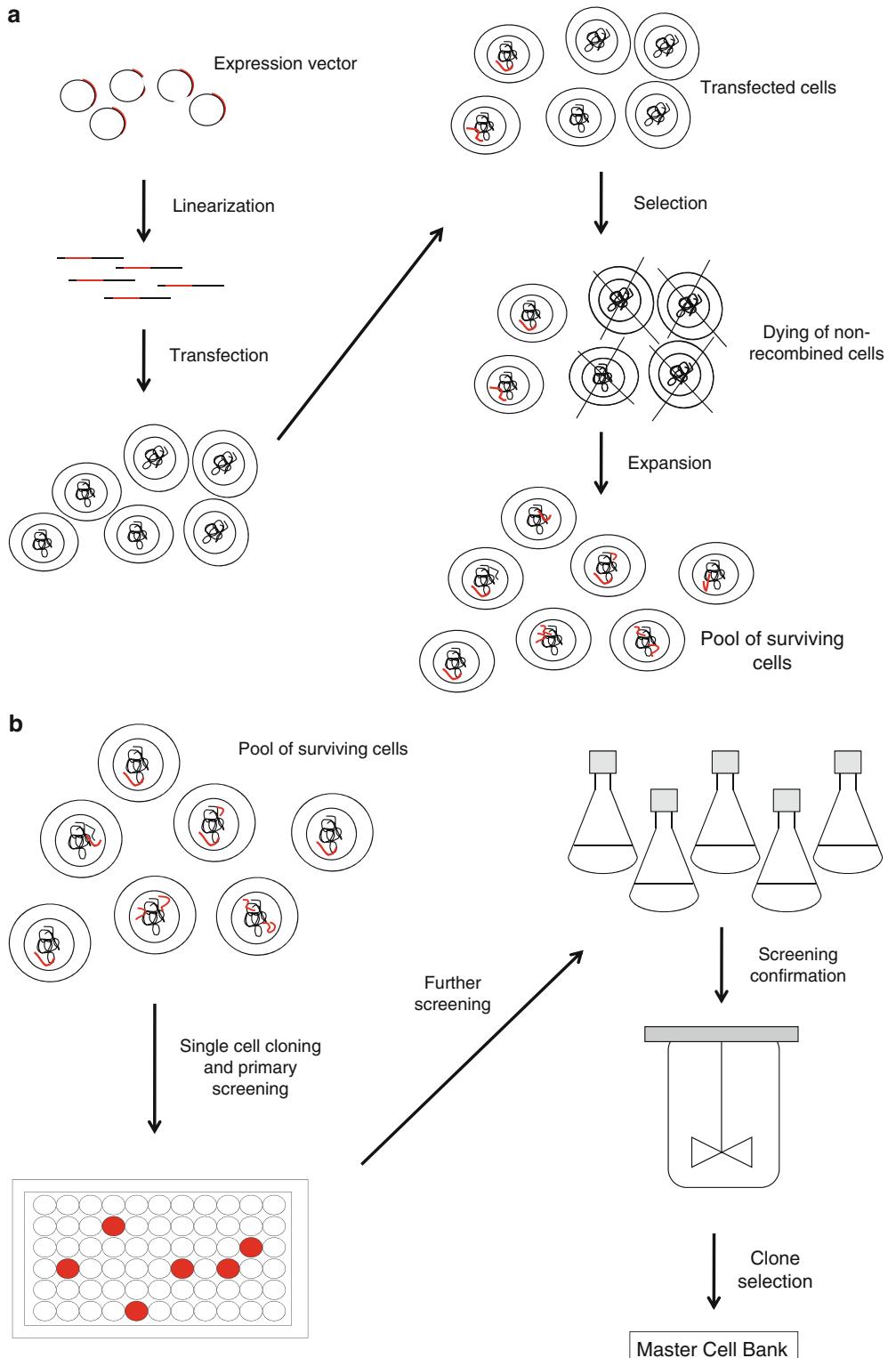


Fig. 1. Overview of a typical cell line development process. (a) Transfection and selection: Cell line development starts with an expression vector which, optionally after linearization, is transfected in the host cell line. Subsequently, one or more selection steps are applied to selectively kill cells that did not stably integrate the expression vector. Depending on the

materials used to prepare the media has to be certified by an official certificate of origin of the manufacturer. Full traceability of substrates used during cell line development is key to ensure regulatory compliance. The end product of the cell line development process, the MCB, has to fulfill also potential regulatory requirements given in up to 10 years in the future, when the protein of interest produced by such cell line is submitted for market approval. Last but not least, the host cell line needs to fulfill viral safety requirements and needs to secrete the protein of interest with a suitable human type of post translational modification showing low immunogenic properties.

Currently, most of the marketed biotherapeutics from mammalian expression are produced in Chinese hamster ovary (CHO) cells. CHO cells generate glycan patterns that are well tolerated by patients and can be adapted to grow to high cell densities in animal component-free and defined media, which makes them an attractive host cell system.

CHO cells initially have been isolated in the 1950s (5) and gave rise to a number of different strains that are available today, including cell lines that are lacking functional dihydrofolate reductase (dhfr) genes (6, 7). Such dhfr negative strains are well suitable for metabolic selection and gene amplification procedures as described below in the selection strategies. Different strategies for metabolic engineering, cell cycle engineering, and anti-apoptotic engineering of CHO cells in order to improve expression and process yields have been described (reviewed in (8–10)). Also, engineering of the glycosylation pathway for improved *in vivo* activity of the protein of interest, like antibodies with enhanced ADCC activity, has been successful in CHO cells (11–13). With increasing sequence information of the CHO genome and recent cell line engineering technologies such as targeted nucleases and shRNA, further room for enhancing host cell lines is arising.

Besides CHO cells, other rodent cells like the mouse myeloma cell lines NS0 and Sp2/0 (14) or Baby hamster kidney cells (BHK21) (15) are established hosts for stable expression. The human embryonic kidney cell line HEK293 is mainly used for transient expression but generally also applicable for stable transfections (15–17). New cell lines for stable protein production that have been described more recently include the human retina cell line Per.C6 (18), human neuronal cells (19), and avian embryonic stem cells (20).

3. Vector Systems

Plasmid vectors are most commonly used to introduce the gene(s) of interest into the host cells. Suitable transfection methods for gene transfer in mammalian cells include lipofection and electroporation.

High-level expression is achieved by using strong viral or mammalian promoters. For nontoxic, robust molecules like antibodies, constitutively active promoters have proven to be well suitable for high-yielding fed-batch processes. Several inducible promoter systems for mammalian expression have been described (21), which may be of special value for products which exhibit negative effects on the host cells during cultivation.

For multimeric proteins like antibodies, co-expression of the different subunits in a single cell has to be achieved. Co-transfection of multiple plasmids, each encoding one subunit, is one possibility. Advantageous in such an approach is that different ratios of the individual plasmids may be used according to the most optimal balance of expression levels. Alternatively, several expression cassettes can be combined in a single plasmid. Tandem and sandwich vectors have been described for expressing antibody light and heavy chains from a single plasmid (22, 23). Different open reading frames (ORFs) can also be combined in a single, multi-cistronic expression cassette (24–26). Internal ribosomal entry site (IRES) motives, thereby, are used to separate the ORFs and to drive translation initiation of the downstream cistrons (27). For antibodies, expression of both chains from a single ORF has been shown to be possible by combining a special 2A peptide sequence in combination with a furin cleavage site as a processing element between both chains. During translation and secretion, the fusion peptide is processed in a way that fully assembled native IgG antibodies are formed (28, 29).

In addition to the expression cassette(s) for the gene(s) of interest, plasmid vectors for stable expression usually contain one or more selection marker cassettes. Selection markers and systems are described in more detail in the next paragraph.

Traditionally, plasmid vector systems rely on random integration of the transgene into the host cell genome to generate stable expressing recombinant cell lines. The site of integration, thereby, is believed to have a major influence on the phenotype and more specifically on the expression performance of the resulting cell line. Thus, comparably high clone screening efforts or highly stringent selection protocols are necessary to identify clones with very high expression levels and sufficiently high expression stability.

Several technologies for targeted integration of expression vectors in mammalian cells have been described, which offer the potential to reduce or eliminate integration site heterogeneity in cell line development. Vector integration at the target site, thereby, can be driven by heterologous recombinases like flip (FLP) from yeast or Cre from phage (30–33), if the corresponding recognition sites are present in the vector and the chromosomal acceptor location. Alternatively, integration can be mediated by nuclelease-induced homologous recombination if suitable homology regions are present in the vector. The recombination frequency at the target

site, thereby, is boosted by the nuclease-induced double-strand break, which triggers cellular repair mechanisms. Suitable targeted meganucleases for such purposes include meganucleases and zinc finger nucleases and should have minimal background activity at other chromosomal locations (34, 35).

Besides plasmid vectors, other gene transfer vehicles such as artificial chromosomes have been evaluated. Artificial chromosomes have been shown to allow transfer of multiple copies of the transgene into the host cell and to be stably inherited in cell divisions over many passages (36).

4. Selection Systems

After gene transfer, typically one or more selection steps are applied to enrich cells having the vector stably integrated in their genome. Selection is based on protective effects mediated by expression of selectable marker genes from the transfected and integrated vector under selective culture conditions.

Popular selection systems include markers mediating resistance to antibiotics like Neomycin or Hygromycin and metabolic markers like dihydrofolate reductase (DHFR) or glutamine synthetase (GS) (14, 37).

Antibiotics resistance marker genes often derive from microbial organisms and are enzymes that can deactivate the antibiotic reagent. In combination with suitable vectors, such systems are particularly suitable to rapidly kill non-transfected and non-expressing cells. Antibiotics resistance markers are regarded as non-amplifiable and have the potential disadvantage that a small subset of high expressing cells may quickly inactivate the selection reagent, allowing also non-expressing cells to survive.

Metabolic selectable marker genes are often endogenous enzymes involved in critical steps of the host cells' metabolism. The most commonly used ones are dihydrofolate reductase (DHFR) and glutamine synthetase (GS) (14, 37).

DHFR is involved in folate metabolism and its activity is crucial for nucleotide synthesis. Host cells lacking functional endogenous dhfr genes like CHO-DG44 or CHO-DXB11 can be used for dhfr-driven selection of recombinant cell lines by culturing transfected cells in the absence of certain nucleotides. Additionally, selection pressure can be further increased by adding methotrexate (MTX), which as a folate analogue is inhibiting the enzymatic activity of dhfr. By stepwise increasing of the concentration of MTX during the selection, gene amplification of the transgene can be achieved, often leading to improved productivities (14, 37). However, due to the long timelines of gene amplification protocols, such approaches are getting less and less popular nowadays. Therefore, the dhfr/MTX

system often is applied in a non-amplifiable mode using high selection pressure from beginning on.

GS is involved in the synthesis of the essential amino acid glutamine. Transfected CHO cells can be selected by adding the inhibitor methionine sulfoximine (MSX) to the culture medium. Recently, a GS deletion CHO cell line has been generated and shown to be efficiently selectable in glutamine-free media without MSX (38).

A rather new approach is to use folate receptor as a dominant metabolic selectable marker. Here, cells over-expressing folate receptor are selected under folate deprivation conditions to enrich high-producing transfectants (39).

Generally, the higher the stringency of the selection system is, the higher is also the average productivity of surviving cells. But, as a disadvantage, the higher the stringency is, the higher is also the time needed to recover cells from selection. High-stringency selection systems can achieve quite high productivities already without single-cell cloning, allowing to quickly producing significant amounts of recombinant protein in early project phases.

5. Clone Isolation Technologies

For large-scale manufacturing, monoclonal recombinant cell lines are used in order to ensure high batch-to-batch consistency. One or more single-cell cloning steps are usually applied in order to generate such cell lines. Several different cloning methods are available, including limiting dilution cloning, flow cytometry, colony picking (manual and automated), and laser-enabled analysis and processing (LEAP).

Generally, selective and nonselective methods can be distinguished. Classical approaches like limiting dilution or manual colony picking are typically nonselective, meaning that the different productivities of individual cells cannot be taken into account, when isolating the clones. Accordingly, high numbers of clones may need to be screened for productivity later on in order to identify very-high-producing ones. In contrast, selective cloning methods allow isolating cells based on their phenotype regarding productivity. This can lead to a much higher abundance of high-producing cell lines in the population after cloning (reviewed in (40, 41)). The most commonly applied selective cloning technologies are either flow cytometry or automated colony picking (e.g., CloniPix FL). In both cases, productivity of cells is typically analyzed using fluorescently labeled staining reagents directed against the protein of interest.

Flow cytometric cell sorters are capable of analyzing and seeding individual cells in multiwell plates at very high throughput and

speed. Thereby, parameters like size, granularity, and fluorescence intensity can be used to selectively isolate cells based on their phenotype. In order to allow surface staining of the protein of interest on living cells, several technologies have been described that lead to some degree of cell surface fixation (42–44). Alternatively, special staining procedures that do not depend on surface display of the protein of interest (37, 41, 44, 45), co-expression of reporter genes (25, 46, 47) and micro droplet encapsulation of cells (48, 49), are applicable.

Today's cell line development processes are typically utilizing suspension cells under serum-free conditions throughout the whole procedure. In order to achieve colony type growth and colony picking with such cells with, e.g., a ClonepixFL system, plating in semisolid medium can be done. If a suitable fluorescently labeled staining reagent is embedded in this medium, diffusion of protein of interest from colonies leads to formation of a halo of fluorescent precipitates which can be detected by an imaging system (40). Based on the shape and/or size of the colony and the fluorescent halo, colonies can be selectively picked and seeded into multiwell plates for further expansion and screening.

In order to assure monoclonality of cell lines, several cycles of cloning may be applied leading to very high statistical probability of monoclonality. In case of limiting dilution or flow cytometry-based cloning, imaging or visual observation of the multiwell plates shortly after seeding allows the documentation of the presence of a single cell only in a well as a proof of monoclonality.

6. Screening Methods

One of the challenges in cell line development is to identify screening methods for assessing key performance parameters early and with sufficiently high throughput. Typically, initial screening of clones is focused on productivity and performed in multiwell formats. However, such small-scale assays are not fully predictive for the fitness of the clones for large-scale manufacturing (50, 51). Therefore, usually several layers of screening are necessary to identify those clones having the highest potential for high-titer bioreactor production processes. Thereby, with every layer the number of clones typically is reduced until the final clone for MCB generation is selected.

Due to the ease of handling and the comparably high throughput that can be achieved, shake flask cultures are widely used as a second-layer screening. Here, inocula with defined cell densities, feeding, monitoring of growth, and other parameters are possible. In such cultures, quite high cell densities and productivities can be achieved, which also enables purification of the protein of interest

from these cultures to analyze important product quality attributes. However, one need to know in detail which of the quality attributes behaves conservative during scale-up, and which might drastically change.

Semi-automated small- to mid-scale culture systems with online analysis capabilities like pH and/or turbidity as a measure of cell density are getting more and more popular. Such systems aim to better simulate large-scale manufacturing processes to support evaluation and screening candidate clones for their suitability to large-scale production conditions.

The screening results from multiwell and shake flask systems are usually confirmed in small- to mid-scale bioreactor experiments to support the final clone selection decision prior to generating an MCB.

7. Clone Characterization

High-performance cell lines for manufacturing of biotherapeutics have to fulfill a number of requirements. Besides high productivity and good growth behavior, clonal production stability is a key requirement for production clones. Ideally, productivity and growth rate of manufacturing cell lines stay more or less constant during cultivation over many passages. However, besides stable clones also unstable clones, which show gradual drop of productivity over time, may be obtained with most expression systems. Especially after extensive gene amplification procedures, clones carrying very high copy numbers may have the tendency to lose transgene copies over time which can translate in a reduction of expression also. Other possible mechanisms behind production instability include silencing of the promoter and changes in the surrounding chromatin structure. Depending on the production scale, a process from thawing the vial of the cell line to harvesting the final bioreactor stage can take 10–12 weeks or more. Thus, for final clone candidates, production stability is usually monitored over a period of time that covers at least the duration of this process.

In order to minimize the risk of product-related impurities, integrity and sequence of the expressed mRNA of the gene of interest are checked. In addition, extensive product quality analysis is done to assure that all target product profile properties are met and no unexpected and/or unwanted, clone-specific product attributes are present. MCBs are typically generated under cGMP conditions and are subject matter of intensive testing for adventitious agents like viruses, according to the guidelines of the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH).

8. Summary

Mammalian stable expression of biotherapeutics is well established and, for glycoproteins, the current industry standard. CHO cells, thereby, are the most frequently used host cells, but an increasing number of alternative cell lines are available. Improvements in host cell lines, culture media, vector, and screening technologies led to a very substantial increase of volumetric and specific productivities during the past 10–15 years. At the same time, cycle times for cell line generation have been drastically reduced. Stable cell line development technologies have reached a high level of robustness and became more and more cost-efficient, which make them also attractive for proteins that traditionally would have been considered for microbial expression systems primarily. However, further improvements are still possible and with biosimilars entering the markets, more competition for cost-efficient manufacturing can be expected. Next-generation sequencing and genome editing technologies that are currently evolving at high speed may open the way for intensive host cell engineering activities and transfer mammalian expression systems to a new performance level.

References

- Takeuchi M, Kobata A (1991) Structures and functional roles of the sugar chains of human erythropoietins. *Glycobiology* 1:337–346
- Jefferis R (2007) Antibody therapeutics: isotype and glycoform selection. *Expert Opin Biol Ther* 7:1401–1413
- Wurm FM (2004) Production of recombinant protein therapeutics in cultivated mammalian cells. *Nat Biotechnol* 22:1393–1398
- Jostock T (2011) Expression of antibody in mammalian cells. In: Al-Rubeai M (ed) *Cell Engineering*. Springer Science and Business Media B.V., Dordrecht
- Puck TT (1958) Genetics of somatic mammalian cells. *J Exp Med* 108:945–955
- Urlaub G, Chasin LA (1980) Isolation of Chinese hamster cell mutants deficient in dihydrofolate reductase activity. *Proc Natl Acad Sci U S A* 77:4216–4220
- Urlaub G, Kas E, Carothers AM, Chasin LA (1983) Deletion of the diploid dihydrofolate reductase locus from cultured mammalian cells. *Cell* 33:405–412
- Fussenegger M, Bailey JE, Hauser H, Mueller PP (1999) Genetic optimization of recombinant glycoprotein production by mammalian cells. *Trends Biotechnol* 17:35–42
- Dinnis DM, James DC (2005) Engineering mammalian cell factories for improved recombinant monoclonal antibody production: lessons from nature? *Biotechnol Bioeng* 91:180–189
- Florin L, Pegel A, Becker E, Hausser A, Olayioye MA, Kaufmann H (2009) Heterologous expression of the lipid transfer protein CERT increases therapeutic protein productivity of mammalian cells. *J Biotechnol* 141:84–90
- Yamane-Ohnuki N, Kinoshita S, Inoue-Urakubo M, Kusunoki M, Iida S, Nakano R, Wakitani M, Niwa R, Sakurada M, Uchida K, Shitara K, Satoh M (2004) Establishment of FUT8 knockout Chinese hamster ovary cells: an ideal host cell line for producing completely defucosylated antibodies with enhanced antibody-dependent cellular cytotoxicity. *Biotechnol Bioeng* 87:614–622
- Yamane-Ohnuki N, Yamano K, Satoh M (2008) Biallelic gene knockouts in Chinese hamster ovary cells. *Methods Mol Biol* 435:1–16
- Umana P, Jean-Mairet J, Moudry R, Amstutz H, Bailey JE (1999) Engineered glycoforms of an antineuroblastoma IgG1 with optimized antibody-dependent cellular cytotoxic activity. *Nat Biotechnol* 17:176–180

14. Birch JR, Racher AJ (2006) Antibody production. *Adv Drug Deliv Rev* 58:671–685
15. Durocher Y, Butler M (2009) Expression systems for therapeutic glycoprotein production. *Curr Opin Biotechnol* 20:700–707
16. Graham FL, Smiley J, Russell WC, Nairn R (1977) Characteristics of a human cell line transformed by DNA from human adenovirus type 5. *J Gen Virol* 36:59–74
17. Shaw G, Morse S, Ararat M, Graham FL (2002) Preferential transformation of human neuronal cells by human adenoviruses and the origin of HEK 293 cells. *FASEB J* 16:869–871
18. Jones D, Kroos N, Anema R, van Montfort B, Vooy A, van der Kraats S, van der Helm E, Smits S, Schouten J, Brouwer K, Lagerwerf F, van Berkel P, Opstelten DJ, Logtenberg T, Bout A (2003) High-level expression of recombinant IgG in the human cell line per.c6. *Biotechnol Prog* 19:163–168
19. Rose T, Winkler K, Brundke E, Jordan I, Sandig V (2005) Alternative strategies and new cell lines for high-level production of biopharmaceuticals. In: Knäblein J (ed) *Modern biopharmaceuticals*, Wiley-VCH, pp 761–777
20. Olivier S, Jacoby M, Brillou C, Bouletreau S, Mollet T, Nerriere O, Angel A, Danet S, Souttou B, Guehenneux F, Gauthier L, Berthome M, Vie H, Beltraminelli N, Mehtali M (2010) EB66 cell line, a duck embryonic stem cell-derived substrate for the industrial production of therapeutic monoclonal antibodies with enhanced ADCC activity. *MAbs* 2:405–415
21. Weber W, Fusenegger M (2004) Inducible gene expression in mammalian cells and mice. *Methods Mol Biol* 267:451–466
22. Kalwy S, Rance J, Young R (2006) Toward more efficient protein expression: keep the message simple. *Mol Biotechnol* 34:151–156
23. Schlatter S, Stansfield SH, Dennis DM, Racher AJ, Birch JR, James DC (2005) On the optimal ratio of heavy to light chain genes for efficient recombinant antibody production by CHO cells. *Biotechnol Prog* 21:122–133
24. Jostock T, Vanhove M, Brepoels E, Van Gool R, Daukandt M, Wehnert A, Van Hegelsom R, Dransfield D, Sexton D, Devlin M, Ley A, Hoogenboom H, Mullberg J (2004) Rapid generation of functional human IgG antibodies derived from Fab-on-phage display libraries. *J Immunol Methods* 289:65–80
25. Li J, Menzel C, Meier D, Zhang C, Dubel S, Jostock T (2007) A comparative study of different vector designs for the mammalian expression of recombinant IgG antibodies. *J Immunol Methods* 318:113–124
26. Li J, Zhang C, Jostock T, Dubel S (2007) Analysis of IgG heavy chain to light chain ratio with mutant Encephalomyocarditis virus internal ribosome entry site. *Protein Eng Des Sel* 20: 491–496
27. Borman AM, Deliat FG, Kean KM (1994) Sequences within the poliovirus internal ribosome entry segment control viral RNA synthesis. *EMBO J* 13:3149–3157
28. Fang J, Qian JJ, Yi S, Harding TC, Tu GH, VanRoey M, Jooss K (2005) Stable antibody expression at therapeutic levels using the 2A peptide. *Nat Biotechnol* 23:584–590
29. Jostock T, Dragic Z, Fang J, Jooss K, Wilms B, Knopf HP (2010) Combination of the 2A/furin technology with an animal component free cell line development platform process. *Appl Microbiol Biotechnol* 87:1517–1524
30. Oumard A, Qiao J, Jostock T, Li J, Bode J (2006) Recommended method for chromosome exploitation: RMCE-based cassette-exchange systems in animal cell biotechnology. *Cytotechnology* 50:93–108
31. Bode J, Schlake T, Iber M, Schubeler D, Seibler J, Snezhkov E, Nikolaev L (2000) The transgeneticist's toolbox: novel methods for the targeted modification of eukaryotic genomes. *Biol Chem* 381:801–813
32. O'Gorman S, Fox DT, Wahl GM (1991) Recombinase-mediated gene activation and site-specific integration in mammalian cells. *Science* 251:1351–1355
33. Fukushige S, Sauer B (1992) Genomic targeting with a positive-selection lox integration vector allows highly reproducible gene expression in mammalian cells. *Proc Natl Acad Sci U S A* 89:7905–7909
34. Arnould S, Delenda C, Grizot S, Desseaux C, Paques F, Silva GH, Smith J (2010) The I-CreI meganuclease and its engineered derivatives: applications from cell modification to gene therapy. *Protein Eng Des Sel* 24:27–31
35. Porteus MH, Carroll D (2005) Gene targeting using zinc finger nucleases. *Nat Biotechnol* 23: 967–973
36. Kennard ML, Goosney DL, Monteith D, Zhang L, Moffat M, Fischer D, Mott J (2009) The generation of stable, high MAb expressing CHO cell lines based on the artificial chromosome expression (ACE) technology. *Biotechnol Bioeng* 104:540–553
37. Cacciatore JJ, Chasin LA, Leonard EF (2010) Gene amplification and vector engineering to achieve rapid and high-level therapeutic protein production using the Dhfr-based CHO cell selection system. *Biotechnol Adv* 28: 673–681

38. Mott J (2011) Cell line development and engineering conference, Munich
39. Jostock T, Knopf H-P, Wilms B, Drori S, Assaraf YGA (2010) Antibody development and production conference. IBCLifeScience, Carlsbad, CA
40. Browne SM, Al-Rubeai M (2007) Selection methods for high-producing mammalian cell lines. *Trends Biotechnol* 25:425–432
41. Carroll S, Al-Rubeai M (2004) The selection of high-producing cell lines using flow cytometry and cell sorting. *Expert Opin Biol Ther* 4:1821–1829
42. Manz R, Assenmacher M, Pfluger E, Miltenyi S, Radbruch A (1995) Analysis and sorting of live cells according to secreted molecules, relocated to a cell-surface affinity matrix. *Proc Natl Acad Sci U S A* 92:1921–1925
43. Holmes P, Al-Rubeai M (1999) Improved cell line development by a high throughput affinity capture surface display technique to select for high secretors. *J Immunol Methods* 230: 141–147
44. Borth N, Zeyda M, Kunert R, Katinger H (2000) Efficient selection of high-producing subclones during gene amplification of recombinant Chinese hamster ovary cells by flow cytometry and cell sorting. *Biotechnol Bioeng* 71:266–273
45. Brezinsky SC, Chiang GG, Szilvasi A, Mohan S, Shapiro RI, MacLean A, Sisk W, Thill G (2003) A simple method for enriching populations of transfected CHO cells for cells of higher specific productivity. *J Immunol Methods* 277: 141–155
46. DeMaria CT, Cairns V, Schwarz C, Zhang J, Guerin M, Zuena E, Estes S, Karey KP (2007) Accelerated clone selection for recombinant CHO CELLS using a FACS-based high-throughput screen. *Biotechnol Prog* 23: 465–472
47. Sleiman RJ, Gray PP, McCall MN, Codamo J, Sunstrom NA (2008) Accelerated cell line development using two-color fluorescence activated cell sorting to select highly expressing antibody-producing clones. *Biotechnol Bioeng* 99:578–587
48. Powell KT, Weaver JC (1990) Gel microdroplets and flow cytometry: rapid determination of antibody secretion by individual cells within a cell population. *Biotechnology (N Y)* 8: 333–337
49. Kenney JS, Gray F, Ancel MH, Dunne JF (1995) Production of monoclonal antibodies using a secretion capture report web. *Bio-technology (N Y)* 13:787–790
50. Porter AJ, Dickson AJ, Racher AJ (2010) Strategies for selecting recombinant CHO cell lines for cGMP manufacturing: realizing the potential in bioreactors. *Biotechnol Prog* 26: 1446–1454
51. Porter AJ, Racher AJ, Preziosi R, Dickson AJ (2010) Strategies for selecting recombinant CHO cell lines for cGMP manufacturing: improving the efficiency of cell line generation. *Biotechnol Prog* 26:1455–1464

Chapter 16

Transgenic Expression of Therapeutic Proteins in *Arabidopsis thaliana* Seed

Cory L. Nykiforuk and Joseph G. Boothe

Abstract

The production of therapeutic proteins in plant seed augments alternative production platforms such as microbial fermentation, cell-based systems, transgenic animals, and other recombinant plant production systems to meet increasing demands for the existing biologics, drugs under evaluation, and undiscovered therapeutics in the future. We have developed upstream purification technologies for oilseeds which are designed to cost-effectively purify therapeutic proteins amenable to conventional downstream manufacture. A very useful tool in these endeavors is the plant model system *Arabidopsis thaliana*. The current chapter describes the rationale and methods used to over-express potential therapeutic products in *A. thaliana* seed for evaluation and definitive insight into whether our production platform, Safflower, can be utilized for large-scale manufacture.

Key words: *Arabidopsis thaliana*, *Agrobacterium*-mediated transformation, Transgenic seed, Therapeutic protein, Insulin

1. Introduction

Recombinant protein production in plants has provided an invaluable tool in linking gene with function, understanding control of expression, identifying functional domains, elucidating active sites in enzymes, delineating protein conformation and interaction, etc. This understanding has accelerated a closely related branch of biotechnology referred to as plant molecular farming (PMF) (1–4). For the pharmaceutical industry, the ability to over-express and accumulate therapeutic proteins holds particular promise for targets requiring large-scale production in a cost-effective manner (5, 6). As a eukaryotic production system, plants possess the cellular machinery associated with posttranslational modifications required for proper folding *in vivo* resulting in biologically active conformations,

and thereby negating the necessity for in vitro refolding required namely in some microbially produced therapeutics. Advancements in the expression and glycol engineering of recombinant proteins in plants (7–10) increase the applicability of plant-based production platforms to include antibodies and/or other “humanized” therapeutics. Products can be targeted or compartmentalized within the cell for stable accumulation. Perhaps the most obvious advantage is the existing agricultural infrastructure and capacity to generate metric tonnes of feedstock, where the efficiency of production relies upon the energy of the sun and nutrients of the soil to convert organic material into therapeutic proteins. In addition, by selectively expressing the therapeutic target in seeds, our system offers the added advantage of long-term, cost-effective, stable storage in comparison to other available systems (11).

SemBioSys Genetics Inc. has developed technologies designed for the large-scale production of therapeutic proteins in oilseeds (11, 12). This technology involves targeting to subcellular organelles known as oilbodies that enables cost-effective recovery of recombinant pharmaceutical proteins for therapeutic applications in humans and animals. A key component of this technology relies on the high-level expression of proteins in oilseeds (e.g., 0.1–5.0% expression of total seed protein in *Arabidopsis* is equivalent to roughly 0.2–5 mg expression target/gram seed). Over-expression of targets in *Arabidopsis thaliana* seed provides assessment of cellular compartmentation coupled with construct configurations designed for oilbody technology (Stratoderm™ and Stratocapture™ (12)). It also provides accelerated access to seed material for biochemical/structural characterization, small-scale process development (protein purification), and biological activity in preclinical studies. These analyses provide insight and predictability before expression and purification within our production platform, *Carthamus tinctorius*, for large-scale clinical evaluation and commercial manufacture.

From an experimental standpoint, *A. thaliana* affords researchers with a model system in which the complete genome has been sequenced (13, 14), and thus provides a sequence-based map which continues to be augmented by markers associated with single-nucleotide polymorphisms and insertion–deletion polymorphisms (15). This can serve as a valuable tool when sequencing the context of transgene(s) insertion by TAIL-PCR (16) and mapping the location(s) of inserts following *Arabidopsis* transformation experiments. As an established research model, collaborative efforts within the *Arabidopsis* community are accompanied by a wealth of publications covering a vast array of research topics. These extensive publications provide a reliable database (The Arabidopsis Information Resource (TAIR) at <http://www.arabidopsis.org/>) from which to exploit information and insight when interpreting expression results. Advancements in transformation coupled with

concerted efforts within this community have also established seed stock centers (NASC: The European Arabidopsis Stock Centre; ABRC: Arabidopsis Biological Resource Center) committed to the preservation and accession of different ecotypes and numerous mutants useful in understanding and/or optimizing the expression of certain therapeutic targets. From a physical standpoint, the small size and rapid development time of *Arabidopsis* provides an affordable plant expression system requiring reduced growth space together with high seed yield. In addition, containment between different transformants can be easily managed and the risk of outcrossing is mitigated because *Arabidopsis* is a self-pollinator.

For the purposes of this chapter, we have selected the over-expression of insulin in *A. thaliana* seeds originally described in Nykiforuk et al. (17) as an example of the methods employed to derive transgenic seed. This includes preparation of the wild-type *Arabidopsis* host for transformation using the floral dip method (18) together with the design, rationale, and preparation of plant binary vectors used in *Agrobacterium*-mediated transformation (19). Finally, the methods used to select recombinant T1 seed for further propagation and analysis are included. Where possible, we provide detailed descriptions to avoid potential pitfalls that may negatively impact the efficiency of transformation and ultimately the successful expression of the target in *Arabidopsis* seeds.

2. Materials

A. thaliana L. cv Columbia (C24) seed (ABRC at The Ohio State University) (see Notes 1–3). Unless otherwise noted, solutions should be prepared with Milli-Q water (Type 1; <http://www.millipore.com/>) and chemicals and reagents should be of the highest grade possible. Typical lab equipment for vector construction (thermocyclers, agarose gel boxes, incubators, etc.) and *Arabidopsis* growth (environmental growth chambers; Percival Scientific, Perry, IA, USA; ArabidSun Lighting System, Lehle Seeds, Round Rock, TX, USA) may or may not be available; so where appropriate, conditions have been supplied for adaptation. Commercial kits related to DNA manipulation, subcloning, and electroporation (plasmid isolation, incubators, sequence analysis, etc.) should adhere to manufacturer's protocols.

2.1. Plant Binary Vector

1. For the plant binary vector, SemBioSys employed a modified version of *pPZP200* described by Hajdukiewicz et al. (20). The genetic elements outside the T-DNA include a Spectinomycin resistance gene (*aadA* encoding amino-glycoside-3'-adenyltransferase) for selection in *DH5α* subclones and for double selection in kanamycin-resistant *EHA101*

Agrobacterium, *ColE1*, and *pVSI* plasmid origins for replication in *Escherichia coli* and *Agrobacterium*, respectively (see Notes 4 and 5).

2. Nykiforuk et al. (17) describes the genetic elements between the right and left T-DNA borders resulting in the seed-specific expression of an oleosin–insulin fusion protein with an intervening cleavage peptide and constitutive expression of a phosphothricin acetyltransferase (PAT) gene for selection by phosphinothricin (PPT; Crescent Chemical Co., Islandia, NY, USA) (see Note 6).
3. Prior to plant binary vector construction, sequence analysis regarding plant-based expression was performed (see Note 7). In addition, particular care and attention was used to ensure expression levels and subsequent oilbody-based purification and maturation in vitro were coupled (see Notes 8 and 9).

2.2. Competent *E. coli* Bacteria and EH101 Agrobacterium

Equipment related to these activities include a visible spectrophotometer, environmental shaking incubator (variable speed), incubator (capable of controlled temperatures ranging from ambient temperature to 37°C), variable speed and temperature-controlled centrifuge(s) with appropriate rotors for centrifuge tubes ranging in size from 15 to 250 ml, and appropriate Pyrex® glassware (autoclavable). Dewar flask and liquid nitrogen are used for the preparation of competent cells and therefore are not required if competent cells are obtained elsewhere as frozen glycerol stocks. For electroporation of *EHA101 Agrobacterium*, a Bio-Rad Gene Pulser® II (Bio-Rad Laboratories Inc) or equivalent can be used. Other materials can be obtained from appropriate vendors including Falcon 2059 tubes (Becton Dickinson Labware, Franklin Lakes, NJ, USA) and Oakridge tubes (Nalgene, Thermo Fisher Scientific, Rochester, NY, USA) for pelleting competent cells.

1. *E. coli* DH5α competent cells can be purchased commercially (available from numerous sources) or prepared.
2. Agrobacterium tumefaciens EHA101 carrying the T1 plasmid (20, 21). Other useful *Agrobacterium* strains and their availability are reported in the literature (22).
3. *Super optimal broth (SOC medium)*. 2% Granulated tryptone (w/v), 0.5% yeast extract (w/v), 10 mM NaCl, 2.5 mM KCl, 10 mM MgCl₂, 10 mM MgSO₄, and 20 mM glucose, adjust to pH 7.0 with 1 M NaOH. The solution must be filter sterilized through a 0.2-μm filter (see Note 10).
4. *Luria broth (LB; also sometimes called Lenox broth)*. Dissolve 1% granulated tryptone (w/v), 0.5% yeast extract (w/v), 0.5% sodium chloride (w/v) in Milli-Q water, adjust to pH 7.0 with 1 M NaOH. Autoclave and cool before storing at 4°C. For antibiotic selection add appropriate amount of filter-sterilized antibiotic (from stock solution), to result in final concentrations

- of 100 µg/ml Spectinomycin ($\text{LB}_{\text{Spec}100}$) and Kanamycin ($\text{LB}_{\text{Kan}100}$) (see Note 11).
5. *Spectinomycin stock (10 mg/ml)* (see Note 12). Dissolve 10 mg/ml Spectinomycin dihydrochloride in Milli-Q water and filter sterilize. Aseptically aliquot Spectinomycin stock solution into blue capped Falcon tubes (BD Falcon™, 15-ml high-clarity polypropylene conical tube OR Falcon® 2070 Blue Max™ 50-ml polypropylene conical tubes; Becton Dickinson Labware, Franklin Lakes, NJ, USA) and store at -20°C.
 6. *Kanamycin stock (10 mg/ml)*. Dissolve 10 mg/ml Kanamycin Monosulfate in Milli-Q water and filter sterilize. Aseptically aliquot Kanamycin stock solution into blue capped Falcon tubes and store at -20°C.
 7. *LB agar*. Mix 1.2% granulated pure agar (Cat# 1.01614, EMD Science) after pH adjustment of LB and prior to final volume adjustment. Autoclave and cool to 54°C in preset water bath. Add antibiotics as required where $\text{LB}_{\text{Spec}100}=100$ µg/ml Spectinomycin, $\text{LB}_{\text{Kan}100}=100$ µg/ml Kanamycin, and $\text{LB}_{\text{Kan}100\text{Spec}100}=100$ µg/ml Kanamycin and Spectinomycin. Stir to distribute agar evenly before pouring into Petri dishes (~30 ml/plate) (Falcon 1005, Optilux™, 100×20 Petri dish, Becton Dickinson Labware, Franklin Lakes, NJ, USA). Let plates dry for 2 days in flow hood at room temperature, after which plates can be bagged and stored at 4°C for up to 3 months.
 8. *Agrobacterium broth (AB)*—(1) *PreAB salt solution*. In a 1 L Pyrex® bottle, measure 462 ml Milli-Q water and 25 ml 20× AB salts (see below). After stirring, sterilize by autoclaving. (2) *ABDex*. Dilute 50 µl of a 1 M CaCl_2 solution (filter sterilized) into 1 ml sterile Milli-Q water by pipetting. Into 7 ml 2 M dextrose solution (filter sterilized), add with stirring 5 ml of $\text{FeSO}_4 \cdot 7\text{H}_2\text{O}$ (25% w/v stock solution filter sterilized), 0.5 ml of 1 M MgSO_4 stock solution (filter sterilized), and diluted CaCl_2 solution. (3) *AB*. Combine PreAB salt solution (487 ml) and *ABDex* and add antibiotics as required. For $\text{AB}_{\text{Kan}100}$, add from Kanamycin stock so that final is 100 µg/ml; for $\text{AB}_{\text{Spec}100}$, add from Spectinomycin stock so that final is 100 µg/ml; and for $\text{AB}_{\text{Kan}100\text{Spec}100}$, add from Kanamycin and Spectinomycin stocks so that each is 100 µg/ml.
 9. *AB selection plates*. Mix and autoclave 925 ml Milli-Q water and 15 g agar. When cool to touch add 50 ml 20× AB salts (see below), 13.9 ml Glucose (2 M stock solution filter sterilized), 10 ml $\text{FeSO}_4 \cdot 7\text{H}_2\text{O}$ (25% w/v stock solution filter sterilized), 1 ml 1 M MgSO_4 stock solution (filter sterilized), and 100 µl of diluted CaCl_2 solution (dilute 50 µl of a 1 M CaCl_2 solution filter sterilized into 1 ml sterile Milli-Q water). For $\text{AB}_{\text{Kan}100}$, add from Kanamycin stock so that final is 100 µg/ml;

for AB_{Spec100}, add from Spectinomycin stock so that final is 100 µg/ml; and for AB_{Kan100Spec100}, add from Kanamycin and Spectinomycin stocks so that each is 100 µg/ml. Pour agar solution with or without appropriate antibiotics into plates (~30 ml/plate).

10. *20× AB salts.* 688.9 mM K₂HPO₄, 333.3 mM NaH₂PO₄, 747.8 mM NH₄Cl, 80.5 mM KCl, sterilized by autoclaving. Aliquot aseptically under fume hood into 50-ml Falcon tubes and store at room temperature.
11. *DMSO.* Dimethyl sulfoxide.
12. *TE buffer.* 10 mM Tris-HCl, 1 mM EDTA, pH 8.0, autoclaved.
13. *YEP broth.* 1% (w/v) Peptone from casein (a.k.a. Tryptone), 1% (w/v) yeast extract, 0.5% NaCl in Milli-Q water, adjust pH to 7.5 with 1 M NaOH, sterilized by autoclaving. For YEP_{Kan100}, add from Kanamycin stock so that final is 100 µg/ml; for YEP_{Spec100}, add from Spectinomycin stock so that final is 100 µg/ml; and for YEP_{Kan100Spec100}, add from Kanamycin and Spectinomycin stocks so that each is 100 µg/ml.

2.3. Preparation of *Arabidopsis* for In Planta Transformation

1. Acid washed sea sand (Cat# VW3358-3, VWR, Arlington Heights, IL, USA).
2. Sunshine #4 soil mixture (Sun Gro Horticulture, Bellevue, WA, USA) is formulated with Canadian sphagnum peat moss, coarse grade perlite, gypsum, dolomitic lime, and long-lasting wetting agent. Dolomitic limestone provides buffering between 5.0 and 7.0, along with calcium and magnesium. Gypsum supplies a source of sulfur and calcium. Peter's (J.R. Peters, Inc, Allentown, PA, USA) fertilizer (Nitrogen:Phosphorus:Potassium, 20:19:18).
3. Ten 4" green pots, ten squares of screen to cover 4" pots, ten elastic bands (size 24, 6" × 1/16" Grand&Toy). For convenience, pots can be held in Aratrays (Lehle Seeds, Round Rock, TX, USA).
4. Growth chamber (Percival Scientific) with variable temperature setting, light cycling alternatively ArabidSun growth racks (Lehle Seeds) can be used with walk-in growth chambers or equivalent.

2.4. In Planta Transformation of *Arabidopsis* C24

1. *½ MS selection plates.* Combine 5 ml 100× NH₄NO₃, 5 ml 100× KNO₃, 5 ml 100× CaCl₂·2H₂O, 5 ml 100× MgSO₄·7H₂O, 5 ml 100× KH₂PO₄, 5 ml 100× micros, 1 ml 1,000× organics, and 10 g glucose. Adjust volume to 950 ml with Milli-Q water. Adjust to pH 5.7–5.8 with 1 M KOH. Add 6.5 g/L Phytablend agar (Caisson Laboratories, North Logan, UT, USA) and bring volume to 1 L with Milli-Q water. Autoclave and cool media to 54°C. Add selection components to media maintained at 54°C; 1 ml/L of Timentin (300 mg/L stock solution

- filter sterilized), 1.5 ml/L PPT (10 mg/ml sterile stock solution), and 5 ml/L Kanamycin (10 mg/ml stock solution, filter sterilized). Aliquot ~40 ml per tissue culture plate, dry in fume hood, and store in sleeve at 4°C until ready to use.
2. 100× NH₄NO₃. 2 M stock solution, filter sterilized and stored at room temperature for up to 2 months.
 3. 100× KNO₃. 1.88 M stock solution, filter sterilized and stored at room temperature for up to 2 months.
 4. 100× CaCl₂·2H₂O. 300 mM stock solution, filter sterilized and stored at room temperature for up to 2 months.
 5. 100× MgSO₄·7H₂O. 150 mM stock solution, filter sterilized and stored at room temperature for up to 2 months.
 6. 100× KH₂PO₄. 125 mM stock solution, filter sterilized and stored at room temperature for up to 2 months.
 7. 100× Micros. 10 mM H₃BO₃, 0.5 mM KI, 10 mM MnSO₄·H₂O, 3 mM ZnSO₄·7H₂O, 0.1 mM Na₂MoO₄·2H₂O, 10 µM CuSO₄·5H₂O, and 10.5 µM CoCl₂·5H₂O. Filter sterilize and aliquot. Good for 1 month at room temperature and 2 months at 4°C.
 8. 1,000× Organics (vitamin component). 8 mM nicotinic acid, 5 mM pyridoxine hydrochloride, and 30 mM thiamine hydrochloride, aliquot and store at -20°C. Keep tubes thawed at 4°C for immediate use.
 9. PPT (10 mg/ml). Phosphinothricin/glufosinate ammonium salt, filter sterilized after completely dissolved in sterile Milli-Q water, stored at 4°C.
 10. Timentin (300 mg/L). Timentin stock solution, filter sterilized and aliquoted (1 ml) into sterile microfuge tubes, stored at -20°C.
 11. Top agar. 0.55–0.6% Phytoblend (Caisson Laboratories, North Logan, UT, USA) agar (w/v) dissolved in ddH₂O water, autoclaved, and stored at 4°C.
 12. ArabiSun (Lehle growing rack, Round Rock, TX, USA).
 13. Aracons (base and tube, seed collection system; Lehle Seeds, Round Rock, TX, USA).
 14. Silwet L-77 (surfactant; OSi Specialties Inc, Danbury, CT, USA) to a final 0.05% (w/v).

3. Methods

For all methods, refer to Fig. 1 and corresponding Table 1 in order to coordinate activities according to the integrated timelines. It is important to track and indicate what steps of the transformation protocol have been performed. Informational databases allow the

identification of transformation experiments, what gene construct is being expressed, and results associated with each transformation experiment. Experiments can also be monitored by preparing a checklist to ensure that all steps were followed.

3.1. Preparation and Transformation of Chemically Competent *E. coli* DH5 α (Adapted from Ref. 23) Cells

E. coli cells are utilized to amplify and purify the pSBS4405 plant binary vector for subsequent transformation into electrocompetent *EHA101 Agrobacterium*. If competent *E. coli* cells are purchased, follow the manufacturer's protocols.

1. Aseptically inoculate a 5 ml LB culture with *DH5 α* (no selection) and incubate overnight at 37°C on shaker. After about 24 h inoculate 500 ml LB broth within a 2-L Erlenmeyer flask and incubate overnight at room temperature on shaking incubator at 225 rpm (inoculate 1:500 if ambient temperature is 22–23°C and 1:1,000 if ambient temperature is higher) (if using different *E. coli* strain, see Note 13).

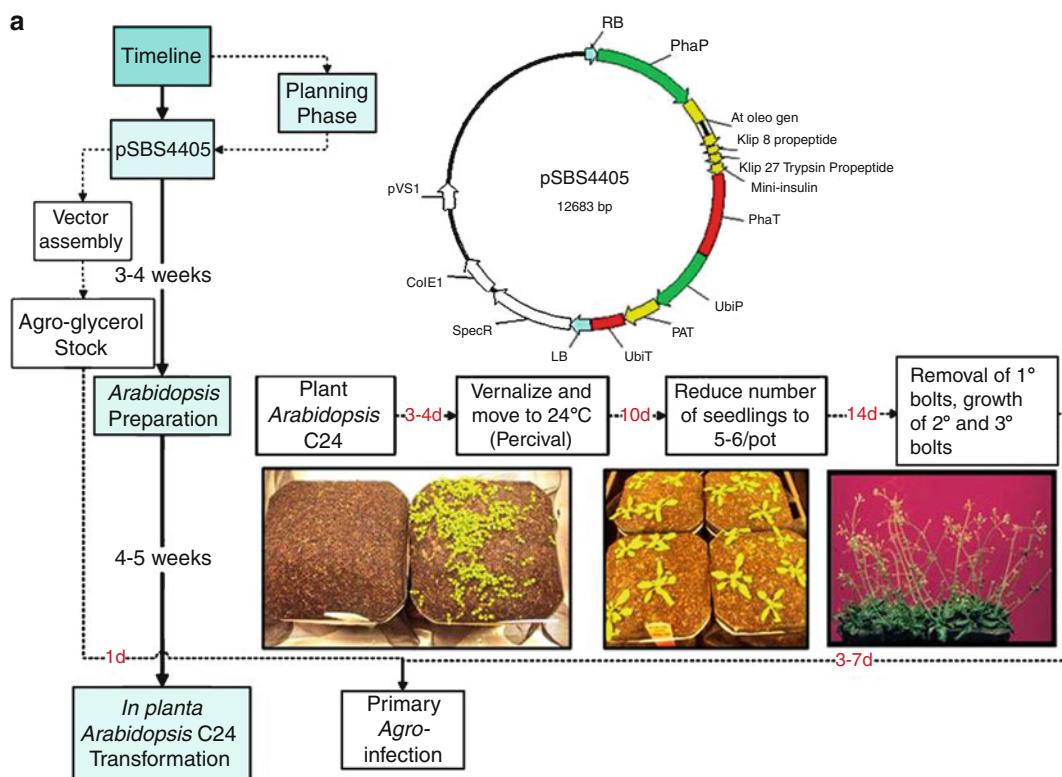


Fig. 1. Integrated timeline of *Arabidopsis thaliana* *in planta* transformation with pSBS4405. The corresponding units of operation with additional comments are provided in Table 1. (a) The plant binary vector preplanning phase, vector assembly, and concurrent preparation of wild-type C24 *Arabidopsis* as they converge for *in planta* transformation. (b) *In planta* transformation of *Arabidopsis* combined with harvest and selection of T1 seeds to generate T2 and subsequent generations.

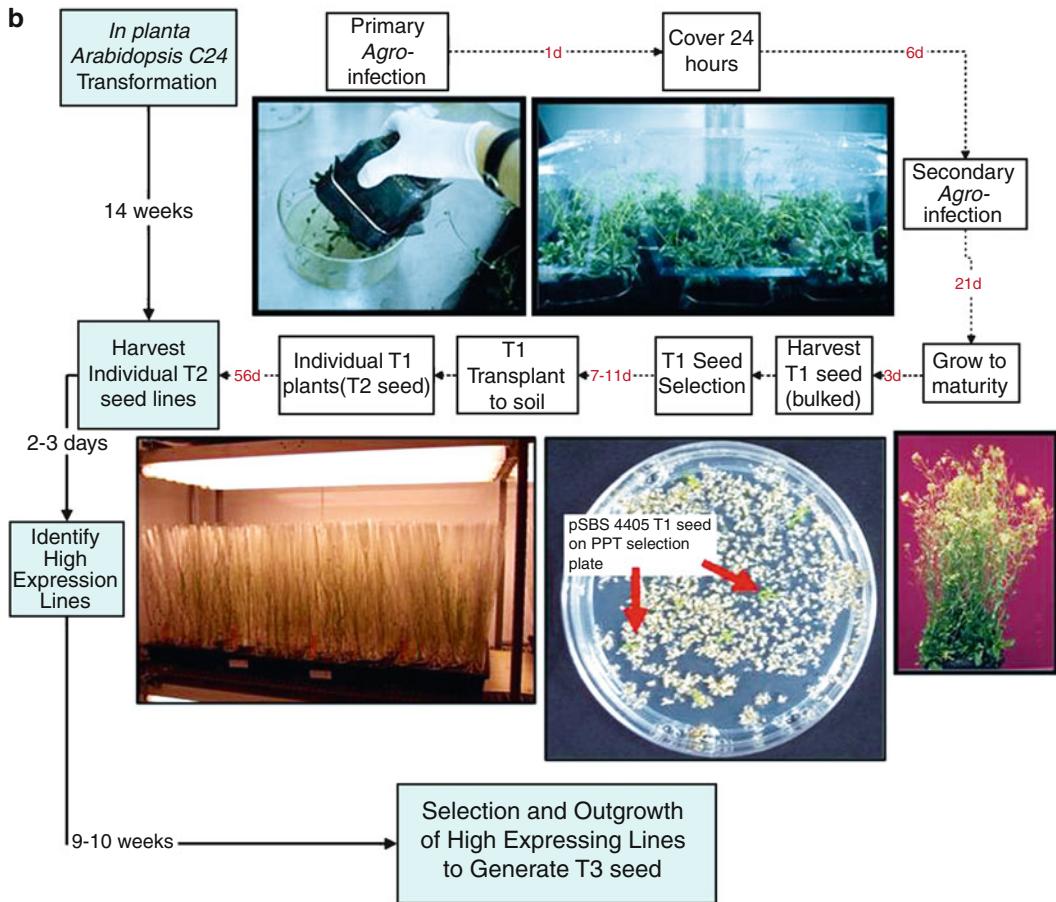


Fig. 1. (continued).

2. The following morning measure the optical density (OD) of the culture and harvest the cells when the $A_{600} = 0.3\text{--}0.4$. From this point onwards maintain the culture/cells on ice. When the appropriate turbidity is obtained transfer the culture to appropriate vessel and pellet cells by centrifugation for 10 min at $3,700 \times g$ while maintaining at 4°C (see Note 14). Depending upon the centrifugation volumes, additional rounds of pelleting can be performed using the conditions above (see Note 15).
3. Resuspend each pellet in 1 part ice-cold TB buffer (filter-sterilized 10 mM PIPES, 15 mM CaCl_2 , 250 mM KCl, pH 6.7, 55 mM MnCl_2) to 16.6 parts of original culture volume and maintain on ice for 10 min.
4. Centrifuge resuspension as above to pellet cells, decant supernatant, and resuspend each TB-washed pellet by gentle swirling in 1 part ice-cold TB buffer to 53.7 parts of original culture. Next, slowly add DMSO in a fume hood to a final concentration of 7% (based on volume of TB buffer) while keeping cells cold. Maintain resuspended cells on ice overnight.

Table 1
Integrated timeline with corresponding comments for transgenic expression of pSBS4405 *Arabidopsis thaliana*

Cumulative timeline	Unit of operation	Subtask (time) Methodology/comments
<0 (Preplanning stage)	pSBS4405 plant binary vector	<p><i>Planning phase (vector design completed prior to initiating transformation)</i></p> <ol style="list-style-type: none"> 1. Ensure that selectable marker for bacterial growth is compatible with <i>Agrobacterium</i> strain (see Note 5). 2. Ensure that origins of replication are suitable for amplification in bacterial subcloning host and <i>Agrobacterium</i> (see Note 5). 3. Should codon usage be optimized for expression in plants? 4. Are there any cryptic splice sites within the transgene? (see Note 7) 5. Are there any internal polyadenylation sites within the transgene? (see Note 7) 6. Are there any nucleotide palindromes or potential hairpins that may impair transcription/translation? 7. If expressing as a fusion protein, what cleavage strategy should be employed? (see Notes 8 and 9) 8. Are appropriate signal peptides required for targeting expression through the secretory pathway (ER and apoplast), or is expression to be targeted to plastid (plastid signal peptide), oilbody (oleosin), or cytosol?
3–4 Weeks (3–4 weeks)	pSBS4405 plant binary vector	<p><i>Vector assembly (3–4 weeks)</i></p> <ol style="list-style-type: none"> 1. For pSBS4405^a, the Klip27-Mini-insulin was synthesized from overlapping oligonucleotides ligated via <i>Bsu36I</i>. The assembled Klip27-Mini-insulin dsDNA fragment was then amplified using high-fidelity PCR with engineered restriction sites. The PCR fragment was subcloned in DH5α by T4 DNA ligation and subsequently amplified for directional in-frame cloning of a <i>XhoI</i>-Klip27-Mini-insulin-<i>HindIII</i> fragment into a pSBS plant binary vector already housing the <i>Arabidopsis</i> 18 kDa oleosin-Klip8 with an MCS under the transcriptional control of <i>Phaseolin</i> promoter-terminator. Alternatively other means of constructing the plant binary vector are available including synthesized cDNAs with incorporated restriction sites for directional cloning^b, using high-fidelity PCR methods (engineered for directional cloning with restriction sites incorporated into amplicon), restriction fragments from existing plasmids (compatible ends for MCS, or blunt end cloning), etc. Whatever methods are used, it is important to subclone the cDNA into a bacterial host for long-term storage, verification of sequence, source of template for labeling probes, or positive control for PCR screening/restriction mapping. 2. Restrict cDNA from purified plasmids and ligate cDNA into precut plant binary vector (disarmed) cassette^c. 3. Transform chemically competent DH5α <i>Escherichia coli</i>. 4. Plate and isolate individual colonies. 5. Screen individual colonies for positive clones (PCR and/or restriction digest) with appropriate positive and negative controls. 6. Generate long-term glycerol stock(s) clearly labeled (see Note 4). 7. Transform electrocompetent <i>Agrobacterium</i> strain with purified plasmid from single colony isolated from glycerol stock (pSBS4405-DH5α). 8. Plate, screen individual clones, verify positive clones, and prepare long-term glycerol stock. Update tracking file.

(continued)

Table 1
(continued)

Cumulative timeline	Unit of operation	Subtask (time) Methodology/comments
7–9 Weeks	<i>Arabidopsis</i> preparation (4–5 weeks)	<p><i>Plant Arabidopsis C24</i></p> <p><i>Vernalization treatment (3–4 days)</i> 1. Helps synchronize germination.</p> <p><i>Transfer tray to percival (10 days)</i> 1. Following vernalization treatment transfer germinated plants to growth chamber at 24°C under continuous light.</p> <p><i>Thin individual pots down to 5–6 plants (7 days)</i> 1. Try and retain rosettes at same stage of development.</p> <p><i>Transfer plants to Lehle rack (7 days)</i></p> <p><i>Remove primary inflorescence to promote growth of secondary and tertiary bolts (7 days)</i> 1. Remove primary inflorescence when about 2-cm tall.</p>
21–23 Weeks	<i>In planta</i> <i>Arabidopsis</i> C24 transformation (14 weeks)	<p><i>Primary Agrobacterium inoculation by floral dip (7 days)</i> 1. Following the floral dip, the trays are covered for 24 h to retain humidity and returned to Lehle racks.</p> <p><i>Secondary Agrobacterium inoculation by floral dip (24 days)</i> 1. A second dip will increase transformation efficiency, targeting susceptible ovules following 1 week's growth after the primary inoculation. 2. Following the second floral dip the trays are returned to Lehle racks and plants are grown to maturity. 3. Irrigation of plants occurs every 3 days with application of fertilizer every 5 days. 4. After 3 weeks, watering is stopped and plants are allowed to dry out for 3–4 days.</p> <p><i>Harvest T1 seed (2 days)</i> 1. Seed can be harvested from excised siliques and bulked.</p> <p><i>T1 seed selection (7–11 days)</i> 1. Seeds are plated in top agar on MS selection media (PPT). 2. Include a negative control (wild-type C24) and if possible a positive control (PPT-resistant seed line); otherwise, consider germinating in top agar in the absence of selection.</p> <p><i>Transfer T1 seedlings to soil (56 days)</i> 1. Transfer only seedlings in which the root has penetrated the MS selection media. 2. Seedlings are transferred to soil placed in trays (Araflats, Lehle seeds). 3. Once rosettes form the individual plants are covered with a seed collection system (Aracons, Lehle seeds) to prevent any cross-pollination and allow individual T2 seed to harvest.</p>

(continued)

Table 1
(continued)

Cumulative timeline	Unit of operation	Subtask (time) Methodology/comments
T2 seed available for analysis of expression levels		<p><i>Harvest individual T2 seed lines (2 days)</i></p> <p><i>Screen T2 seed lines for expression of transgene (2–3 days)</i></p> <p>1. Total seed soluble protein can be extracted and separated on Coomassie-stained gels and corresponding Western blots probed with antibodies specific against target (insulin for pSBS4405) or fusion partner (oleosin for pSBS4405).</p> <p>2. Alternative means of assessing transgene expression may involve developed ELISA.</p>
30–33 Weeks	T3 seed	<p><i>Selection and Outgrowth of High Expression lines to generate T3 seed (9–10 weeks)</i></p> <p>1. T2 seed is plated in top agar on MS selection plates, transferred to soil, and individually matured as outlined above and in Fig. 1. Selection performed on this basis will allow assessment of segregation (negative seed will be culled during selection) and potentially isolate homozygous lines (“genetically stable”) depending upon segregation of events.</p> <p>2. Alternatively, T2 seed can be pooled and seeded en masse to bulk seed for analyses related to process development or providing sufficient materials for biochemical characterization and/or biological function. In this instance selection would require spraying plants with herbicide and culling negative lines.</p>
<i>Additional outgrowths (9–10 weeks)</i>		
1. At this point the ability to outgrow, bulk, or develop homozygous lines is at the discretion of the researcher.		

^aRefer to Nykiforuk et al. 2006 (17) for cloning details and vector assembly

^bSynthesized clones are typically supplied as purified plasmid or subclone in which the sequence has been verified. If provided as a purified plasmid, subclone into an appropriate *E. coli* strain for long-term storage and manipulation. If provided as a subclone, amplify on selection media for long-term glycerol stocks and a source of purified plasmid for excision of the cDNA for cloning into disarmed plant binary vector

^cThe current strategy assumes that the plant binary vector contains a multiple cloning site (MCS) for cloning in the correct orientation (directional cloning) and within the right context of promoter/terminator cassette and/or in-frame with fusion protein partner

5. The following day, aseptically aliquot 100 µl into individual sterile microfuge tubes and flash freeze in liquid nitrogen before long-term storage at –80°C. Competent cells aliquoted and stored in this manner should be used only once, so do not remove and return cells once thawed for use in laboratory.
6. For transformation of competent *E. coli* DH5α, retrieve a frozen 100 µl aliquot and thaw on ice. Transfer the 100 µl aliquot by pipette to 17×100-ml round-bottomed polypropylene Falcon 2059 tubes and return to ice. After 10 min, pipette 5–10 µl of purified pSBS4405 (25–50 ng plant binary vector diluted in TE buffer), mix by flicking tube (do not pipette up and down to mix) with finger, and return to ice for an

additional 10 min. Thereafter, heat shock by submersing the bottom of the tube in a 42°C water bath for 45 s and return to ice for 2 min. Add 0.9 ml room-temperature SOC medium and incubate at 37°C shaking at 225 rpm for 1 h. At this point 50–100 µl is spread aseptically on LB_{Spec₁₀₀} agar plates and incubated at 37°C overnight (see Note 16). Confirmation of positive clones was performed by amplifying single colonies in LB_{Spec₁₀₀} broth, followed by plasmid purification (plasmid preparation kits are readily available commercially), restriction mapping, PCR and/or sequencing, etc. Long-term DH5α-pSBS4405 clones were maintained in amplified culture mixed with an equal volume of 50% glycerol (glycerol:sterile Milli Q water v/v) at -80°C.

3.2. Preparation and Transformation of Electrocompetent EHA101 Agrobacterium with pSBS4405 Plant Binary Vector

Long-term glycerol stocks should be clearly labeled with date, construct ID, and other appropriate tracking strategies. Complete a corresponding permanent tracking record/file describing the date, construct ID, storage location, genetic elements (annotated plasmid map), bacterial host, antibiotic selection, and composition of glycerol stock. The importance of maintaining and preserving an accurate database for glycerol stocks (bacterial, *Agrobacterium*, etc.) can prevent loss of biological assets, provide modular-based backbones for subsequent expression strategies (i.e. different expression cassettes) and preserve “institutional” memory.

1. Inoculate an overnight LB_{Kan100} culture with *EHA101* and incubate at 28°C with shaking (225 rpm). Prepare 3 × 100 ml LB_{Kan100} cultures in 500-ml Erlenmeyer flasks and inoculate with 100 µl of the overnight *EHA101* culture at the end of the following day. Incubate again overnight at 28°C with shaking (225 rpm).
2. Harvest *EHA101* cells from the culture when the OD at 600 nm is 0.5–1.0. First chill the flasks on ice for 30 min, then transfer to sterile Oakridge tubes, and centrifuge at 3,900 × g for 15 min at 4°C (see Note 17) to pellet. Decant supernatant and resuspend *Agrobacterium* pellet in ~40 ml ice-cold sterile Milli-Q water (see Note 18) and centrifuge/pellet as above. Decant the water and resuspend the remaining *Agrobacterium* pellet in 20 ml cold sterile Milli-Q water. Re-pellet the cells and perform a final wash and centrifugation in 5 ml ice-cold sterile 10% glycerol. Decant the glycerol wash (using serological pipette) and resuspend each pellet in a final volume of 1.0–1.5 ml ice-cold sterile 10% glycerol. Pool final resuspensions into a Falcon 2059 tube and place on ice for 1 h.
3. Aliquot the *Agrobacterium* (40 µl) into sterile microfuge tubes, flash freeze in liquid nitrogen, and store at -80°C.
4. For electro-transformation, thaw several aliquots of *EHA101* competent cells (several are prepared at once in the event that

arching during electroporation occurs) on ice for 5–10 min. Add 25–50 ng of purified plant binary vector (pSBS4405 derived from DH5 α glycerol stock; see Note 19) diluted in low-ionic-strength buffer (TE) in a volume of 5 μ l and mix well by fingertip. Maintain on ice for 2 min.

5. Transfer the mixture to a 0.2-cm Gene Pulser® cuvette pre-chilled on ice (see Note 20) and shake the suspension to the bottom of the cuvette by hand, avoiding the formation of air bubbles. Pulse cuvette once at 25 μ F capacitor, 2.5 kV and 200 Ω . If a “popping” sound is heard, this indicates that arcing of the sample occurred (indicating excessive salt, air bubbles, or insufficiently chilled cuvette) and should be repeated. Otherwise, remove the cuvette from the pulse chamber and immediately add 1 ml of SOC media to resuspend the cells. Transfer the cell suspension to Falcon 2059 polypropylene tubes and incubate at 28°C for 2 h with shaking (225 rpm).
6. Following the recovery period, aseptically plate 100–300 μ l of the cell suspension (see Note 21) onto AB_{KanSpec100} plates (or appropriate selection for positive clones), wrap plates with Parafilm to prevent drying out, and incubate plates at 28°C.
7. After approximately 48 h, the appearance of individual clones should be visible. Dilute single isolates in 100 μ l sterile water. Using a loop, streak out the colony on fresh AB_{KanSpec100} plates using the quadrant technique (see Note 22) to ensure single colony growth. Return streaked plates to incubate at 28°C in Parafilm-wrapped plates.
8. After an additional 2 days, transfer a single colony to 100 μ l sterile water. Inoculate 5 ml (in Falcon 2059 tube) of AB_{KanSpec100} with 25 μ l of the suspension. Inoculate a separate 10 ml (in Falcon 2059 tube) of YEP_{Kan100Spec100} media with 25 μ l of the suspension (see Note 23). Return the inoculums to 28°C with shaking (225 rpm).
9. Monitor the OD of the AB_{Kan100Spec100} culture and when the OD A600 is between 0.6 and 0.8 prepare glycerol stocks by mixing one part *Agrobacterium* with one part 50% glycerol and store at –80°C. Isolate plasmid from YEP_{Kan100Spec100} for verification of *Agrobacterium* clone by PCR, restriction mapping, etc.

3.3. Preparation of *Arabidopsis* for In Planta Transformation

1. For ten pots (40–60 seeds per pot) of plants, add three scoops (~15 mg or ~600 *Arabidopsis* C24 seeds) of seed to a microfuge tube. Add a volume of acid-washed sea sand equivalent to about 4 \times the level of seed and shake.
2. Prepare pots by first breaking up Sunshine #4 soil mixture by hand to ensure uniformity and moisten. Fill 10 \times 4" green pots with moistened soil mixture to overfilling to form a small mound. This allows for settling of the soil, but allows contact with the screen over a smooth surface. Do not overpack/compact the soil to the point that root penetration would be impaired.

3. Sprinkle the seed/sand mixture evenly over the ten pots (see Note 24). Cover each pot with a piece of screen and apply an elastic band around each pot to secure.
4. Place all pots in a tray (Aratrays), and add ~2 L of water to the tray. Cover the tray with a plastic dome to maintain humidity and vernalize by placing tray in dark at 4°C for 3–4 days (see Note 25).
5. After vernalization/germination, place trays in 24-h light (incandescent and fluorescent light at 150 $\mu\text{E m}^{-2}/\text{s}$) at 24°C. Once seedlings appear, fertilize plants once per week with 1% Peter's fertilizer (Nitrogen:Phosphorus:Potassium, 20:19:18).
6. Irrigate plants every 2–3 days interval.
7. After approximately 2 weeks following planting, remove seedlings with tweezers so that only 5–6 seedlings/pot remain.
8. When plants are about 2 cm in height, the primary (1°) inflorescence is cut/removed to promote the growth of secondary (2°) and tertiary (3°) bolts. Following 4–5 days after removing the primary bolts, the plants are ready to be infected with *Agrobacterium*.

**3.4. In Planta
Arabidopsis
Transformation by
Floral Dipping Method
(Adapted from Ref. 18)**

1. Prepare 500 ml LB (in a 2-L flask) with addition of antibiotic (Kanamycin and Spectinomycin, or as required) to a final concentration of 100 $\mu\text{g/ml}$ to result in LB_{Kan100/Spec100}.
2. Inoculate LB_{Kan100/Spec100} with 50 μl of *Agrobacterium* glycerol stock (pSBS4405 in this example). Place the inoculated flask in a 28°C shaker overnight (~24 h). When the OD at 600 nm is 0.5–0.6, harvest the *Agrobacterium*.
3. Harvest Agrobacteria from LB by pelleting cells by centrifugation at 4°C (15 min at 6,829 $\times g$) (see Note 26).
4. Decant the supernatant as waste and dispose of accordingly (autoclave). Resuspend the *Agrobacterium*-pSBS4405 pellet(s) in a 500 ml solution of 5% sucrose by hand shaking. Pour the resuspended *Agrobacterium*-pSBS4405 into a crystallizing bowl (Pyrex® crystallizing bowl, 170 × 90 mm) and then add Silwet L-77 to a final 0.05% (w/v).
5. Dip each pot of prepared *Arabidopsis* plants by immersion (inverted into the *Agrobacterium* solution) for 20 s (see Note 27).
6. Return the pots to a tray and cover with a plastic dome for up to 24 h to retain elevated humidity.
7. Approximately 1 week after the initial dipping, a second dipping is performed as above to infect new florets.
8. Grow all infected plants to maturity (~3 weeks after second dipping) and harvest T1 seed (will be a combination of both transformed and untransformed seed).

**3.5. Selection
of Recombinant T1
A. thaliana Seed**

1. Measure approximately 0.8 g of the putative transformed T1 seed obtained in Subheading 3.4, step 8. Place seed in a 50-ml Falcon tube. For small amounts of seed, use 15-ml Falcon tube.
2. Soak seed for 1 h in sterile dH₂O on a shaker set at 130 rpm. Allow seeds to settle at bottom (pour off any seeds that remain floating and destroy) and decant water. Remove surface lipids by washing the seeds for 30 s in 70% ethanol (v/v) by hand shaking. Surface sterilize the seeds with 20% commercial bleach (v/v) and 0.1% Tween 20 (v/v) and return to shaker at 130 rpm for 15 min. Decant bleach/Tween20 solution and wash seeds five times with sterile dH₂O for 10-min intervals on shaker set at 130 rpm. Remove floating impurities, immature seeds, or dirt particles in between washes.
3. During the seed washes prepare selection plates (1/2 strength MS plates containing 0.8% Agar, 3% sucrose, and 80 µM PPT) with clearly visible labels for construct ID and date. If prepared plates show signs of condensation or are unusually wet, then dry in a hood prior to plating. For each 0.8 g of seed prepared, use 20 selection plates. If including controls (see Note 28), a single plate each is sufficient.
4. Melt 0.55–0.6% top agar (microwave, hot plate but ensure that it does not boil over) and incubate in preset 60°C water bath until ready to use. Use 50 ml top agar for every 20 selection plates (~2.5 ml top agar/plate).
5. When the seed has been sufficiently rinsed to remove bleach (very weak bleach smell and no visible soapsuds in tube), combine rinsed seed with 50 ml top agar. Aliquot ~2.5 ml top agar seed mix onto appropriately labeled plates and swirl gently by hand to evenly distribute the seed. Maintain the seed agar mix in warm water retrieved from the 60°C water bath during plating to keep top agar warmed. Allow plates half covered to cool in a fume hood for about 1 h or until no sign of condensation on lids. Parafilm each plate and place under constant light (80–100 µE m⁻²/s) at 24°C.
6. Putative positive seedlings (green and growing versus non-transformed seeds which do not germinate or are bleached of color) are transferred to soil 7–10 days later. Individual transformants are seeded in trays (38 slots/tray) (see Note 29) and covered with a plastic dome for 5–7 days to retain humidity. After 7 days each seedling is covered with a seed collector system (base and tube) from Lehle Seeds to prevent inadvertent seed loss and cross-pollination. The seedlings are allowed to grow to maturity (~8 weeks) and each plant is harvested individually to obtain T2 lines for further analysis. Plants are irrigated after removal of plastic domes in 3-day intervals and fertilized in 5-day intervals with 1% Peter's fertilizer.

7. Individual T2 lines are harvested by first drying the plants down (stop watering) to complete maturity for about 1 week. Seeds are collected by removing siliques and thrashing between fingers to release seed. Remove debris from seed and combine with cleaned seed retained in base of seed collector system (Aracons). Harvested seed clean of debris can be stored in glass vials (Kimble opticlear 15 × 45 mm screw threaded vials with vented lids).
8. Following the identification of high expressors within the T2 lines, further grow outs can be performed as outlined above.

3.6. Expression Analysis of Recombinant T2 *A. thaliana* Seed Lines

Once T2 seed is obtained from individual lines, the analysis of expression is straightforward. Typically seeds are homogenized in neutral buffers (with and without protease inhibitors) and solubilized under reduced and non-reduced conditions for analysis by Coomassie-stained SDS-PAGE with corresponding Western blotting. Other techniques including capillary electrophoresis, size exclusion chromatography, and ELISA can also be applied if appropriate controls are available. After the initial expression screening of lines is complete, more quantitative and qualitative measurements can be applied following oilbody partitioning for pSBS4405 as described in Nykiforuk et al. (17). The liquid–liquid phase separation of oilbodies allows for enrichment of transgenic oleosin-klip8-klip27-insulin to a high degree. This simplifies maturation in vitro with trypsin, followed by a second round of oilbody separation to derive the cleaved insulin from the aqueous phase. Further purification by conventional chromatography provides final purification and isolation of the insulin from plants. Thereafter, assays (cell based in vitro and glucose tolerance tests in vivo) demonstrated that biologically active DesB30-insulin was purified to high homogeneity.

4. Notes

1. For accession to *Arabidopsis* ecotypes in Asia, Australia, and the Americas, the ABRC at <http://abrc.osu.edu/> is available. For accession in Europe, Africa, and other regions, the European Arabidopsis Stock Centre (NASC) at <http://arabidopsis.info/> is available.
2. The materials and methods described herein have been optimized for *A. thaliana* C24 ecotype and therefore may perform suboptimally when other ecotypes are employed. Therefore, investigators may want to incorporate different ranges of sucrose and surfactant to the transformation media unless described elsewhere in the literature. In addition, if negative

selection is used for identifying transformed seed, one should perform kill curves with the non-transformed seed to establish useful levels of selection.

3. The primary target of transformation using these techniques is the ovule (the female gametophytic tissues) within the locule (24). The timing of the Agro-inoculation is a critical component, and therefore transformation should be performed during the late stage of floral development after the divergence of individual pollen or egg cells (hemizygous), but before fertilization has occurred. Access of *Agrobacterium* to the ovules within the locule can be obtained about 3 days prior to anthesis. If different *Arabidopsis* ecotypes are employed, it may be necessary to identify the inflorescence developmental stage that is most efficient for transformation (18).
4. Whether cDNAs for expression are derived from endonuclease restriction fragments, generated by PCR, or synthesized from commercial venues, it is always advisable to confirm the integrity of the sequence in the final plant binary vector. Typically sequence analysis can be performed after subcloning in a bacterial host (sufficient plasmid obtained) to ensure that during vector construction or amplification no frameshifts, deletions, or insertions of nucleotides have occurred prior to transformation experiments. Other pitfalls in expression can be avoided by employing sequence analysis tools described in Note 7.
5. Plant binary vectors that are disarmed means the tumor-inducing (Ti) or rhizogenic-inducing (Ri) genes (*vir* genes) required for infection are harbored in the companion *Agrobacterium* strain (*vir* helper strain) as the helper tumor-inducing or rhizogenic-inducing plasmid, respectively. Many plant binary vectors are described in the literature (25), but aside from the T-DNA border repeat sequences, the common genetic elements required for transformation include antibiotic resistance genes for positive selection and growth in media and origins of replication for amplification in bacteria and *Agrobacterium*. Therefore, based upon the choice/purchase of bacterial host for subcloning and *Agrobacterium* host for transformation, the proper combination of selectable markers and origin of replication can be crucial. Care must be taken to ensure that the antibiotic resistance markers on the plant binary replicon are compatible for selection once expressed in the *Agrobacterium* strain being used. For instance *EHA101* *Agrobacterium* possesses both rifampicin and kanamycin resistance and redundant resistance in a plant binary vector would make selection very difficult, whereas Spectinomycin allows for the easy identification of recombinant clones by double selection (i.e., Kanamycin and Spectinomycin agar plates). When using Spectinomycin resistance, investigators should perform kill curves with the *vir*

helper strain (*Agrobacterium*) lacking the binary vector (untransformed) to ensure that effective killing occurs. Other considerations should be taken into account when using binary vectors containing tetracycline resistance gene (26) or ampicillin resistance (B-lactamase gene, (27). The elements for origin of replication provide the ability to replicate in bacteria and *Agrobacterium*. For instance *pVSI* allows broad host replication in both *E. coli* and *Agrobacterium*. *ColE1* alone will allow maintenance of larger vectors in *E. coli*, but not *Agrobacterium*. The presence of both *pVSI* and *ColE1* provides high-level replication in *E. coli* during subculture (between 10 and 200 copies/cell) while *pVSI* allows for good replication in *Agrobacterium* (7–10 copies/cell).

6. As described in Nykiforuk et al. (17), the pSBS4405 plant binary vector encodes a fusion protein, oleosin–human insulin, comprising a recombinant human mini-insulin (OB-hIN) with an N-terminal trypsin cleavable pro-peptide (Klip27-mini-insulin) (28) fused to the C terminus of the *A. thaliana* 18 kDa oleosin-Klip8 polypeptide. The transgene is under the tight seed-specific transcriptional control of the β-phaseolin promoter/terminator from *Phaseolus vulgaris* (29). This promoter ensures specific temporal- and tissue-specific expression during seed development. The fusion protein is targeted to the oilbody by expression of an *Arabidopsis* 18 kDa oleosin (19) as a function of the inherent hydrophobic domain topology (30, 31). The intervening klip sites are pro-peptide regions to allow enzymatic cleavage by chymosin (klip8) and/or trypsin (klip27). The mini-insulin encoded for Des-B30 insulin evolves after enzymatic removal of the mini-C peptide, AAK, with trypsin. For selection of transgenic *Arabidopsis*, a pat gene conferring phosphinothricin resistance (32) under the control of the constitutive ubiquitin promoter/terminator from *Petroselinum crispum* (33) was used. This construct represents one of many tested in *Arabidopsis* with the rationale of determining what combination of elements was necessary for (1) high levels of expression, (2) proper conformation (folding and posttranslational formation of disulfide bonds) for biological activity, (3) ease of extraction leveraging SemBioSys technology, and (4) testing in vitro maturation following extraction. When testing expression of therapeutic proteins in *Arabidopsis*, the choice of promoter can affect both levels of expression and tissue specificity (11, 34). Our technology uses nascent oil-bodies to recover target proteins by either Stratoderm™ (covalent attachment of fusion directly to oleosin with subsequent targeting to the oilbody *in vivo*) or Stratocapture™ (the nascent oleosins on the oilbody act as a ligand for single-chain antibodies fused to target) (11, 12), and therefore seed-specific

promoters with high levels of expression overlapping oilbody ontogeny are employed. Aside from the expression strength of the promoter, other issues regarding its choice may be influenced by freedom to operate if commercial applications are being sought (suitable seed-specific promoters can be found in 34 and 11). Aside from ubiquitin, another constitutive promoter commonly used to drive the selectable marker includes the 35-S CaMV promoter (35).

7. In the current example the pSBS4405 was expressed as a contiguous fusion protein (oleosin-Klip8-Klip27-mini insulin) which resulted in high levels of expression localized to the oilbodies upon extraction (17). Prior to binary vector construction the fusion protein nucleic acid sequences were codon-optimized for expression in plants. In addition, nucleic acid sequences were assessed for cryptic splice sites (using NetPlantGene Server; <http://www.cbs.dtu.dk/services/NetPGene/>) and internal poly adenylation sites (numerous online resources or programs available for download). Other sequence-based mining (both nucleic acid and protein sequence) can be performed to identify putative glycosylation sites, signal peptide targeting, potential hairpins, etc. to aid in the design of your target with the intention of promoting heterologous protein expression and analysis in *Arabidopsis*.
8. Expression as a fusion protein can influence the level of accumulation and solubility (36), but can also hamper efforts to generate an authentic product because cleavage away from the fusion protein is required. The fidelity of fusion partner removal is dependent upon a number of factors, and therefore construct design is paramount before assembling/synthesizing expression constructs. Chemical cleavage often requires elevated temperatures under acidic/toxic conditions (acid labile Asp-Pro bond, cyanogen bromide cleavage at methionine residues, or hydroxylamine at Asn-Gly bond) often in the presence of denaturants (urea or guanidine hydrochloride). For example, we have used acid cleavage for the removal of fusion partner based on an intervening acid labile bond under high temperature and denaturing conditions (37). This was possible because the therapeutic target lacked internal acid labile bonds (Asp-Pro), had a high thermal denaturation temperature, and did not require typical folding associated with biological function (requires association with phospholipids). In the case of insulin fusion protein (pSBS4405), the milder enzymatic cleavage approach was adopted to reduce the complexity of downstream purification (no refolding reaction required). The enzymatic cleavage of the Klip8 site was based on the heterologous cleavage design described by Kühnel et al. (38) with removal of the Klip27 during final maturation in vitro with trypsin to result in

an authentic insulin N-terminal B chain (and coincidentally an authentic N-terminal A chain after removal of the mini-C peptide). However, during the course of investigation, the use of Klip8 was not required for removal from oleosin/oilbodies as cleavage proceeded specifically and efficiently with trypsin alone (17). The main factors to consider regarding enzymatic cleavage include primary (amino acid sequence of the scissile bond) and secondary (avoid steric hinderance) specificity of the cleavage site which can influence precision and efficiency. The primary specificity encompasses the stretch of amino acids that interact with the active site of the protease. The amino acids involved in this interaction are designated S_n on the protease site and P_n on the substrate target site. Counting from the scissile bond, N-terminal residues have the suffix 1, 2, ... n and on the C-terminal side 1', 2' ... n' (39, 40). Some proteases have very narrow specificity towards the sequence of the scissile bond, while others exhibit broad specificity (e.g., subtilisin) and should be considered in design. For example, the trypsin S1 binding pocket only accepts the positively charged arginine and lysine residues in the P1 position, due to the presence of a negatively charged aspartic acid residue at the base of the deep binding pocket. The efficiency of cleavage with trypsin is further influenced by amino acids occupying the P2 and P1' positions (41). As a comparison, subtilisin exhibits a broad specificity accepting Phe, Leu, Ile, Val, and Ala (nonpolar residues) at the P1 position. Therefore, criteria used in selecting a protease for enzymatic cleavage based on primary specificity will include the following: (1) does the therapeutic protein possess internal scissile bonds (precision versus potential for product degradation); (2) is the stability of the fusion protein compatible with enzymatic conditions required for optimal protease activity (pH, ionic strength, and temperature can all affect structure and therefore susceptibility of the product to proteolysis); (3) authenticity of N terminus following enzymatic cleavage (while extra amino acids may lend themselves to subsequent removal with aminopeptidases/carboxypeptidases, the lack of N-terminal amino acids (*Des*-polypeptide) cannot be reversed); and (4) if the therapeutic is expressed as a naturally occurring pro-peptide, is the protease responsible for maturation commercially available. In addition, when designing a cleavage strategy attempts to model the putative cleavage site should be made using predicted structure ascertained *in silico* and/or crystal/liquid structures when available. Although there is less flexibility regarding secondary specificity, useful criteria when modeling include the following: (1) is the linker between the fusion partner and therapeutic target compatible with access of the protease to the scissile bond (length, structure, and unnecessary stretches of highly

charged or incompatible amino acids that would repel or impair binding of the protease); (2) may consider introducing proline residues to act as a molecular hinge or accessible loops to increase cleavage efficiency/specificity, but this may also increase in vivo lability; (3) if other labile sites are present in the primary amino acid sequence, are these sites buried or exposed to proteolytic activity; (4) are there any disulfide bonds in close proximity to the scissile bond which may impair or block access of the protease; and (5) if access to the scissile bond is blocked or impaired, can access be acquired by relaxing structure through the use of detergents, ionic strength, increasing hydrophobicity with chaotropes, etc. All of these factors can be considered and aid in the experimental design of cleavage strategies, but empirical evidence remains the ultimate test.

9. Many commercially available strategies for enzymatic cleavage are now available including those based on Factor Xa (42), enterokinase (43), thrombin (44), kex2 (45), subtilisin (46), and furin (47). Also available are protease-free self-splicing intein sequences (48) and self-recognizing split-SUMO (49).
10. Alternatively, the tryptone, yeast extract, NaCl, and KCl can be mixed to around 95 ml and autoclaved. After the solution has cooled, the addition of magnesium salts can be performed from a 2 M stock solution (filter-sterilized 20.33 g MgCl₂·6 H₂O plus 24.65 g MgSO₄·7 H₂O in 100 ml MilliQ) and 1 ml 2 M glucose stock (filter sterilized). Ensure or adjust pH to 7.0 and bring to 100 ml with sterile, distilled water. Filter the completed medium through a 0.2-μm filter.
11. Antibiotics at the appropriate concentration can be added to LB and AB broths prior to inoculation under aseptic technique. Antibiotic can also be applied to plates by spreading stock solutions, but uneven spreading (differential concentrations of antibiotic) or negative patches on the plate is a possibility, which in turn could result in selecting false positive clones.
12. Some antibiotics are not 100% active or “potent.” If the potency of the antibiotic is known, the stock solutions can be compensated by adding the correct amount of antibiotic to account for the lack of potency in a stock solution. This is a straightforward calculation, where the potency (known in μg/mg) is converted to the desired concentration. For example if the potency is 815.35 μg/mg, then it is actually only 81.535% active for every milligram of powder dissolved, and therefore if a 10 mg/ml stock solution with 100% activity is desired the actual concentration of the under-potent antibiotic should be $100/81.535 = 1.22 \times 10 \text{ mg/ml} = 12.2 \text{ mg/ml}$.
13. Inoculum amounts will vary for different *E. coli* strains such as *DH10b*, *BL21*, *EXL1Gold*, etc. and therefore may want to test

- growth rates (log phase growth over time) to attain appropriate densities for preparation of competent cells.
14. Pelleting of *DH5 α* cells was performed using Nalgene 250-ml centrifuge tubes (Thermo Fisher Scientific, Rochester, NY, USA) on an Avanti J-25 centrifuge using a JLA 16.250 fixed angle rotor (Beckman Coulter, Brea, CA, USA).
 15. Using 250-ml centrifuge tubes, two rounds of pelleting are performed; so the final pellet yield is equivalent to 500 L of culture. To the final pellet, resuspension in 30 ml of cold TB buffer is added (1:16.7 v/v). Following the centrifugation 9.3 ml of ice-cold TB buffer (1:53.7 v/v) and 0.7 ml DMSO was added before incubation on ice overnight.
 16. Retain a small amount of transformation mixture at 4°C overnight. If there is overgrowth of clones on LB selection plates the following day, plate less or dilute appropriately with TE buffer and replate to generate individually accessible colonies.
 17. Using 50-ml Oakridge centrifuge tubes, *EHA101* cells are pelleted in Avanti J-25 centrifuge using JS 13.1 swing out bucket rotor (Beckman Coulter, Brea, CA, USA).
 18. Pre-store sterile ultrapure water (Milli-Q) and 10% glycerol at 4°C. Use as needed.
 19. Although many established procedures for plasmid isolation from recombinant *DH5 α* cells are available within the literature, we typically employ the Qiagen plasmid preparation kits available commercially (QIAprep Spin Miniprep Kit, Qiagen Inc., Valencia, CA, USA).
 20. The 0.2-cm Gene Pulser® cuvettes are individually packed and therefore can remain within the plastic bag while precooling on ice. If the cuvette accidentally gets wet, be sure to dry off with kim-wipe. Avoid direct contact with hands or gloves to the bottom half of the cuvette (electrode contacts) until electroporation is performed.
 21. Increasing amounts of transformed suspension from 100 to 300 μ l provide different efficiencies of transformation to grow to different densities on AB_{KanSpec100} plates. In other words, if the efficiency was high, plating 300 μ l may result in a blanket of cells on the plate, whereas the 100 μ l plating may allow the identification of individual clones aiding in their retrieval and subsequent amplification/storage.
 22. Typically use quadrant streaking of plates to ensure that single clones can be isolated. Simply flame sterilize the loop until it is red hot and allow it to cool. Dip and remove a loopful from the 100 μ l suspension in sterile water. Immediately streak the inoculating loop back and forth very gently over a third of

the plate. Flame the loop again and allow it to cool. Turn the plate a quarter turn counterclockwise and repeat streaking across a second quadrant of the plate being sure to catch the edge of the previous streaks. Flame loop again and allow it to cool. Turn the plate another quarter turn counterclockwise and repeat. Flame loop again and allow to cool. This time, after turning another quarter turn counter clockwise, streak the remaining area being sure to catch the last quadrant. This technique allows for serial dilutions of the original suspension of colonies over an entire plate. Following incubation of the plates, individual clones can typically be isolated usually from the third or fourth quadrant streaked across the plate.

23. While the AB_{KanSpec100} culture serves as the source for the glycerol stock, the YEP_{KanSpec100} culture can be used to verify recombination occurred by isolating the plasmid and performing restriction digest and/or confirmation by PCR. Alternatively the diluted colony in water can also be used to verify the recombinant clone by PCR using primers against the TDNA expression construct.
24. One should practice the packing of soil and sprinkling of *Arabidopsis* seeds to ensure proper growth and distribution of seeds within pots is established.
25. There is no need to fertilize the seed/pots at this point (seeds will germinate using stored reserves). The vernalization treatment will help synchronize germination.
26. Pelleting of recombinant *Agrobacterium*-pSBS4405 cells was performed using Nalgene 250-ml centrifuge tubes (Thermo Fisher Scientific, Rochester, NY, USA) on an Avanti J-25 centrifuge using a JLA 16.250 fixed angle rotor (Beckman Coulter, Brea, CA, USA) at 6,000–6,500×*g* for 15 min at 4°C.
27. Immersion in the resuspended *Agrobacterium* solution allows infection of the plants' flowers, and more specifically the ovules. Normally five pots per 500 ml resuspension solution are used, and therefore if dipping all ten pots then 1 L of *Agrobacterium* should be prepared or conversely the other five pots could be used for a different construct configuration.
28. During selection of T1 transformants, it is advisable to include a negative control (wild-type C24) and positive control (herbicide-resistant line).
29. Do not transplant T1 seedlings with roots that have not penetrated through the top agar and into the selection medium.

References

1. Obembe OO et al (2010) Advances in plant molecular farming. *Biotechnol Adv* 29(2): 210–222
2. Sourrouille C et al (2009) From Neanderthal to nanobiotech: from plant potions to pharming with plant factories. *Methods Mol Biol* 483:1–23
3. Ma JK et al (2003) The production of recombinant pharmaceutical proteins in plants. *Nat Rev Genet* 4:794–805
4. Twyman RM et al (2003) Molecular farming in plants: host systems and expression technology. *Trends Biotechnol* 21(12):570–578
5. Davies HM (2010) Commercialization of whole-plant systems for biomanufacturing of protein products: evolution and prospects. *Plant Biotechnol J* 8:854–861
6. Dunwell JM (2011) Foresight project on global food and farming futures. *Crop biotechnology: prospects and opportunities*. *J Agric Sci* 149: 17–27
7. Castilho A et al (2010) *In planta* protein sialylation through overexpression of the respective mammalian pathway. *J Biol Chem* 285(21):15923–15930
8. De Muynck B et al (2010) Production of antibodies in plants: status after twenty years. *Plant Biotechnol J* 8:529–563
9. Loos A et al (2010) Production of monoclonal antibodies with a controlled N-glycosylation pattern in seeds of *Arabidopsis thaliana*. *Plant Biotechnol J* 9(2):179–192
10. Loos A et al (2011) Expression of antibody fragments with a controlled N-glycosylation pattern and induction of ER-derived vesicles in seeds in *Arabidopsis thaliana*. *Plant Physiol* 155:2036–2048
11. Boothe J et al (2010) Seed-based expression systems for plant molecular farming. *Plant Biotechnol J* 8:588–606
12. Markley N et al (2006) Producing proteins using transgenic oilbody oleosin technology. *Biopharm Int* 19:34–47
13. The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408:796–815
14. Koornneef M, Meinke D (2010) The development of *Arabidopsis* as a model plant. *Plant J* 61:909–921
15. Rounseley SD, Last RL (2010) Shotguns and SNPs: how fast and cheap sequencing is revolutionizing plant biology. *Plant J* 61:922–927
16. Liu Y-G et al (1995) Efficient isolation and mapping of *Arabidopsis thaliana* T-DNA insert junctions by thermal asymmetric interlaced PCR. *Plant J* 8(3):457–463
17. Nykiforuk CL et al (2006) Transgenic expression and recovery of biologically active recombinant human insulin from *Arabidopsis thaliana* seeds. *Plant Biotechnol J* 4:77–85
18. Clough SJ, Bent AF (1998) Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana*. *Plant J* 16(6):735–743
19. Van Rooijen GJH, Moloney MM (1995) Structural requirements of oleosin domains for subcellular targeting to the oil body. *Plant Physiol* 109:1353–1361
20. Hajdukiewicz P et al (1994) The small, versatile *pPZP* family of *Agrobacterium* binary vectors for plant transformation. *Plant Mol Biol* 25:989–994
21. Hood EH et al (1986) The hypervirulence of *Agrobacterium tumefaciens* A281 is encoded in a region of pTiBo542 outside of T-DNA. *J Bacteriol* 168(3):1291–1301
22. Hellens R et al (2000) A guide to *Agrobacterium* binary Ti vectors. *Trends Plant Sci* 5(10): 446–451
23. Inoue H et al (1990) High efficiency transformation of *Escherichia coli* with plasmids. *Gene* 96:23–28
24. Schneitz K et al (1995) Wild-type ovule development in *Arabidopsis thaliana*: a light microscope study of cleared whole-mount tissue. *Plant J* 7(5):731–749
25. Lee L-Y, Gelvin SB (2008) T-DNA binary vectors and systems. *Plant Physiol* 146: 325–332
26. Luo ZQ, Farrand SK (1999) Cloning and characterization of a tetracycline resistance determinant present in *Agrobacterium tumefaciens* C58. *J Bacteriol* 181:618–626
27. Cheng AM et al (1998) Timentin as an alternative antibiotic for suppression of *Agrobacterium tumefaciens* in genetic transformation. *Plant Cell Rep* 17:646–649
28. Kjeldsen T et al (2001) Expression of insulin in yeast: the importance of molecular adaptation for secretion and conversion. *Biotechnol Genet Eng Rev* 18:89–121
29. Slightom JL et al (1983) Complete nucleotide sequence of a French bean storage protein gene: phaseolin. *Proc Natl Acad Sci USA* 80:1897–1901
30. Abell BM et al (2002) Membrane protein topology of oleosin is constrained by its long hydrophobic domain. *J Biol Chem* 277(10): 8602–8610

31. Abell BM et al (2004) Membrane topology and sequence requirements for oilbody targeting of oleosin. *Plant J* 37:461–470
32. Wohlleben W et al (1988) Nucleotide sequence of the phosphinothricin N-acetyltransferase gene from *Streptomyces viridochromogenes* TA1/4494 and its expression in *Nicotiana tabacum*. *Gene* 70:25–37
33. Kawalleck P et al (1993) Polyubiquitin gene expression and structural properties of the ubi4-2 gene in *Petroselinum crispum*. *Plant Mol Biol* 21:673–684
34. Stoger E et al (2005) Sowing the seeds of success: pharmaceutical proteins from plants. *Curr Opin Biotechnol* 16:167–173
35. Rothstein SJ et al (1987) Promoter cassettes, antibiotic-resistance genes, and vectors for plant transformation. *Gene* 53:153–161
36. Georgiou G, Valax P (1996) Expression of correctly folded proteins in *Escherichia coli*. *Curr Opin Biotechnol* 7:190–197
37. Nykiforuk CL et al (2011) Expression and recovery of biologically active recombinant Apolipoprotein AI_{Milano} from transgenic safflower (*Carthamus tinctorius*) seeds. *Plant Biotechnol J* 9(2):250–263
38. Kühnel B et al (2003) Precise and efficient cleavage of recombinant fusion proteins using mammalian aspartic proteases. *Protein Eng* 16(10):777–783
39. Schechter I, Berger A (1967) On the size of the active site in proteases. I. Papain. *Biochem Biophys Res Commun* 27(2):157–162
40. Schechter I, Berger A (1968) On the active site of proteases. III. Mapping the active site of papain; specific peptide inhibitors of papain. *Biochem Biophys Res Commun* 32(5):898–902
41. Keil B (1992) Specificity of proteolysis. Springer-Verlag, Berlin/Heidelberg/New York, p 335
42. Nagai K et al (1985) Oxygen binding properties of human mutant hemoglobins synthesized in *Escherichia coli*. *Proc Natl Acad Sci USA* 82:7252–7255
43. LaVallie ER et al (1993) Cloning and functional expression of a cDNA encoding the catalytic subunit of bovine enterokinase. *J Biol Chem* 268:23311–23317
44. Geng Y et al (2010) Expression of active recombinant human tissue-type plasminogen activator by using *in vivo* polyhydroxybutyrate granule display. *Appl Environ Microbiol* 76(21):7226–7230
45. Bader O et al (2008) Processing of predicted substrates of fungal Kex2 proteinases from *Candida albicans*, *C. glabrata*, *Saccharomyces cerevisiae* and *Pichia pastoris*. *BMC Microbiol* 8:1–16
46. Nilsson CP et al (1989) Engineering substilisin BPN' for site-specific proteolysis. *Proteins* 6(3):240–248
47. Kahle NA et al (2010) Furin cleavage of bacterial expressed glutathione-S-transferase-pro-transforming growth factor β1 fusion protein *in vitro*. *Protein Pept Lett* 17(4):416–418
48. Chong S et al (1997) Single-column purification of free recombinant proteins using a self-cleavable affinity tag derived from a protein splicing element. *Gene* 192:277–281
49. Butt TR et al (2005) SUMO fusion technology for difficult-to-express proteins. *Protein Expr Purif* 43:1–9

Chapter 17

Methods for Chromatographic Removal of Endotoxin

Adam J. Lowe, Cameron L. Bardliving, and Carl A. Batt

Abstract

Endotoxin removal is critical when producing therapeutic proteins in gram-negative bacterial systems. This hydrophobic compound can be removed through chromatography or filtration, but presents unique challenges dependent upon protein composition as well as production scale. Here we present a robust method for endotoxin removal at the pilot production scale using fast protein liquid chromatography and buffers specifically engineered for endotoxin removal.

Key words: Lipopolysaccharides, LPS, Endotoxin, Lipoglycan, Chromatography, Triton X-114

1. Introduction

Endotoxin, also known as lipopolysaccharides, is a component of the gram-negative bacterial cell wall that can cause septic shock ([1–3](#)) in humans. Endotoxin is composed of three main elements: Lipid A, core oligosaccharides containing sugars such as keto-deoxyoctulosonate and heptose, and a variable O-antigen region that can be used to serotype different bacterial strains. The Lipid A portion of endotoxin causes an immunogenic response from the body, but does not directly attack cells ([4, 5](#)). In higher doses, it causes a massive innate immune response, inducing chemokine and cytokine release ([6, 7](#)) as well as inflammation and fever. This cascade of immune responses sometimes leads to shock and death ([1](#)).

These health risks directly impact production of recombinant proteins in gram-negative bacterial systems, such as *Escherichia coli*, since endotoxin inevitably contaminates the protein preparation when the cells are lysed. Due to endotoxin's effect on the immune system, the FDA has set a limit of five endotoxin units

per kilogram patient weight per hour of infusion (8). Endotoxin is a challenging substance to remove from protein preparations due to its tendency to irreversibly bind to proteins, especially those derived from inclusion bodies or with strong hydrophobicity indices. It maintains both a hydrophobic characteristic from its Lipid A component as well as a net negative charge from its phosphorylation. To reach the low levels mandated by the FDA, various purification methods have been developed such as ultrafiltration (9, 10), anion exchange chromatography (11), cation exchange chromatography (12, 13), affinity resins (14), histidine (15), two-phase micellar extraction (16, 17), and Polymyxin B (18, 19). Ion exchange chromatography uses the ability of charge titration to separate LPS from the protein of interest. By leaving either the protein or LPS charged while the other remains neutral, selective separation and elution are possible on a chromatography column. The success of this technique largely depends on the pKa of the protein and the ion exchange resin used. Histidine and other affinity resins can also nonspecifically bind the LPS, but successful purification again largely depends on the protein properties as well as the form of LPS in solution. LPS may be in supermolecular aggregates or in small micelles depending on LPS concentration and buffer conditions. Ultrafiltration does not scale well due to relatively low flow rates and often leads to large product loss. Polymyxin B is also not suitable for products destined for intravenous use as the antibiotic is physiologically active in humans (20).

Every protein presents different problems during purification and endotoxin removal, but we have developed a general method that worked well in endotoxin reduction for several 6x-His-tagged proteins, even using the C41DE3 *E. coli* strain, which has an unusually thick cell wall and LPS layer (21). We found that pre-treatment with deoxycholic acid and Triton-114 prior to loading onto the first affinity column greatly reduces the endotoxin load of the protein preparation. Additionally, using a similar wash while the protein is bound to an immobilized metal affinity chromatography (IMAC) column eliminates most of the endotoxin, and levels can be reduced to below the limit of detection on an anion exchange chromatography (AXC) column. We then use a final polishing column to further purify the protein (Fig. 1).

2. Materials

Prepare all solutions using the highest purity water available. All buffers should be passed through a 0.2-μm filter prior to use. β-mercaptoethanol should be added aseptically to solutions requiring it after filtration, as β-mercaptoethanol often sticks to filter membranes.

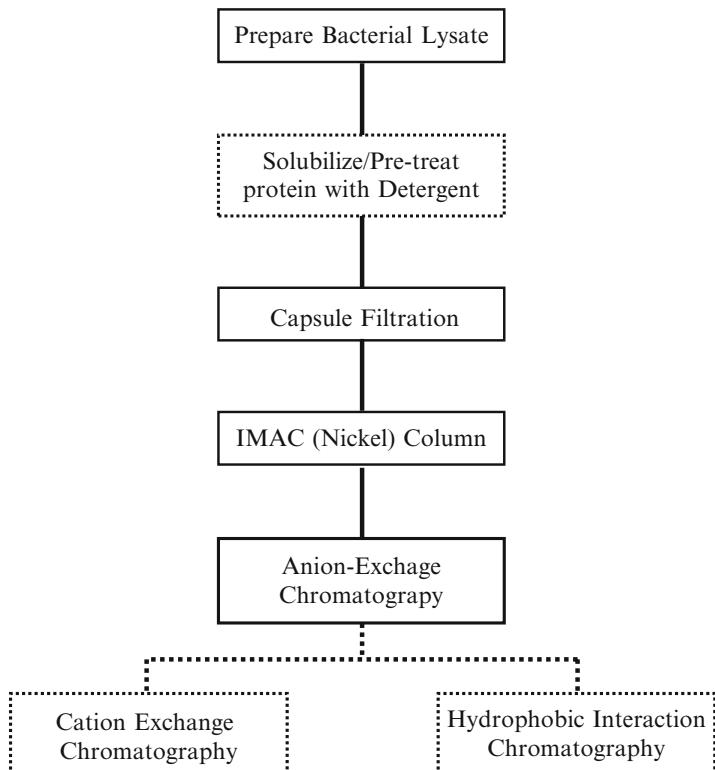


Fig. 1. Flowchart demonstrating general method of endotoxin removal. *Solid boxes and lines* denote mandatory steps while *dotted lines* are optional but recommended steps.

2.1. Buffer Composition

Note that our protein of interest is denatured and urea is included in all buffers. For proteins that will retain their native conformation, urea should *not* be included. It is possible to denature the protein for initial processing steps and to refold it later in the absence of other proteins, which has been described extensively (22–24).

1. Solubilization buffer: 2% w/v Deoxycholate (sodium salt), 1% v/v Triton-114, 8 M urea, 50 mM sodium phosphate, 200 mM NaCl, 100 mM KCl, 10 mM imidazole, 2.5 mM β -mercaptoethanol at pH 7.5. This solution should be prepared using very hot or boiling water, adding the urea to the solution first in a steady pour. Only 40% of the total volume of the solution should be used as a starting water volume since the urea will displace a significant amount of the solution.
2. Urea buffer: 4 M Urea, 50 mM sodium phosphate, 2.5 mM β -mercaptoethanol at pH 7.5.
3. Imidazole buffer: Formulated as urea buffer with 500 mM imidazole.

4. Carbonate buffer: 4 M Urea, 10 mM sodium carbonate, 1 mM 2-Mercaptoethanol at pH 10.5.
5. Carbonate elution buffer: Formulated as carbonate buffer with 1 M sodium chloride at pH 10.5.
6. Final bulk buffer: 4 M Urea, 50 mM sodium phosphate, 145 mM NaCl, 50 mM glycine at pH 6.5.
7. Prepare 1 M NaOH solution by dissolving 40 g of NaOH into 1 L of WFI water. Filter solution into a clean 5-L bottle through a 0.2- μm vacuum filter unit and immerse all inlet lines into solution.

2.2. Resins Used

1. Chelating Sepharose Fast Flow resin (GE Healthcare, Piscataway, NJ).
2. Q Sepharose XL resin (GE Healthcare, Piscataway, NJ).
3. SP Sepharose XL resin (GE Healthcare, Piscataway, NJ).

2.3. Equipment Required

Akta Purifier fast protein liquid chromatography (FPLC) (GE Healthcare, Piscataway, NJ) supported by Unicorn Software (GE Healthcare, Piscataway, NJ) or Equivalent FPLC.

3. Methods

These methods are for pilot-scale purification using a high cell density 20 L fermentation, 5 L IMAC column, and 1 L polishing columns. The purification process can be scaled up linearly from the benchtop scale other than flow rates, which must be adjusted based upon column dimensions. For benchtop experimentation, a 1 ml/min flow rate through a 5 ml IMAC column often gives desirable results. Note that your protein of interest must be His-tagged to be compatible with the IMAC column.

3.1. Lysate Pretreatment (Solubilization)

1. Obtain approximately 15 L of *E. coli* lysate from a high cell density fermentation expressing your protein of interest. If present, inclusion bodies may be washed before solubilization using tangential flow filtration (TFF) to eliminate host protein contamination (see Note 1).
2. Dilute the lysate into 35 L of solubilization buffer (30% of its original concentration).
3. Allow the lysate to mix with the solubilization buffer via stir bar for 18 h.
4. Prepare a 10-in. (25.4 cm) 1.2- μm low-protein-binding cartridge filter by passing 2 L of water through it. This is best accomplished using a peristaltic pump, making sure to purge as much air as possible from the filter to eliminate headspace.

5. Pass the solubilized lysate through the filter and collect the clarified lysate using a peristaltic pump or other positive pressure source.
6. Repeat steps 4–5 using a 0.5- μm low-protein-binding cartridge filter.

3.2. FPLC System Preparation

1. Prepare FPLC system for column evaluation by immersing all inlet lines into 2.5 L of WFI water (see Note 2) and flushing system with 2 L of water at a flow rate of 20 ml/min.
2. Flush system with 1 L of NaOH at a flow rate of 10 ml/min. NaOH hydroxide is used to remove endotoxin from the system and is highly effective, especially on stainless steel components.
3. Rinse out the NaOH following the procedure in step 1. After the flush step, manually run 200 ml of WFI water through each of the outlet lines and collect effluent into pyrogen-free containers. Determine endotoxin levels in each sample. If each sample passes, the FPLC system is cleared for use (see Note 3).

3.3. General Column Packing Procedure for 1 or 5 L Column (see Note 4)

1. Dispense 6 L of resin slurry into a clean container (see Note 5). Rinse resin with WFI water three times to remove storage buffer using a 0.2- μm vacuum filter unit.
2. Assemble column stand. Secure end piece to stand and secure bottom valve to end piece. Wet column end piece with a few centimeters of WFI water using a syringe connected to the bottom valve. Place support net on end piece and remove any air bubbles trapped in the mesh.
3. Wet column end piece filter net with WFI water and attach to column end piece. Eliminate all air bubbles (see Note 6).
4. Position PTFE guide ring in grove on the end piece. Place o-ring inside guide ring. Carefully place column tube on top of o-ring. Screw support rods into end piece and secure flange onto the column tube with support rods with four bolts. Check leaks by filling column with a few cm of WFI water (see Note 7).
5. Check the column standing horizontally using a spirit level. If column is not level, adjust stand wheels until column is level.
6. Carefully pour resin into column. Ensure that no air bubbles are trapped in resin. Do not pour resin directly into the column. Allow resin to flow down the side of the column tube.
7. While resin is settling, assemble column adaptor (see Note 8). Attach column adaptor to pressure gauge. Connect adaptor to pump and flow WFI through the adaptor at a low flow rate such that a thin film of water covers the adaptor plate. Wet adaptor support net and place on top of the adaptor plate making sure to remove all bubbles. Attach adaptor filter net to

adaptor plate. Make sure that no air bubbles are trapped in the filter net (see Note 9).

8. After resin has settled, insert adaptor. Secure adaptor in place to the top plate with four bolts (torque wrench not required).
9. Use height adjuster handle to lower adaptor below the top of the liquid surface. Eliminate any air bubbles that may form at the liquid surface.
10. Seal the adaptor o-ring. Again using the height adjuster handle, lower the adaptor to expel any air in the adaptor line.
11. Attach column to pump and remove any air in the line through the top safety valve.
12. Direct flow into column and immediately open bottom valve.
13. Adjust flow rate to slowly increase pressure to until it stabilizes at 30% of maximum packing pressure of resin. Let media pack until a constant bed height is reached. Maintain packing flow rate for at least 3 column volumes (CVs). Adjust flow rate if pressure deviates from set point.
14. Mark the top of resin bed and turn off pump. Immediately close top and bottom valves.
15. Lower adaptor to mark expelling excess water through the top safety valve. Run pump and expel any excess air in line through top valve.
16. Direct flow into column and slowly adjust flow rate until pressure valve reads 70% of maximum packing pressure of resin. Allow the media to pack until a constant bed height is reached. Maintain the packing flow rate for an additional 3 column volumes. Mark position of resin bed and turn off pump. Close top and bottom valves.
17. Lower adaptor to resin bed mark. Do not lower adaptor for more than 5 mm into resin bed.
18. Expel any air in line and direct flow into column. Adjust flow rate to reach 70% packing pressure. Let run for 15 min and check for any compression of resin.
19. If no compression occurs, column is packed. If resin bed compresses, repeat steps 17 and 18. Close valves and prepare column for evaluation.

3.4. General Procedure for Column Evaluation

1. Prepare 2 M NaCl solution for column evaluation (see Note 10). Filter and degas solution using a 0.2- μm vacuum filter unit.
2. Connect packed column to FPLC system. Immerse one inlet line into 2 M salt solution and another line into WFI water (see Note 11).

3. Equilibrate column for 2 column volumes with 0.5 M NaCl solution.
4. Spike column with 0.7 column volumes of 2 M NaCl solution.
5. Chase spike with 2 column volumes of 0.5 M NaCl solution.
6. Flush system with 2 column volumes of WFI water.
7. Monitor salt peak using conductivity curve. Measure peak elution distance (V_e) and peak height.
8. Evaluate column efficiency in terms of height equivalent to a theoretical plate (HETP), asymmetry factor, and reduced plate height (see Note 12).

3.5. Running IMAC Column

1. Program your FPLC to run the following program or administer the following program manually.
 - (a) Pre-equilibrate the 5 L IMAC column with 2 CVs of solubilization buffer at a flow rate of 100 ml/min (see Note 13).
 - (b) Load the solubilized, clarified lysate at 100 ml/min until all of the lysate has passed through the column. A sample of load flow through should be retained for testing but the rest should be directed to waste (see Note 14).
 - (c) Pass 7 CVs of solubilization buffer through the column at 70 ml/min. This fraction should be sampled and the rest sent to waste. This detergent-containing buffer removes most of the endotoxin.
 - (d) Pass 5 CVs of urea buffer through the column at 70 ml/min. This fraction should be sampled and the rest sent to waste. This buffer removes excess detergent from the protein preparation. Detergent removal can be noted by the significant drop in 280 nm absorbance.
 - (e) Pass 5 CVs of 15% imidazole buffer diluted into urea buffer through the column at 70 ml/min. The FPLC should be able to mix this for the user as part of the program. Retain this elution fraction, though the His-tagged protein should not be in this fraction.
 - (f) Pass 3 CVs of imidazole buffer through the column at 70 ml/min. Retain this fraction separately as it should have the eluted protein.
2. Test the protein preparation for endotoxin levels using a commercially available kit as per the manufacturer's directions. The endotoxin levels should have dropped significantly from the loaded material to the eluted material. See Table 1 for a typical change in endotoxin throughout the purification process.

Table 1
Sample endotoxin reduction during protein purification

Sample	Endotoxin concentration (EU/ml)
IMAC load	4913800
IMAC elution	316.1
QXL elution	10.9
AXC load	<5
AXC elution	<5
Final bulk	<5

3.6. Running Anion Exchange Chromatography Column (AXC)

- Pack a 1 L column containing Q Sepharose XL resin as described above.
- Program your FPLC to run the following program or administer the following program manually.
 - Equilibrate with 3 CVs of carbonate buffer at 50 ml/min, sending the fraction to waste (see Note 15).
 - Load the retentate with carbonate buffer at a 70:30 ratio at 50 ml/min until the entire fraction has been loaded.
 - Wash the column with 2.5 CVs of carbonate buffer at 50 ml/min, directing this fraction to waste.
 - Elute your protein of interest with 5 CVs of carbonate elution buffer. Your protein may require less elution volumes, depending on its characteristics.

3.7. Running Cation Exchange Chromatography Column (CXC)

- Pack a 1 L column containing SP Sepharose XL resin as described above (see Note 16).
- Program your FPLC to run the following program or administer the following program manually.
 - Equilibrate the CXC with carbonate elution buffer for 2 CVs at 50 ml/min, sending this fraction to waste.
 - Load the AXC elution onto the column at 50 ml/min until the entire fraction has been loaded. *Retain this fraction* as it will contain your protein of interest.
 - Wash out the column with 1.5 CVs of carbonate elution buffer to obtain the rest of your protein. Combine this fraction with the fraction obtained in step 2b, Subheading 3.7.
 - Resolve your fractions on an SDS-PAGE gel. You should have a purified protein with little to no endotoxin.

4. Notes

1. TFF washing of inclusion bodies can be achieved using a 0.2- μm membrane. We have found that keeping inlet pressure less than 25 kPa prevents inclusion body loss for a variety of proteins.
2. WFI: Water for Injection. Consult US Pharmacopeia Manual for specific details. For “research”-grade material WFI water is not necessary. Use of WFI water is necessary for cGMP-grade material and prevents pyrogen contamination from the water source.
3. We find that it is best to replace all inlet tubing on the FPLC system prior to cleaning. In certain cases it may be necessary to replace all the tubing connected to the FPLC if endotoxin test fails. Use clean glass bottles to store cleaning solutions as plastic carboys may not be pyrogen free.
4. Column packing method described is based on a constant flow rate throughout the column. Optimum packing flow rate can be empirically determined by developing a pressure/flow rate curve. Flow rate is linear flow rate (cm/h).
5. It is recommended that resin be mixed with packing media (WFI water) to form a 75% slurry, where the packed volume/slurry volume = 0.75.
6. We find that for larger columns (column diameter > 200 mm) attaching filter net to column takes two personnel. Some air bubbles can be suctioned out of the filter net with a pipette.
7. Use torque wrench to tighten bolts. Reference BPG columns instruction manual for torque wrench settings.
8. For BPG 100 and 200 columns reference column adaptor assembly section in *Instructions for Use*.
9. Small air bubbles can be suctioned out with a pipette. If air bubbles are large, remove filter and support net and reassemble on adaptor plate.
10. For a 1 and 5 L column make 2 and 6 L of 2 M NaCl, respectively.
11. Take care to have enough WFI water for the entire method. If water level in container gets low, pause evaluation program and refill carboy.
12. HETP = L/N , where L = bed height (cm) and N = number of theoretical plates. N is commonly defined by the equation $5.54(V_c/W_h)^2$. W_h = peak width at half height (ml).

Peak asymmetry (A_s) is calculated as b/a , where b is distance between peak apex and 10% of peak height on the descending side of peak and a is the corresponding measurement on the ascending side of the peak. Reduced plate height = HETP/ d_p , where d_p is the mean particle diameter. Peak asymmetry factor should be as close to unity (1) as possible. Generally column is considered passing if A_s is between 0.8 and 1.5. A theoretically well-packed column should have a reduced plate height of 1.5–3 based on theoretical efficiency derived from a Van Deemter analysis of a packed bed column.

13. Column diameter can become critical at the 5-L scale. We used a 12" diameter column that provided adequate flow rate. Experimentation may be required, but we have found a larger diameter column to be very advantageous at the IMAC step.
14. Note that 100 ml/min may have to be changed based upon your protein. If your protein is very "sticky" or you have a high back-pressure, you should use a slower flow rate. Additionally it may be advantageous to further dilute the protein prior to loading on the IMAC column.
15. The IMAC elution fraction must be buffer exchanged into pH 10.5 carbonate buffer prior to loading onto the AXc column. Use a commercially available TFF apparatus, making sure that your membrane pore size is at least 20 kDa smaller than your protein of interest.
16. The CXc functions as a flow-through, polishing column. The protein does not bind, but some contaminating endotoxin does bind the column. It is possible to conduct another buffer exchange after the AXc into a low pH buffer and bind the CXc if greater protein purity is desired.

References

1. Bone RC (1991) The pathogenesis of sepsis. Ann Intern Med 115:457–469
2. Glauser MP, Zanetti G, Baumgartner JD, Cohen J (1991) Septic shock: pathogenesis. Lancet 338:732–736
3. Venet C, Zeni F, Viallon A, Ross A, Pain P, Gery P, Page D, Vermesch R, Bertrand M, Rancon F, Bertrand JC (2000) Endotoxaemia in patients with severe sepsis or septic shock. Intensive Care Med 26:538–544
4. Morrison DC, Ulevitch RJ (1978) The effects of bacterial endotoxins on host mediation systems. A review. Am J Pathol 93:526–618
5. Kotani S, Takada H, Tsujimoto M, Ogawa T, Takahashi I, Ikeda T, Otsuka K, Shimauchi H, Kasai N, Mashimo J et al (1985) Synthetic lipid A with endotoxic and related biological activities comparable to those of a natural lipid A from an *Escherichia coli* re-mutant. Infect Immun 49:225–237
6. Hack CE, Aarden LA, Thijss LG (1997) Role of cytokines in sepsis. Adv Immunol 66:101–195
7. Lukacs NW, Hogaboam C, Campbell E, Kunkel SL (1999) Chemokines: function, regulation and alteration of inflammatory responses. Chem Immunol 72:102–120
8. United States Pharmacopeia—National Formulary (2010) Vol. USP 34-NF29
9. Jang H, Kim HS, Moon SC, Lee YR, Yu KY, Lee BK, Youn HZ, Jeong YJ, Kim BS, Lee SH, Kim JS (2009) Effects of protein concentration and detergent on endotoxin reduction by ultrafiltration. BMB Rep 42:462–466
10. Yamamoto C, Kim ST (1996) Endotoxin rejection by ultrafiltration through high-flux, hollow fiber filters. J Biomed Mater Res 32:467–471

11. Chen RH, Huang C Jr, Newton BS, Ritter G, Old LJ, Batt CA (2009) Factors affecting endotoxin removal from recombinant therapeutic proteins by anion exchange chromatography. *Protein Expr Purif* 64:76–81
12. Kunioka M, Choi HJ (1995) Properties of biodegradable hydrogels prepared by gamma-irradiation of microbial poly(epsilon-lysine) aqueous-solutions. *J Appl Polym Sci* 58: 801–806
13. Morimoto S, Sakata M, Iwata T, Esaki A, Hirayama C (1995) Preparations and applications of polyethyleneimine-immobilized cellulose fibers for endotoxin removal. *Polym J* 27:831–839
14. Lowe AJ, Anderson KA, Bardliving CL, Huang C Jr, Teixeira LM, Damasceno LM, Ritter G, Old LJ, Batt CA (2011) Expression and purification of cGMP grade NY-ESO-1 for clinical trials. *Biotechnol Prog* 27(2):435–441
15. Matsumae H, Minobe S, Kindan K, Watanabe T, Sato T, Tosa T (1990) Specific removal of endotoxin from protein solutions by immobilized histidine. *Biotechnol Appl Biochem* 12:129–140
16. Liu CL, Nikas YJ, Blankschtein D (1996) Novel bioseparations using two-phase aqueous micellar systems. *Biotechnol Bioeng* 52: 185–192
17. Nikas Y, Liu CL, Abbott N, Blankschtein D (1992) Experimental and theoretical investigations of protein partitioning in 2-phase aqueous micellar systems. *Abstr Pap Am Chem Soc* 203:140-IEC
18. Anspach FB, Spille H, Rinas U (1995) Purification of recombinant human basic fibroblast growth factor: stability of selective sorbents under cleaning in place conditions. *J Chromatogr A* 711:129–139
19. Karplus TE, Ulevitch RJ, Wilson CB (1987) A new method for reduction of endotoxin contamination from protein solutions. *J Immunol Methods* 105:211–220
20. Damais C, Jupin C, Parant M, Chedid L (1987) Induction of human interleukin-1 production by polymyxin B. *J Immunol Methods* 101: 51–56
21. Chen RH (2009) Increasing inclusion body extractability and recoverability by altering fermentation conditions in high cell density *Escherichia coli* cultures. Thesis, Cornell University, Department of Chemical Engineering. <http://dspace.library.cornell.edu/handle/1813/13895>
22. Burgess RR (2009) Refolding solubilized inclusion body proteins. *Methods Enzymol* 463:259–282
23. de Marco A (2011) Molecular and chemical chaperones for improving the yields of soluble recombinant proteins. *Methods Mol Biol* 705: 31–51
24. Eiberle MK, Jungbauer A (2010) Technical refolding of proteins: do we have freedom to operate? *Biotechnol J* 5:547–559

Chapter 18

Effectiveness of Various Processing Steps for Viral Clearance of Therapeutic Proteins: Database Analyses of Commonly Used Steps

Dana Cipriano, Michael Burnham, and Joseph V. Hughes

Abstract

The successful implementation of any biologically derived product in human clinical trials and as a marketed biopharmaceutical requires the critical utilization of effective viral clearance steps. As biologic products have inherent risks of potentially carrying and/or amplifying adventitious viruses that may be present in or introduced into the original materials, a number of processing steps are needed to provide adequate virus removal. Some common process steps are introduced into downstream purification schemes that provide a physical means to separate and/or remove viruses from the therapeutic protein. The physical steps often include virus-removing filters and chromatographic resins in column or membrane configurations, but can also include the introduction of irradiation, high heat steps, or other means for destroying the infectivity of a virus. Chemical treatment steps are often utilized as a means to inactivate a wide variety of virus types.

A general overview is provided that describes the most commonly used techniques for virus removal or inactivation for the validation of virus clearance. Data sets from studies performed at WuXi AppTec for a wide variety of biologics reveal a number of steps that provide guidance for the design of process steps dedicated to viral clearance. The overall efficiency of several process steps reveals a number of efficient, robust steps, such as nanofiltration which can be designed for removal of almost all viral species. Exposure to a low pH or solvent detergent is also a robust step for inactivating enveloped virus. Steps with greater variances in predictability include chromatography steps such as capture columns and anion exchange resins. A lower removal capacity is typically expected for other chromatography steps such as cation exchange steps.

Key words: Viral clearance, Virus inactivation, Virus filtration, Log reduction values, Database, Virus removal, Virus safety, Process development, Downstream purification, Virus validation

1. Introduction

Biopharmaceutical manufacturers and regulatory agencies recognize that any product derived from a biological source is at risk for the potential of a viral contamination event. Such products can

include monoclonal antibodies, recombinant proteins, vaccines, gene therapy vectors, cell therapies, blood- or plasma-derived products, animal- and human-derived products, or medical device products derived from biological materials such as animal tissue or serum. A viral contamination can be a serious issue for the safety of these products. Regulatory agencies have established laws and guidelines for safety testing at multiple stages throughout the drug development process, starting at the Investigational New Drug (IND) or Investigational Device Exemption (IDE) stage through post-commercialization (1–6). The guidelines provide a framework for executing viral clearance studies to address the known level of retrovirus particles that are present in most cells used for recombinant protein production as well as any potential adventitious contaminants.

There have been documented cases of viral contaminations of biological products throughout the years starting as early as the 1940s. More recent viral contaminations of biological products continue to demonstrate the need for focused attention on supply chain testing and risk assessments, as well as designing and implementing efficient process steps to remove and/or inactivate adventitious viruses and other agents. In the last 15 years there have been several documented cases of Minute Virus of Mice (MVM) contamination for companies using large-scale cell culture to produce monoclonal antibodies (7–9). There have also been documented contaminations in biologic manufacturing with other small, non-enveloped viruses like Calcivirus (Vesivirus 2117 strain) and porcine circovirus type 1 (PCV-1) (10, 11).

Many of the typical steps that are evaluated for viral clearance are outlined in Table 1. Except for the Virus Removal Filtration, most of these steps are utilized primarily to achieve high purity for the product as well as remove other process-related impurities such as residual host cell DNA and proteins or product aggregates. For plasma- or blood-derived protein products, separation steps for plasma fractionation also have proven to be effective virus removal steps. As raw materials have been implicated as the root cause for many of the documented virus contamination events, evaluation of the production processes has also become important to evaluate for virus removal. Different process steps are typically examined for virus clearance for various raw materials including gamma irradiation, ultraviolet light treatments, and high-temperature, short-time (HTST) technologies.

Viral clearance studies are undertaken by biopharmaceutical manufacturers to validate that a process can remove or inactivate known or potentially unknown viral contaminants. The process involves adding different viruses at certain process steps (termed “spiking”) and evaluating the inactivation or removal once the

Table 1
Process steps evaluated for viral clearance based on biologic product type

Product	Removal steps	Inactivation steps
Monoclonal antibodies and recombinant proteins	Filtration—Virus removal filters Chromatography—Affinity/capture; anion and cation exchange; hydrophobic interaction; mixed mode	Low pH Solvents/detergents
Therapeutic proteins from plasma	Filtration—Virus removal filters Chromatography Separation processes—Cohn fractionation; precipitation	Low or high pH Solvents/detergents Alcohols Heat—Pasteurization; high temperature, short time (HTST)
Raw materials	Filtration—Virus removal filters	Heat inactivation—HTST UV or gamma irradiation

process step is performed. A panel of viruses that provides a broad range of different physicochemical characteristics (Table 2) is typically used to demonstrate the robustness of the process to clear viruses. The viruses selected for a clearance study are based on the source material of the product and raw materials used in the process. The evaluation of a wide range of viruses (DNA, RNA, enveloped, and non-enveloped) is assumed to increase the assurance for removing known viral particles as well as previously unrecognized contaminants or new and emerging viruses.

Since viruses should not be introduced into a GMP manufacturing facility, virus studies are typically conducted at a small scale or bench laboratory scale (1–6). Regulatory documents expect that process steps chosen for viral clearance evaluation should be orthogonal process steps with a defined mechanism for virus removal or inactivation, without overlap between the mechanisms or steps. Each step is then independently spiked with virus, the step is performed as per standard manufacturing procedure at reduced scale, and samples are collected throughout the process step and evaluated for a virus titer. Each sample is then assessed by either an infectivity assay or quantitative PCR to measure virus particle levels. Virus reduction factors or log reduction values (LRVs) are then calculated for the initial material after spiking and for all intermediate and final samples selected from the process. LRVs are calculated following guidelines specified in the ICH Q5A regulations (3) by taking the \log_{10} of the ratio of the virus concentration (load) in the pre-purification material and the virus

Table 2
Typical virus panel evaluated in viral clearance studies mammalian cell-derived processes

Virus family	Virus	Envelope	Genome	Approximate size (nm)
Retroviridae	Xenotropic murine leukemia (X-MuLV) Amphotrophic murine leukemia (A-MuLV) Human immunodeficiency virus (HIV)	Yes	RNA	80–130
Herpesviridae	Herpes simplex (HSV-1) Pseudorabies virus (PrV) Infectious bovine rhinotracheitis virus (IBR)	Yes	DNA	120–225
Parvoviridae	Minute virus of mice (MMV) Porcine parvovirus (PPV) Canine parvovirus (CPV)	No	DNA	18–26
Reoviridae	Reovirus type 3	No	RNA	60–85
Picornaviridae	Poliovirus Encephalomyocarditis virus (EMC) Hepatitis virus (HAV)	No	RNA	27–32
Flaviviridae	Bovine viral diarrhea virus (BVDV)	Yes	RNA	40–60
Rhabdoviridae	Vesicular stomatitis virus (VSV)	Yes	RNA	45–100 × 100–430
Papovaviridae	Simian virus type 40 (SV40)	No	DNA	45–50

concentration in the post-purification material. The formula takes into account both the titers and the volumes of the materials before and after the purification step.

Generally steps that result in LRVs of greater than 3 or 4 logs are considered to be robust steps for viral clearance. Conversely, process steps that exhibit an LRV less than 1 Log₁₀ are not generally considered by the regulatory agencies to be effective virus removal or inactivation steps and are assumed to be within the variability of the viral detection method.

As a contract testing organization WuXi AppTec has been performing viral clearance studies for sponsor companies over the last 20 years and has collected data from more than 2,000 studies. Utilizing summaries of this database, this chapter reviews viral log reduction data for the most common process steps evaluated for therapeutic proteins with an emphasis on those used for monoclonal antibodies and other recombinant proteins.

2. Virus Titer Determination and Log Reduction Recording

2.1. Plaque Assay Titer Determination

To determine the potency of each virus stock and virus spiked samples, serial dilutions (usually tenfold) of each viral sample are prepared. Samples are titrated on the appropriate indicator cells using 3 wells of 6-well tissue culture plates for each dilution or larger plates (150 cm^2) for large-volume testing. The plates are inoculated, then overlaid with agarose overlay medium, and incubated at $37 \pm 2^\circ\text{C}$ with $5 \pm 2\%$ CO_2 for 1–14 days dependent upon the virus. Following incubation, the plates are fixed and stained to detect viral cytopathic effects (CPEs). Infectious centers are identified visually as an area devoid of cells (or with foci) and recorded, and the final infectious titer determined from the serial dilutions.

2.2. qPCR Analysis Titer Determination

A quantitative PCR (qPCR) assay is used to quantitate virus titer for some of the removal steps. RNA or DNA (virus dependent) is extracted using commercial kits and then concentrated into a small volume of buffer for testing. For each sample, 5 μL of appropriately diluted sample is assayed in triplicate. A standard line of the appropriate virus RNA or DNA (as needed) provides a known linear range for quantitation. Virus titers are determined for each sample to quantitate the amount of virus genome copies in each sample.

2.3. Database

The viral log reduction data, presented in the chapter, is extracted from the database of viral clearance studies performed at WuXi AppTec, focusing on recombinant proteins and monoclonal antibodies produced in mammalian cell culture. For the filtration steps, the filter type, nominal pore size, virus spike percentage, and other parameters are recorded. For the column chromatography steps, data is examined to summarize the resin types, bind and elute versus flow-through mode, and in some cases the protein concentrations, buffer conditions, pH, and conductivity. For inactivation steps, the incubation times, pH, temperatures, chemical composition, and concentrations are examined.

2.4. Data Sorting

The data is sorted by unit operation and then by virus. Following the ICH guidelines Q5A (3), for mammalian cell culture-derived products, four virus families are usually evaluated. To file for an IND to get into Phase I clinical studies, two viruses are generally recommended, including a retrovirus and a parvovirus (5). Retroviruses are classified by the regulations as a relevant virus, as the most commonly used cells producing biologic products (CHO and murine cells) have retrovirus particles. Parvoviruses are classified as model viruses that demonstrate a challenge for most purification processes as they are resistant to many chemical inactivation steps and are of

such a small size (18–26 nm). Viral clearance studies conducted for a commercial filing just before marketing evaluate typically four viruses (1–4). In addition to the retrovirus and parvovirus, a DNA enveloped virus, usually from the herpes virus family, and an RNA non-enveloped virus, such as Reovirus, are evaluated in these studies. Other viruses may be evaluated if there is a reason to conclude that additional virus risks may be present, such as the presence of serum during manufacturing, for which additional viruses would be included. This chapter concentrates mostly on results for four commonly used viruses (retro-, parvo-, herpes-, and reoviruses) but in a few cases also examines a broader range of viruses.

3. Virus Clearance Methods and Effectiveness

3.1. Nanofiltration or Virus Removal Filters

Nanofilters or virus removal filters are utilized in most antibody and other protein purification processes to primarily remove potential viral contaminants. They are traditionally inserted towards the end of the purification process, where the feed stream or load material is highly purified. The mechanism of action of these filters is primarily size exclusion. The product is allowed to flow through the filter and virus is held back or retained by the pores and the structure of the filter. There are different pore size filters available on the market and they are generally grouped into large virus removal filters (>30 nm) and small virus removal filters (20 nm or smaller). Depending on the molecular weight and size (nm) of the antibody or protein product, this drives the pore size selection appropriate to each protein. The approximate size of monoclonal antibodies is 8–12 nm, and many other therapeutic proteins are less than 15 nm, but IgM antibodies are over 30 nm. Filters can be operated in either tangential flow (TFF) or dead end filtration mode (16, 17, 20). Product concentration at this stage of processing is typically around 1–10 g/L, while the robustness of the virus removal filter is sometimes challenged with product concentrations as high as 40 g/L.

Data from several small virus filters extracted from the database includes studies performed with the Planova 15 N, Planova 20 N, Planova BioEX, DV20, Virosart CPV, NFP, and Viresolve Pro filters (Fig. 1). For larger sized viruses, such as retroviruses, herpesviruses, and reoviruses, each of these filters demonstrate robust viral clearance with most processes achieving over 5 LRVs. With all of these larger viruses, the complete removal of virus is achieved, even for those studies where only 2–3 logs of clearance is demonstrated. Other factors, such as virus titer, amount of virus in the load material, as well as toxic and interfering effects of the load material on the indicator cells, can contribute to the different LRVs.

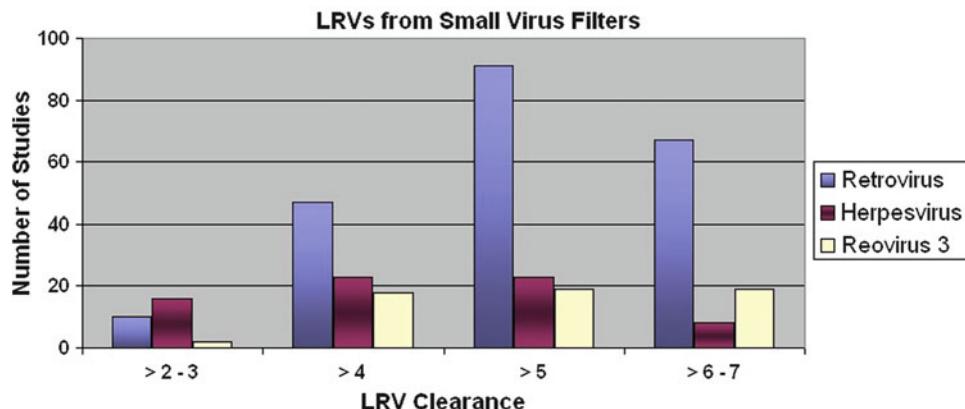


Fig. 1. LRVs from small virus filter. Clearance (LRVs) for retroviruses (X-MuLV and A-MuLV), herpesviruses, and reovirus type 3 was determined for several studies using nanofilters of 20 nm or less.

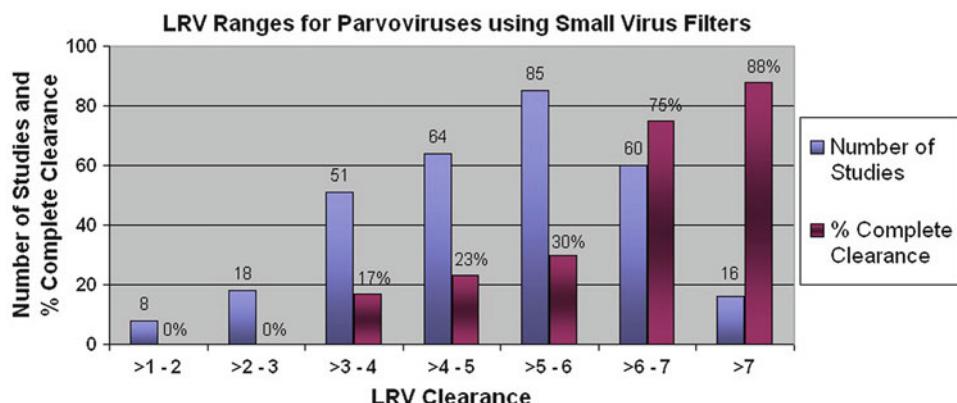


Fig. 2. LRV ranges for parvoviruses using small virus filters. Clearance (LRVs) for parvoviruses (PPV and MVM) was determined for several studies using nanofilters of 20 nm or less.

The parvoviruses at 18–26 nm in size are difficult to remove from a therapeutic protein source, so nanofiltration is typically considered a critical manufacturing step to evaluate for viral clearance as the filter pore sizes in the 20 nm range are close to the known size of the parvoviruses. A separate data analysis performed for the parvoviruses (Fig. 2) demonstrates that these filters result in more than 4 LRVs in most studies (over 75%). Overall these filters reduce virus to non-detectable levels in 78% of the studies and thus are considered a robust step for virus clearance even for one of the smallest virus types. It is also noted that the percentage of studies that demonstrated complete removal increases as the total LRV increases (Fig. 2).

The large virus filtration data includes studies utilizing the Planova 35 N, Millipore NFR, and DV50 filters. Evaluation of the data shows that 80% of the studies provides greater than 5 logs of clearance, with complete removal of larger sized virus types (data

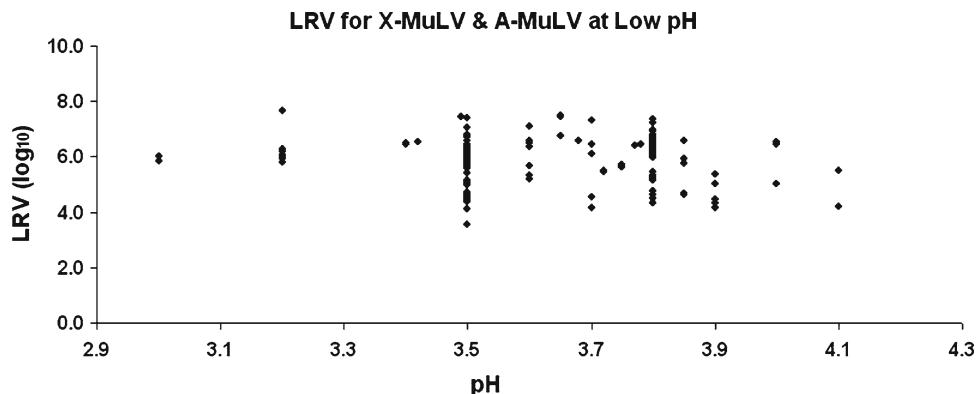


Fig. 3. LRV for X-MuLV and A-MuLV inactivation by incubation at low pH, for 60 min. Data was summarized from several studies, where the pH of samples was reduced and then incubated for 60 min.

not shown). The remaining 20% yields LRVs of >3 logs of clearance, and two studies demonstrating incomplete clearance. The larger pore-sized nanofilters are not sufficient for the removal of small viruses like parvovirus, but would be beneficial to include in purification processes for therapeutic proteins larger than 20 nm. Overall, these larger pore-sized nanofilters are robust for retroviral (80–130 nm) removal, but are not as effective for parvoviruses.

3.2. Low pH Inactivation

Process steps that employ low pH incubations are frequently performed during the purification of monoclonal antibodies and occasionally with other therapeutic proteins. Most steps are typically conducted at pH levels in the range from 3.0 to 4.2 pH by adding various acid solutions to the protein matrix, and then holding the solution for incubation times typically from 30 to 60 min (with a few studies extending to 120 min). The exposure time is based on the stability of the protein product under the low pH conditions. Kinetics of inactivation experiments demonstrate that complete inactivation of infectious virus is typically obtained in low pH processes with incubation times of 60 min or greater, while infectious virus may remain after 30 min (data not shown). While some virus may remain after 30 min, LRVs of greater than $4 \log_{10}$ are typically obtained and thus this condition is still considered robust for retroviral inactivation. Several studies employing low pH inactivation and utilizing the most common time frame of 60 min of exposure demonstrate good inactivation for the enveloped retroviruses: X-MuLV and A-MuLV (Fig. 3). These viruses are readily inactivated at pH 3.0–3.8 with LRVs ranging from 4 to 8 logs. The LRVs from these studies are often dependent on how much the sample needs to be diluted before adding a nontoxic–non-interfering amount to the indicator cells, as well as the input titer of virus, and the volume of sample available for large volume plating. At higher pHs (greater than pH 4.0), the inactivation of retrovirus was more variable.

Most low pH steps are conducted at ambient temperature, but some processes are also designed for colder temperatures, in the range of 2–8°C. Inactivation kinetics differ from those conducted at room temperature and inactivation is not as robust (data not presented). For the non-enveloped viruses, such as parvovirus and reovirus, there is no inactivation of these viruses in this pH range (pH 3.0–4.50).

3.3. Solvent Detergent

For antibody products, the solvent detergent step usually occurs just prior to downstream processing. This step is added by some manufacturers to help clarify the bulk harvest and break up cell debris and cell membranes that remain when the culture process is complete. The process step is also reported as a good method for inactivation of enveloped viruses. Harvest cell culture is treated with solvent detergent and incubated with or without mixing for a period of time, dependent on the process. The mechanism of action of the detergent is saturation of the membrane and full solubilization of the lipid envelope into mixed micelles (12). Therefore, processing parameters that are important for virus inactivation include detergent concentration, incubation time, and temperature.

Data for treatment with Triton X-100, one of the most common detergents utilized, is presented in Fig. 4 which represents information for several studies using either an X-MuLV or A-MuLV retrovirus. In general, solvent detergent treatment is a very robust step typically yielding LRVs of greater than 4 logs for inactivation of enveloped viruses like the retroviruses, as well as herpesviruses, rhabdoviruses, and flaviviruses. A cutoff point for robust virus inactivation is noted at concentrations below 0.025% Triton X-100 (Fig. 4) and this is seen for all enveloped viruses (data not shown). Solvent Detergent steps are not typically evaluated for non-enveloped viruses

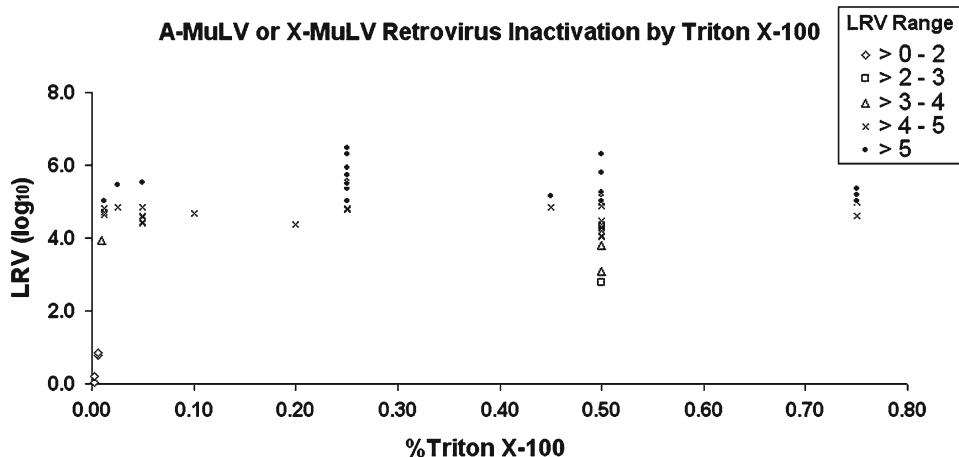


Fig. 4. Inactivation of murine retroviruses by Triton X-100. Data was summarized from several studies, where Triton X-100 was added to the indicated concentration and then incubated for 60 min at 18–25°C.

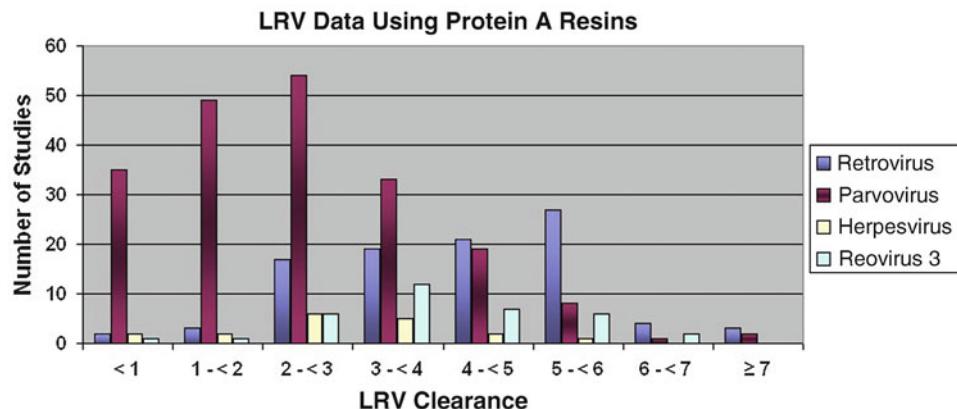


Fig. 5. LRV data for four virus families using Protein A resins. Data was collected from a number of column studies where the column loads were spiked with one of the four commonly used viruses, and the eluted protein fraction was assayed for virus levels. The data for the MuLV retrovirus and herpesvirus was collected as qPCR data, while the parvovirus and reovirus data was determined by infectivity assays.

(like parvoviruses or reoviruses) as this step is not effective at inactivation of these viruses. Even their use on tissue-based medical device products results in low or no substantial effect on virus infectivity for non-enveloped viruses.

3.4. Chromatography Resins

3.4.1. Capture or Affinity Chromatography

The capture or affinity chromatography step is typically the first step in the downstream purification process of an antibody or recombinant protein. The load material is usually the cell culture harvest from the upstream cell culture process. This step serves to capture the product by binding to the specific resin, and most often impurities such as host cell DNA and protein, process-related impurities, and sometimes adventitious viral contaminants (17) are in the flow through and wash fractions. In general, the protein product is typically eluted from an affinity column by a pH or conductivity shift.

Some of the most commonly used capture resins, especially for monoclonal antibodies, are the Protein A resins, and the data analyses in Fig. 5 evaluated data from studies using standard Protein A, Mab Select, and Mab Sure resins. Four virus families that represent the most commonly used viruses for these studies demonstrate a wide range of LRVs from 1 to >7 LRV. The retrovirus clearance is typically seen to peak in the range of 3–5 LRVs, but there are some instances of outliers where there is less than 1 or 2 LRVs. Some of these examples are likely the result of underloading the resin, i.e., using less than 50% of the maximum capacity. The clearance of parvoviruses peaks more in the 1–3 LRVs range, but there have been some studies where 4 or more LRVs is achieved. For the other two virus families, herpesviruses and reoviruses, this resin type appears to represent a more robust removal step as the majority of the studies peak with 3–5 LRVs, and sometimes 6 LRVs.

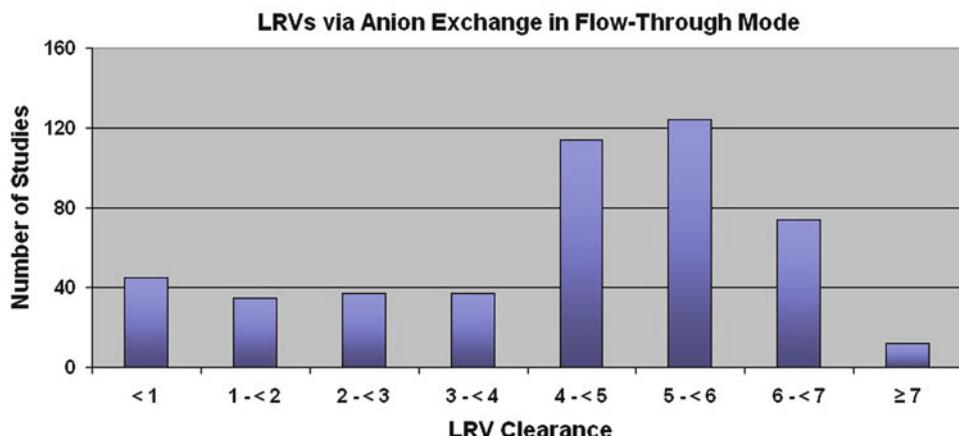


Fig. 6. LRVs for multiple virus families (21 virus types) evaluated via anion exchange in Flow-Through mode. Resins and membrane adsorbers were included in the data set. Multiple resin types were included in the analysis of anion exchange resins such as but not limited to Q Sepharose Fast Flow, Q Sepharose XL, Capto Q, Fractogel TMAE, DEAE, Fractogel, and Toyopearl QAE. Membrane anion exchangers, Sartobind Q and Mustang Q, were also included in the analysis.

This data suggests that Protein A chromatography is not a robust step for all viruses, but it may be utilized to provide extra LRVs for some viruses (18).

3.4.2. Anion Exchange Chromatography

Charged resins or membranes such as anion exchange (AEX) are routinely used in downstream purification processes for monoclonal antibodies and many other recombinant proteins, and are used as polishing steps to remove impurities such as DNA and host cell proteins (HCPs). These steps can be operated in two modes: Flow-Through mode where the product does not bind to the resin, and Bind and Elute mode where the product is bound and then released with changes in the conductivity or pH of the elution buffer. Many studies examine the virus clearance for AEX columns and membrane absorbers (that have AEX properties) that are run in the Flow-Through mode (Fig. 6). In this summary for the Flow-Through mode, studies from over 21 different virus families are presented. Multiple resin types are included in the analysis of AEX resins as well as some membrane AEXs (like Sartobind Q and Mustang Q). For the majority of these viral clearance studies the AEX step demonstrates an LRV range of 4–7 logs of clearance in flow-through mode, though a number of studies achieve less than 1 or 2 LRVs (Fig. 6). For some of the more common viruses in viral clearance studies this step is robust for virus removal: parvoviruses demonstrate >65% of studies with >4 logs of clearance, while the retroviruses, herpesviruses, and reoviruses demonstrate >75% of studies with >4 logs of removal (data not shown).

The Bind and Elute mode for the AEX step is typically used less than the Flow-Through mode, and the viral clearance data also reveals a different pattern for the multiple virus families (Fig. 7).

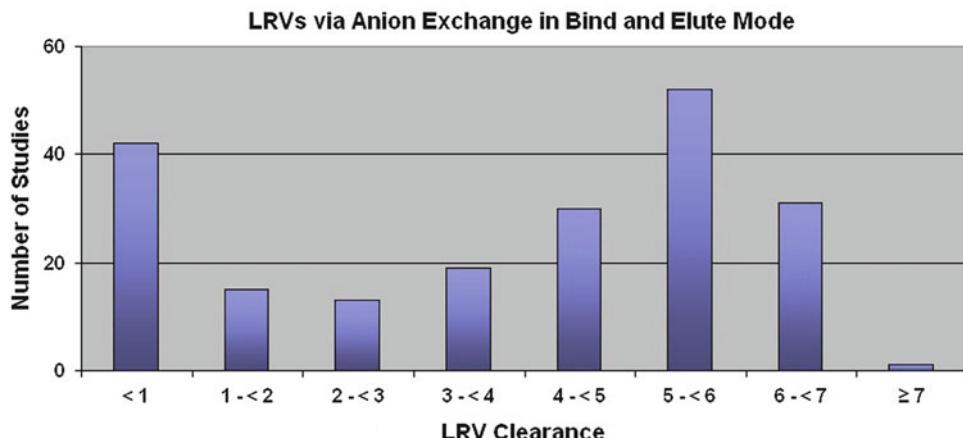


Fig. 7. LRVs for multiple virus families (21 virus types) evaluated via anion exchange in Bind and Elute mode. Resins and membrane adsorbers were included in the data set.

Two trends of log reduction are detected where the majority of the studies demonstrate 4–6 LRVs clearance in a very robust fashion, but there is also a second data set with some process steps not working efficiently to remove viruses (less than 1 LRV). For some of the more common viruses used in viral clearance studies this step is more variable than the flow-through mode: parvoviruses demonstrate only >50% of studies with >4 logs of clearance while the retroviruses are only 65%, but the herpesviruses and reoviruses demonstrate >80% of studies with >4 logs of removal (data not shown).

3.4.3. Cation Exchange Chromatography

Cation exchange resins are negatively charged resins that can also be used in downstream purification processes as polishing steps to remove impurities and sometimes to separate monomeric or aggregated forms of the product. As most antibodies and other proteins have a high pI, the cation step is commonly run in the Bind and Elute mode (22). Data analyses for over 18 virus families demonstrate that the cation resins most often result in a low capacity to remove viruses (Fig. 8). More than 50% of the studies result in less than two logs of removal for both enveloped and non-enveloped viruses. The outliers with this method are those instances where the step resulted in over 4 LRVs. This data suggests that the cation exchange step is not generally considered a robust step for virus removal as there are very few studies yielding greater than 4 LRVs.

4. Further Considerations and Conclusions

Viral Clearance evaluation is a critical step in establishing the safety of biological products. As virus contamination events continue to be a risk for the biopharmaceutical industry, establishing robust

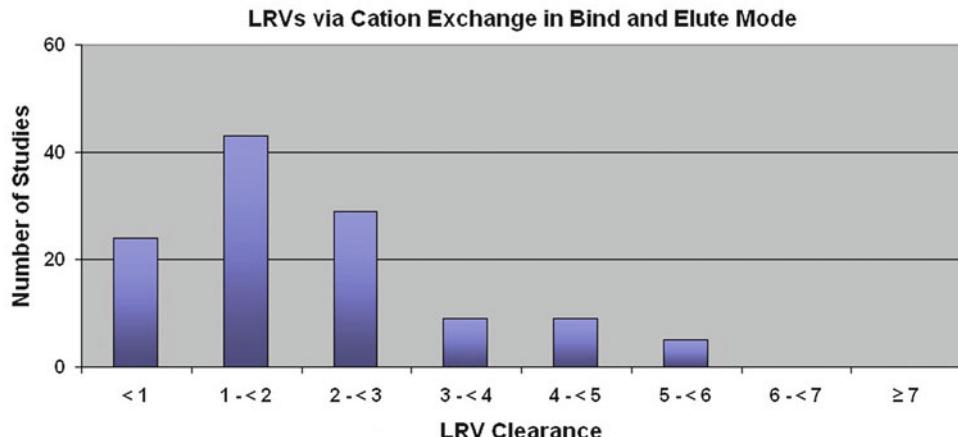


Fig. 8. LRVs for multiple virus families (18 virus types) evaluated via cation exchange chromatography in Bind and Elute mode. Resins evaluated included SP Sepharose, SP Sepharose XL, SP 650, Fractogel, CM Sepharose, Poros 50HS, and other cation exchange resins.

steps to clear viruses is very important to guide drug manufacturers during process development activities. Generic steps for therapeutic protein production that can work on many or all virus families are important for successful viral clearance programs as these can be expected to work on previously known contaminants as well as unrecognized or emerging viruses (25). In addition, many companies are now working on platform processes if they are developing several similar proteins (such as with monoclonal antibodies), where robust clearance steps are advantageous for effective product development.

The analysis of the database at WuXi AppTec helps to identify some common downstream processing steps used during antibody or protein manufacturing which are robust for virus removal or inactivation. In addition, some commonly used steps that provide only limited or more variable results for log reduction of viruses are identified. Mechanisms of action for virus clearance were also defined that could be employed to provide orthogonal viral clearance steps. The LRVs from orthogonal steps can be summed together by drug manufacturers to establish the overall reduction factor for the safety of a process.

The use of viral removal filters is identified as one of the most consistent and robust steps that works on all types of viruses. While there are nanofilters that were initially designed for removal of larger viruses and have pore sizes >35 nm, most of the protein processes now target the use of the smaller pore sizes of 15 and 20 nm unless these will limit the protein recovery. For larger viruses that are commonly evaluated and are over 50 nm in size, the use of “small virus” filters results in complete removal of virus and often 4–7 or more LRV. For the parvoviruses, which

are only 18–26 nm, the vast majority of clearance studies on small filters result in high LRV with only limited breakthrough of small amounts of virus. Several factors may play a role in contributing to this data. In some cases the final filtrate sample is collected and added to the “Chase” sample, which is essentially a buffer flush of the membrane post filtration to recover the remaining product held up in the filter. With this processing, there is a potential to flush through virus that may be normally retained in the membrane. In addition, large volume plating of the fractions is routinely performed to increase the LRV that can be claimed for this step, but this increased sensitivity can also increase the chance of detection of a small amount of virus in the sample matrix, if present.

Virus purity is another factor to consider (19). Flux decay or membrane fouling can be an issue when high-titer virus is added to the feed stream, particularly if there are contaminants in the virus preparation or there are aggregates with the antibody or other protein, causing plugging of the filter. In evaluations of purer parvovirus preparations, an improvement is seen in the performance of the filtration runs with less flux decay or plugging. Filter performance can be enhanced during the virus spiking trial with purer virus and this translates into more product capacity per surface area of membrane in manufacturing. While some limited amounts of virus may pass through the membrane, one consistently sees 4 logs or more of clearance, making the nanofiltration step a robust step for parvovirus clearance.

The two chemical inactivation steps, low pH or the addition of solvents/detergents, are efficient at inactivating envelope-containing viruses. The mechanism of action of the low pH step is solubilization of the virus envelope (12). Transmission electron microscopy (TEM) studies conducted at the FDA at pH 3.8 in either acetate or citrate buffer demonstrated impact to the capsid and membrane structure of the retrovirus, X-MuLV. Changes in pH to acidic conditions were found to induce surface conformational changes that interfered with the infectivity and membrane fusion (15). The detergent steps also affect the structure of the membrane of enveloped viruses. While data in this chapter was presented only for Triton X-100, similar inactivation was seen for Tween and Tween with Tri (*N*-butyl) phosphate (TnBP). These two steps both result in over 4 LRVs each and sometimes up to 7 or 8 LRVs for all of the envelope-containing viruses. But as robust as these steps are for enveloped viruses, the steps do not affect non-enveloped viruses. Most of the non-enveloped viruses, such as the parvoviruses, reoviruses, etc., are typically not affected by the chemical inactivation conditions that are routinely used in processing proteins, and would only be inactivated at extreme levels that would destroy the protein product.

Several chromatographic steps are available for protein purification, but the efficient processing of the therapeutic protein on various chromatographic resins does not always result in a complete separation of the spiked virus from the protein fractions. Nevertheless some processes are being designed that can result in >4 LRV, though the conditions may not work for all viruses nor to the same extent across all virus types.

AEX is documented as a robust step for virus removal (17, 23, 24), and the data presented here supports that the majority of Flow Through processes will lead to >4 LRV. AEX resins are positively charged, therefore binding negatively charged process impurities such as DNA, HCPs, and viruses (22, 24). Many impurities will bind to the resin and are removed through this step. The mode of action of these resins is based on the pI of product and the pH and conductivity of the buffers to perform the step. Products with a higher pI (pH 7.0–8.0) do not typically bind to AEX resins. However, the pI of viruses, particularly non-enveloped viruses (parvoviruses and reoviruses), is much lower. The pI of reovirus is in the range of 3.0–4.0 (13) and the pI of parvovirus is 5.0–5.3 (14). These viruses bind tightly to the resin and do not typically co-elute with the product. At WuXi AppTec, we also demonstrated that AEX technology under the appropriate conditions is very efficient for purifying virus stocks. Likely the pI of the protein product and the pI of the different viruses play a significant role in the different pattern of separation on this resin type in either the Bind and Elute or Flow-Through modes.

In most cases cation exchange chromatography is not efficient for viral clearance, though there are a few exceptions for some proteins. For monoclonal antibodies this is not routinely useful for more than a few LRVs. As most antibodies are basic, this impacts the binding capacity of these products to the negatively charged resin. In acidic conditions, most antibodies become protonated, potentially increasing the binding capacity for antibodies to the resin. Since viruses are typically negatively charged, binding conditions that would be optimal for products binding to a positively charged resin may also be optimal for virus binding, further decreasing the likelihood that this step can be robust for virus removal, as viruses may co-elute with the product.

WuXi AppTec is continuing to analyze the data in the database for trends with certain proteins, buffer conditions, or other parameters that result in an impact on the LRVs for common steps. Key processing parameters are also being identified that will be important for Design of Experiments (DOE) studies or Quality by Design for viral clearance studies (21). These parameters will assist with process validation set points and ranges, especially for manufacturers that employ platform processes. Future studies will analyze critical conditions for common steps, with focus on those parameters that impact viral clearance and overall product safety.

References

1. US Food and Drug Administration (FDA) Center for Biologics Evaluation and Research (CBER) (1997). "Points to consider in the manufacture and testing of monoclonal antibody products for human use." 94D-0259
2. US Food and Drug Administration (FDA) Center for Biologics Evaluation and Research (CBER) (1993) "Points to consider in the characterization of cell lines used to produce biologicals"
3. Committee for Proprietary Medicinal Products (CPMP) (1997) "International conference on harmonization (ICH) Topic Q 5 A. Quality of biotechnological products: viral safety evaluation of biotechnology products derived from cell lines of human or animal origin." Consensus Guideline ICH Viral Safety Document: Step 4. CPMP/ICH/295/95
4. Committee for Proprietary Medicinal Products (CPMP) (1996) "The design, contribution and interpretation of studies validating the inactivation and removal of viruses." CPMP/BWP/268/95
5. European Medicines Agency (EMEA) (2009) "Guideline on virus safety evaluation of biotechnological investigational medicinal products." EMEA/CHMP/BWP/398498
6. American National Standard (2007). "Medical devices utilizing animal tissues and their derivatives—Part 3: validation of the elimination and/or inactivation of viruses." ANSI/AAMI/ISO 22442-3
7. Garnick RL (1996) Experience with viral contamination in cell culture. *Dev Biol Stand* 88:49–56
8. Skrine J (2010) A biotech production facility contamination case study—Minute virus of mice. Presented at PDA/FDA adventitious viruses in biologics: detection and mitigation strategies workshop. Bethesda, MD, 1–3 Dec
9. Moody M (2010) MVM contamination—A case study: detection, root cause determination and corrective actions, at merrimack pharmaceuticals. Presented at PDA/FDA adventitious viruses in biologics: detection and mitigation strategies workshop. Bethesda, MD, 1–3 Dec
10. Jones N (2010) Identification and remediation of a cell culture virus contamination, at genzyme. Presented at PDA/FDA adventitious viruses in biologics: detection and mitigation strategies workshop. Bethesda, MD, 1–3 Dec
11. Pierard I (2010) Contamination by porcine circovirus: findings, investigations and learning, at GSK. Presented at PDA/FDA adventitious viruses in biologics: detection and mitigation strategies workshop. Bethesda, MD, 1–3 Dec
12. Norling L (2011) Virus inactivation and application of the modular approach. Presented at IBC's 8th international conference viral safety for biologicals. Orlando, FL, 24–25 Feb
13. Floyd R, Sharp DG (1978) Viral aggregation: effects of salts on the aggregation of poliovirus and reovirus at low pH. *Appl Environ Microbiol* 35:1084–1094
14. Zhou J (2008) Methods for removing viral contaminants during protein purification. United States Patent Application Publication. Pub. No.: US 2008/0132688 A1. Pub. Date: 5 Jun
15. Brorson K, Krejci S, Lee K, Hamilton E, Stein K, Yuan Xu (2003) Bracketed generic inactivation of rodent retroviruses by low pH treatment for monoclonal antibodies and recombinant proteins. *Biotechnol Bioeng* 82: 321–329
16. Brough H, Antoniou C, Carter J, Jakubik J, Xu Y, Lutz H (2002) Performance of a novel Viresolve NFR virus filter. *Biotechnol Prog* 18: 782–795
17. Gottschalk U (2009) Process scale purification of antibodies (Chapters 4, 7, 8, 9). Wiley, Hoboken, NJ
18. Lute S et al (2008) Robustness of virus removal by Protein A chromatography is independent of media lifetime. *J Chromatogr A* 1205:17–25
19. Technical Report No. 47 (2010) Preparation of virus spikes used for viral clearance studies. Parenteral Drug Association (PDA) Bethesda, MD
20. Technical Report No. 41 (2008) Virus filtration. Parenteral Drug Association (PDA) Bethesda, MD
21. Cherney B (2010) Application of quality of design in the control of adventitious viruses: gaps in the current processes in the prevention of virus contaminants. PDA/FDA adventitious viruses in biologics: detection and mitigation strategies workshop. Bethesda, MD, 1–3 Dec
22. Norling L et al (2005) Impact of multiple reuse of anion exchange chromatography media on virus removal. *J Chromatogr A* 1069:79–89
23. Strauss D et al (2010) Strategies for developing design spaces for viral clearance by anion exchange chromatography during monoclonal antibody production. *Biotechnol Prog* 26: 750–755
24. Strauss D et al (2009) Understanding the mechanism of virus removal by Q Sepharose fast flow chromatography during the purification of CHO-cell derived biotherapeutics. *Biotechnol Bioeng* 104:371–380
25. Miesegaes G, Lute S, Brorson K (2010) Analysis of viral clearance unit operations for monoclonal antibodies. *Biotechnol Bioeng* 106:238–246

Chapter 19

High-Throughput Quantitative N-Glycan Analysis of Glycoproteins

Margaret Doherty, Ciara A. McManus, Rebecca Duke,
and Pauline M. Rudd

Abstract

N-linked oligosaccharides are complex non-template-derived structures that are attached to the side chains of asparagine, via the nitrogen atom. Specific changes in the N-glycans of serum glycoproteins have been associated with the pathogenesis of many diseases. The oligosaccharides present on the C_H2 domain of immunoglobulins are known to modulate the effector functions of the molecule. These glycans provoke various biological effects, necessitating the development of robust high-throughput technology in order to fully characterize the N-glycosylation profile. This chapter describes in detail four methods to release N-glycans from the glycoprotein of interest. Two of these protocols, referred to as the “In-Gel Block” and “1D sodium dodecyl sulfate-polyacrylamide gel electrophoresis” methods, require immobilization of the glycoprotein prior to analysis. An automated method is also described, involving the purification of immunoglobulins directly from fermentation media, and, finally, an “In-solution method” is detailed, which directly releases the N-glycans into solution. HILIC and WAX-HPLC are used to analyze the N-glycan profile. Exoglycosidase enzymes digestion arrays, in combination with computer-assisted data analysis, are used to determine both the sequence and linkage of the N-glycans present.

Key words: N-glycan analysis, HPLC, High throughput, Automation, Glycomics, Robotics, Oligosaccharide

1. Introduction

Many proteins exist as glycoproteins, meaning that their surface is functionalized with sugar molecules, commonly referred to as glycans. The glycan composition typically consists of monosaccharides, such as D-mannose, D-glucose and D-galactose (as well as their N-acetylated derivatives), L-fucose, and 5-N-acetyl- α -neuraminic acid (see Fig. 1). Most glycosylation can be divided into N- and O-linked depending on whether the glycan is attached

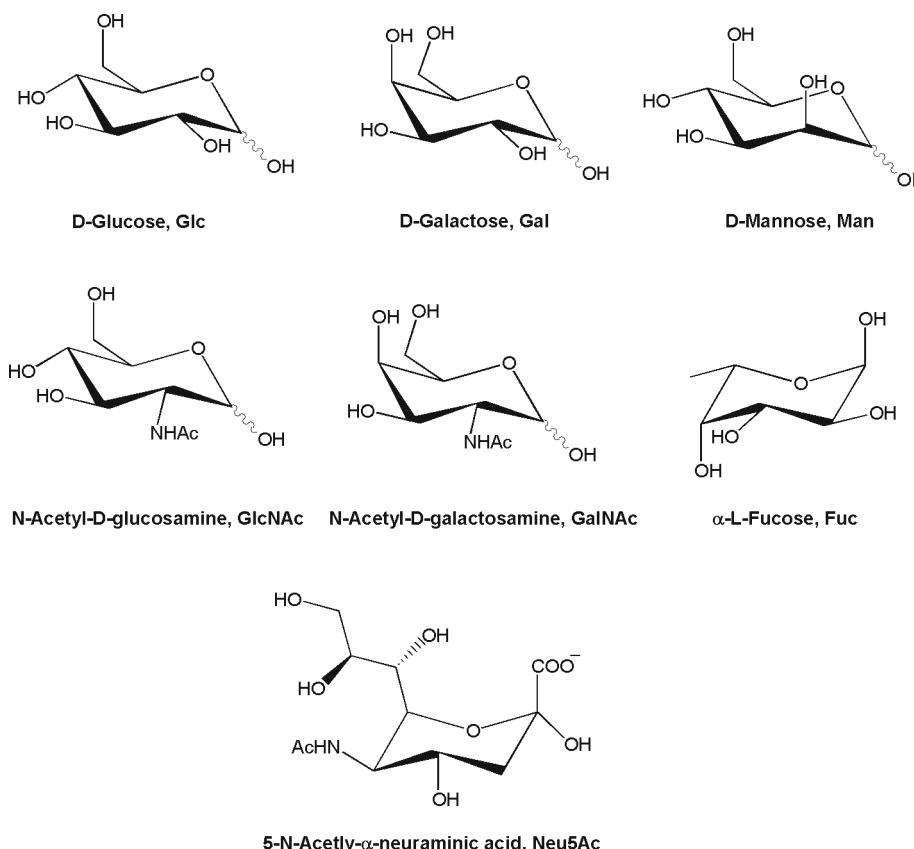


Fig. 1. Chemical structure, name, and abbreviation of selected monosaccharides present in N-glycans. Structures are presented in their chair-like pyranose conformations, with the wavy line representing variability at the anomeric center (α or β). Fucose occurs naturally in the L-confirmation. Neu5Ac is one of more than 50 sialic acids and often terminates glycan chains in mammalian cells.

via the amide group of an asparagine residue (N-linked) or the hydroxyl group of a serine or threonine residue (O-linked) (see Fig. 2). Less frequently, serine residues may be modified with a single fucose, mannose, or galactosamine residue (1).

It is evident that glycosylation plays a fundamental role in the pathogenesis of many different biological processes. Therefore, it is important to be able to fully analyze the glycosylation pattern from body fluids such as serum, as well as individual proteins such as antibodies, in order to ascertain the effect that the glycans may have on their biological functions (2). Glycoproteins now constitute important bio-therapeutics; therefore, in order to support quality by design (QbD) and process analytical technology (PAT), rapid detailed quantitative analyses of the glycans are fundamental to the biotechnology and pharmaceutical industries. During therapeutic protein production, cellular glycosylation may differ as a result of varying production parameters (3). Thus, when utilizing different cell lines, it is necessary to monitor these changes closely (4).

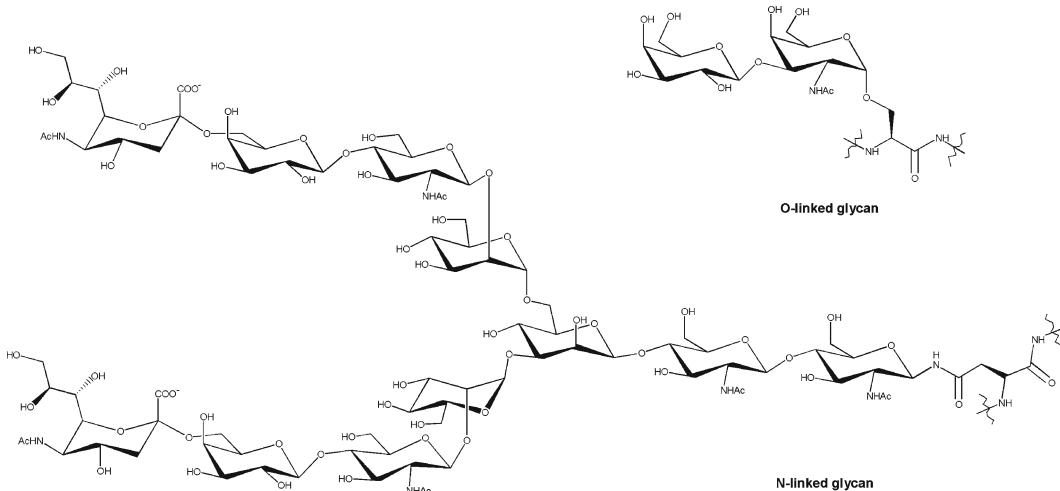


Fig. 2. Structures of typical N-linked and O-linked glycans attached to asparagine and serine, respectively.

Glycoproteins usually display a diverse heterogeneous population of glycoforms, in which different glycans are located at each glycosylation site, further complicating their analyses. This heterogeneity can depend on a number of factors including the levels of glycosyltransferases (5), the nucleotide-activated sugars present (6), as well as the tertiary structure of the protein at the glycosylation site (7). All these factors play a role in determining the glycosylation pattern of the glycoprotein.

In this chapter we outline the methods required for the analysis of N-glycans utilizing a Hydrophilic Interaction Liquid Chromatography (HILIC) strategy. HILIC separates compounds by passing a hydrophobic mobile phase across a neutral hydrophilic stationary phase, and solutes then elute in order of increasing hydrophilicity. The N-glycans are initially cleaved from the protein backbone using Peptide-N-Glycanase F (PNGase F). As glycans lack a chromophore, the released oligosaccharides are fluorescently labeled with 2-aminobenzamide (2-AB), providing very high sensitivity (10 fmol) for HPLC analyses (Fig. 3). A dextran ladder (2-AB-labeled glucose homopolymer) is used as an external reference standard. The data are fitted to a fifth-order polynomial distribution in order to assign glucose units (GUs) to each peak resolved on the chromatograph (8). The monosaccharides that comprise a glycan chain are predominantly linked via the hydroxyl group at the 1' position of one sugar to the 2,3,4,6, or 8 hydroxyl position of the sequential sugar, forming a glycosidic bond, which may have either α or β -stereochemistry (Fig. 4) (9). These structural complexities are recognized by a range of exoglycosidase enzymes, which allow the sequence and linkage of N-glycans to be determined (10). Each enzyme is specific for a particular monosaccharide in a particular

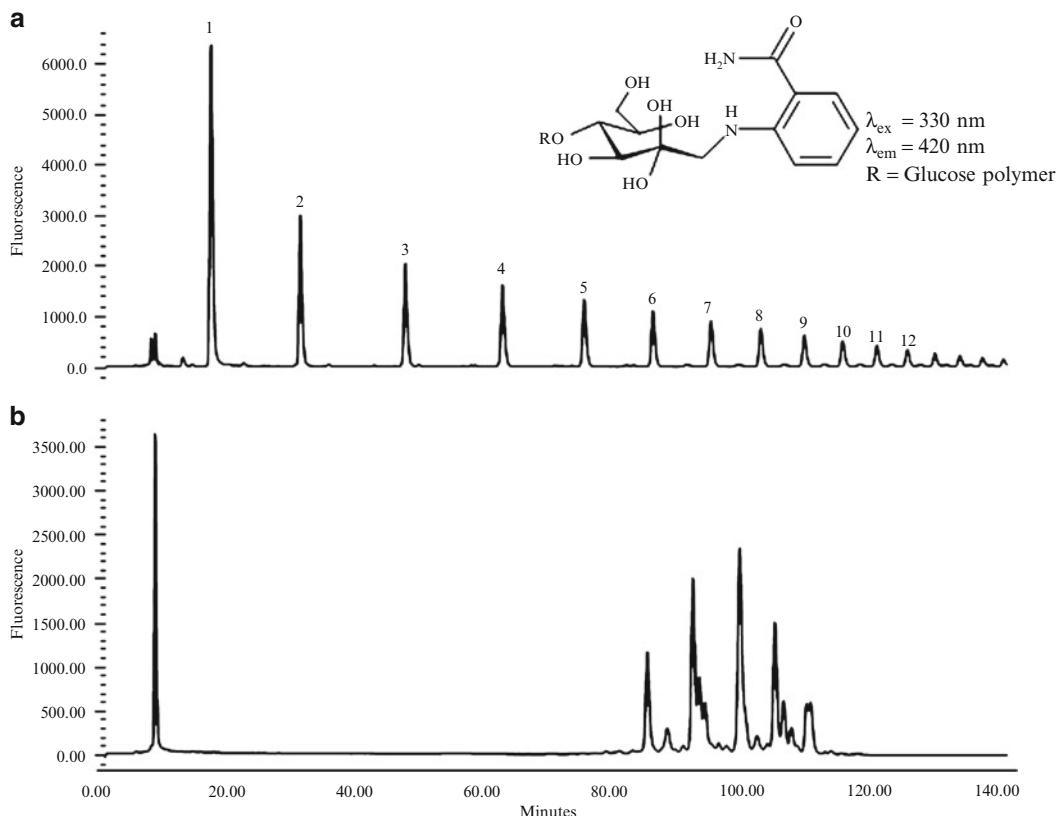


Fig. 3. (a) HILIC chromatographic profile showing 2-AB-labeled dextran standard and corresponding glucose unit (GU) values. Insert shows the chemical structure of 2-AB-labeled glucose polymer and excitation and emission wavelengths for 2-AB. (b) HILIC profile of 2-AB-labeled N-glycans released from human serum IgG.

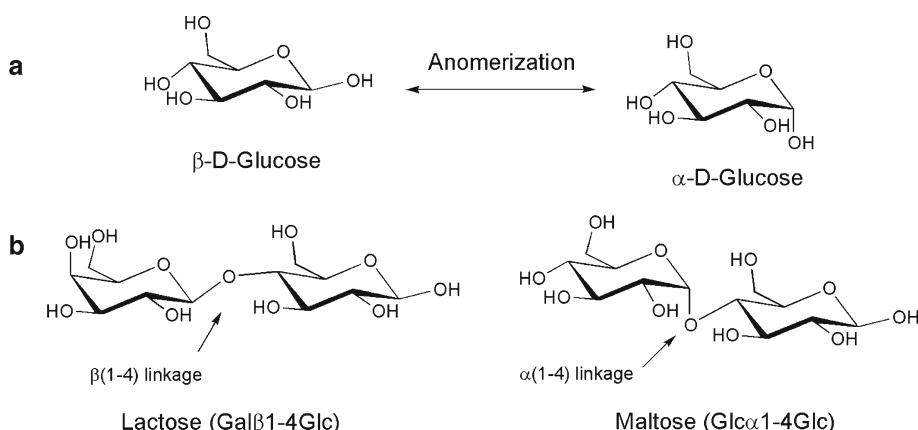


Fig. 4. (a) The two anomeric forms of D-glucose. (b) Chemical structures of two disaccharides, lactose and maltose, showing the different glycosidic bonds, $\beta(1,4)$ and $\alpha(1,4)$, respectively.

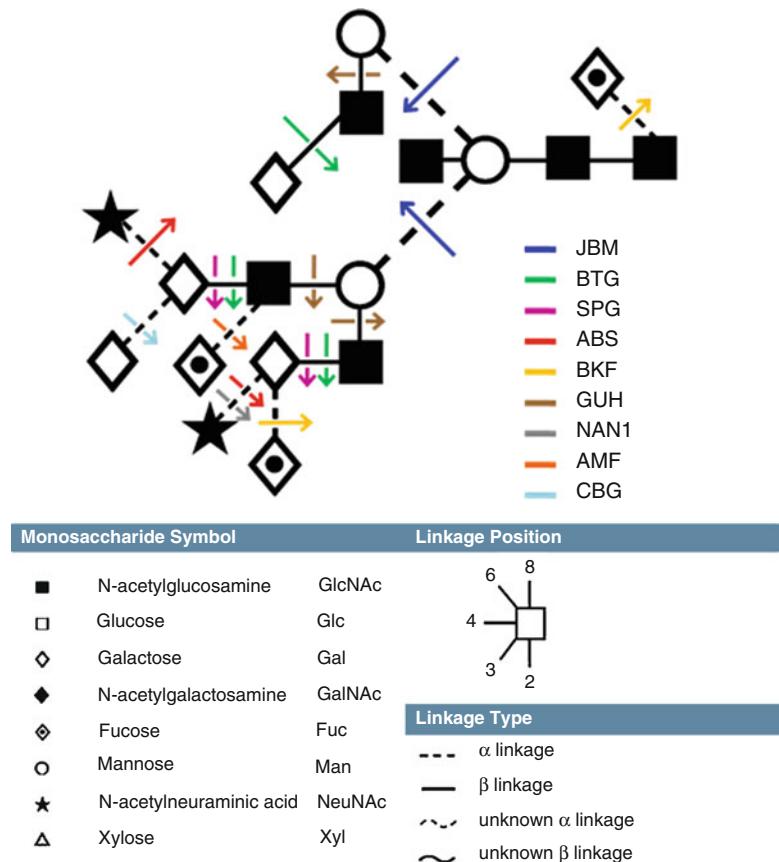


Fig. 5. Illustration of exoglycosidase enzyme specificities for N-glycans, using Oxford nomenclature. See Table 1 for more detail of the enzyme specificities.

linkage (Fig. 5). After treatment with PNGase F, aliquots of the released glycan pool are subjected to arrays of exoglycosidase digestions (Fig. 6), followed by HPLC analyses. Weak Anion Exchange (WAX) HPLC (see Fig. 7) is also used to separate glycans by their charge, enabling more information to be gleaned in relation to any sialic acid, phosphate, or sulfate groups present.

In addition to enzyme digestion arrays, structures are assigned using computational tools such as GlycoBase (database) and AutoGU (analytical tool). GlycoBase, developed in house, is a relational database for glycan structures and contains important information, such as HPLC GU values for over 380 2-AB-labeled N-glycans, as well as exoglycosidase digest data. These informatic tools provide an invaluable resource in glycan analysis and are discussed in further detail in a separate chapter (11).

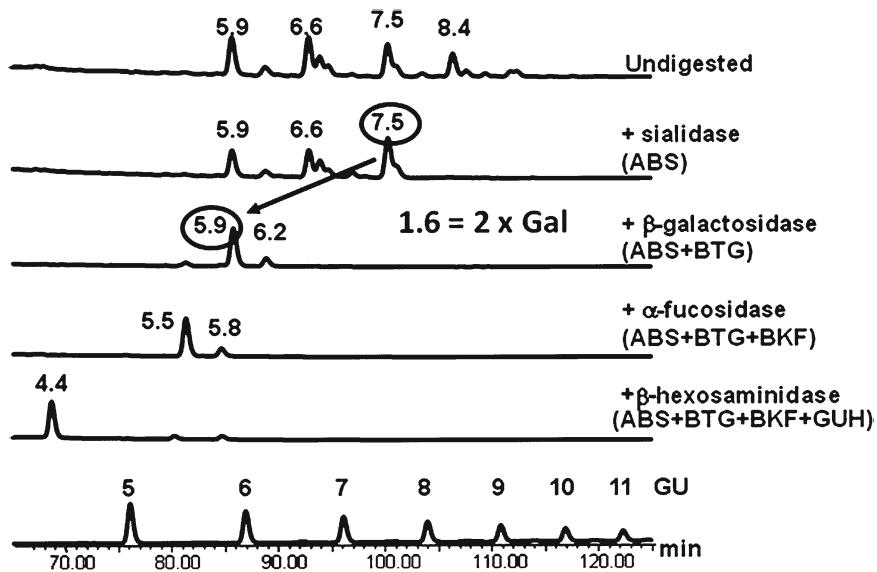


Fig. 6. Chromatographic HPLC profiles of 2-AB-labeled human IgG N-glycans. The first glycan profile displays undigested whole pool glycans, each additional profile a series of exoglycosidase digestions.

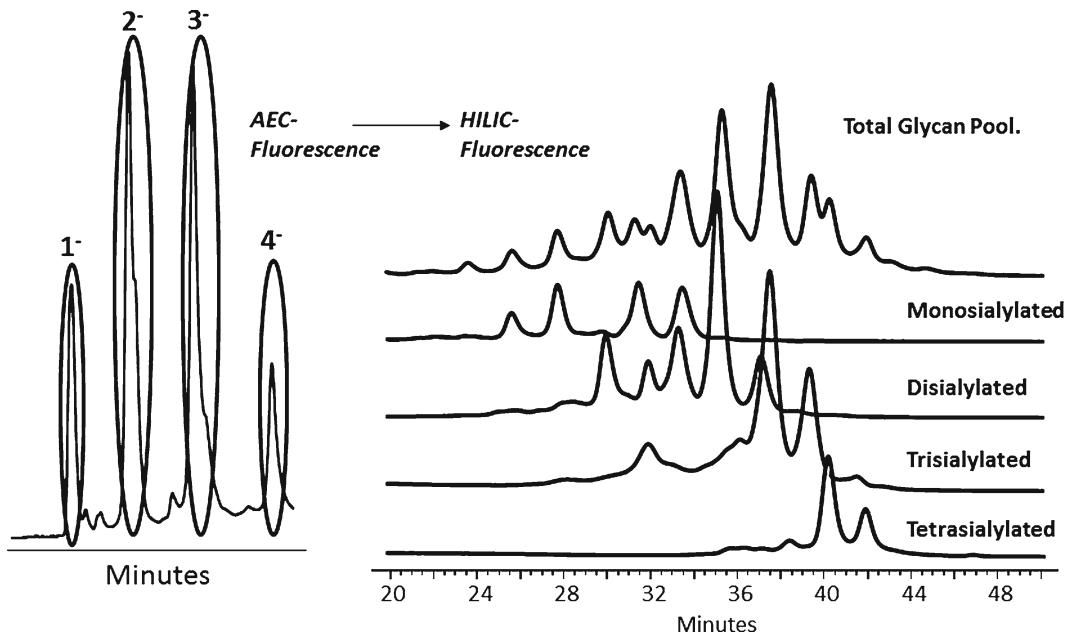


Fig. 7. WAX-HPLC 2D chromatography profiles of bovine serum fetuin N-glycans showing separation of the mono-, di-, tri-, and tetrasialylated glycans.

The following chapter details four methods for the analysis of N-glycans from glycoproteins. The main focal point of this chapter is a high-throughput method, which is referred to as the “In Gel Block” method and utilizes 96-well plates (Fig. 8). A second rapid

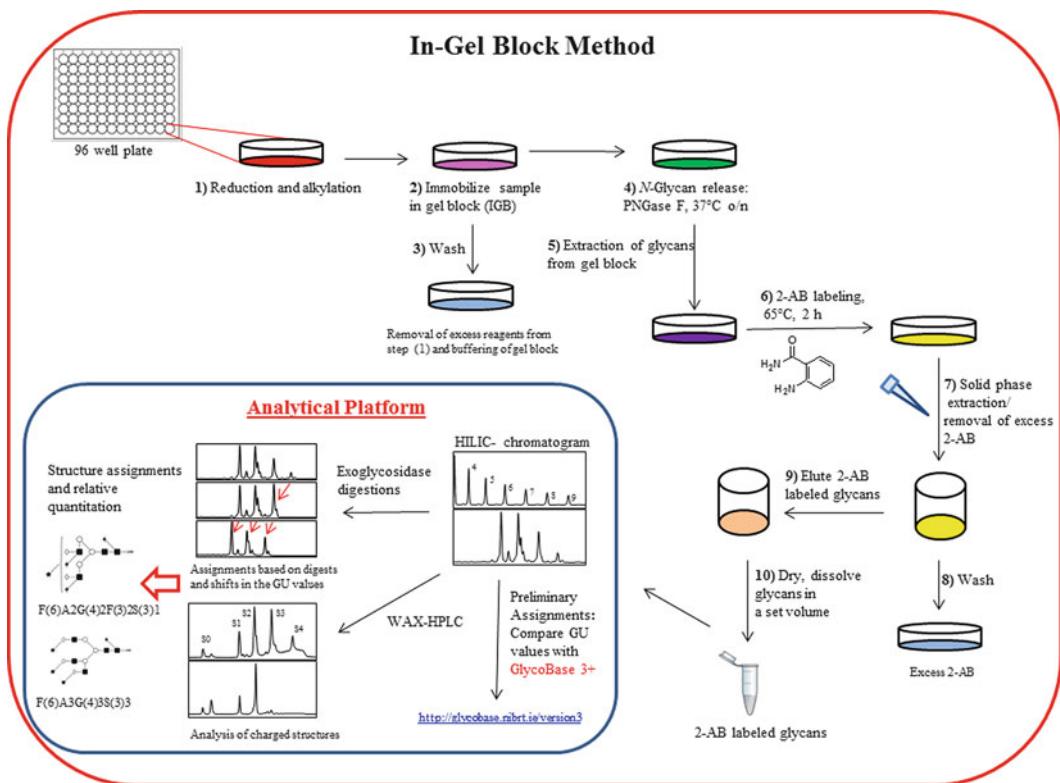


Fig. 8. Schematic summarizing the In Gel Block method for N-glycan analysis and the subsequent data analysis.

automated method focuses on purifying IgG directly from fermentation media. This particular method also uses 96-well plates and has been fully automated on a liquid-handling robotic system, hence increasing the throughput further. Scripts for the automated method were written using Hamilton software, Vector (Reno, NV, USA). The Hamilton STAR robotic platform was adapted fit for purpose and contains an integrated vacuum manifold, a plate shaker, and incubators to facilitate the N-glycan preparation and release method. However, this 96-well plate format is easily amenable to automation on any robotic platform. A third method utilizing 1D sodium dodecyl sulfate (SDS)-polyacrylamide gel electrophoresis (PAGE) is also described which isolates glycans directly from gel bands, and finally an “In-Solution” method is described in detail. In the case of the latter, the sample is not immobilized prior to N-glycan release; as a result the sample is filtered after PNGase F digestion in order to separate the protein and N-glycans. The method of choice depends on both the sample and the information required, i.e., 1D SDS-PAGE can be used to isolate glycoproteins from a mixture of proteins and subsequently perform N-glycan analysis. In comparison, the other three methods characterize all N-glycans present in the sample (12).

2. Materials

2.1. Sample Preparation—In Gel Block and 1D SDS-PAGE

1. 0.5 M Dithiothreitol (DTT): Dissolve 7.7 mg of DTT in 100 µL of water (see Note 1) and immediately freeze in single-use (20 µL) aliquots at -20°C.
2. 100 mM Iodoacetamide (IAA): Prepare 18.5 mg in 1 mL of water and use immediately.
3. Sample buffer: 100 µL of 10% SDS + 62.5 µL of stacking buffer + 337.5 µL of water.
4. 5× Laemmli sample buffer: 0.04 g of Bromophenol blue, 0.625 mL of stacking buffer, pH 6.6, 1 mL of 10% SDS, and 0.5 mL of glycerol in 2.875 mL of water.

2.2. Automated Sample Preparation—IgG Purification

1. Protein A plates, Pierce (Rockford, IL, USA).
2. Phosphate-buffered saline solution (pH 7.2).

2.3. Sample Preparation—In Solution

1. Milli Q water.
2. 1.0 M DTT: Dissolve 15.42 mg of DTT in 100 µL of water and immediately freeze in single-use (20 µL) aliquots at -20°C.
3. 100 mM IAA (see Subheading 2.1, step 2).

2.4. Immobilization of Sample—In Gel Block

1. Protogel: 30% (w/w) Acrylamide: 0.8% (w/v) bis-acrylamide stock solution (37.5:1) (Protogel ultrapure protein and sequencing electrophoresis grade, gas stabilised, National Diagnostics, Hessle, Hull, UK).
2. 10% APS: 1 g of Ammonium peroxisulphate (AnalaR; BDH) in 10 mL of water. Store as 20 µL aliquots at -20°C. Do not refreeze.
3. TEMED: *N,N,N',N'*-Tetramethyl-ethylenediamine.
4. 10% SDS: 10 g of SDS (specially pure; BDH) in 100 mL of water.
5. Gel buffer 1.5 M Tris pH 8.8: 18.2 g of Tris in 100 mL of water. Adjust pH to 8.8 using HCl.

2.5. 1D SDS-PAGE

1. XCell Sure Lock™ Minigel system and gel cassettes, Invitrogen, Paisley, UK.
2. Resolving gel buffer: 1.5 M Tris–HCl pH 8.8. Weigh 18.2 g of Tris in 100 mL of water and adjust to pH 8.8 with HCl (see Note 2).
3. Stacking gel buffer: 0.5 M Tris–HCl, pH 6.6. Weigh 6.6 g of Tris in 100 mL of water and adjust to pH 6.6 with HCl.

4. Protogel: 30% (w/w) acrylamide/0.8% (w/v) bis-acrylamide stock solution (37.5:1.0, Protogel, National Diagnostics, Hessle, Hull, UK).
5. 10% Ammonium persulfate (APS): 1 g of APS in 10 mL of water.
6. TEMED (see Note 3).
7. 10% SDS: 1 g of SDS in 10 mL of water.
8. SDS-PAGE running buffer: 0.025 M Tris-HCl, pH 8.3, 0.192 M glycine, 0.1% SDS (see Note 4).
9. Water-saturated butanol: Shake equal volumes of water and butanol together in a glass bottle. Use the top layer.
10. Coomassie stain: 1.25 g of Coomassie R250 (brilliant blue), 250 mL of methanol, 50 mL of concentrated acetic acid, and 200 mL of water.
11. Destain 1: 50% (v/v) methanol, 7% (v/v) concentrated acetic acid, 43% water.
12. Destain 2: 5% (v/v) methanol, 7% (v/v) concentrated acetic acid, 88% water.
13. Molecular-weight markers: Sigma wide-range (molecular weight 6,500–205,000, Sigma Aldrich, Ireland).

2.6. N-Glycan Release and Extraction—In Gel Block, 1D SDS-PAGE, and Automated Method

1. PNGase F buffer: 20 mM of NaHCO₃, pH 7.0 (0.168 g in 100 mL of H₂O adjusted to pH 7.0 with HCl). Store frozen at -20°C in 10 mL aliquots. PNGase F stock solution: PNGase F (Prozyme, Carlsbad, CA) made up in water to 1,000 U/mL.
2. 1% Formic acid (Sigma) in water.

2.7. N-Glycan Release and Preparation for Derivatized—In Solution

1. PNGase F buffer and PNGase F stock solution (see Subheading 2.6, step 1).
2. PALL Nanosep 10 K MWCO membrane filters (see Note 5).
3. 1% Formic acid (Sigma) in water.

2.8. Fluorescent 2-AB Labeling and Removal of Excess

Underderivatized Material

1. LudgerTag 2-AB labeling kit (Ludger Ltd., Abingdon, Oxon, UK). Acetonitrile (E Chromasolv® for HPLC far UV, Riedel-de Haën, Sigma).
2. LudgerClean A cartridges (Ludger Ltd., Abingdon, Oxon, UK) in a 96-well plate format.

2.9. Fluorescent 2-AB Labeling and Removal of Excess

Underderivatized Material—In Solution

1. LudgerTag 2-AB labeling kit (see step 1 in Subheading 2.8).
2. Normal phase resin tips, Phytip, available from Phynexus.
3. 20% Acetonitrile.
4. 95% Acetonitrile.

2.10. HILIC Profiling of 2-AB-Labeled N-Glycans

1. HILIC stock solution: 2 M Ammonium formate, pH 4.4. Weigh 184.12 g of formic acid into a 2-L glass beaker. Place the beaker in an ice bath to which salt has been added to take the temperature down to -10°C , add 1 L of water, and stir with a glass rod. Adjust the pH by adding $4 \times 50\text{ mL}$ 25% ammonia solution making sure that the temperature drops between each addition; then add in 5 mL aliquots until pH 4.4 is reached at room temperature. This causes a rapid rise in temperature; thus, make sure that the ammonia is added in small amounts.
2. Transfer the solution to a 2-L volumetric flask and make up to 2 L with water. Store this stock solution in a brown Winchester bottle at room temperature. Waters 2695 (Waters Ltd., Elstree, Herts, UK) separations module with a Waters 474 or 2475 fluorescence detector (or another HPLC system which delivers a reproducible shallow gradient and has a fluorescence detector).
3. HILIC column: TSK Amide-80 3 μm (4.6-mm ID \times 15 cm L) (Anachem, Luton, UK).
4. 2-AB-dextran ladder (2-AB-glucose homopolymer, Ludger).

2.11. Array of Exoglycosidase Digestions

1. An array of exoglycosidase enzymes such as those listed in Table 1 is required. These can be purchased from a variety of companies such as Prozyme, Ludger, Merck Biosciences (Nottingham, UK), New England Biolabs (Hitchin, Herts, UK), or Sigma. However, the standard and specificity of these enzymes may vary, so it is important to check their specificity occasionally against standard N-glycans (obtainable from Ludger or Prozyme).
2. Incubation buffers: 10 \times 50 mM sodium acetate, pH 5.5, for mixed enzyme incubations. 5 \times incubation buffers at the optimal pH are usually supplied with individual enzymes.
3. Protein-binding filter plates for enzyme removal before HPLC (PALL Corporation, AcroPrep PN 5034. Ann Arbor, MI, USA).

2.12. WAX-HPLC Profiling of 2-AB-Labeled N-Glycans

1. 1 M ammonium acetate stock buffer solution: Add 57 mL of glacial acetic acid (Merck pro analysis grade or equivalent) into a glass beaker containing 500 mL of Milli-Q water. Rinse the measuring cylinder with Milli-Q water and add these washings to the beaker. Adjust the pH of the solution to pH 7 using 25% ammonia solution. Once a stable pH has been achieved adjust the solution to volume using Milli-Q water in a 1-L prerinced volumetric flask. Transfer to a 1-L Duran bottle and store at room temperature. This stock solution is then used for the preparation of Mobile phase B.

Table 1
Exoglycosidase enzyme array

Enzyme	Full name and activity
JBM	<i>Jack Bean Mannosidase</i> Releases α (1-2)- and α (1-6)- more efficiently than α (1-3)-linked mannose residues
BTG	<i>Bovine testes β-Galactosidase</i> Hydrolyzes non-reducing terminal β (1-3)- and β (1-4)-linked galactose
SPG	<i>Streptococcus pneumoniae β-Galactosidase</i> Hydrolyses non-reducing terminal β (1-4)-linked galactose residues
ABS	<i>Arthrobacter ureafaciens Sialidase</i> Releases α (2-6)- and α (2-8)-linked non-reducing terminal sialic acids (NeuNAc and NeuNGc)
BKF	<i>Bovine kidney α-Fucosidase</i> Releases α (1-2)- and α (1-6)-linked non-reducing terminal fucose residues more efficiently than α (1-3)- and α (1-4)-linked fucose. Used for release of core fucose residues
GUH	<i>Streptococcus pneumoniae Hexosaminidase</i> Recombinantly expressed in <i>Escherichia coli</i> . Releases β -linked GlcNAc but not bisecting GlcNAc β (1-4) Man
NAN1	<i>Recombinant Sialidase</i> Releases α (2-3)-linked non-reducing terminal sialic acids (NeuNAc and NeuNGc)
AMF	<i>Almond Meal α-Fucosidase</i> Releases α (1-3)- and α (1-4)-linked non-reducing terminal fucose residues <i>Does not</i> release core α (1-3)- and α (1-6)-linked fucose
CBG	<i>Coffee Bean α-Galactosidase</i> Hydrolyzes α (1-3)- and α (1-4)-linked terminal galactose residues

This table displays a range of exoglycosidase enzymes which can be used to cleave monosaccharides at specific linkage positions from a larger glycan structure

2. Mobile phase A consists of 20% v/v acetonitrile (Sigma-Aldrich Acetonitrile E CHROMASOLV for HPLC, far UV) in Milli-Q water. Prepare 2 L of this solution by diluting 400 mL of acetonitrile to a final volume of 2 L with Milli-Q water. Ensure sufficient mixing and venting of released gas.
3. Mobile phase B consists of 0.1 M ammonium acetate buffer pH 7.0 in 20% v/v acetonitrile. Prepare 1 L of this solution by diluting 200 mL of acetonitrile and 100 mL of the 1 M ammonium acetate buffer stock solution to 1 L with Milli-Q water. Ensure sufficient mixing and venting of released gas.
4. The standard used for WAX analysis is a 2% v/v fetuin N-glycan solution. To prepare this standard, dilute 2 μ L of the fetuin N-glycan stock standard, stored at -20°C, to a final volume of

100 μL with Milli-Q water. Scale accordingly for performing multiple injections. Aliquot the standard in a 150- μL glass insert in an autosampler vial for analysis.

3. Methods

Carry out all procedures at room temperature unless otherwise stated. Make sure that all water used is Milli-Q water.

3.1. In Gel Block Method

3.1.1. Sample Preparation

1. Using a 96-well microtiter polypropylene plate, add the following to each well:
 - 5 μL of sample (e.g., serum, purified IgG).
 - 2 μL of water.
 - 2 μL of sample buffer.
 - 1 μL of 0.5 M DTT (reductant).
2. Mix gently using pipette action and incubate at 65°C for 15 min.
3. Add 1 μL of 100 mM IAA (alkylating agent) and mix again using pipette action.
4. Incubate for 30 min at room temperature in the dark.

3.1.2. Immobilization of Sample

Add the following ingredients of the gel to each of the wells of the 96-well plate:

- 22.5 μL of Protogel.
- 11.25 μL of gel buffer.
- 1 μL of 10% SDS.
- 1 μL of 10% APS.

1. Mix gently, then add 1 μL of TEMED, then mix gently again. Any bubbles present should disappear.
2. Leave gels to set for approximately 15 min.
3. Once all gels have set, transfer them to a Whatman Protein Precipitation FF plate for washing. Wearing non-powdered gloves (see Note 6), carefully remove the gel from the plate well with the aid of a pipette tip and place it into the appropriate well of the precipitation plate.

3.1.3. N-Glycan Release and Extraction

1. Add 1 mL of acetonitrile to each of the gels, and place the plate on a plate mixer for 10 min. Vacuum eluant to waste. Repeat the procedure with 1 mL of 20 mM NaHCO₃, followed by 1 mL of acetonitrile, 1 mL of 20 mM NaHCO₃, and finally 1 mL of acetonitrile, in that order.

2. Place the protein precipitation plate on top of a deep well collection plate and add 25 µL of 100 mU/ml PNGase F solution to each gel. Incubate the plate at room temperature for 5 min to ensure enzyme uptake by the gel piece, before addition of an additional 25 µL of PNGase F solution. Cover the surface of the gel piece with a further 50 µL of 20 mM NaHCO₃. Seal the plate with a SealPlate adhesive film, ensuring that all edges are firmly secure, and incubate at 37°C overnight.
3. Add 200 µL of deionized water to each gel, and place on a shaker (still on top of the deep well collection block) for 10 min. Vacuum the eluant to a deep well collection block using a vacuum manifold. Repeat in the following order: 200 µL of water, 200 µL of water, 200 µL of acetonitrile, 200 µL of water and 200 µL of acetonitrile. The acetonitrile dehydrates the gel ensuring that all released glycans are eluted. Concentrate eluants to dryness in a centrifugal evaporator.
4. Add 20 µL of 1% formic acid (made up fresh) to each dried sample and incubate at room temperature for 15 min. Dry completely in a centrifugal evaporator. If some samples are to be analyzed by mass spectrometry, retain some sample prior to labeling with a fluorophore.

3.1.4. Fluorescent 2-AB Labeling and Removal of Excess Underivatized Material

1. Add 5 µL of 2-AB labeling mixture to each well, and agitate for 5 min to ensure mixture with the sample. Incubate at 65°C for 2 h.
2. Excess fluorescent label can be sufficiently removed using LudgerClean A glycan cleanup cartridges (Ludger) in a microplate format (see Note 7). In order to equilibrate the cartridges add 1 mL of deionized water followed by 1 mL of 96% acetonitrile.
3. Prepare the sample with the addition of 200 µL of 96% acetonitrile and mix gently either manually by pipette action or if on a robotic platform use the integrated plate shaker. Load each sample onto pre-equilibrated cartridges and apply a gentle vacuum.
4. Add 1 mL of 96% acetonitrile and vacuum off the eluant gently. Repeat this twice to wash away any non-glycan contaminants.
5. Add a deep well collection block below the microplate containing the cartridges to collect eluants. Add 500 µL of deionized water to elute the glycans. Vacuum off very gently and repeat an additional three times to ensure that all of the glycans are eluted.
6. Dry the eluants in a centrifugal evaporator and reconstitute the sample in a known volume prior to HPLC analysis (see Subheading 3.5).

3.2. Robotic Platform Method

3.2.1. Sample Preparation

3.2.2. Immobilization of IgG on Protein A

3.2.3. N-glycan Release

3.2.4. Fluorescent 2-AB Labeling and Removal of Excess Underivatized Material

3.3. 1D SDS-PAGE Method

3.3.1. Sample Preparation

3.3.2. Preparation of Gel Bands for Glycan Extraction

1. Dispense 200 µL of PBS to a protein A plate. Vacuum the eluants to waste and repeat once more.

2. Clarify the fermentation media by centrifuging at 1,000 rpm for 5 min. Discard pellet.

1. Dispense 200 µL of the clarified fermentation media supernatants to the protein A plate and allow to adsorb for 10 min.

2. Wash any unbound immunoglobulin with 200 µL of PBS and vacuum to waste.

1. Release the N-glycans from the immobilized IgG by dispensing 50 µL of 0.25 U/ml PNGase F in 20 mM NaHCO₃ (pH 7.2). Allow the samples to incubate at room temperature for 10 min.

2. Incubate the protein A plate at 37°C for 60 min in the integrated incubator with the aid of the robotic gripper tools.

3. Release the N-glycans by dispensing 100 µL of Milli-Q water and collect the eluants in a collection block placed in the integrated vacuum manifold. Repeat this process four times.

4. Concentrate the released N-glycans in a centrifugal evaporator (Thermo, Basingstoke, Hampshire, UK).

5. Dispense 20 µL of 1% formic acid (made up fresh) to each dried sample and incubate at room temperature for 15 min. Dry once again in a centrifugal evaporator.

1. Follow procedure as detailed in Subheading [3.1.4](#).

1. 5–10 µg of sample is loaded to each gel lane. The sample is reduced and alkylated prior to SDS-PAGE separation to ensure maximum release of glycans by PNGase F. Add 4 µL of 5× Laemmli sample buffer, 2 µL of 0.5 M DTT, and make to a total volume of 20 µL with water. Incubate at 70°C for 10 min. For non-reduced sample incubate without DTT and do not alkylate.

2. Alkylation of reduced sample: Add 2 µL of 100 mM IAA to the reduced sample and incubate in the dark for 30 min at room temperature.

1. These instructions assume the use of XCell SureLock Mini-Cell apparatus for SDS-PAGE with freshly prepared gels (80×80×1 mm).

2. For one 10% gel, mix 1.5 mL of resolving gel buffer, 2 mL of protogel, 2.5 mL of water, and 60 µL of 10% SDS. Add 60 µL

of 10% APS and 6 μ L of TEMED to the mix immediately prior to pouring the gel. Allow space for stacking gel, gently overlay with water-saturated butanol, and leave to set for approximately 15–20 min.

3. Prepare the stacking gel by mixing 625 μ L of stacking gel buffer, 333 μ L of protogel, 1.525 mL of water, and 25 μ L of 10% SDS. Add 25 μ L of 10% APS and 2.5 μ L of TEMED. Mix well and fill the top of the cassette with stacking gel. Insert the comb immediately without introducing air bubbles. Leave to set for 15–20 min.
4. Carefully remove the combs and rinse the wells three times with running buffer. Peel off the tape from the bottom of the gel cassette.
5. Load the samples into the wells of the gel using gel loading pipette tips and load 5 μ L of molecular weight markers.
6. Assemble the unit according to the manufacturer's instructions with a magnetic stirrer at the bottom. Fill the inner compartment with SDS-PAGE running buffer and then fill the outer compartment with the remaining running buffer until it is about three quarters full. Electrophoresis at 25 mA per gel until the dye front has reached the bottom of the cassette.
7. Following electrophoresis, open the cassette with a spatula. The gel will remain on one side of the cassette. Gently drop the gel into a plastic box containing sufficient Coomassie stain to cover the gel. Leave to stain on a platform shaker for 2 h or alternatively leave overnight.
8. Tip out the Coomassie stain and cover with destain 1 for 5 min on a shaking platform.
9. Tip out destain 1 and replace with destain 2 and leave on a platform shaker for several hours or overnight until the gel has been sufficiently destained (see Note 8).
10. Photograph the gel.
11. On a clean glass plate over a light box, cut out the Coomassie-stained bands from the gel with a clean scalpel. Cut each gel band into approximately 1–3-mm pieces and transfer to a 1.5-mL Eppendorf tube. Freeze for at least 30 min or overnight (see Note 9).
12. Wash the gel pieces thoroughly with the addition of 1 mL of acetonitrile, vortex, and then mix on a roller mixer for 30 min at room temperature. Pipette off and discard the liquid. Repeat the procedure with 1 mL of 20 mM NaHCO₃, followed by 1 mL of acetonitrile, 1 mL of 20 mM NaHCO₃, and finally 1 mL of acetonitrile, in that order.
13. Dry the gel pieces in a vacuum centrifuge.

3.3.3. Fluorescent 2-AB Labeling and Removal of Excess Underderivatized Material

1. Add 20 µL of 1% formic acid to the dried sample and incubate at room temperature for 15 min.
2. Dry the sample in a speed vacuum.
3. Add 5 µL of 2-AB labeling mixture to each well, and agitate for 5 min, to ensure mixture with sample. Incubate at 65°C for 2 h.
4. Using a 1.5-mL Eppendorf tube, bring the sample to a volume of 1 mL so that the final composition is 90% acetonitrile.
5. Condition the Phytip with 3×500 µL in-out cycles of 95% acetonitrile.
6. Follow this with 3×500 µL aspirate–dispense cycles of 20% acetonitrile.
7. Perform one final wash with 3×500 µL of 95% acetonitrile. The tips are now ready for sample loading.
8. Apply the 1 mL sample to the tip using at least ten in-out cycles.
9. To remove the excess 2-AB, wash the chromatographic bed of the tip with 10×1 mL of 95% acetonitrile.
10. Elute the retained N-glycans using 5×200 µL aspirate–dispense cycles of 20% acetonitrile.
11. Dry the sample in a speed vacuum. The sample is now ready for analysis using HPLC.

3.4. In Solution Method

3.4.1. Sample Preparation

Using a 1-mL Eppendorf tube:

1. Add water to the sample (e.g., IgG) to give a final volume of 200 µL.
2. Add 5 µL of 1 M DTT.
3. Incubate for 10 min at 80°C.
4. Add 10 µL of 100 mM IAA and keep the sample in the dark.
5. Incubate for 10 min at room temperature.

3.4.2. N-Glycan Release and PNGase F Removal

1. Add 260 µL of PNGase F buffer.
2. Add 25 µL of PNGase F.
3. Incubate the sample overnight at 37°C.
4. Wash the PALL membrane with 200 µL of water.
5. Centrifuge for 5 min at 14,000×*g*.
6. Discard the filtrate.
7. Add the sample (500 µL) to the PALL membrane.
8. Centrifuge for 5 min at 14,000×*g*.
9. Wash the sample tube with 100 µL of water and add to the PALL membrane.

10. Centrifuge for 5 min at 14,000 $\times g$.
11. Discard the filter and dry the sample in a speed vacuum.

3.4.3. Fluorescent 2-AB

*Labeling and Removal
of Excess Underivatized
Material*

3.5. HILIC-HPLC Profiling of 2-AB Labeling N-Glycans

1. Any HPLC system which has a fluorescence detector can be used. Set the excitation and emission wavelengths of the fluorescence detector to 330 and 420 nm, respectively, with a bandwidth of 16 nm, set at maximum sensitivity.
2. Samples are injected in 80% acetonitrile. Take an aliquot of the sample (or standard dextran ladder) and make up to 20 μ L with water; then add 80 μ L of acetonitrile. It is a good idea to use only a small percentage of your sample in the first run in order to get some idea of how much must be loaded to produce a good chromatogram.
3. HILIC-HPLC running buffer is made by diluting 50 mL of HILIC stock solution to 2 L with water (solvent A). Solvent B is acetonitrile. Gradient conditions are outlined in Table 2.
4. Run a dextran ladder standard followed by the samples, and set the injection volume to 95 μ L. Ensure that a dextran standard is run with each batch of samples.
5. Calibration and allocation of GU: The dextran ladder is used to calibrate the HPLC runs against any day-to-day or system-to-system changes. The GU value is calculated by fitting a

Table 2
Gradient method for HILIC-HPLC analysis

HILIC-HPLC gradient method

Time (min)	Flow (mL/min)	% Solvent A	% Solvent B
0	0.48	35	65
48	0.48	47	53
49	0.48	100	0
53	0.48	100	0
54	0.48	35	65
60	0.48	35	65

HILIC gradient method used for the separation of glycans with a runtime of 60 min. Solvent A is ammonium formate, pH 4.4, and Solvent B is acetonitrile

fifth-order polynomial distribution curve to the dextran ladder (usually glucose 1–15), and then using this curve to allocate GU values from retention times (Empower GPC software from Waters can be used to calculate GU values). This facilitates direct comparison with database values (11).

3.6. N-Glycan Exoglycosidase Digestion and Cleanup

1. Concentrate 5 µL aliquots of the 2-AB-labeled glycans in a centrifugal evaporator.
2. Add 1 µL of 10× buffer (50 mM sodium acetate) pH 5.5, the required enzymes, and make up to a final volume of 10 µL with Milli-Q water. Incubate samples overnight at 37°C.
3. Pre-equilibrate the 96-well enzyme removal plates (10 kDa, PALL Corporation, Ann Arbor, USA) with the addition of 200 µL Milli-Q water. Vacuum eluant to waste. If a smaller number of samples are being analyzed then individual protein binding spin filters are also available from PALL Corporation.
4. Insert a collection plate in the base of the vacuum manifold to collect the eluants. Add the digested samples to the various wells of the enzyme removal plate and allow to adsorb. Ensure that all digested samples are transferred to the plate by washing the digest container with 20 µL of Milli-Q water. Vacuum the samples to the collection block. Elute the glycans with 100 µL of Milli-Q water and again vacuum to the collection block.
5. Dry the eluants in a centrifugal evaporator and reconstitute the samples in 20 µL of Milli-Q water. Prepare the samples for HPLC as indicated in Heading 3.5.

3.7. Weak Anion Exchange

This analysis assumes performance using a Waters Alliance 2695 separations module complete with a Waters 474 or 2475 fluorescence detector or equivalent under the control of Empower Chromatography Workstation. The analytical column used is a Prozyme GlycoSep C polymeric DEAE anion exchange column, 75 × 7.5 mm i.d., 10-µm particle size.

1. For sample analysis take a 2 µL aliquot of sample made up to a final volume of 100 µL with Milli-Q water. Place the sample in an autosampler vial for instrumental analysis. For sample pre-fractionation use 10–20 µL of sample made up to a final volume of 100 µL with Milli-Q water. The quantity of sample used for pre-fractionation can be scaled accordingly based upon the quantity of sample used for your provisional HILIC analysis. Gradient conditions are outlined in Table 3.
2. Compare the elution positions of peaks in the sample to those of the fetuin standard. The larger sialylated triantennary glycans elute prior to the sialylated biantennary glycans with the

Table 3
Gradient conditions for WAX fractionation

WAX method

Time (min)	Flow (mL/min)	% Solvent A	% Solvent B
0	0.75	100	0
5	0.75	100	0
40	0.75	0	100
42.50	0.75	0	100
43.00	0.75	100	0
50.00	0.75	100	0

Weak anion exchange (WAX) gradient method used for the separation of mono-, di-, tri-, and tetrasialylated glycans with a runtime of 50 min. Mobile phase A consists of 20% v/v acetonitrile. Mobile phase B consists of 0.1 M ammonium acetate buffer pH 7.0 in 20% v/v acetonitrile

same charge (see Fig. 7). If excess 2-AB label has not been sufficiently removed, a large curve in the baseline can occur which can obscure the glycan peaks.

3.8. Structural Assignment of Glycan Pool

1. The GU value of the glycan is directly related to the number and linkage of its monosaccharide components, the higher the glucose unit value the larger the N-glycan. Glycan structures can be predicted using the GU values as each value represents a monosaccharide in a particular linkage and adds a given amount to the structure. This information is outlined in Table 4.
2. Preliminary assignments of the undigested glycan pool can be made using GlycoBase, a repository for glycan structures which denotes GU values for glycan structures. This requires further confirmation by dissecting the glycan structures into their individual monosaccharide components via exoglycosidase digestions.
3. The fluorescence intensity on the chromatogram is directly related to the number of moles of labeled glycans present. As a result quantitation of the amount of glycans in the sample can be calculated by measuring the areas of the resolved peaks. The relative percentage areas of different glycans can then be directly compared within a glycan pool.

Table 4
**Incremental values for specific monosaccharides
 and linkages**

Monosaccharide	Linkage	To	GU increment
Mannose	α 1-2,3,6	Mannose	0.7–0.9
GlcNAc	β 1-2,4,6	α -Mannose	0.5
GlcNAc (Bisect)	β 1-4	β -Mannose	0.2–0.4
Galactose	α or β 1-3,4	GlcNAc or Gal	0.8–0.9
Core fucose	α 1-6	Core GlcNAc	0.5
Outer arm fucose	α 1-3,4	GlcNAc	0.8
Outer arm fucose	α 1-2	Gal	0.5
Neu5Ac	α 2-3,6	Gal	0.7–1.2

The GU value for a glycan is directly related to the number and linkage of its constituent monosaccharides; the larger the glycan, the higher its GU value

4. Notes

1. All reagent water used in these experiments was obtained from a Milli-Q Gradient A10 Elix system (Millipore, Bedford, MA, USA) and was $18.2\text{ M}\Omega$ or greater with a total organic carbon (TOC) content less than 5 parts per billion (ppb).
2. All solutions are stored at room temperature unless otherwise stated.
3. Purchasing small quantities of TEMED is recommended as it may decline in quality after opening which affects the length of time needed for polymerization of the gel.
4. Prepare 5× running buffer (0.25 M Tris, 1.92 M glycine, 1% SDS). Weigh 30.3 g of Tris and 144 g of glycine, mix, and dissolve in 1.8 L of water. Add 10 g of SDS and mix. Add water to a final volume of 2 L.
5. 96-Well plates from PALL may be utilized here as an alternative depending on sample number as well as sample volume.
6. Non-powdered gloves should be used at all stages of experimentation as powdered gloves may cause a contaminating polysaccharide ladder which will obscure sample analysis.
7. Alternatively, HyperSep-96 Diol cartridges (10 mg, Thermo Scientific) can be used to remove excess 2-AB. Phytips (Phynexus) can also be used, however, for 96 samples, this is more labour intensive.

8. Add a small piece of clean polyurethane foam to destain 2, as this greatly enhances the destain process.
9. Freezing gel pieces helps to break down the matrix so that more of the gel pieces are accessible to PNGase F.

References

1. Varki A, Cummings RD, Esko JD, Freeze H, Hart G, Marth J (2009) Essentials of glycobiology. Cold Spring Harbour Laboratory, Cold Spring Harbour, NY
2. Alavi A, Axford JS (2008) Sweet and sour: the impact of sugars on disease. *Rheumatology (Oxford)* 47:760–770
3. Rodriguez J et al. (2010) High productivity of human recombinant beta-interferon from a low-temperature perfusion culture. *J Biotechnol* 150:509–518
4. Chung CH et al. (2008) Cetuximab-induced anaphylaxis and IgE specific for galactose-alpha-1,3-galactose. *N Engl J Med* 358:1109–1117
5. Pacis E et al. (2011) Effects of cell culture conditions on antibody N-linked glycosylation—what affects high mannose 5 glycoform. *Biotechnol Bioeng* 108(10):2348–2358
6. Wong NS et al. (2010) An investigation of intracellular glycosylation activities in CHO cells: effects of nucleotide sugar precursor feeding. *Biotechnol Bioeng* 107:321–336
7. Ferrara C et al. (2006) The carbohydrate at Fc gamma RIIIa Asn-162. An element required for high affinity binding to non-fucosylated IgG glycoforms. *J Biol Chem* 281: 5032–5036
8. Royle L et al. (2006) Detailed structural analysis of N-glycans released from glycoproteins in SDS-PAGE gel bands using HPLC combined with exoglycosidase array digestions. *Methods Mol Biol* 347:125–143
9. Gabius H-J (2009) The sugar code: fundamentals of glycosciences. Blackwell, London
10. Royle L et al. (2008) HPLC-based analysis of serum N-glycans on a 96-well plate platform with dedicated database software. *Anal Biochem* 376:1–12
11. Campbell MP et al. (2008) GlycoBase and autoGU: tools for HPLC-based glycan analysis. *Bioinformatics* 24:1214–1216
12. Marino K et al. (2010) A systematic approach to protein glycosylation analysis: a path through the maze. *Nat Chem Biol* 6:713–723

Chapter 20

High-Throughput Multimodal Strong Anion Exchange Purification and N-Glycan Characterization of Endogenous Glycoprotein Expressed in Glycoengineered *Pichia pastoris*

Sujatha Gomathinayagam, Erik Hoyt, Alissa M. Thompson, Eric Brown, Khanita Karaveg, Stephen R. Hamilton, and Huijuan Li

Abstract

The secretory pathway of the yeast *Pichia pastoris* has been engineered to produce complex human-type N-glycans (Choi et al., Proc Natl Acad Sci USA 100:5022–5027, 2003; Hamilton et al., Science 301:1244–1246, 2003; Hamilton et al., Science 313:1441–1443, 2006). In contrast to the heterogeneous glycans produced on the therapeutic glycoproteins expressed in mammalian cell lines, glycoengineered *P. pastoris* can be designed to produce a specific, preselected glycoform. In order to achieve glycan uniformity on the target protein, No Open Reading Frame (NORF) yeast cell lines are screened extensively during various stages of glycoengineering. In the absence of the target protein of interest, screening the NORF yeast cell lines for glycoform uniformity becomes a challenge. The common approach so far has been to analyze the total cell glycan pool released from glycoproteins of the NORF yeast cells to predict the N-glycan uniformity. As this does not always accurately predict the N-glycan end product, we describe in this chapter a detailed protocol for a non-affinity-based high-throughput purification of an endogenous glycoprotein. This protein of interest has been introduced during the early stages of glycoengineering process and its N-glycan profile is utilized as a tool for glycoengineering screening.

Key words: High-throughput purification, Non-affinity protein purification, Multimodal strong anion exchange chromatography, CattoTM adhere, N-glycosylation, *Pichia pastoris*

1. Introduction

The methylotrophic yeast *Pichia pastoris*, as a single-celled microorganism, is not only easy to manipulate and culture but also capable of many of the posttranslational modifications performed by higher eukaryotic cells, including proteolytic processing, folding, disulfide bond formation, and glycosylation (1). Thus, many proteins that

end up as inactive inclusion bodies in bacterial systems are expressed as biologically active molecules in *P. pastoris* (1). However, the N-linked glycan structures on *P. pastoris* are often referred to as hyperglycosyl- or hypermannosyl-type structures which are typically nonhuman glycoforms (2). In addition to our laboratory, several other labs have reported work on engineering glycosylation pathways in *P. pastoris* (3, 4). We have successfully engineered the N-linked glycosylation pathway of *P. pastoris* by deleting several yeast-specific glycosylation pathway genes combined with the introduction of 14 heterologous genes, which allows for the production of complex terminally sialylated human glycoproteins (5–7).

One of the main advantages of glycoengineered *P. pastoris* is that it can successfully produce recombinant target proteins with highly uniform N-linked glycosylation profiles. However, during different stages of engineering of the secretory pathway in *P. pastoris*, the cell lines need to be screened for glycan uniformity. In the absence of a target protein, glycan screening up to now relied on analyzing glycans isolated from a total cell glycoprotein pool. Since this involves lysis of the cells, the intermediate glycans in the secretory pathway have the potential to be released and significantly impact the N-linked glycosylation profile. Cellular glycans released by highly stringent salt concentrations at high temperatures can also significantly affect the profile of the released N-linked glycans. Thus, as an alternative path for glycoengineering analysis, we designed a high-throughput non-affinity purification of an endogenous glycoprotein which had been engineered into *P. pastoris* during glycoengineering as a pseudo reporter protein for screening. The screening was based on high-throughput purification employing CaptoTM adhere media followed by enzymatic N-glycan release and matrix-assisted laser desorption ionization time of flight (MALDI-TOF) analysis.

High-throughput purification of affinity-tagged (hexahistidine tag (His-tag) or glutathione S-transferase (GST)-tag) proteins has been well established (8, 9). A fully automated, robust, and cost-effective method has been developed for the purification of affinity-tagged proteins that can be used to quickly characterize expression clones in a high throughput manner (10). But the purification of non-tagged proteins introduces a greater challenge because of the host cell protein contaminants that could be co-purified. In this chapter, we describe a high-throughput purification protocol employing CaptoTM adhere media in a flow-through mode as a primary capture step for a non-tagged endogenous glycoprotein from *P. pastoris* for glycoengineering screening. CaptoTM adhere is a strong anion exchanger with the multimodal functionality ligand, *N*-benzyl-*N*-methyl ethanol amine, which exhibits ionic interaction, hydrogen bonding, and hydrophobic interactions. It has proved to improve yield, productivity, and process economy with high capacity in a flow-through mode. CaptoTM adhere has been

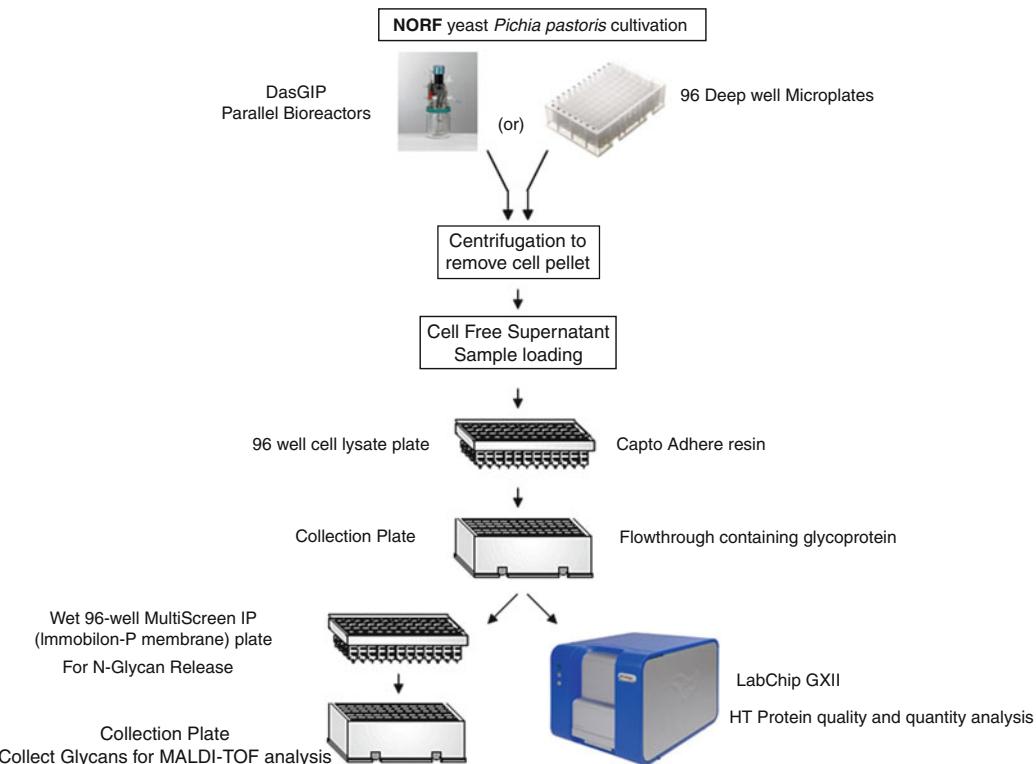


Fig. 1. Schematic representation of high-throughput purification employing multimodal strong anion exchange chromatographic step and subsequent N-glycan screening.

designed primarily for post-Protein A purification of monoclonal antibodies in a flow-through mode, where the antibodies pass directly through the column while the contaminants are adsorbed. The main advantage of Capto™ adhere media is the wide operational window of pH and conductivity. The medium is based on a rigid high flow agarose matrix that allows high flow velocities to be used. The highly cross-linked agarose base matrix gives the medium high chemical and physical stability. Employing Capto™ adhere media in a flow-through mode, a high-throughput purification and subsequent N-glycan characterization (Fig. 1) were designed for glycan screening of No Open Reading Frame (NORF) yeast cell lines to screen for glycan uniformity.

2. Materials

2.1. Non-affinity Protein Purification

1. Equilibration buffer: 50 mM MOPS (pH 7.0) (see Note 1).
2. Wash buffer 1: 50 mM MOPS, 1 M NaCl (pH 7.0).
3. Wash buffer 2: 1 M Tris-HCl (pH 8.0).

4. Strip buffer: 1 M Sodium hydroxide.
5. Storage buffer: 20% Ethanol.
6. Capto™ adhere media Cat. No. 17-5444-01 (GE healthcare, Piscataway, NJ).
7. 96-Well lysate clearing plate (Wizard SV96, Promega Corp, Madison, WI).
8. Water, high-performance liquid chromatography (HPLC) grade.
9. 96-Well collecting plates.
10. Beckman Biomek® FX robot (Beckman Coulter, Fulerton, CA).

2.2. N-Glycan Release and Analysis

1. Centrifugal evaporator (Thermo Savant, Holbrook, NY).
2. N-glycosidase F (New England BioLabs, Beverly, MA).
3. Voyager DE PRO linear MALDI-TOF (Applied Biosystems, Foster City, CA).

2.3. LabChip GXII Analysis

1. LabChip GXII (Caliper Life Sciences, Hopkinton, MA).
2. All buffers for LabChip GXII analysis were from LabChip Caliper Kit.
3. 250 mM Iodoacetic acid (IAA) (0.046 g of IAA, 1 mL of distilled water).
4. 500 mM Dithiothreitol (DTT) (0.077 g of DTT, 1 mL of distilled water).
5. Nonreducing solution: 700 µl of sample buffer (Caliper pre-made buffer) and 70 µl of 250 mM IAA.
6. Reducing solution: 700 µl of sample buffer (Caliper premade buffer) and 70 µl of 500 mM DTT.

3. Methods

3.1. Non-affinity Purification

Purify endogenous glycoprotein from the cell-free supernatant medium by multimodal strong anion exchange chromatography utilizing a 96-well format on a Beckman Biomek® FX robot. The robotic purification is a scale down model of the protocol optimized using an AKTA Explorer system (see Note 2).

Robotic protocol

1. Prepare a 96-well purification plate by manually transferring 400 µl of Capto™ adhere media to each well to yield 200 µl of packed media and place the plates on the deck position of the Biomek® FX robot as shown in Fig. 2.
2. Wash three times with 200 µl of water.

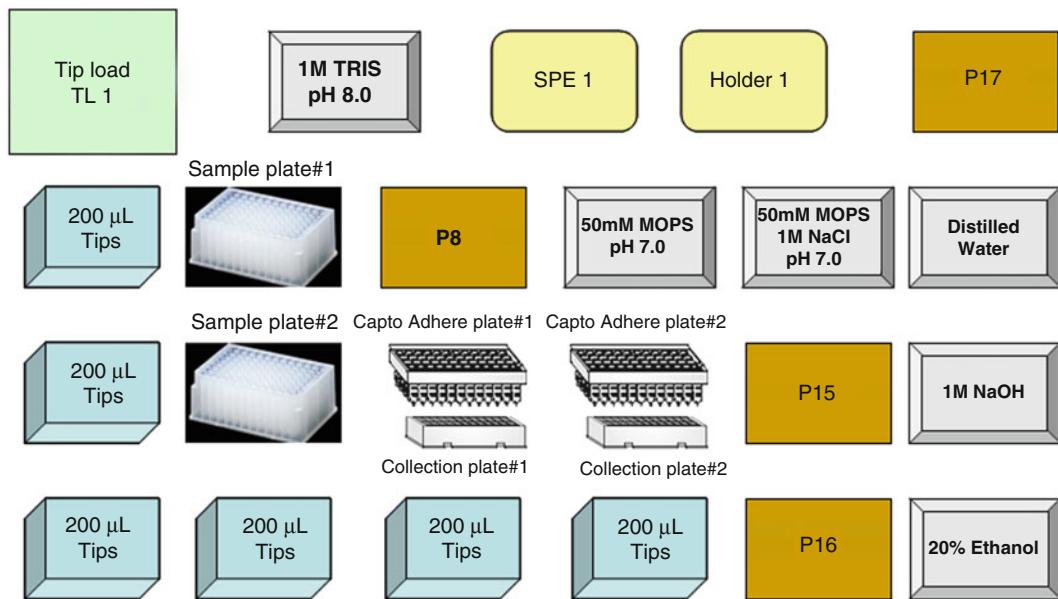


Fig. 2. Deck configuration of the Biomek® FX for high-throughput purification employing multimodal strong anion exchange chromatography. Two 96-well plate purification setup.

3. Wash five times with 200 µl of equilibration buffer.
4. Load samples five times with 200 µl of cell-free culture supernatant (see Note 3).
5. Collect the flow-through sample in the collection plate and aliquot 25 µl on a separate collection plate for LabChip GXII analysis (see Subheading 3.3). Evaporate the remaining sample to dryness in a centrifugal evaporator (Thermo Savant, Holbrook, NY) for N-glycan release (see Subheading 3.2).
6. Wash three times with 200 µl of equilibration buffer (see Note 4).
7. Wash three times with 200 µl of wash buffer 1.
8. Strip resin four times with 200 µl of strip buffer.
9. Wash three times with 200 µl of wash buffer 2.
10. Wash three times with 200 µl of water.
11. Store the resin in 20% ethanol at 4°C.
12. This resin can be reused approximately for ten sets of purification, without loss of efficiency.

3.2. N-Linked Glycan Release and MALDI-TOF Mass Spectrometry

Release the N-linked glycans from the glycoprotein using previously reported methods by Papac et al. (11) and Li et al. (12). For the MALDI-TOF analysis, use a Voyager DE PRO linear MALDI-TOF (Applied Biosystems, Foster City, CA) and generate spectra with the instrument in the positive and/or negative ion mode for neutral and/or charged glycans, respectively (see Note 5).

MALDI-TOF spectra of N-linked glycans released from high-throughput purified endogenous glycoprotein from different

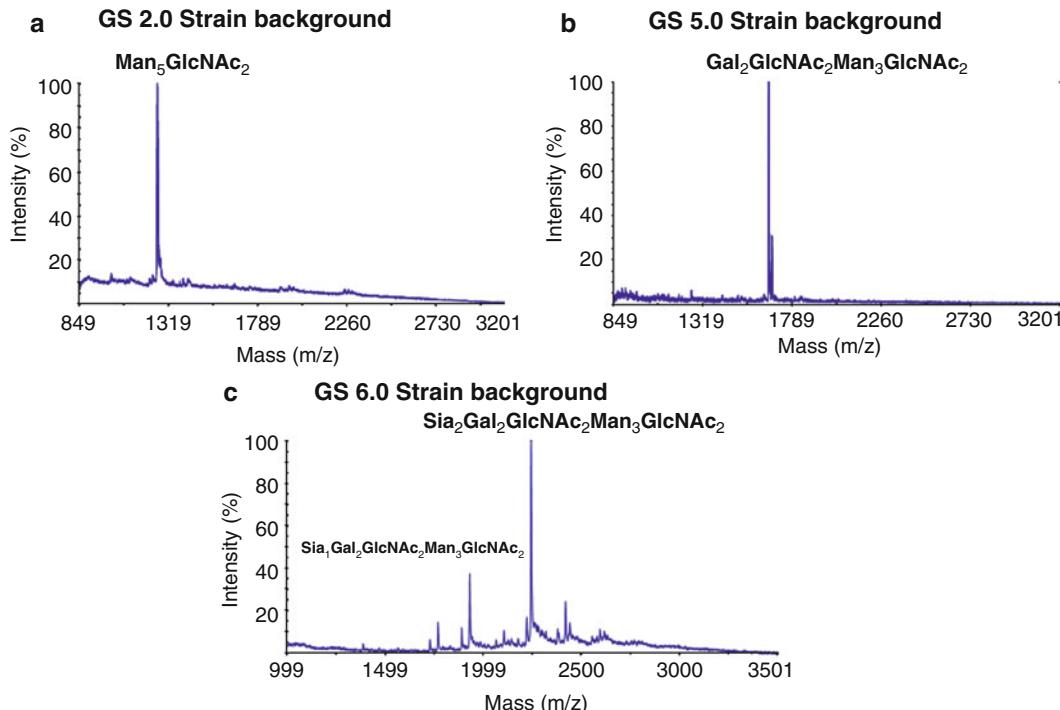


Fig. 3. MALDI-TOF analysis of NORF yeast *P. pastoris* strains. (a, b) Positive-ion MALDI-TOF mass spectra of N-linked glycans with neutral glycoforms. (c) Negative-ion MALDI-TOF mass spectra of N-linked glycans with terminally sialylated (charged) glycoforms.

P. pastoris NORF strains with various glycosylation machinery are illustrated in Fig. 3.

3.3. High-Throughput Protein Quality and Quantity Analysis Using LabChip GXII

The LabChip GXII platform (Caliper Life Sciences, Hopkinton, MA), a high-throughput CE-based analytical technique, was used to determine the purity and quantity of the purified endogenous glycoprotein. The microfluidics technology-based LabChip GXII eliminates the need to handle SDS-PAGE gels. This high-throughput screen not only allows to quickly screen for protein quality, but also ensures protein quantity which is critical for N-glycan characterization. This instrument has a sample acquisition time under 40 s for each sample and thus can analyze 90 samples in an hour, which eliminates throughput bottlenecks to improve efficiency. The LabChip GX software version 3.0.618.0 was used for data analysis. LabChip GXII protocol for high-throughput purified samples

1. Prepare the dye and destain by adding 520 μ l of gel to a spin column for destain and 1,040 μ l of gel and 38 μ l of dye to a 5-mL tube, and vortex. Add 540 μ l to two spin columns. Centrifuge at 9.2 rcf for 5 min.
2. Prepare samples by adding 7 μ l of either nonreducing or reducing solution to each well and 2 μ l of sample. Incubate samples

at 72°C for 15 min. Add 35 µl of water to each well, and centrifuge at 835 ×*g* for 2 min.

3. Prepare ladder by adding 15 µl of ladder to a PCR tube and incubate at 99°C for 5 min. Add 135 µl of water and move to a caliper ladder tube.
4. Prepare Chip by washing all active wells with water. Add prepared buffers to appropriate wells (see below).

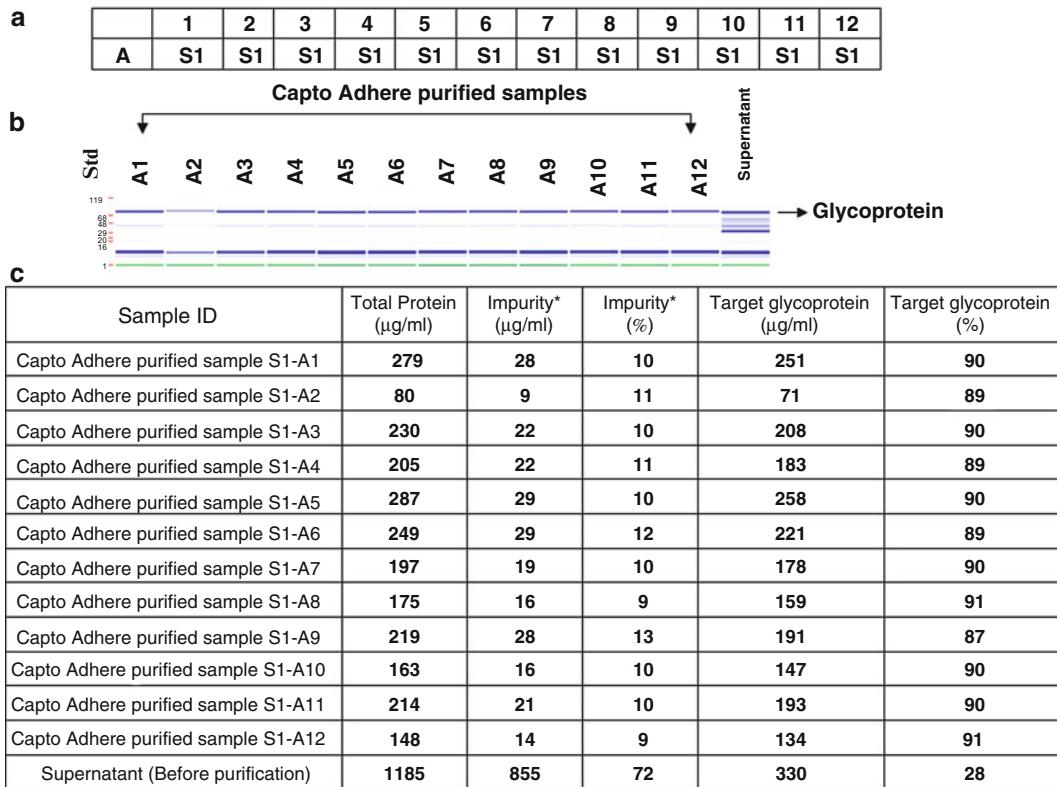
Waste	Detection window	Gel dye 120 µl
Destain 75 µl		Destain 75 µl
Gel dye 75 µl		Gel dye 75 µl
Lower marker 120 µl		Gel dye 75 µl

5. Add 750 µl of wash buffer to the wash buffer vial.
6. Add the chip, ladder vial, wash buffer vial, and sample plate to the machine. Run Protein 200 method, choosing either 384-well or 96-well plate sipping middle.
7. Once processed, label samples with names and analyze data.

High-throughput analysis of the purified glycoprotein using LabChip GXII is exemplified in Fig. 4. A step recovery of 90% can be achieved for the target glycoprotein using this high-throughput purification method with a significant percentage (62%) reduction of host cell protein impurities (see Note 6). Since the non-glycosylated proteins present in the host cell protein impurities are a major concern for glycosylation analysis, removing these contaminant proteins and at the same time enriching for the target glycoprotein provides a useful tool for glycosylation screening of NORF yeast cell lines for glycan uniformity.

4. Notes

1. Stock solutions of 1 M MOPS pH 7.0, 5 M sodium chloride, and 1 M Tris–HCl pH 8.0 are used to make the reagents.
2. Several other mixed mode resins were evaluated in the initial scouting experiments and based on the experimental results Capto™ adhere was chosen. The main advantages were that most of the host cell proteins bind to Capto™ adhere resin, leaving the protein of interest in the flow through, and that the supernatant samples can be loaded directly on to the column without any sample preparation or pH adjustments.
3. The pH of the culture supernatant should be between 6.5 and 7.0 in order to ensure optimum binding of the endogenous glycoprotein to the Capto™ adhere media.



* Impurity - Host cell proteins other than the target glycoprotein

Fig. 4. High-throughput analysis of the purified glycoprotein using LabChip GXII. (a) A 96-well plate layout for the high-throughput purification of the supernatant from the NORF yeast cell lines that are capable of expressing glycoproteins with terminally sialylated N-linked glycan structures. (b) High-throughput SDS-PAGE analysis of purified glycoproteins using LabChip GXII. (c) High-throughput protein quantification of the target glycoprotein present in the purified samples.

4. Even though the Capto™ adhere medium is based on a rigid high flow agarose matrix that allows high flow velocities, when using this media for high-throughput Biomek® FX purification sufficient time needs to be built in the protocol so that sample and the buffers can flow through the media completely.
5. High concentrations of salt and media components in the flow-through sample did not interfere with PNGaseF digest on the 96-well plate. Still, in order to remove the salt, extensive washing of the membranes with HPLC-grade water is necessary before incubating the protein with the enzyme in volatile buffer of 10 mM NH_4HCO_3 .
6. LabChip GXII analysis shows that the high-throughput purification designed here employing Beckman Biomek® FX robot can produce reliable and reproducible data with minimal errors when the same sample was purified through multiple wells.

References

1. Higgins DR, Cregg JM (1998) Introduction to *Pichia pastoris*. Methods in Molecular Biology: *Pichia* Protocols, Volume 103, Humana Press, Totowa, NJ
2. Brethauer RK, Castellino FJ (1999) Glycosylation of *Pichia*-derived proteins. Biotechnol Appl Biochem 30:193–200
3. Koji M, Narutoshi S, Tomoyasu R (1996) Sugar chain-elongating protein and DNA derived from yeast *Pichia* genus. Jpn Patent JP 8336387
4. Callewaert N, Laroy W, Cadirgi H et al (2001) Use of HDEL-tagged *Trichoderma reesei* mannosyl oligosaccharide 1,2- α -mannosidase for N-glycan engineering in *Pichia pastoris*. FEBS Lett 503:173–178
5. Choi BK, Bobrowicz P, Davidson RC et al (2003) Use of combinatorial genetic libraries to humanize N-linked glycosylation in the yeast *Pichia pastoris*. Proc Natl Acad Sci USA 100:5022–5027
6. Hamilton SR, Bobrowicz P, Bobrowicz B et al (2003) Production of complex human glycoproteins in yeast. Science 301:1244–1246
7. Hamilton SR, Davidson RC, Sethuraman N et al (2006) Humanization of yeast to produce complex terminally sialylated glycoproteins. Science 313:1441–1443
8. Lin CT, Moore PA, Auberry DL et al (2006) Automated purification of recombinant proteins: combining high-throughput with high yield. Protein Expr Purif 47:16–24
9. Lin CT, Moore PA, Kery V (2009) Automated 96-well purification of hexahistidine-tagged recombinant proteins on MagneHis Ni(2)+-particles. Methods Mol Biol 498: 129–141
10. Scheich C, Sievert V, Büssow K (2003) An automated method for high-throughput protein purification applied to a comparison of His-tag and GST-tag affinity chromatography. BMC Biotechnol 3:12
11. Papac DI, Briggs JB, Chin ET et al (1998) A high-throughput microscale method to release N-linked oligosaccharides from glycoproteins for matrix-assisted laser desorption/ionization time-of flight mass spectrometric analysis. Glycobiology 8:445–454
12. Li H, Miele RG, Mitchell TI et al (2003) N-linked glycan characterization of heterologous proteins. Methods Mol Biol 389:139–149

Chapter 21

Databases and Tools in Glycobiology

**Natalia V. Artemenko, Andrew G. McDonald, Gavin P. Davey,
and Pauline M. Rudd**

Abstract

Glycans are crucial to the functioning of multicellular organisms. They may also play a role as mediators between host and parasite or symbiont. As many proteins (>50%) are posttranslationally modified by glycosylation, this mechanism is considered to be the most widespread posttranslational modification in eukaryotes. These surface modifications alter and regulate structure and biological activities/functions of proteins/biomolecules as they are largely involved in the recognition process of the appropriate structure in order to bind to the target cells. Consequently, the recognition of glycans on cellular surfaces plays a crucial role in the promotion or inhibition of various diseases and, therefore, glycosylation itself is considered to be a critical protein quality control attribute for commercial therapeutics, which is one of the fastest growing segments in the pharmaceutical industry.

With the development of glycobiology as a separate discipline, a number of databases and tools became available in a similar way to other well-established “omics.” Alleviating the recognized shortcomings of the available tools for data storage and retrieval is one of the highest priorities of the international glycoinformatics community. In the last decade, major efforts have been made, by leading scientific groups, towards the integration of a number of major databases and tools into a single portal, which would act as a centralized data repository for glycomics, equipped with a number of comprehensive analytical tools for data systematization, analysis, and comparison. This chapter provides an overview of the most important carbohydrate-related databases and glycoinformatic tools.

Key words: Glycoinformatics, carbohydrate structure databases, glycoinformatic tools, Carbohydrates, High-throughput analysis

1. Introduction

With recent advances in glycoscience the scientific community is entering a new era in glycobiology and glycoinformatics, where carbohydrates and glycoconjugates are becoming highly important in therapeutics development alongside standard targets, such as proteins and nucleotide-based compounds (1, 2). In 2003, MIT’s Technology Review nominated glycomics as one of the ten

emerging technologies that will change the world (3). Today, glycans are recognized as the third major class of biomolecules and the largest class of posttranslational modification of proteins, whose content is highly dependent on cell type and physiological environment.

The therapeutic potential of glycoproteins and polysaccharides, together with the growing understanding of glycan structure and functionality, is set to change the drug development industry by using the benefits of engineered glycosylation (1, 4), for example, in monoclonal antibodies or bacterial and viral vaccines. In cell cultures, the potential formation of immunogenic glycoforms, containing nonhuman glycans, has a major impact on clinical efficacy and safety. Non-fucosylated glycoforms of IgG1 human Fc γ RIIIA showed a 50-fold increase in their efficacy in comparison with fucosylated analogues (5–7). Thus, glycosylation is a critical quality control attribute for commercial therapeutics, which is one of the fastest growing segments in the pharmaceutical industry. Regulatory agencies around the world develop and regularly update guidelines and requirements for high-quality biosimilar medical products. These requirements aim to reduce risks in biopharmaceutical production by controlling glycosylation (ICH Q6B and ICH Q5E) within acceptable limits (8, 9). Of the top 15 US pharmaceutical products for 2008 and 2009, 4 were protein-based drugs (10, 11). These were Enbrel and Remicade, which are fusion proteins used to treat autoimmune diseases; Epoprostenol, which is a cytokine (erythropoietin), for the treatment of anemia associated with chronic renal disease, and as it possesses an erythropoiesis-stimulating activity, it can be used as a performance-enhancing drug (12, 13); and Neulasta used as an immunostimulator in patients undergoing chemotherapy. During 2010 Enbrel and Remicade sales values increased and the products moved up in the top 15 (14). According to new research reports on analysis of the global protein therapeutics market (15), this market is expected to grow at a Compound Annual Growth Rate (CAGR) of around 12% during the next 2 years.

This chapter explores the bioinformatics resources available within this important field.

2. Overview of and Need for Glycoinformatics Resources

The search for biologically relevant structure–function relationships of sugars, produced by either a single organism or an individual, is becoming highly dependent on the superposition and integration of available resources, scientific disciplines/methods, and largely disconnected data sets and tools to form a complete

picture of the glycome. Such integration would largely benefit the modern biotechnology and pharmaceutical industries.

With the development of glycobiology as a separate discipline, a number of databases and tools became available in a similar way to the other well-established “omics.” However, due to the much higher level of complexity of carbohydrate structures, many popular algorithms developed and applied to linear biopolymer sequences are not applicable. In addition, glycoinformatics is often faced with a shortage of well-designed training data sets for development and testing purposes. The inapplicability of the standard methods and techniques, used in genomics and proteomics, could also explain a partial disconnect and frequent incompatibility of the collections of experimental data and resources that are available to the glycoscience community today. Alleviating the recognized shortcomings in comprehensive tools for data storage and retrieval is one of the highest priorities of the international glycoinformatics community.

In the last decade major efforts have been made, by leading scientific groups, towards the integration of a number of major databases and tools into a single portal which would act as a centralized data repository for glycomics, equipped with a number of comprehensive analytical tools for data systematization, analysis, and comparison. The first step in this direction was made by a number of recent initiatives such as *GLYCOSCIENCES.de* (16), the Kyoto Encyclopedia of Genes and Genomes (*KEGG*) collection (17), *RINGS* (18), the Consortium for Functional Glycomics (*CFG*) (19), *GlycomeDB* (20, 21), and *EUROCarbDB* (22). Table 1 lists a number of Web portals, which offer a range of carbohydrate-related tools and databases. More details on each are provided in the next section “Glycan structure databases and tools.”

Recent initiatives have highlighted the importance of standardization when storing structural and analytical carbohydrate-related data to facilitate cross-discipline research and data exchange between different repositories. A number of approaches have been used by several groups to develop robust and reliable formats for unique identification of glycan structures, aimed at facilitating data exchange and unifying queries across multiple databases. Among those are *GLYDE* (23), *GLYDE-II* (24), and *GlycoCT* (25). *GlycoCT* is the main unifying format for compact structure encoding, which was developed as a part of the *EUROCarbDB* project. It is also used as the main encoding format in the *GlycomeDB* database (20). It has two forms, a condensed format, mainly used for structure encoding, and an *XML*-based format, designed to facilitate data exchange. Another commonly used format for carbohydrate-related data exchange is *GLYDE-II* (24, 26). Both *GlycoCT* and *GLYDE-II* were a step towards full glycome data integration.

However, data exchange is still a serious bottleneck for the commercial software packages that support HPLC/UPLC analytical equipment. Data management and extraction of large volumes

Table 1
Popular carbohydrate-related databases and glycoinformatic tools

Name	URL	Data available and description
<i>Databases</i>		
CCSD (CarbBank)	http://www.boc.chem.uu.nl/ sugabase/carbbank.html	the first carbohydrate database. It contains: published data on carbohydrate structures, their taxonomy and citations for a variety of species
GLYCOSCIENCES.de	http://www.glycosciences.de/	published data on carbohydrate structures, their taxonomy, citations, MS- and NMR-experimental data, 3D-structures for a variety of species
CFG Glycan	http://www.functionalglycomics.org/glycomics/molecule/jsp/ carbohydrate/carbMolecule- Home.jsp	data on carbohydrate structures and their taxonomy mainly for mammalian species
KEGG GLYCAN	http://www.genome.jp/ kegg/glycan/	data on carbohydrates structures and their metabolic pathways; linked to other KEGG databases
Bacterial Carbohydrate Structure Database (BCSDB)	http://www.glyco.ac.ru/bcsdb3/	published data on carbohydrate structures, their taxonomy, citations, and NMR-experimental data, for bacteria
EUROCarbDB	http://www.ebi.ac.uk/ eurocarb/home.action	published data on carbohydrate structures, their taxonomy, citations, MS-, NMR- and HPLC-experimental data, for a variety of species
GlycomeDB	http://www.glycome-db.org/	Meta-database, which cross-links major carbohydrate-related databases through a single portal
GlycoBase 3.0	http://glycobase.nibrt.ie/ database/	published data on carbohydrate structures, their taxonomy, citations and HPLC experimental data for a variety of species
ExplorEnz and Reaction Explorer	http://www.enzyme-database.org/ http://www.reaction-explorer.org/	relational database which contains data on known enzymes classified according to the IUBMB enzyme nomenclature and linked to the associated database on the enzyme reactions. Each list of associated reactions can be presented in the form of pathway maps.
Glyco3D	http://glyco3d.cermav.cnrs.fr/ glyco3d/index.php	collection of 3D-structures of glycans, polysaccharides, lectins, glycosyltransferases and glycosaminoglycan binding proteins

(continued)

Table 1
(continued)

Name	URL	Data available and description
JCGGDB Database collection	http://riodb.ibase.aist.go.jp/ rcmg/glycodb/Top	MS spectra for carbohydrate structures, data on affinity constant for glycan-lectin interactions, glycoproteins, glycosyltransferases, etc.
<i>Tools</i>		
GlycoForm and Glycologue	http://www.boxer.tcd.ie/gf/	GlycoForm is a fast interactive glycan viewer, which accepts direct user input; all visualised structures can be stored in an associated list of libraries. Glycologue is a tool, which is used to manipulate/print high quality digital images of glycans stored in GlycoForm libraries.
GlycoExtractor	http://glycobase.nibrt.ie/ demoglycoextractor	tool for extracting the data into commonly used formats to facilitate data exchange
GlycoWorkBench	http://www.ebi.ac.uk/eurocarb/ gwb/home.action	tool for manual annotation and interpretation of MS-spectra of carbohydrates
GlycoPeakFinder	http://www.ebi.ac.uk/eurocarb/ gpf/Introduction.action	tool for annotation of MS-spectra of carbohydrates
PDB-care	http://www.glycosciences.de/ tools/pdbcare/	tool for structural validation of 3D-carbohydrate structures

of data are mainly supported by specialized data management software, which is often offered as a separate package and not freely available. Moreover, manufacturers of analytical laboratory equipment have developed various proprietary data formats, which are often not compatible with each other (*AIA/ANDI* (Analytical Instrumentation Association/ANalytical Data Interchange), *ANDI/NetCDF* (ANalytical Data Interchange/network Common Data Form), the Analytical Information Markup Language (AnIML), *GAML* (Generalized Analytical Markup Language), *mzXML* (27), and *mzData* (28)). The last two formats are very popular in mass spectrometry. Some *HPLC* manufacturers have made a step forward by developing data converters (29); however, they have never adopted any of the formats to be used as a unified standard. As the result the existing tools for *HPLC*-glycan data collections' export and exchange are very time-consuming and cumbersome.

The proposal of the *guXML* format (30), designed to store *HPLC/UPLC*-related data, aimed to stimulate discussions between

manufacturers on standardized open formats for data exchange to facilitate the integration of new repositories and tools for high-throughput *HPLC* data analysis.

3. Glycan Structure Databases

3.1. Complex Carbohydrate Structure Database

The need for a centralized large-scale database, that combines structural information with experimental properties, similar to those developed in genomics and proteomics, was recognized in the late 1980s (31). As the result, the first publicly available structural database was developed at the Complex Carbohydrate Research Center in the University of Georgia. The Complex Carbohydrate Structure Database (CCSD) became better known as *CarbBank* (32), the name of its querying tool. The last release of the CCSD contains 50,000 records with data about the primary structure, its source, the analytical methods used to determine the structure, and literature references. *CarbBank* utilized an extended International Union of Pure and Applied Chemistry (*IUPAC*) nomenclature for carbohydrates based on a three-letter code for monosaccharides combined with a structural representation across multiple lines. The development and maintenance of the database were discontinued in 1997; nevertheless, the database (33, 34) remains available to the scientific community from the Bijvoet Center server and as part of *SweetDB* (35, 36), which also contains *SUGABASE*'s content (37). The CCSD still remains one of the largest repositories of carbohydrate-related data and served as the main source and foundation for the majority of the open-access database developments in glycoinformatics.

With the discontinuation of *CarbBank*, a few other initiatives were established which used some of the data from *CarbBank* and combined it with their own data collections. The major currently available carbohydrate databases are presented below.

3.2. GLYCOSCIENCES.de Database

GLYCOSCIENCES.de is an integrated portal for glycosciences, containing a combination of carbohydrate-related data and tools for glycome analysis. The *GLYCOSCIENCES.de* (16) [formerly *SweetDB* (36)] database inherited the content of CCSD and *SUGABASE* (37), which mainly comprised *NMR* spectra for carbohydrates. The database provides the following information for each individual structure: 2D structure of the glycan in *IUPAC* format, chemical formula, molecular weight, number of atoms, composition of known glycan structural motifs found in the given structure, experimental *NMR*, *MS* and crystallographic data, citations, and taxonomy (Fig. 1). The *IUPAC* nomenclature is used for encoding monosaccharide structures. The encoding format for

(a) Substructure / Search / Beginner

Click here to reset input.

a-D-Manp
1-6

a-D-Manp 1-3 b-D-Manp 1-4 b-D-GlcNAc

1-4

b-D-GlcNAc

with 3D-Co-ordinates (Sweet2) | with NMR data | max # residues | min # residues

with PDB entries | min. resolution | all chains | all methods

species human

(use NCBI-Taxon-ID e.g. 9606 or name)

Search Glycosciences
 Search in BCSDB

Search now

Structure Search
You can enter from monosaccharid to pentasaccharid. (For monosaccharides)

Advanced mode

(b)

Explore LinusID 243:
• Structure • Motifs • General Structure Info • Composition • NMR Data • Theoretical Masspeaks • References • Taxonomy
• Expand all • Collapse all

Structure for LinusID 243

b-D-GlcNAc-(1-1)-a-D-Manp-(1-6)+
|
b-D-GlcNAc-(1-4)-b-D-Manp-(1-4)-b-D-GlcNAc-(1-4)-b-D-GlcNAc-(1-4)-a-Aeo-(1-2)-a-Aeo
|
a-D-Manp-(1-3)+
|
b-D-GlcNAc-(1-2)+

[+] Carbohydrate Components: Found 7 Corresponding Structures
[+] Found 1 Structure Motif for LinusID 243
+ N-glycan core (Show) [Search Database]
[+] General Structure Data for LinusID 243
[+] Composition for LinusID 243
Hex
Hex
Hex
HexNAc

[+] NMR Proton Chemical Shift Info for LinusID 243

Fig. 1. Substructural search in *Glycoscience.de*. (a) Glycan structure specification form for querying *Glycosciences.de*. (b) Illustration of the glycan description page containing structural description and properties for one of the output structures.

more complex oligosaccharides is Linear Notation for Unique description of Carbohydrate Structures (38) (*LINUCS*).

In addition to the standard molecular information, *GLYCOSCIENCES.de* provides a range of very useful tools for performing various quality checks and structural analysis. These include the following:

- *GlycoFragment*—Calculates the main fragments of complex carbohydrates that are expected to be present in *MS* spectra.
- *Pdb2linucs* and *LINUCS*—Convert structural formats from Protein Data Bank (*PDB*) and *IUPAC* to *LINUCS*.
- *GlyTorsion*—Performs statistical analysis of carbohydrate torsion angles derived from *PDB*.
- *GlyVicinity*—Statistically analyzes the presence of amino acids in the vicinity of carbohydrate residue.
- *Pdb-care*—Checks carbohydrate residues in *pdb* files for errors.
- *PubFinder*—Searches for relevant literature in *PubMed* using keywords.

3.3. CFG Glycan Database

The *CFG* was established to define the role of protein–carbohydrate interactions in cell surface communication. The central *CFG* database, available to the glycoscientific community, consists of several complex relational sub-databases, which are interfaced between each other to facilitate the search for relevant data from different resources. This database comprises the following modules: glycan structural database (19), glycan-binding proteins (*GBPs*) molecule database, and glycosyltransferase database.

The glycan structures are derived from *CCSD*, Glycominds (a commercial database developed by Glycominds Ltd.), and glycans synthesized by the Consortium. The mono- and oligosaccharide encoding scheme, used in the *CFG* database, was adopted from Glycominds Ltd. (39). The *CFG* database provides the following information for individual glycan entries: composition, molecular weight, and class of the carbohydrate structure; a pictorial structure representation using *CFG* and *IUPAC* extended notation; a string structure representation encoded using linear *IUPAC* format and *LinearCode*; reference citations; biological source of the glycan; *PDB* entries containing a particular glycan in a form of ligand; and corresponding *GLYCOSCIENCES.de* database entries (Fig. 2).

3.4. KEGG GLYCAN Database

KEGG (40, 41) is an integrated set of knowledge base resources covering systems biology-, genomics-, proteomics-, and chemistry-related data for biological sciences. This data bank consists of 16 main databases, where the *KEGG GLYCAN* database (17) is presented as an additional module of the *KEGG LIGAND* database. The *KEGG GLYCAN* database includes about 11,000 unique glycan structures, which mainly derived from 40,000 entries of *CarbBank*, literature, and *KEGG PATHWAY*. The *KEGG GLYCAN* database offers a range of tools and information pages regarding the available glycan structures. This includes similarity search (*KCaM*), an expression analysis tool which links transcriptomic data to glycan chemical structures; manually drawn metabolic, regulatory, and structure pathway maps; a *GBPs* database; a glycosyltransferases database and their reactions; a composite structure mapping and glycan structure editing tool (*KegDraw* drawing tool). *KegDraw* provides two drawing modes, one of which is a standard chemical structure editor, and the other, which is a glycan mode, allows the user to draw glycan structures using monosaccharide units as building blocks. The input encoding format in *KEGG* database is *KEGG Chemical Function* (*KCF*) format (42). More formats are supported for the output function: *KCF*, Portable Network Graphics (*PNG*), and *LINUCS*.

3.5. Bacterial Carbohydrate Structure Database

Bacterial carbohydrate structure database (*BCSDB*) (43, 44) is an open-access database that contains carbohydrate-related data on structures originated from bacteria. It contains almost 10,000

(a) Screenshot of the CFG functionalglycomicsgateway Glycan Structures Database. The interface shows a search results table with columns: Oligosaccharide, Molecular Wt., IUPAC, Composition, Family, Sub Family, and Source. The source column lists biological sources for glycan structures with Hexose (HEX) = 4 and Hexosaminidase N-acetyl (HEXNAc) = 4. Some entries include: *Torpedo californica* (pacific electric ray); *Familial Hypercholesterolemia*; *Gangliosidosis*; *Oryctolagus cuniculus* (Pika); *Familial Hypercholesterolemia*; *Gangliosidosis*; *Bos taurus* (Bovine); *Familial Hypercholesterolemia*; *Gangliosidosis*; and *Homo sapiens* (Human); Liver; Serum; Milk; *Familial Hypercholesterolemia*.

(b) Screenshot of a specific glycan description page. It shows a detailed glycan structure diagram (Carbon Representation), its IUPAC text format, and various links and details such as IUPAC Code, Linear Code, and Carb Bank Links.

Fig. 2. Composition search in CFG Glycan database. (a) List of glycan structures, retrieved by the chosen criteria (HEX = 4; HEXNAC = 4). (b) Glycan description page with pictorial representation of the glycan structure, structure encoded in text format, and other relevant information.

structures, 3,500 of which are from *CarbBank*. Each structure in this database is characterized by data on its biological source, methods of structure elucidation, experimental and theoretical spectral data, biological activity and genetics, conformational data, and its synthesis. Each record is linked with PubMed and the *GLYCOSCIENCES.de* database. All structures are encoded using *BCSD* text format, which is similar to *IUPAC* extended format and is supported by automatic convertors to *IUPAC* and from *GlycoCT* formats for oligosaccharides.

3.6. EUROCARBDB Database

The *EUROCARBDB* project is designed to provide resources and bioinformatic tools for the deposition, storage, and annotation of structural and experimental analytical data obtained by different analytical methods (*HPLC*, *MS*, and *NMR*) for a wide range of carbohydrate structures (Fig. 3). It was launched in 2005 to address the emerging issues in supporting analytical processes of glycan

(a) Browse structures

13,471 distinct sequences. Indefinite sequences are highlighted.

Structure	Evidence	Biological contexts	References
			1. glycosciences.de entry 12 2. Carbbank entry 43757 3. Albersheim et al., 1992
ID 12, entered 15-Apr-2008 by Carbbank			
			1. Carbbank entry 8978 2. Carbbank entry 21724 3. glycosciences.de entry 23 4. Carbbank entry 8319 5. Zopf et al., 1979
ID 23, entered 15-Apr-2008 by Carbbank			
ID 8, entered 15-Apr-2008 by Carbbank			
ID 10, entered 15-Apr-2008 by Carbbank			
			1. species: Vigna radiata 2. species: Vigna radiata 3. species: Pisum sativum 4. Vigna radiata var. radiata 5. Vigna radiata var. radiata 6. Vigna radiata var. radiata (and 17 more...)
ID 7, entered 15-Apr-2008 by Carbbank			

(b) Glycan sequence detail

EurocarbDB Glycan Sequence ID: 12

Entered 15-Apr-2008 14:05:18, by Carbbank
(No evidence for this sequence has been contributed yet)

Biological contexts in which this sequence has been observed:

References:

- Albersheim P, Darvill A, Angier C, Cheung J, Eherhard S, Hahn M, Harfa V, Huhnen D, Koenig K, Meissner M, Neubauer S, Reiter R. Oligosaccharide regulatory molecules. *Acc Chem Res* (1992) 25; 77-83 (Review).
- Carbbank entry 43757
- glycosciences.de entry 12

Composition:

- 4 alpha-D-Gluc: 1
D-Gal: 4
D-Gluc: 1

Sequence:

```

S23
4 alpha-D-Gluc(1->6)-D-Gal(1->3)-D-Gluc(1->6)-D-Gal(1->3)
23(1->6)-D-Gal(1->3)-1,15
93(1->6)-D-Gal(1->3)-1,15
93(1->6)-D-Gal(1->3)-1,15
93(1->6)-D-Gal(1->3)-1,15
L23
1,15(1->15)24
  
```

Fig. 3. *EUROCcarbDB* Database. (a) List of glycan structures available within the database generated using the “browse structures” search option. (b) Individual glycan Web page, which provides the user with all information available for this particular structure.

structure determination and data storage with the aim to simplify new data entry at the user level. This project aims to establish a foundation for a new infrastructure of distributed open-access databases and open source tools, which will be cross-linked to provide the user with exhaustive information about any glycan, which is registered within the database.

The *EUROCcarbDB* consists of several domains which communicate with each other through the central “core” domain, which plays the role of a central repository of carbohydrate structures, their references, and associated biological context. All structures are encoded using *GlycoCT* format—a comprehensive format that allows unique identification of either fully or partially determined structures where some information on linkage type, location of terminal residues, or distributed sulfate substituents might be missing. The available tools enable fast and reliable semi-automatic interpretation of experimental mass and *NMR* spectrograms and chromatograms. These are *GlycanBuilder* (45), a carbohydrate structure editor; *Glyco-Peakfinder* (46), a tool that

allows fast annotation of carbohydrate *MS* spectra, using de novo approach; *GlycoWorkbench* (47), a semi-automated tool for manual interpretation of *MS* spectra; *AutoGU* (48), a tool for interpretation and semi-automated assignment of *HPLC* glycan profiles; *ProSpectND*, a tool for integrated *NMR* data processing and inspection; and *CASPER* (49, 50), an *NMR*-spectra simulator which is designed to assist with *NMR* peak assignments.

3.7. GlycomeDB Database

GlycomeDB (20, 21) is the latest of the carbohydrate structure databases. It is cross-linked with major publicly available databases, such as *GLYCOSCIENCE.de*, *CFG Glycan*, *KEGG GLYCAN*, *BCSDB*, *PDB*, and *EUROCarbDB*. In 2010, the database contained 35,873 unique carbohydrate sequences, supported by taxonomic annotations, approximately 33% of which were fully determined. Due to cross-linkage orientation, *GlycomeDB* automatically updates its content on a weekly basis by retrieving the latest additions from the associated databases. Complementary software, *GlycoUpdateDB*, with a satellite relational database automatically downloads and converts new sequences from associated open-access databases into *GlycoCT*, which is used as the main structure encoding format in *GlycomeDB*. Another tool, *pdb2linux*, is used to extract available carbohydrate structures from *PDB*, which also runs automated checks on the correctness of the registered structures.

GlycomeDB has four standard structural search options, which allow the querying of structures based on the substructural fragments or similarity. There are two carbohydrate structure editors available as a structural input option: *GlycanBuilder* (45) from *EUROCarbDB*, or *DrawRings* (51, 52) from the *RINGS* portal. In addition, a text input format is available too. Pictorial representations of glycans can be visualized using one of the three popular notations in addition to *GlycoCT* format style: *CFG*, *Oxford*, and *IUPAC* notations style.

3.8. GlycoBase3 Database

GlycoBase 3.0 (48) is a relational database which contains experimental *HPLC* elution positions for over 380 2-aminobenzamide (2-AB)-labeled N-glycans together with predicted products of exoglycosidase digestions. Despite a very recent first publication date (2008), this database was compiled over a 10-year period.

The name of each glycan entry is presented using *OXFORD* notation (53) and is comprehensively annotated with its *HPLC* retention time expressed as the glucose unit (GU) value, its monosaccharide composition, and citation references. The GU value of a carbohydrate is an additive parameter and can be defined as a combination of the relative shifts in the retention time of monosaccharides in respect to their linkage type and confirmation. A 2-AB-labeled dextran ladder is used as an internal standard to calibrate each chromatographic profile using the

fifth-order polynomial equation/curve. This technique (54) is very robust and allows the elimination of any equipment-related variation errors. The standard deviation is within 0.2 for over 95% of the collected GU values registered in the database.

Each glycan entry is typically characterized by the following information: pictorial representation of a glycan structure, its name, HPLC retention time expressed as an average GU value, associated standard deviation, NCBI PubMed citations, monosaccharide composition, map of a digest pathway, and links to exoglycosidase digest products, a list of subgroups in which glycan can be found (Fig. 4).

GlycoBase uses *GlycoCT* as its structure encoding format, which can be used to convert pictorial representations of glycan structures dynamically between a number of widely used visualization styles, such as the *Oxford*, *CFG*, and *IUPAC* text notations.

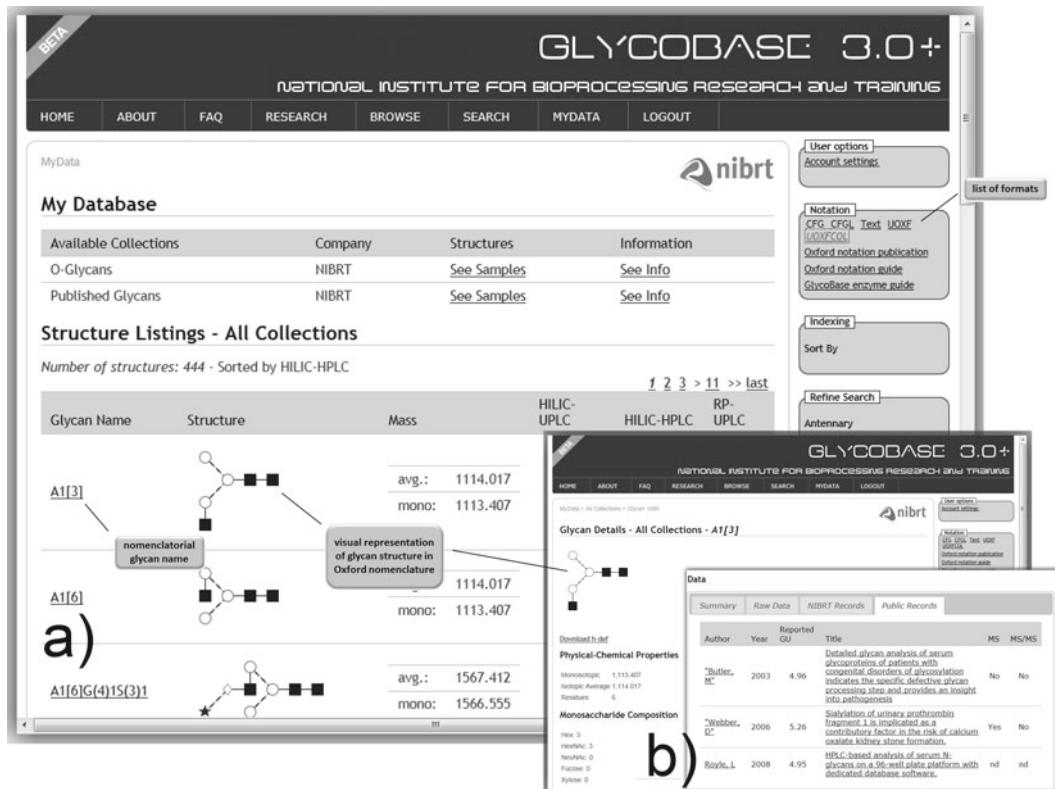


Fig. 4. *GlycoBase 3.0*—relational database containing HPLC/UPLC retention times for a series of 2-aminobenzamide (2-AB)-labeled N- and O-glycans. Each glycan entry is represented by images using one of the three main visualization styles (*CFG* monochrome and color styles, *Oxford* notation—both monochrome and color styles, and *IUPAC*) and is comprehensively annotated. The list of displayed entries can be filtered by using substructural search options.

3.9. ExplorEnz and Reaction Explorer Databases

The *IUBMB* Enzyme Nomenclature is a classification of enzymes based on the reactions they catalyze. *ExplorEnz* (55) is a combination of a MySQL database of the *IUBMB* Enzyme Nomenclature (the Enzyme List) and an associated Web interface, which provides comprehensive search facilities of the enzyme data. Substring searching with all fields is the default behavior, but searches can be limited to fields specified by the user and the display of the results can be customized similarly. Boolean searches are also supported.

Searching *ExplorEnz* using chemical synonyms or common abbreviations is currently limited to those entries that possess a glossary field. A derivative work, *Reaction Explorer*, is a database of the reaction equations of the Enzyme List that can be searched through either a simple Web interface or a desktop application (56). The user can search the reactions using chemical names not found in the Enzyme List. For example, a search for “G6P” will return three primary terms: “D-glucose 6-phosphate,” “ α -D-glucose 6-phosphate,” and “ β -D-glucose 6-phosphate.” Each term in the search results links to a Web page listing the reactions in which that compound participates. Each reactant within the reaction list is itself linked to a similar page, thus enabling the user to step through a “pathway.” The desktop application version of *Reaction Explorer*, which can be downloaded from the main site, provides a way to graph the data as the pathway is traversed (see Fig. 5). In order to prevent an explosion in the number of possible reactions presented to the user, the desktop application excludes common cofactor pairs and small molecules from consideration.

3.10. Glyco3D Database

Glyco3D (57) is an integrated set of databases containing 3D structures of carbohydrates, lectins, glycosyltransferases, and glycosaminoglycan-binding proteins. Carbohydrates are subdivided into four categories: mono-, di-, oligo-, and polysaccharides, which *Glyco3D* organizes into separate databases. The polysaccharide database contains mainly polymer carbohydrates which are usually out of the focus of glycobiology.

The monosaccharide database comprises 18 categories of monosaccharides, which are often used as standard “building blocks” for glycan structures. Each monosaccharide is represented by a number of isomers, where an *HTML* page for each isomer provides information on the monosaccharide family, anomery, cycle type, its configuration, a 2D view of the molecular structure, and 3D structural information in the form of *pdb* format. The latter can be downloaded as an image or a *pdb* file, which can be visualized using *Jmol* viewer (58).

The di- and oligosaccharide databases have a slightly different structure from the monosaccharide database. As the level of complexity is much higher, the user is required to select several options from drop-down menus to specify the structure of interest. When a specific glycan structure is selected, the user is provided with a list

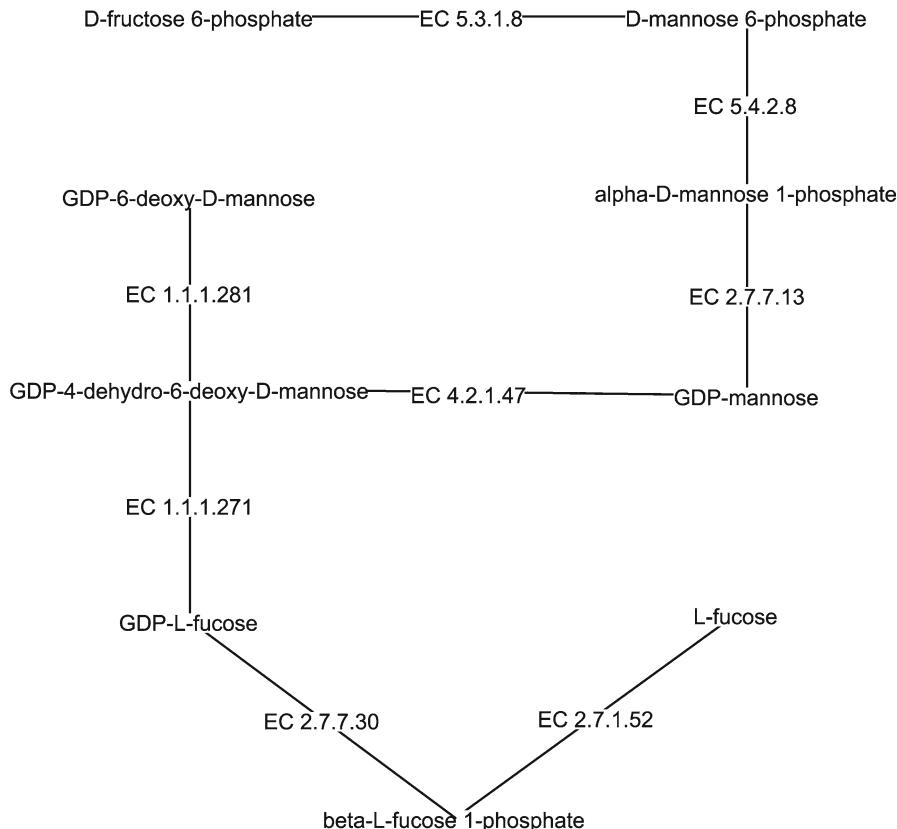


Fig. 5. A Reaction Explorer graph of GDP-L-fucose and GDP-mannose biosynthesis from d-fructose-6-phosphate and the l-fucose salvage pathway. The enzymes involved are GDP-l-fucose synthase (EC 1.1.1.271), GDP-4-dehydro-6-deoxy-d-mannose reductase (EC 1.1.1.281), fucokinase (EC 2.7.1.52), mannose-1-phosphate guanylyltransferase (EC 2.7.7.13), fucose-1-phosphate guanylyltransferase (EC 2.7.7.30), GDP-mannose 4,6-dehydratase (EC 4.2.1.47), mannose-6-phosphate isomerase (EC 5.3.1.8), and phosphomannomutase (EC 5.4.2.8).

of low-energy conformers, where each is characterized by an iso-potential energy map (Ramachandran plot), energy value, torsion angles, force field type used to model the glycan, and its 3D structure via Jmol in *pdb* format.

3.11. JCGGDB Databases Collection

A more recent initiative was the launch of a Japan Glycoscience Integrated DataBase (*JCGGDB*) Web portal containing a range of databases and tools for glycan structure analysis and synthetic technology (59–62) by the Glyco-Biomarker Discovery Team at the Research Center for Medical Glycoscience, which is part of the National Institute of Advanced Industrial Science and Technology (*AIST*). This portal provides four comprehensive collections covering mass spectrometry (*GMDB*) (60), the genes associated with glycan synthesis (GlycoGene Database or *GGDB*) (59), the affinity constants between a number of lectins and pyridylaminated

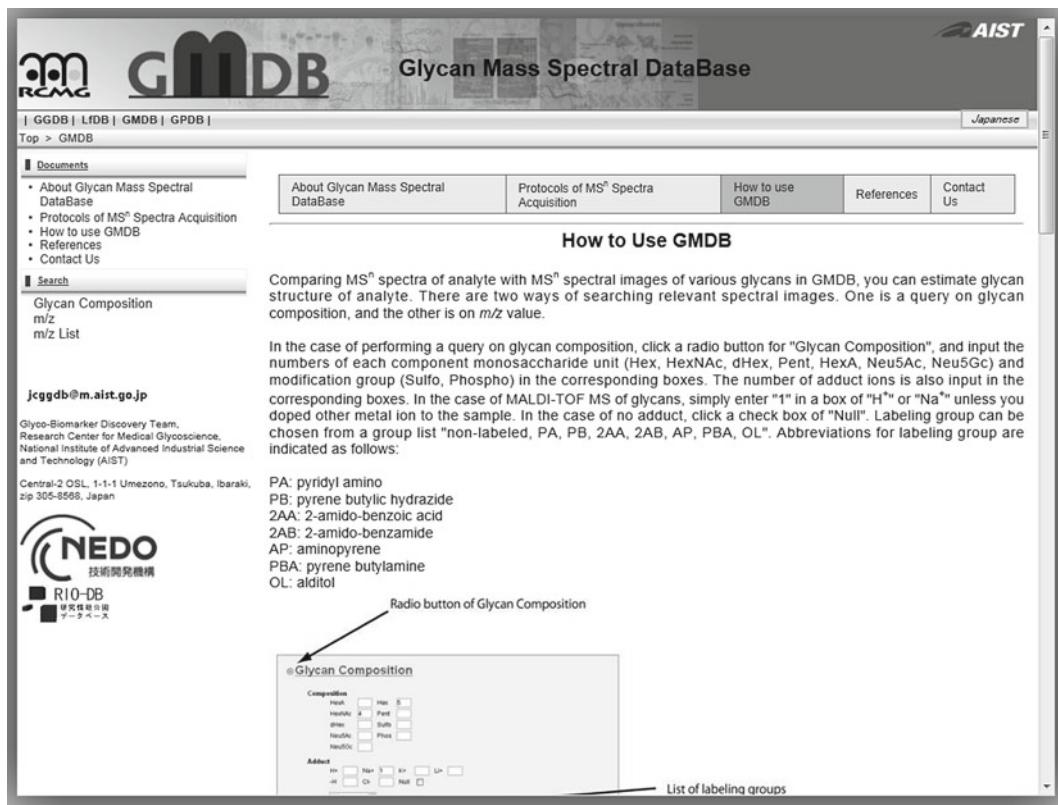


Fig. 6. Glycan Mass Spectral Database (*GMDB*).

glycans (Lectin *Frontier* DataBase or *LfDB*) (61, 63), and N-glycoproteins, which are identified experimentally from *Caenorhabditis elegans* N2 and mouse tissues (strain C52BL/6J, male)—(*GlycoProtDB* or *GPDB*) (62, 64–66).

The Glycan Mass Spectral Database (*GMDB*) is a multi-stage tandem mass spectral database, which contains a variety of structurally defined glycans (Fig. 6). It currently stores mass spectra of N- and O-linked glycans, and glycolipid glycans as well as the partial structures of these glycans. N-glycans and glycolipid glycans are mostly tagged with 2-aminopyridine (PA), which can be used for fluorescence detection in *HPLC*.

There are two types of search implemented in *GMDB*. When querying glycan composition, the user should define the number of each component monosaccharide units and modification groups. To query the equivalent mass (*m/z*) value the user defines the equivalent weight of a precursor ion of the spectrum of interest. Output results are presented in the form of a list of structures that can be characterized by an equivalent weight value with a tree-structured list of equivalent weights. Each spectrum can be visualized in the form of a spectral diagram.

Fig. 7. GlycoGene Database (GGDB).

GGDB (67) was the first database to provide information on substrate specificity (Fig. 7). This database comprises a collection of 180 human glycogenes associated with glycosyltransferases, sugar nucleotide synthases, sugar nucleotide transporters, and sulfotransferases. Each glycogene is characterized by the following information: gene names (gene symbols), enzyme names, DNA sequences, tissue distribution (gene expression), substrate specificities, homologous genes, EC numbers, and external links to associated databases. This information can be easily extracted as *XML*. Cross-linkage between the four databases of the portal will be implemented in the future.

LfDB provides quantitative interaction data between a number of lectins with various PA glycans. The data are provided on two pages: the lectin information page and the interaction page. The first page contains basic information on lectins, while the second displays the interaction data obtained by the automated frontal affinity chromatography with fluorescence detection. A substructural search function, which uses predefined structural elements, is also available to facilitate structural queries.

GPDB contains experimental data on N-linked glycoproteins. The typical information page comprises the following data: protein

(gene) ID, protein name, glycosylated sites, and types of lectins used to capture glycopeptides. At the moment the data are restricted to glycoproteins from *C. elegans* and mouse liver. An extension of the list of tissues covered is among future plans for the database development.

4. Glycoinformatics Tools

4.1. GlycoForm and Glycologue

Glycosylation of proteins is an important aspect of postprocessing, and must, therefore, be given due consideration when developing bioengineered therapeutic proteins. Prediction of the glycans associated with recombinant proteins is increasingly being done through mathematical modeling of different aspects of glycosylation (68–72). One such endeavor (71) has used a numerical, nine-digit, code to describe N-glycans most frequently obtained with Chinese hamster ovary cell lines.

The system is based upon the reactions of the enzymes involved; each code uniquely describes a structure by taking into account the number of occurrences of each sugar residue as well as the linkage type, and also the degree of branching (antennarity).

Two software tools have been developed (73) to allow the use of the nine-digit code, the first of which, *GlycoForm*, reads a nine-digit code entered by the user and displays the result using commonly used symbols for the sugar units: *CFG* and *Oxford* (*UOXF*). A text-only rendering method is also available, which shows explicitly the linkage types between residues. Structures can be added to a library stored locally on the user's hard disk. The nine-digit code can also be represented by *GlycoForm* as a string in *Oxford* notation, which can then be used as a search term in *GlycoBase* (48). Examples of the different forms of output provided for the structure code 310333344 are shown in Fig. 8. A companion program, *Glycologue*, accepts as input files containing lists of such codes, displaying them in a grid. In combination, these software tools provide a rapid display and ready comparison of large numbers of structures simultaneously.

4.2. GlycoExtractor

GlycoExtractor (30) is a Web-based tool developed to automate the routine extraction of glycan profile data from *HPLC* systems and its visualization. The tool allows the end user to merge multiple data sets and to store them in a single file, rather than a collection of files. A selection of widely used file formats (*XML*, *JSON*, and *CSV*) opens the field to new approaches for high-throughput data interpretation and storage. The overall performance significantly reduces the amount of manual data manipulations required to export and prepare data for further analysis.

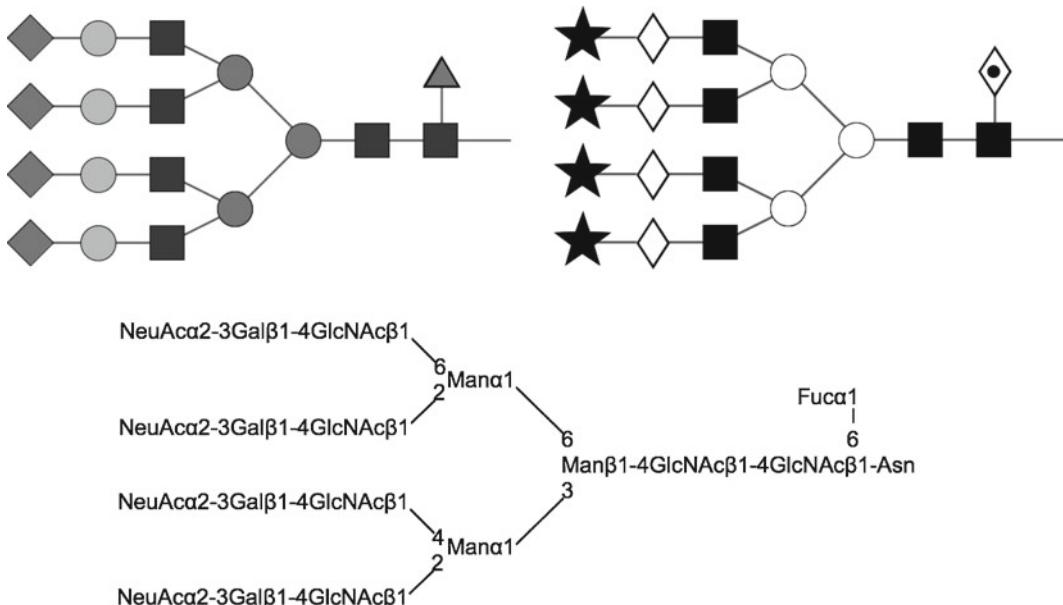


Fig. 8. Three alternative GlycoForm renderings of the N-glycan F(6)A4G(4)S4, represented by the nine-digit code 310333344. *Upper left:* CFG grayscale symbols; *upper right:* Oxford symbols; *lower:* text-only rendering.

The data selection algorithm follows the general hierarchy of the experimental database, which contains multiple subsets of the experimental data. Each subset is characterized by a set of specific features shared by its members, where each member can be identified by a number of unique parameters such as peak number, peak area, and glucose unit values (Fig. 9).

GlycoExtractor includes the following features: (a) the ability to specify which data sets to export; (b) the ability to export samples across multiple sample sets; (c) specification of the order in which items will be exported; (d) loading and saving of settings to allow the user to specify the data over multiple sessions or to automate commonly used sets; (e) an undo/redo function; (f) export in the following file formats: *guXML* (*an XML schema tailored to glycans*), *guJSON* (*a JSON format tailored to glycans*), CSV, and HTML; and (g) a generated HTML page, which interactively allows user to select the samples and dates to be displayed.

The first variant of the tool was presented to the scientific community in 2010 to initiate the creation of new data interchange resources for *HPLC*-glycan technologies. The application uses a modular architecture, which allows different data sources and different output file formats to be added. Therefore, it can be easily integrated with different systems and software used in high-throughput *HPLC* data analysis and can be tailored to meet future requirements and advances in *HPLC* technology. It is anticipated that this tool will set the foundations for new bioinformatic

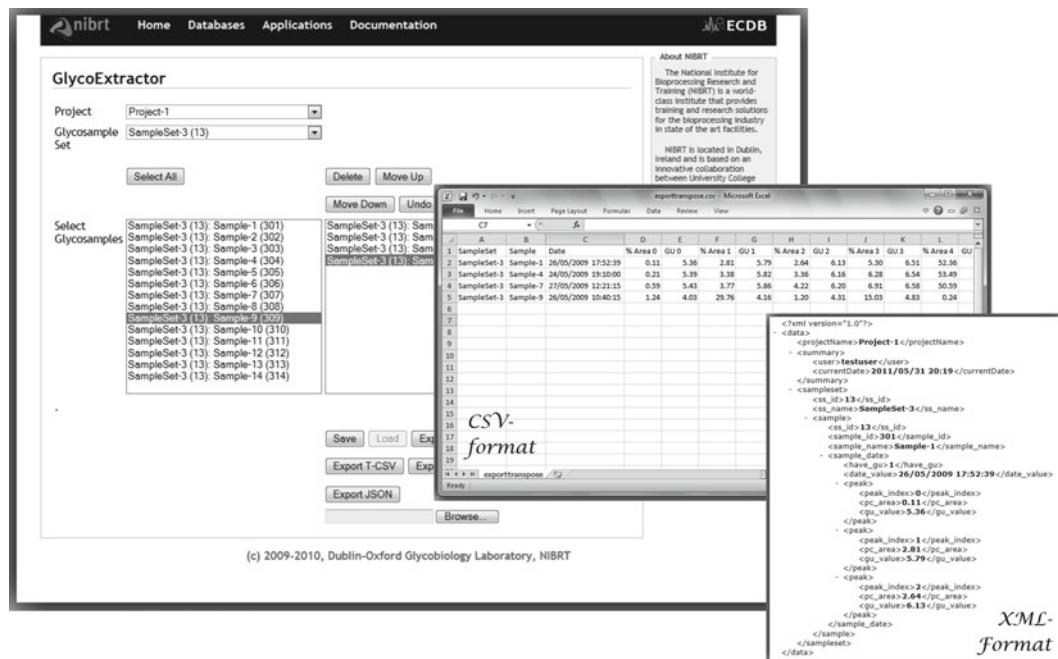


Fig. 9. GlycoExtractor User Interface with two examples of output formats (CSV type and XML type).

resources for HPLC-glycan technologies and the development of new file formats in the field. This will further facilitate the implementation of a high-throughput analytical platform and will benefit a number of research activities including biomarker discovery, validation, and monitoring of online bioprocessing conditions for next-generation biotherapeutics.

4.3. GlycanBuilder

GlycanBuilder (45) is a tool that allows the drawing and displaying of glycan structures in symbolic notation. It was developed for the *EUROCarbDB* project. Due to the high complexity of glycan structures at the atomic level, standard chemical structure editors are often of no use. *GlycanBuilder* was one of the first user-friendly input/output tools for carbohydrates and can be used in a similar way to various structure drawing editors.

GlycanBuilder provides a graphical representation of glycans using one of the three commonly adopted symbolic notations, which are *CFG* notation (both black and white and color schemes); black and white *Oxford* notation; and *IUPAC* notation. The rendering algorithm that was implemented for this builder uses a set of rules to determine the style and location of the residues and linkages for the visual representation. The biggest advantage of this tool is that it reduces user input to the minimum while still allowing the generation of high-quality images. *GlycanBuilder* uses the

main *EUROCarbDB* format as its input and its output can be easily handled by *EUROCarbDB* framework.

4.4. GlycoWorkBench

Mass spectrometry is one of the main analytical methods, used in glycobiology, for glycan structure identification. One of the major bottlenecks in high-throughput glycomics is the interpretation of experimental spectral data. *GlycoWorkbench* (47) was developed within *EUROCarbDB* framework to facilitate routine manual interpretation of *MS* spectra. The interpretation of spectra in mass glycospectrometry is severely limited by the fact that there is a limited availability of experimental data on mass fragmentation of glycans and their assignments. The main application of the *Glyco Workbench* is to assist the user in evaluation of the proposed structures by automatically checking the corresponding theoretical fragment masses against the peaks derived from the spectrum. *Glyco Workbench* uses *GlycanBuilder* as a glycan structure editor. It also allows the import of structures encoded using structural formats of the major carbohydrate-related databases, such as *LINUCS*, *LinearCode* (CFG and Glycominds), and *GlycoCT* (*EuroCarbDB* and *GlycoBase 3.0*). This facilitates the integration of the tool with relevant databases.

Each structure is characterized by a number of parameters, which specify the type of persubstitution, identities, and quantities of ion adducts and the neutral exchange used in computation of mass-to-charge ratios. The “*PeakList*” allows some manual manipulation of the list of labeled picks for further annotation. Experimental values of the peak list can be imported from either a tab-separated text file or commonly used *XML* formats for storage of raw mass spectrometric data.

4.5. Glyco-PeakFinder

Glyco-Peakfinder has been widely used for calculation of monosaccharide compositions and annotation of the mass spectra of various oligosaccharides (46). As mass spectrometry is one of the key analytical methods used for glycan analysis, the biggest advantage of this tool is in the use of the *de novo* algorithm for generation of glycan composition without knowledge of the biological system and the fragmentation technique. By aiming to solve some limitations of its predecessors, *Glyco-PeakFinder* was developed to assign all types of fragmentation, including monosaccharide cross-ring cleavage products, ions with multiple charges, persubstituted products, commonly used reducing-end mutations, and glycoconjugates. To handle such a variety of fragmentation types, full user control over the imposed constraints, when performing combinatorial search, is provided.

The mass accuracy level is determined by the method used to obtain the spectrum. Data on experimental peaks can be uploaded from a file or manually entered using the “mass page.” Calculation preferences of the masses and the source of data (original compound

or its fragments) should be selected to appropriately define the processing method. Minimal and maximal occurrence level for the monosaccharides and substituents can be set up using the “residue page.”

The default settings for the processing methods were designed to allow the fast identification of native carbohydrates that are usually associated with glycans common for mammalian organisms. Analysis of any other type of biological system requires modification of the processing method settings. A variety of possible substitutions and mutations are provided on the “modification page,” where the user can define the appropriate ensemble for the analyzed system. Fragmentation of non-carbohydrate part of the system is ignored. To avoid combinatorial explosion, a number of restrictions, governing the calculations, are imposed, for instance the maximum mass threshold obtained from the list of experimental peaks. It results in the interruption of further calculations for every subbranch, where the calculated mass exceeds given threshold. A stepwise technique, with allowable user input for any corrections/modifications/restrictions, also has the potential to speed up the final step of the combinatorial search.

4.6. PDB-Care

The current methods used in carbohydrate-recognition studies often rely on known protein folds to identify active glycoprotein domains. Often these methods fail to find active sugar-binding domains as this type of activity evolves in different structural contexts. As a result, crucially important glyco-receptors could easily be missed in genome screening processes, which are based on known protein families. This requires the development of novel strategies to identify such glycoproteins. Screening of glycan arrays has huge potential to provide a critical bridge between our knowledge of glycomics and the function of specific glycans as target ligands for receptors (74).

It is anticipated that conformational analysis will play a key role in revealing the function of various carbohydrates on a molecular level. Looking at the progress in genomics and proteomics to date, glycomics is still lagging in terms of development and the knowledge available. Only a limited number (5%) of crystallographic structures with well-resolved complex oligosaccharides have been reported, which corresponds to about 30% of the carbohydrate-containing *PDB* entries (75, 76). The absence of carbohydrate coordinates in *PDB* files taken from X-ray crystallographic experimental data often means that they cannot be considered to be reliable.

Due to the complexity of carbohydrate structures in comparison to the linear sequences used to encode proteins and genes, the existing tools and algorithms that are used in proteomics and genomics cannot be easily reapplied in glycomics. *PDB-care* (77) closed the gap in automatic identification and assignment of carbohydrate structures using the 3D atom coordinates and reported

atom types from *PDB* files. *PDB-care* uses *LINUCS* notation to decode carbohydrate structures. This format was also used to create a translational table, which aligns both systematic names from the *LINUCS* format and the corresponding *PDB* residue codes. This program automatically checks for inconsistencies and errors related to the reported atom connectivity and completeness. The most common type of error was a mismatch between residue nomenclature and the determined structure.

PDB-care plays a similar role in glycomics to that of *WhatCheck* (78) and *ProCheck* (79, 80) in proteomics. More frequent use of *PDB-care* as a standard checkup procedure, when depositing a new structure into the *PDB*, should significantly improve the quality of the reported crystallographic structures.

5. Conclusions

The importance of carbohydrates, one of the major classes of post-translational modifications of proteins, cannot be underestimated. While it is difficult to arrive at the exact figures, it is very well known that the glycome is much more complex than the proteome and genome together. For instance, analysis of 650 completely sequenced organisms in *CAZy* (carbohydrate-active enzymes) database suggests that about 5% of the genome encodes enzymes that are involved in glycan synthesis, degradation, or recognition. About 1–2% of genes of an organism encode glycosyltransferases.

The glycomics international community is progressing very fast towards the creation and maintenance of integrated worldwide resources on carbohydrate-related data and cross-linking these resources with other “*omics*” in an effort to fill the existing gaps in molecular and systems biology. As was mentioned beforehand, alleviating the recognized shortcomings of the available tools for data storage and retrieval is one of the highest priorities.

While a first step was made towards linking the major carbohydrate-related databases through a single portal by the launch of *GlycomeDB*, there is still the need to develop an integrated data repository with a simplified user interface instead of cross-linking different databases, which do not have much in common except their main subject. The biggest issue is differences in formats used to store chemical structures and annotations. This often complicates the whole process of formulating queries. A first step in this direction was made by agreeing on using *CLYDE-II* and *GlycoCT* as common formats for encoding carbohydrate structures to simplify data exchange and user access to data (26).

After the discontinuation of the maintenance of *CarbBank*, there were no further efforts made towards providing a single unified repository of published carbohydrate-related data, with no redundancies or incorrect information. Small proprietary databases

are often limited in the scope of their registered data and their annotation. Often information that is provided by the existing databases is partially overlapped and redundant. More attention needs to be paid to the curation of the existing databases to eliminate errors and incorrect links. Citing references should include abstracts as well as the publication sources. Complexity of the database interface is also a common problem.

Furthermore, the absence of user manuals is often listed as one of the recognized shortcomings of the small databases. This should be developed to improve the overall usability of integrated repositories.

Finally, a few key issues to consider are data entry, data annotation, and database updates. It was highlighted that the automated “text-mining” approaches are often not applicable to carbohydrate-related data (81) as structural information about glycans is often presented in the form of images due to their complexity. The ability to deposit new data directly into publicly available databases, in a similar way to how it is organized in genomics and proteomics, would significantly increase the progress in this field.

Ideally, the final product of the integrated efforts should not be dependent on funding. Life sciences would greatly benefit by including future integrated carbohydrate databases into the larger existing infrastructures for biologists and chemists, where glycomics data could be cross-linked with genomics, proteomics, and metabolomics databases. Initial discussions about the integration of glycoinformatic resources with one of the major biological portals have already taken place between members of the glycoscience community (82). The existence of a centralized repository will significantly improve the quality of the query results and decrease the average time spent on searching for the available published data relating to a specific structure or class of structures.

Acknowledgments

The authors would like to thank Prof. Keith F. Tipton from The School of Biochemistry and Immunology at Trinity College Dublin for valuable suggestions during the preparation of the manuscript.

References

- Shriver Z, Raguram S, Sasisekharan R (2004) Glycomics: a pathway to a class of new and improved therapeutics. *Nat Rev Drug Discov* 3(10):863–873
- Gerlach JQ, Cunningham S, Kane M et al (2010) Glycobiomimetics and glycobiosensors. *Biochem Soc Trans* 38(5):1333–1336
- van der Werff TJ (2003) 10 emerging technologies that will change the world. Global Future Report™. (<http://www.globalfuture.com/mit-trends2003.htm>)
- Solá RJ, Griebenow K (2010) Glycosylation of therapeutic proteins. An effective strategy to optimize efficacy. *BioDrugs* 24(1):9–21

5. Arnold JN, Wormald MR, Sim RB et al (2007) The impact of glycosylation on the biological function and structure of human immunoglobulins. *Annu Rev Immunol* 25:21–50
6. Shields RL, Lai J, Keck R et al (2002) Lack of fucose on human IgG1 N-linked oligosaccharide improves binding to human Fc γ RIII and antibody-dependent cellular toxicity. *J Biol Chem* 277:26733–26740
7. Ferrara C, Stuart F, Sondermann P et al (2006) The carbohydrate at Fc γ RIIa Asn-162. An element required for high affinity binding to non-fucosylated IgG glycoforms. *J Biol Chem* 281:5032–5036
8. European Medicines Agency/Committee for Medicinal Products for Human Use (2005) Guideline on similar biological medicinal products containing biotechnology-derived proteins as active substance: nonclinical and clinical issues. (http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003920.pdf)
9. European Medicines Agency/Committee for Medicinal Products for Human Use (2007) Guideline on production and quality control of monoclonal antibodies and related substances. (http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003073.pdf)
10. Kemsley JN (2009) Analyzing protein drugs. *Chem Eng News* 87:20–23
11. 2009 US sales and prescription information. Top 15 US pharmaceutical products by sales. (http://www.imshealth.com/deployedfiles/imshealth/Global/Content/StaticFile/Top-Line_Data/Top%202015%20Products%20by%20U.S.Sales.pdf)
12. Amgen announces modifications to US. Prescribing information for use of erythropoiesis-stimulating agents in chronic kidney disease. (<http://checkorphan.org/grid/news/treatment/amgen-announces-modifications-to-u-s-prescribing-information-for-use-of-erythropoiesis-stimulating-agents-in-chronic-kidney-disease>)
13. Erythropoiesis-stimulating agents (ESAs): Procrit, Epogen and Aranesp: drug safety communication. (<http://www.fda.gov/Safety/MedWatch/SafetyInformation/SafetyAlertsforHumanMedicalProducts/ucm200391.htm>)
14. Larkin C (2011) Lipitor topped worldwide drug sales in 2010; Crestor gains most. Bloomberg. (<http://www.bloomberg.com/news/2011-02-10/lipitor-topped-worldwide-drug-sales-in-2010-crestor-gains-most.html>)
15. Wood L (2011) Research and markets: this global protein therapeutics market analysis forecasts the industry to grow at a CAGR of 12% during 2011–2013. FierceBiotech. (<http://www.fiercebiotech.com/press-releases/research-and-markets-global-protein-therapeutics-market-analysis-forecasts->)
16. Lütteke T, Bohne-Lang A, Loß A et al (2006) GLYCOSCIENCES.de: an internet portal to support glycomics and glycoproteomics research. *Glycobiology* 16(5):71R–81R
17. Hashimoto K, Goto S, Kawano S et al (2006) KEGG as a glycome informatics resource. *Glycobiology* 16(5):63R–70R
18. Aoki-Kinoshita KF, Ichikawa M, Ikeda S et al (2006) A web-based resource for glycome informatics. In: 17th International conference on genome informatics (GIW 2006), Yokohama Pacifico, Japan
19. Raman R, Venkataraman M, Ramakrishnan S et al (2006) Advancing glycomics: implementation strategies at the Consortium for Functional Glycomics. *Glycobiology* 16(5):82R–90R
20. Ranzinger R, Hergert S, Wetter T et al (2008) GlycomeDB—integration of open-access carbohydrate structure databases. *BMC Bioinformatics* 9:384
21. Ranzinger R, Hergert S, von der Lieth C-W et al (2011) GlycomeDB—A unified database for carbohydrate structures. *Nucl Acids Res* 39(Database Issue): D373–D376
22. von der Lieth C-W, Freire AA, Blank D et al (2011) EUROCARBDB: an open-access platform for glycoinformatics. *Glycobiology* 21(4):493–502
23. Sahoo SS, Thomas C, Sheth A et al (2005) GLYDE—an expressive XML standard for the representation of glycan structure. *Carbohydr Res* 340(18):2802–2807
24. Packer NH, von der Lieth C-W, Aoki-Kinoshita KF et al (2008) Frontiers in glycomics: bioinformatics and biomarkers in disease. *Proteomics* 8(1):8–20
25. Herget S, Ranzinger R, Maass K et al (2008) GlycoCT—A unifying sequence format for carbohydrates. *Carbohydr Res* 343(12): 2162–2171
26. Lütteke T (2008) Web resources for the glycoscientist. *Chembiochem* 9:2155–2160
27. Pedrioli PGA, Eng JK, Hubley R et al (2004) A common open representation of mass spectrometry data and its application to proteomics research. *Nat Biotechnol* 22(11):1459–1466
28. Orchard S, Taylor CF, Hermjakob H et al (2004) Current status of proteomic standards development. *Expert Rev Proteomics* 1(2): 179–183
29. Waters: innovation in separations science 2011, Dublin, Ireland, 5–6 Apr 2011

30. Artemenko NV, Campbell MP, Rudd PM (2010) GlycoExtractor: a web-based interface for high-throughput processing of HPLC-glycan data. *J Proteome Res* 9(4):2037–2041
31. Doubet S, Bock K, Smith D et al (1989) The complex carbohydrate structure database. *Trends Biochem Sci* 14(12):475–477
32. Doubet S, Albersheim P (1992) Letter to the Glyco-Forum: CarbBank. *Glycobiology* 2(6):505
33. Carbohydrate Databases (CarbBank/Sugabase) (<http://www.boc.chem.uu.nl:8081/sugabase/databases.html>).
34. Berteau O, Stenutz R (2004) Web resources for the carbohydrate chemist. *Carbohydr Res* 339:929–936
35. McNaught AD (1997) International Union of Pure and Applied Chemistry and International Union of Biochemistry and Molecular Biology. JointCommissiononBiochemicalNomenclature. Nomenclature of carbohydrates. *Carbohydr Res* 297(1):1–92
36. Loß A, Bunsmann P, Bohne A et al (2002) SWEET-DB: an attempt to create annotated data collections for carbohydrates. *Nucleic Acids Res* 30(1):405–408
37. van Kuik JA, Vliegenthart JFG (1992) Databases of complex carbohydrates. *Trends Biotechnol* 10:182–185
38. Bohne-Lang A, Lang E, Forster T et al (2001) LINUCS: LInear Notation for Unique description of Carbohydrate Sequences. *Carbohydr Res* 336:1–11
39. Banin E, Neuberger Y, Altshuler Y et al (2002) A novel linear code nomenclature for complex carbohydrates. *Trends Glycosci Glycotechnol* 14(77):127–137
40. Ogata H, Goto S, Sato K et al (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 27(1):29–34
41. Kanehisa M, Goto S, Kawashima S (2004) The KEGG resource for deciphering the genome. *Nucl Acids Res* 32(Database issue): D277–D280
42. Hattori M, Okuno Y, Goto S et al (2003) Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *J Am Chem Soc* 125(39):11853–11865
43. Toukach FV, Knirel YA (2005) New database of bacterial carbohydrate structures. In: Proceedings of the XVIII international symposium on glycoconjugates, Florence, Italy. pp 216–217
44. Toukach PV (2011) Bacterial carbohydrate structure database 3: principles and realization. *J Chem Inf Model* 51(1):159–170
45. Ceroni A, Dell A, Haslam SM (2007) The GlycanBuilder: a fast, intuitive and flexible software tool for building and displaying glycan structures. *Source Code Biol Med* 2:3
46. Maass K, Ranzinger R, Geyer H et al (2007) “Glyco-peakfinder”—de novo composition analysis of glycoconjugates. *Proteomics* 7(24):4435–4444
47. Ceroni A, Maass K, Geyer H et al (2008) GlycoWorkbench: a tool for the computer-assisted annotation of mass spectra of glycans. *J Proteome Res* 7(4):1650–1659
48. Jansson PE, Kenne L, Widmalm G (1991) CASPER: a computer program used for structural analysis of carbohydrates. *J Chem Inf Comput Sci* 31(4):508–516
49. Jansson PE, Stenutz R, Widmalm G (2006) Sequence determination of oligosaccharides and regular polysaccharides using NMR spectroscopy and a novel web-based version of the computer program CASPER. *Carbohydr Res* 341(8):1003–1010
50. Aoki KF, Yamaguchi A, Ueda N et al (2004) KCaM (KEGG carbohydrate matcher): a software tool for analyzing the structures of carbohydrate sugar chains. *Nucleic Acids Res* 32: W267–W272
51. Akune Y, Hosoda M, Kaiya S et al (2010) The RINGS resource for glycome informatics analysis and data mining on the web. *OMICS J Integrative Biol* 14(10):475–486
52. Campbell MP, Royle L, Radcliffe CM et al (2008) GlycoBase and AutoGU: tools for HPLC-based glycan analysis. *Bioinformatics* 24(9):1214–1216
53. Harvey DJ, Merry AH, Royle L et al (2009) Proposal for a standard system for drawing structural diagrams of N- and O-linked carbohydrates and related compounds. *Proteomics* 9(15):3796–3801
54. Royle L, Campbell MP, Radcliffe CM et al (2008) HPLC-based analysis of serum N-glycans on a 96-well plate platform with dedicated database software. *Anal Biochem* 376(1):1–12
55. McDonald AG, Boyce S, Tipton KF (2009) ExplorEnz: the primary source of the IUBMB enzyme list. *Nucl Acids Res* 37(Database Issue): D593–D597. (<http://www.enzymedatabase.org>)
56. McDonald AG, Tipton KF, Boyce S (2009) Tracing metabolic pathways from enzyme data. *Biochim Biophys Acta* 1794:1364–1371, <http://www.reaction-explorer.org>
57. Imberty A, Gerber S, Tran V et al (1990) Database of 3-dimensional structures of disaccharides, a tool to build 3-D structures of

- oligosaccharides. Part I. Oligo-mannose type N-glycans. *Glycoconjugate J* 7(1):27–54
58. Jmol: an open-source Java viewer for chemical structures in 3D. (<http://www.jmol.org/>)
 59. Narimatsu H (2004) Construction of a human glycogene library and comprehensive functional analysis. *Glycoconjugate J* 21(1):17–24
 60. Kameyama A, Kikuchi N, Nakaya S et al (2005) A strategy for identification of oligosaccharide structures using observational multistage mass spectral library. *Anal Chem* 77(15):4719–4725
 61. Nakamura S, Yagi F, Totani K et al (2005) Comparative analysis of carbohydrate-binding properties of two tandem repeat-type Jacalin-related lectins, *Castanea crenata* agglutinin and *Cycas revoluta* leaf lectin. *FEBS J* 272(11):2784–2799
 62. Kaji H, Saito H, Yamauchi Y et al (2003) Lectin affinity capture, isotope-coded tagging and mass spectrometry to identify N-linked glycoproteins. *Nat Biotechnol* 21(6):667–672
 63. Tateno H, Nakamura-Tsuruta S, Hirabayashi J (2007) Frontal affinity chromatography: sugar-protein interactions. *Nat Protoc* 2:2529–2537
 64. Shinkawa T, Taoka M, Yamauchi Y et al (2005) STEM: a software tool for large-scale proteomic data analyses. *J Proteome Res* 4(5):1826–1831
 65. Kaji H, Yamauchi Y, Takahashi N et al (2006) Mass spectrometric identification of N-linked glycopeptides using lectin-mediated affinity capture and glycosylation site-specific stable isotope tagging. *Nat Protoc* 1(6):3019–3027
 66. Kaji H, Kamiie J, Kawakami H et al (2007) Proteomics reveals N-linked glycoprotein diversity in *Caenorhabditis elegans* and suggests an atypical translocation mechanism for integral membrane proteins. *Mol Cell Proteomics* 6(12):2100–2109
 67. Taniguchi N, Suzuki A, Ito Y et al (2008) A database system for glycogenes (GGDB). Springer, Japan, pp 423–425
 68. Shelikoff M, Sinskey AJ, Stephanopoulos G (1996) A modeling framework for the study of protein glycosylation. *Biotechnol Bioeng* 50:73–90
 69. Umaña P, Bailey JE (1997) A mathematical model of N-linked glycoform biosynthesis. *Biotechnol Bioeng* 55(6):890–908
 70. Murrell MP, Yarema KJ, Levchenko A (2004) The systems biology of glycosylation. *Chembiochem* 5:1334–1347
 71. Krambeck FJ, Betenbaugh MJ (2005) A mathematical model of N-linked glycosylation. *Biotechnol Bioeng* 92(6):711–728
 72. Hossler P, Mulukutia BC, Hu W-S (2007) Systems analysis of N-glycan processing in mammalian cells. *PLoS One* 1(8):e713
 73. McDonald AG, Tipton KF, Stroop CJM et al (2010) GlycoForm and Glycologue: two software applications for the rapid construction and display of N-glycans from mammalian sources. *BMC Res Notes* 3:173, <http://www.boxer.tcd.ie/gf>
 74. Laurent N, Voglmeir J, Flitsch S (2008) Glycoarrays-tools for determining protein-carbohydrate interactions and glycoenzyme specificity. *Chem Commun* (37):4400–4412
 75. Lütteke T, Frank M, von der Lieth C-W (2004) Data mining the protein data bank: automatic detection and assignment of carbohydrate structures. *Carbohydr Res* 339:1015–1020
 76. Crispin M, Stuart DI, Jones Y (2007) Building meaningful models of glycoproteins. *Nat Struct Mol Biol* 14:354
 77. Lütteke T, von der Lieth C-W (2004) pdb-care (PDB CArbohydrate REsidue check): a program to support annotation of complex carbohydrate structures in PDB files. *BMC Bioinformatics* 5:69
 78. Hooft RWW, Vriend G, Sander C et al (1996) Errors in protein structures. *Nature* 381:272
 79. Morris AL, MacArthur MW, Hutchinson EG et al (1992) Stereochemical quality of protein structure coordinates. *Proteins* 12:345–364
 80. Laskowski RA, MacArthur MW, Moss DS et al (1993) PROCHECK: a program to check the stereochemical quality of protein structures. *J Appl Cryst* 26:283–291
 81. Ranzinger RR, Hergert S, Lütteke T et al (2009) Carbohydrate structure databases. In: Cummings RD, Pierce JM (eds) *Handbook of glycomics*. Academic, London, UK, pp 211–233
 82. In proceedings: Charles Warren workshop III. 2010, Hindås, Gothenburg, Sweden, 27–30 Aug

Chapter 22

Characterization of PEGylated Biopharmaceutical Products by LC/MS and LC/MS/MS

Lihua Huang and P. Clayton Gough

Abstract

PEGylation of peptide and proteins is an important method of improving their pharmacokinetic, pharmacodynamic, and immunological profiles, and thus enhancing their therapeutic effect. However, PEGylation of peptides and proteins creates significant challenges for detailed structural characterization, such as PEG heterogeneity, site of PEG addition, and number of attached PEG moieties. Here, we present two methodologies for the structural characterization of PEGylated peptides and proteins. LC/MS methodology utilizing post-column addition of amines was developed to obtain accurate masses of PEGylated peptides and proteins, which can be used to assign the structures and number of attached PEGs. The PEGylated sites in PEGylated products could be elucidated with the tandem LC/MS methodology combining in-source fragmentation with CID-MS/MS. Both methodologies are applied to model PEGylated peptides to obtain the accurate masses and identify PEGylated sites.

Key words: PEGylation, Post-column addition, Diethylmethylamine, Triethylamine, PEGylation site mapping, In-source fragmentation, CID

1. Introduction

Proteins and peptides have experienced tremendous growth in the pharmaceutical industry as potential therapeutics since the inception of recombinant DNA technologies and advances in solid-phase synthetic methods (1, 2). Despite these advances in production capability, the inherent chemical and physical instability of peptides and proteins, as well as their susceptibility to proteolytic degradation and rapid clearance in vivo, complicates formulation development, and generally limits delivery options to daily injections. One approach for extending half-life of proteins and peptides is attachment (PEGylation) of water-soluble polyethylene glycol (PEG) (3).

In addition to improving pharmacological properties, PEGylation offers many potential advantages to a therapeutic biologic including increased solubility, improved chemical and physical stability, and decreased risk of immunogenicity (4–7). Since the first PEGylated biopharmaceutical, “Adagen,” was approved by the FDA in 1990, there have been at least eight PEGylated proteins approved and marketed, and still more in clinical trials, that target a broad range of diseases including cancer, hepatitis, gout, and diabetes (8).

PEGylation of peptides and proteins creates significant challenges for detailed structural characterization. For example, characterizing the PEG heterogeneity, site of addition, and the number of attached PEGs is rarely facile, and yet these types of data can be required for registering these molecules as pharmaceuticals. Unfortunately, there are only a limited number of examples in the scientific literature describing characterization approaches for PEG or PEGylated peptides or proteins by mass spectrometry. This chapter discusses two methodologies for detailed structural characterization of PEG or PEGylated protein or peptide products. One method (9) involves LC/MS analysis coupled with a post-column addition of amines for characterization of large-molecular-weight PEGs and PEGylated peptides or proteins. The second method (10) is related to PEGylation site mapping procedure that is based on direct observation from in-source fragmentation (ISF) in combination with conventional CID MS/MS.

2. Materials

1. Mobile phase A: 0.05% trifluoroacetic acid (TFA) aqueous solution.
2. Mobile phase B: 0.04% TFA acetonitrile.
3. 0.5% triethylamine (TEA) or diethylmethylamine (DEMA) in 50% acetonitrile aqueous solution.
4. 20 kDa PEGylated glucagon: The material was obtained through the reaction of a para-nitrophenylcarbonate-20 kDa PEG (PNP-20 kDa PEG) with glucagon under the pH 7 conditions. The reaction with PNP-20 kDa PEG was performed to target PEGylation to occur at a primary amine, in example the N-terminus of the peptide, and/or the side chain of a lysine residue.
5. Doubly 20 kDa PEGylated peptide: The material was obtained through the reaction of maleimide-20 kDa PEG (MAL-20 kDa PEG) with the sulphydryl group of two cysteine residues.

6. Analytical high-performance liquid chromatography (HPLC) Column: PLRP-S, 2.1×50-mm, 1,000 Å pore size, and 8 µm particle size column (Polymer Laboratories/Varian).
7. Analytical HPLC Column: Acuity ultra-performance liquid chromatography (UPLC) BEH C8, 2.1×50-mm, 1.7 µm particle size column (Waters).
8. Micromass LCT Premier™ electrospray/time-of-flight mass spectrometer (Waters).
9. Synapt™ High Definition MS-ion mobility electrospray/quadrupole time-of-flight tandem mass spectrometer (Waters).
10. HPLC (Waters Alliance Model 2795).
11. UPLC (Waters Acuity UPLC).

3. Methods

The methods described below outline the (1) RP-HPLC/MS analysis with a post-column addition of DEMA for a doubly 20 kDa PEGylated peptide, (2) deconvoluted mass spectrum of intact PEGylated peptide, (3) mass spectral data assignments, (4) RP-HPLC/MS/MS analysis for 20 kDa PEGylated glucagon with ISF and collision-induced dissociation (CID), (5) mass and tandem mass spectra, and (6) mass spectral assignments and PEGylation site identification.

3.1. LC/MS Analysis of Intact PEGylated Peptide

When the PEGylated peptide or protein contains a large PEG (\geq 20 kDa), its reverse-phase retention profile is established on the basis of the PEG rather than the peptide or protein. That is, the retention time of large PEGylated peptide or protein on RP-column is usually longer when the peptide or protein is smaller, which is generally the opposite of the retention behavior of a non-PEGylated peptide or protein. This is because the PEG portion of a large PEGylated product is more hydrophobic than is a peptide or protein (see Note 1). This phenomenon will assist us in assigning the PEGylated product mass spectrum.

1. Equilibrate a Waters Acuity UPLC BEH C8 2.1×50 mm, 1.7 µm particle size column with 5% B at 0.2 mL/min flow, and wait for the column temperature to reach 60°C.
2. Inject approximately 5 µL (containing 0.5 µg peptide) (see Note 2) of doubly 20 kDa PEGylated peptide solution onto the HPLC column.

3. Elute the sample using the following gradient (see Note 3).

Time (min) Flow rate (mL/min) Mobile phase (%) Phase B (%) UV (nm)

0.0	0.200	95.0	5.0	214
2.00	0.200	95.0	5.0	214
8.00	0.200	55.0	45.0	214
8.05	0.200	54.0	46.0	214
18.00	0.200	48.0	52.0	214
22.00	0.200	42.0	58.0	214
22.50	0.400	10.0	90.0	214
23.50	0.400	10.0	90.0	214
24.00	0.500	95.0	5.0	214
28.00	0.500	95.0	5.0	214

4. Route the HPLC effluent through a UV detector, and then to one port of a PEEK static mixing tee mixer (see Note 4).
5. Configure a single HPLC pump or syringe pump to deliver a post-column addition solution (0.5% DEMA or TEA in 50% acetonitrile aqueous solution) at a flow rate of 50–60 µL/min (see Note 5).
6. Make connections routing the tubing containing the post-column addition solution first to a switching valve, and subsequently to the second port on the PEEK static mixing tee (see Note 6).
7. Route the output of third port on the PEEK static mixing tee (combined HPLC effluent and post-column neutralization buffer) and connect to a switch valve and then the Micromass LCT premier mass spectrometer.
8. Set a switch valve program to let HPLC stream to go directly to the mass spectrometer between 2 and 22 min, and go to waste for the remaining times.
9. Set mass spectrometer conditions as shown below (see Note 7):

Electrospray polarity	Positive
Analyzer	V Mode
Capillary (V)	3,200.0
Sample cone (V)	120.0/150.0
Desolvation temperature (°C)	300.0
Source temperature (°C)	90.0
Cone gas flow	30.0

(continued)

Desolvation gas flow	700.0
Aperture 1 voltage	30.0/150.0
Retention window (min)	2.000–9.000
Scan duration (s)	1.00
Mass range	600–12,000
Cycle time (s)	2.010
Scan duration (s)	2.00
Interscan delay (s)	0.01
Retention window (min)	9.000–24.000

3.2. Deconvoluted Mass Spectrum of Intact PEGylated Peptide

The procedure for LC/MS data analysis is to (1) obtain a combined average mass spectrum of doubly 20 kDa PEGylated peptide from a total ion chromatogram (TIC, see Fig. 1); (2) deconvolute the combined average mass spectrum to obtain a zero-charge

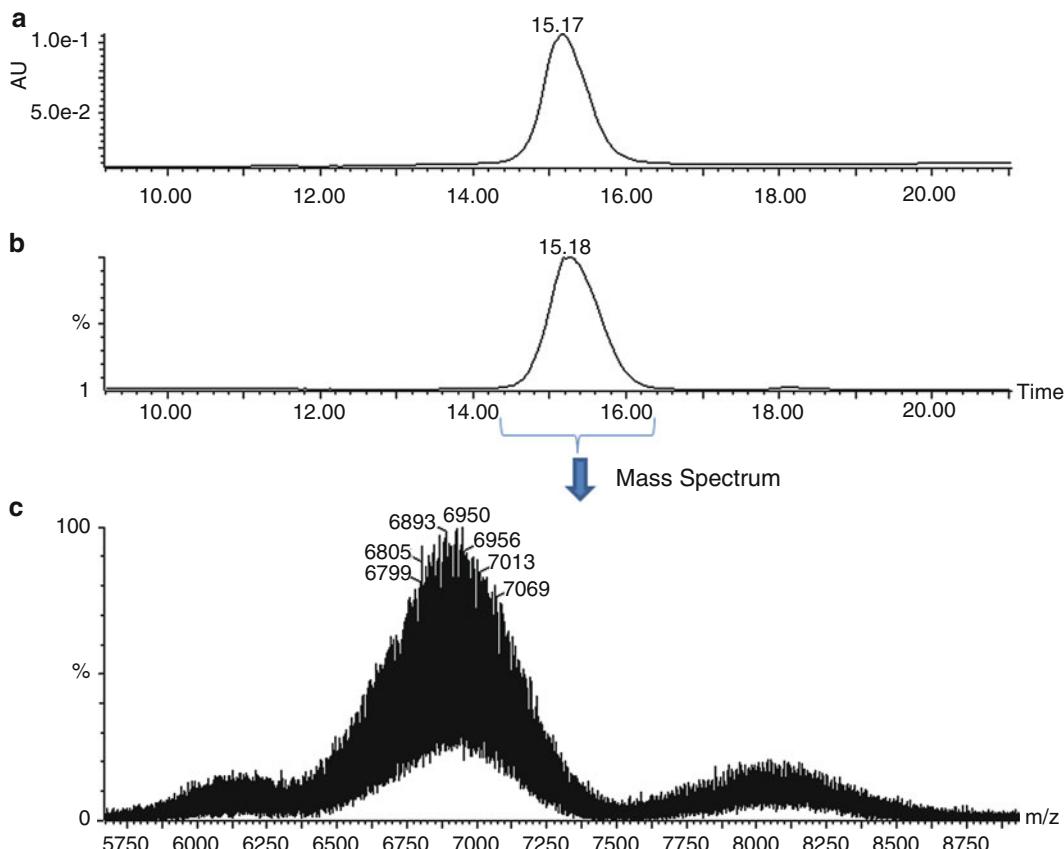


Fig. 1. UV (a) and total ion chromatogram (TIC) (b) of LC/MS analysis for doubly 20 kDa PEGylated peptide with a post-column addition of DEMA and mass spectrum (c) of the main peak.

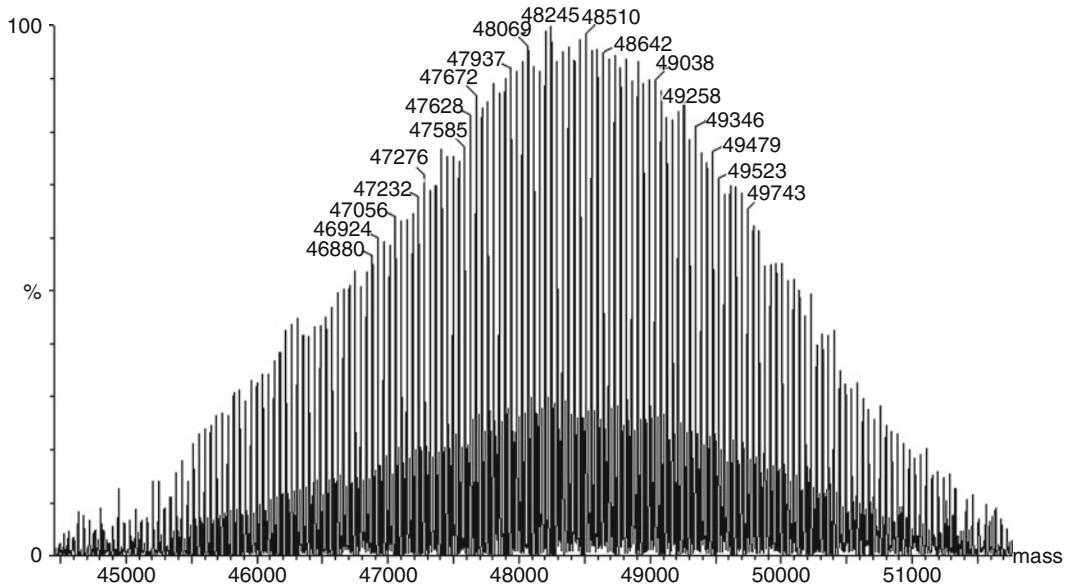


Fig. 2. Deconvoluted mass spectrum of doubly 20 kDa PEGylated peptide with a post-column addition of DEMA.

mass spectrum representing the main peak (Fig. 2); (3) assign mass spectrum; and (4) calculate the polymer properties.

1. Open a TIC (Fig. 1b) containing a main peak representing Doubly 20 kDa PEGylated peptide with “Chromatogram” function available in Waters/Micromass MassLynx™ 4.1 software.
2. Obtain an average mass spectrum that covers the whole main peak (e.g., from 14.5 to 16.2 min).
3. Deconvolute the average mass spectrum with the MaxEnt1 function available in MassLynx™ 4.1 software to give zero-charged mass spectrum for doubly 20 kDa PEGylated peptide with the following conditions (Fig. 2) (see Note 8):

Output Mass: Range, 45,000–52,000; Resolution, 1 Da/channel

Damage Model: Uniform Gaussian with a half height, 5.00

Minimum Intensity Ratio: Left, 15%; Right, 15%

Completion options: Maximum number of interactions, 20

Set Adduct Mass: Hydrogen

3.3. Mass Spectral Data Assignment

More than 100 average masses can generally be detected for a ≥20 kDa PEGylated peptide or protein. The mass heterogeneity is mainly due to the PEG (polymer) containing a different number of oxyethylene (repeating monomer) units. All observed average masses represent the PEGylated peptide or protein with a unique number of monomer units. Thus, the structure of PEGylated

peptide or protein could be assigned using a single mass. As an example, with good mass accuracy, it is possible to calculate the number of oxyethylene units corresponding to a given observed mass present in a deconvoluted mass spectrum. The theoretical mass for the PEGylated peptide or protein can be calculated based on the expected peptide or protein mass with the known PEG structures. Thus, the number of monomer repeating units (again, oxyethylene units) corresponding to a given mass can be obtained as follows:

1. Note a single observed mass in the deconvoluted mass spectrum for the doubly 20 kDa PEGylated peptide: 48,245.8 Da.
2. The expected peptide mass: 5,117.93 Da (average mass).
3. The expected maleimide linker mass for two linkers: $2 \times 209.2249 = 418.450$ Da.
4. The expected mass of two methoxyl groups: $31.0342 \times 2 = 62.0684$ Da.
5. The remaining mass: $48,245.8 - 5,117.93 - 418.45 - 62.07 = 42,647.3$ Da.
6. Oxyethylene units = $42,647.3 / 44.0532 = 968.09 = 968$ monomer units.
7. The calculated mass for doubly 20 kDa PEGylated peptide with 968 oxyethylene units is 48,241.9 Da (see Note 9).
8. From the deconvoluted mass spectrum, one can also obtain the intensity of each observed average mass, for use in quantification when necessary.
9. Since PEG is, by definition, a polymer, several polymer variables, such as number average molecular weight (M_N), weight average molecular weight (M_w), and polydispersity index (PDI), can be calculated once the observed average masses and their intensities of PEGylated peptide or protein with different oxyethylene units are obtained. The definitions and calculations of these polymer variables are readily available in published sources (11–13).

3.4. LC/MS/MS Analysis to Determine PEGylation Site

PEG will be fragmented in source during ionization when LC/MS analysis is performed on a large PEGylated peptide or protein without a post-column addition of amine. The fragmentation of PEG occurs even when very gentle ionization conditions are employed. Thus, if one selects the appropriate mass spectrometric acquisition parameters, PEG ISF is a tool which can be used as a dePEGylation process to aid in elucidation of the PEGylation site.

1. Equilibrate a PLRP-S, 2.1×50 mm, 1,000 Å pore size, and 8 µm particle size column with 5% B at 0.2 mL/min flow rate, and wait for the column temperature to reach 60°C.

2. Inject approximately 50 µL (containing 5.0 µg peptide) of 20 kDa PEGylated glucagon solution onto the HPLC column.
3. Elute the sample using the following gradient (see Note 10).

Time (min) Flow rate (mL/min) Mobile phase (%) Phase B (%) UV (nm)

0.0	0.200	80.0	20.0	214
8.00	0.200	60.0	40.0	214
10.00	0.200	54.0	46.0	214
24.00	0.200	50.0	50.0	214
25.00	0.200	10.0	90.0	214
26.50	0.200	10.0	90.0	214
27.00	0.200	80.0	20.0	214
35.00	0.200	80.0	20.0	214

4. Connect the HPLC stream to a UV detector.
5. Connect HPLC tubing (effluent from UV detector) to a switching valve, then, subsequently, to a Waters Synapt mass spectrometer.
6. Set a switch valve program to direct HPLC stream to flow directly into the mass spectrometer between 2 and 24 min, and then divert to waste for the remainder of the gradient program.
7. Set mass spectrometer conditions as shown:

Polarity	ES+
Analyzer	V Mode
Capillary (kV)	2.5
Sampling cone	60.0
Extraction cone	4.0
Source temperature (°C)	95
Desolvation temperature (°C)	105
Cone gas flow (L/h)	30.0
Desolvation gas flow (L/h)	700.0
LM resolution	5.0
HM resolution	15.0
Transfer collision energy	4.0
Trap gas flow (mL/min)	1.50
Source gas flow (mL/min)	0.00
IMS gas flow (mL/min)	24.00

(continued)

Function parameters—Function 1—TOF MS FUNCTION	
Scan time (s)	1.000
Interscan time (s)	0.020
Start mass	50.0
End mass	2,990.0
Start time (min)	5.00
End time (min)	34.00
Data format	Continuum
Use tune page cone voltage	NO
Cone voltage (V)	60.0
Function parameters—Function 2—TOF MS/MS FUNCTION	
Scan time (s)	1.000
Interscan time (s)	0.020
Set mass	1,396.7
Start mass	50.0
MSMS end mass	2,990.0
Start time (min)	10.00
End time (min)	20.00
Data format	Continuum
Use tune page cone voltage	NO
Cone voltage (V)	60.0
Use tune page collision energy	NO
Collision energy (eV)	50.0

3.5. Mass and Tandem Mass Spectra

The large (>10 kDa) PEG is labile during electrospray mass spectrometry analysis, when LC/MS analysis is conducted in the absence of a post-column neutralization buffer. However, after this ISF of the PEG, the resultant smaller PEG fragments that remain attached to the peptide are stable during CID. Thus, we can elucidate the PEGylation site with CID for a PEGylated peptide after initial ISF (10).

1. Open a TIC (Fig. 3b) containing a main peak of a 20 kDa PEGylated peptide with “Chromatogram” function on MassLynx™ 4.1 software (see Note 11).
2. Obtain an average mass spectrum (Fig. 3c), combining spectra sufficient to cover the whole main peak (i.e., from 12.5 to 15.0 min).
3. Obtain an average mass spectrum (Fig. 3d) combining spectra sufficient to cover the whole main peak (i.e., from 12.5 to 15.0 min) with a background subtraction for this combined spectrum (from 15.0 to 18.0 min).
4. Open a TIC of CID mass spectrum for ion at m/z 1,396.7 with the “Chromatogram” function on MassLynx™ 4.1 software.
5. Obtain an average tandem mass spectrum (Fig. 3e) that covers the whole peak (from 12.5 to 15.0 min) with a background subtraction (from 15.0 to 18.0 min).

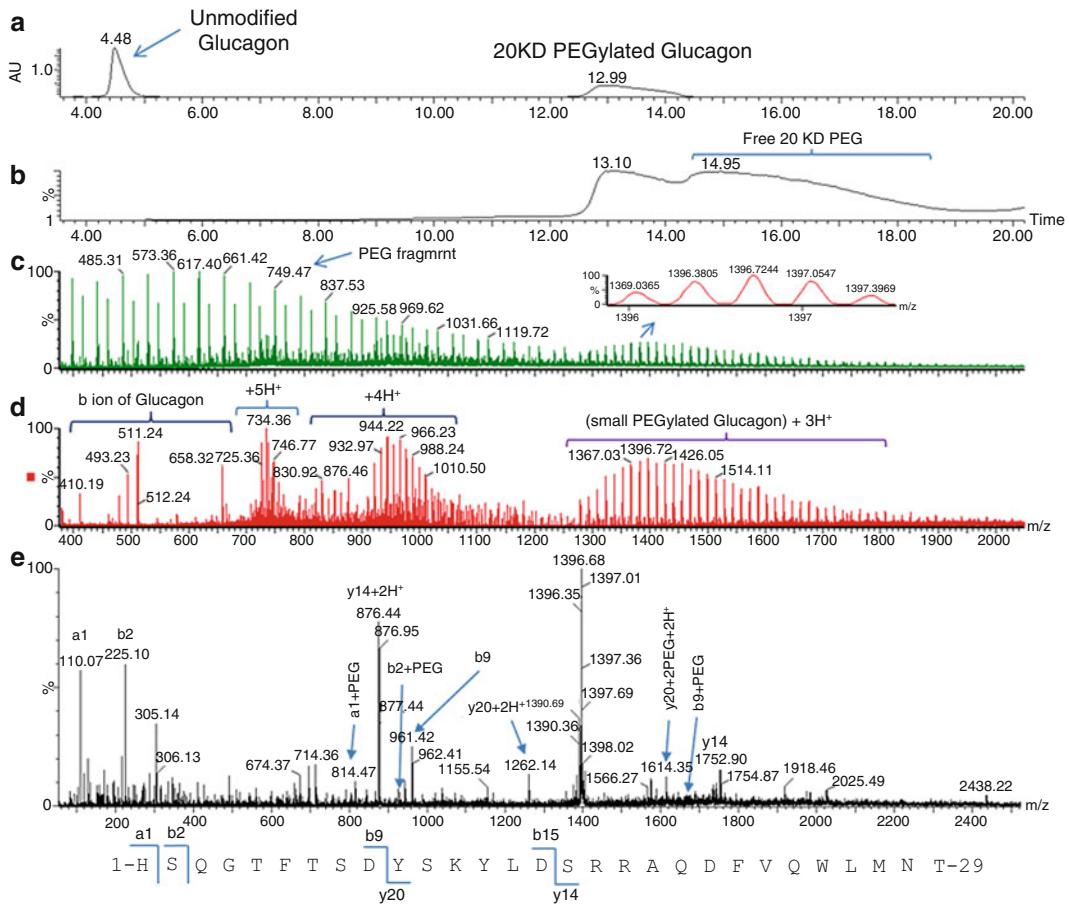


Fig. 3. UV (a) and TIC (b) of LC/MS/MS analysis of 20 kDa PEGylated glucagon with high cone voltage (60 V), mass spectrum (c) of 20 kDa PEGylated glucagon with in-source fragmentation with the cone voltage at 60 V, mass spectrum (d) of 20 kDa PEGylated glucagon with in-source fragmentation with the cone voltage at 60 V after the background subtraction, and tandem mass spectrum (e) the ion at m/z 1,396.7 obtained after in-source fragmentation.

3.6. Mass Spectral Assignments

In general, an intact peptide attached with small PEG after ISF contains multiple charged ions while small PEG itself and peptide fragments are represented by singly charged ions. Thus, one can determine which ions represent the intact peptide attached to a small PEG fragment, even without the specified background subtraction (see Note 12). After the background subtraction, only those ions related to glucagon will be observed. From the mass spectrum, with known mass of the glucagon peptide, it is possible directly to determine the mass of each small PEG attached to the glucagon peptide. For example, there is monoisotopic ion at m/z 1,396.04 (see the inset figure in Fig. 3c) that is a triply charged ion, meaning the intact (non-charged) calculated monoisotopic mass is 4,185.09 Da. Thus, the PEG mass is 704.47 Da (monoisotopic mass of intact glucagon is 3,480.62 Da), which matches 15 oxyethylene units plus the linker (the expected mass: 704.38 Da).

Masses for the other ions can be assigned in the same manner. Once a small PEG mass is determined, PEGylation sites could be elucidated based on the tandem mass spectrum and glucagon sequence.

According to the sequence of glucagon and the active group of PEG, along with PEGylation reaction conditions employed, it is known that two amino acid residues could be attached with PEG moieties through the reaction of a primary amine functional group. One is the histidine 1 (glucagon peptide N-terminus) and the other is the side chain of lysine 12. The tandem mass spectrum of the ion at 1,396.7 (again, small PEG fragment mass is 704 Da) is displayed in Fig. 3e. The major mass assignments are labeled in the figure. The expected a1 and b2 ions without PEG modification are at m/z 110.07 and 225.10 while the ions with the modification are at m/z 814.4 and 929.5 (see Note 13). All the mass ions were detected, indicating that His1 was partially modified. The expected y_{20} ion without and with the modification is at m/z 1,261.1 ($z=2$) and 1,613.3 ($z=2$). Both ions were also detected, indicating that lysine 12 was also partially modified with PEG. These data indicated that both histidine 1 and lysine 12 were partially PEGylated when the PEGylated glucagon was prepared in pH 7 solution.

4. Notes

1. When the peptide or protein is partitioned preferentially in the liquid phase over the solid phase at a given percent of organic solvent in the HPLC mobile phase, a large MW PEG is still partitioned preferentially in the solid phase. The larger peptide or protein will pull PEG from the solid phase to the liquid phase more effectively than will a smaller peptide or protein, which has the net effect that the retention time of large PEGylated product with the smaller peptide or protein is longer.
2. Amount of PEGylated product for LC/MS analysis depends on instrumental sensitivity. If too much material is injected, the obtained mass spectrum will be deviated from Gaussian distribution for some type of instruments (for example, old version of Waters' TOF mass spectrometers).
3. Starting with low percent of acetonitrile is to detect the non-PEGylated peptide or protein in the forced degradation samples.
4. The mixer was placed in tubing post UV detector. In this way, UV detection will not be affected. However, back pressure should be monitored, as excessive back pressure can damage UV detector's flow cell.

5. Once sufficient amine is added to neutralize TFA ion pairing agent, increasing the amount of amine will not further affect the charge distribution of PEGylated product. TEA is more efficient to reduce the charges of PEGylated product (please see ref. 9 for a detailed discussion).
6. The switching valve is included in instrumental setup to enable detection of non-PEGylated peptide or protein. Before the PEGylated product is eluted from the HPLC column, the switching valve can be used to exclude addition of the neutralizing amine to the HPLC stream. The switch valve is not necessary if peptide detection is not required.
7. Typically, the observed mass spectra representing a PEGylated product contain ions $>5,000\text{ }m/z$ after the post-column addition of DEMA or TEA. In order to obtain a good mass spectrum, it can be useful to employ “harsh” ionization conditions with very high Cone and Aperture 1 voltages, and high desolvation temperature to optimize ionization. No PEG or protein fragmentation was detected at the conditions set.
8. There is now a facile, well-documented procedure to obtain a deconvoluted mass spectrum for PEGylated peptide or protein with use of the post-column neutralization buffer. Please see the ref. 9 for the theory and procedure.
9. Under the condition of the post-column addition of DEMA, the protein or peptide portion is still protonated (+1 Da) while PEG portion is adducted with DEMA (88 Da). Hydrogen adduct was selected for MaxEnt1 parameter to deconvolute doubly 20 kDa PEGylated peptide. Thus, the obtained masses are $n \times 87$ Da higher than actual masses. Please see ref. 9 for the detailed explanation of this phenomenon.
10. 20 kDa PEGylated glucagon was prepared in pH 7 solution. The reaction was not complete. Three major components existed in the PEGylated glucagon mixture: glucagon, inactive PEG, and mono-20 kDa PEGylated glucagon. The HPLC gradient was established to resolve these three components well.
11. With the parameters set, an LC/MS raw data file contains three TICs and one UV chromatogram. Function 1 contains a TOF MS TIC with full scan and high cone voltage; Function 2 contains an MSMS with a set ion at m/z 1,396 and high cone voltage. Functions 3 and 4 contain a TOF MS TIC of lock-spray (reference probe) and UV chromatogram.
12. A lot of singly charged PEG ions were generated after PEGylated product was fragmented with in-source fragmentation, which does make the spectrum more complicated. The background spectrum used for subtraction represented the free PEG spectrum with ISF.

13. When PEGylated glucagon was prepared in pH 9 or 10 solution, only lysine 12 was PEGylated. Only 110.07 and 225.10 ions were detected and ions at m/z 814.4 and 929.5 did not exist (see ref. 10). The results indicated that small PEG fragments attached to the peptide were stable passing the CID cell.

References

1. Walsh G (2000) Biopharmaceutical benchmarks. *Nat Biotechnol* 18:831–833
2. Bruckdorfer T, Marder O, Albericio F (2004) From production of peptides in milligram amounts for research to multi-tons quantities for drugs of the future. *Curr Pharm Biotechnol* 5:29–43
3. Abuchowski A, McCoy JR, Palczuk NC, van Es T, Davis FF (1977) Effect of covalent attachment of polyethylene glycol on immunogenicity and circulating life of bovine liver catalase. *J Biol Chem* 252:3582–3586
4. Roberts MJ, Bentley MD, Harris JM (2002) Chemistry for peptide and protein PEGylation. *Adv Drug Deliv Rev* 54:459–476
5. Stigsnaes P, Frokjaer S, Bjerregaard S, van de Weert M, Peter Kingshott P, Moeller EH (2007) Characterisation and physical stability of PEGylated glucagon. *Int J Pharm* 330:89–98
6. Cunningham-Rundles C, Zhou Z, Griffith B, Keenan J (1992) Biological activities of polyethylene-glycol immunoglobulin conjugates resistance to enzymatic degradation. *J Immunol Methods* 152:177–190
7. Chapman AP (2002) PEGylated antibodies and antibody fragments for improved therapy: a review. *Adv Drug Deliv Rev* 54:531–545
8. Veronese FM, Mero A (2008) The impact of PEGylation on biological therapies. *BioDrugs* 22:315–329
9. Huang L, Gough PG, DeFelippis MR (2009) Characterization of poly(ethylene glycol) and PEGylated products by LC/MS with post-column addition of amines. *Anal Chem* 81: 567–577
10. Lu X, Gough PG, DeFelippis MR, Huang L (2010) Elucidation of PEGylation site with a combined approach of in-source fragmentation and CID MS/MS. *J Am Soc Mass Spectrom* 21:810–818
11. Robinson EW, Garcia DE, Leib RD, Williams ER (2006) Enhanced mixture analysis of poly(ethylene glycol) using high-field asymmetric waveform ion mobility spectrometry combined with Fourier transform ion cyclotron resonance mass spectrometry. *Anal Chem* 78: 2190–2198
12. Hanton SD (2001) Mass spectrometry of polymers and polymer surfaces. *Chem Rev* 101:527–569
13. Cooper AR (1989) Determination of molecular weight. In: Winefordner JD (ed) *Chemical analysis*, vol 103. Wiley, New York, p 526

Chapter 23

Identification of Asp Isomerization in Proteins by ^{18}O Labeling and Tandem Mass Spectrometry

Jennifer Zhang and Viswanatham Katta

Abstract

Isomerization of aspartic acid (Asp) to isoaspartic acid (isoAsp) via succinimide intermediate is a common route of degradation for proteins that can affect their structural integrity. As Asp/isoAsp is isobaric in mass, it is difficult to identify the site of modification by LC-MS/MS peptide mapping. Here, we describe an approach to label the Asp residue involved in isomerization at the protein level by hydrolyzing the succinimide intermediate in H_2^{18}O . Tryptic digestion of this labeled protein will result in peptides containing the site of isomerization being 2 Da heavier than the ^{16}O -containing counterparts, due to ^{18}O incorporation during the hydrolysis process. Comparison of tandem mass spectra of isomerized peptides with and without ^{18}O incorporation allows easy identification of the Asp residue involved. This method proved to be especially useful in identifying the sites when isomerization occurs in Asp-Asp motifs.

Key words: Succinimide, Isomerization, ^{18}O incorporation, Tandem mass spectrometry, Asp-Asp motifs

1. Introduction

Isomerization of aspartic acid (Asp) and deamidation of asparagine (Asn) are degradation processes that are difficult to characterize due to the minor changes that occur with the modification (1–3). Isomerization of aspartic acid proceeds via nucleophilic attack by the amide backbone nitrogen on the Asp side chain carbonyl to generate a cyclic imide intermediate (Succinimide, *Suc*). The *Suc* ring is unstable at alkaline pH and can rapidly hydrolyze to the native L-amino acid form or to the isomerized (isoAsp) form, typically at a ratio of approximately 1:3 (4–7). Isoaspartic acid formation is a major source of protein microheterogeneity during storage and, subsequently, may have an impact on protein structure and/or function (8–13). It has been reported that the formation of

isoaspartate sometimes leads to the loss of biological activity (14, 15) and may cause an immunologic response (16–18).

Susceptibility of a particular Asp residue to isomerization depends in part on the chemical nature of the neighboring amino acid on the C-terminal side. Asp-Gly motifs were reported to be involved in majority of the cases, though Asp-Ser motifs were also reported in a few cases (4, 19–22). Other contributing factors are the solvent accessibility and the ability to form stable cyclic intermediates of the particular Asp residue (5, 7, 23). Recently, Asp-Asp motifs in antibodies have been reported to be susceptible to isomerization under mildly acidic conditions (24, 25).

IsoAsp has the same residue mass as the parent aspartic acid, therefore, molecular mass measurement at the protein or peptide level is not sufficient to detect this degradation directly. Such analysis when combined with chromatographic separation may identify the presence of isoAsp in certain peptides, but getting site-specific information is challenging. Edman sequencing of isolated peptides can be used to identify isoAsp linkages indirectly, because the sequencing process stops at the isopeptide bond (26–29). Another approach is the use of protein-L-isoaspartyl methyltransferase (PIMT), an enzyme that specifically methylates the side chains of isoAsp residues only (30) to quantify isoAsp levels in the isolated peptides. Zhang et al. (28) demonstrated that Asp-N protease does not cleave at the N-terminus of isoAsp, and this approach combined with LC/MS analysis of the digest was employed to identify the site of isomerization (31). More recently, O'Connor et al. demonstrated the use of electron capture dissociation (ECD) (32–34) and electron-transfer dissociation (ETD) (35) for the differentiation of isoaspartic acid and aspartic acid residues by the presence of $c+57$ or z^*-57 peaks. However, wider use of these techniques is limited by insufficient sensitivity and the availability of ECD or ETD based instruments.

This chapter describes a method, originally developed by Xiao et al. (7) and Chu et al. (6), to identify the isoAsp in proteins (see Fig. 1). Hydrolyzing the *Suc*-containing protein in ^{18}O water will incorporate ^{18}O in a site-specific manner at the Asp involved in the *Suc* formation. This ^{18}O label on Asp/isoAsp is maintained during the tryptic digestion normally done at pH ~8.3 and during subsequent reverse-phase HPLC peptide mapping. A control experiment is carried out in regular water. The HPLC elution differences of Asp/isoAsp peptides can be complemented with MS/MS data comparison to identify the isoAsp residue in a site-specific fashion. Here we show an application of this method to a protein that was subjected to thermal stress (40°C for 4 weeks at pH 5.4). Under these conditions there was a significant amount of *Suc* formation in this protein (verified by LC/MS analysis of intact and reduced proteins). Detailed step-by-step procedures show how the Asp residues involved in *Suc* formation were identified, even when multiple Asp-Asp motifs were present in close proximity in this protein (25).

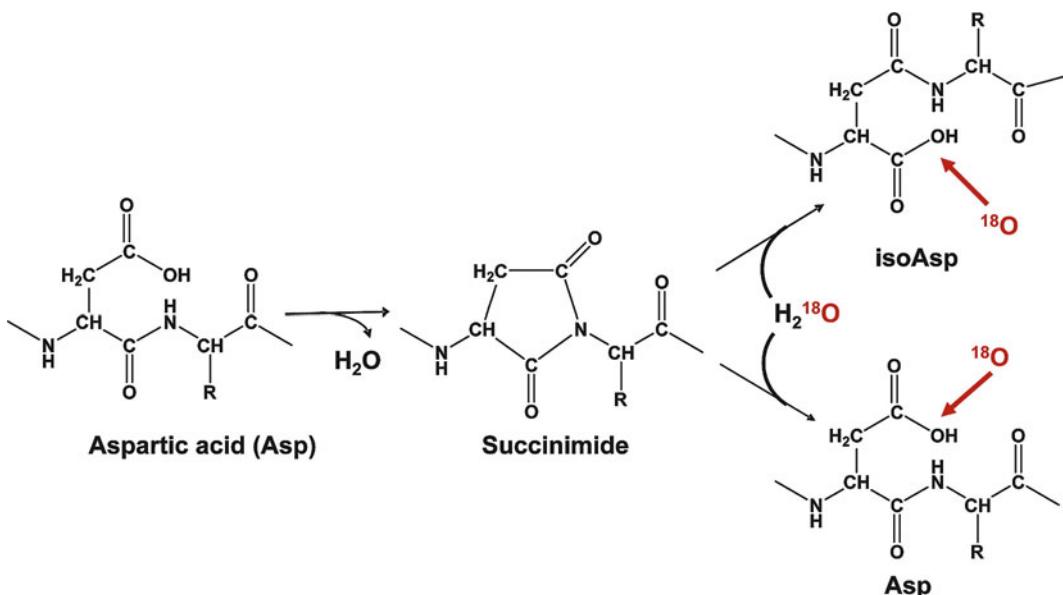


Fig. 1. Scheme of the isomerization of Asp residues and the incorporation of ^{18}O to isoAsp and Asp during succinimide hydrolysis used in this method. Reproduced with permission from ref. 7.

2. Materials

Prepare all solutions using ultrapure water (purify deionized water to attain a conductivity of $18 \text{ M}\Omega^{-1} \text{ cm}$ at 25°C) and analytical grade reagents. Prepare and store all reagents at room temperature (unless otherwise indicated).

2.1. Solutions

1. ^{18}O Labeling buffer: 100 mM NH_4HCO_3 , 8 M guanidine hydrochloride, at pH 8.2:

Dissolve 79.06 mg of ammonium bicarbonate and 7.64 g of guanidine hydrochloride in approximately 8.0 mL of ^{18}O water (purchased from Cambridge Isotope Laboratories, Inc., 97% enriched). Adjust pH to 8.2 by adding small amount of sodium bicarbonate powder using spatula (see Note 1). Add ^{18}O water to 10.0 mL.

2. Denaturing buffer: 6 M guanidine hydrochloride, 0.36 M Tris-HCl, and 2 mM EDTA, pH 8.6:

Add 26.7 g of guanidine hydrochloride, 1.45 g of Tris Base, 0.95 g Tris-HCl, and 2.0 mL of 0.1 M EDTA solution into approximately 950 mL purified water. Stir until dissolved. Adjust pH with 1 N HCl or 1 N NaOH if necessary to 8.6 ± 0.1 .

Add purified water to 1.0 L. Vacuum-filter the buffer through a 0.45- μm nylon membrane.

3. Dithiothreitol (DTT) (1.0 M):
Dissolve 0.154 g of DTT into 1.0 mL of purified water. Prepare fresh.
4. Iodoacetic acid (IAA) (3.5 M):
Dissolve 0.651 g of IAA into 1.0 mL of 1 N NaOH. Prepare fresh and protect from light.
5. Trypsin digestion buffer: 10 mM Tris–HCl, 0.1 mM CaCl₂, pH 8.3:
Add 1.2 g of Tris base and 0.1 mL of 1 M calcium chloride to approximately 950 mL purified water. Stir until dissolved. Adjust pH with 1 N HCl or 1 N NaOH if necessary to 8.3±0.1. Add purified water to 1.0 L. Vacuum-filter the buffer through a 0.45-μm nylon membrane.
6. Trypsin solution:
Reconstitute the lyophilized trypsin (100 μg/vial from Promega) to a concentration of 1 μg/μL by adding 100 μL of 50 mM acetic acid into one vial of trypsin.
7. 10% Trifluoroacetic acid (TFA).
8. Acetonitrile (Baker, Phillipsburg, NJ).

2.2. Other Supplies

1. PD-10 columns, Sephadex® G-25 medium (GE Healthcare).
2. Microcentrifuge tubes, polypropylene, natural, with attached cap, 1.5 mL.
3. Sterile Falcon tubes, polypropylene, round-bottom tube (snap cap), 5.0 mL.
4. Polypropylene autosampler vials with caps.
5. pH paper (pH 7.5–9.5).
6. Bottle-top filters, 0.45 μm, cellulose acetate or nylon.

2.3. Equipment

1. Agilent 1200 capillary HPLC system (Agilent, Palo Alto, CA).
2. LTQ Orbitrap mass spectrometer (Thermo Fisher Scientific, Bremen, Germany) equipped with an IonMax electrospray ionization source.
3. Water bath capable of heating at 37±2°C.
4. SpeedVac centrifuge.
5. pH meter.

2.4. Software

1. Mascot search engine (version 2.1.0, Matrix Science, London, UK).

3. Methods

Carry out all procedures at room temperature unless otherwise specified.

3.1. ^{18}O Incorporation During Succinimide (Suc) Hydrolysis

1. Dry 1 mg of control ($t=0$) and thermally stressed (4 weeks at 40°C) protein samples in formulation buffer (pH 5.4) using a SpeedVac centrifuge and resolubilize in 100 μL of ^{18}O water.
2. Repeat step 1 once and dry samples again to ensure the removal of the residual ^{16}O water completely.
3. Resolubilize the dried samples in 0.5 mL of ^{18}O labeling buffer (100 mM ammonium bicarbonate, 8 M guanidine hydrochloride, pH 8.2 in ^{18}O water).
4. Incubate the mixtures at 37°C in water bath overnight.
5. Dry down the resulting mixtures by SpeedVac centrifuge.

3.2. Reduction, Alkylation, and Tryptic Digestion

1. Following ^{18}O incorporation, suspend the dried samples in 1.0 mL of denaturing buffer (Subheading 2.1, step 2).
2. Add 10 μL of 1.0 M DTT solution to the samples then incubate the solution at 37°C for 1 h.
3. After adding 12 μL of the freshly prepared 3.5 M IAA solution, place the samples at room temperature for 20 min in the dark.
4. Add 30 μL of 1 M DTT to quench the excess IAA immediately. Vortex briefly. Stand the sample vials in a dark area at room temperature for 5 min.
5. Buffer exchange all the samples into 2 mL of digestion buffer (Subheading 2.1, step 5) on a PD-10 column using the following steps. Equilibrate PD-10 columns with at least 25 mL of trypsin digest buffer. Load the entire volume of samples onto equilibrated columns and discard eluent. Add 1.5 mL of digest buffer and discard eluent. Add 2.0 mL of digest buffer and collect eluent in a 5-mL sterile Falcon tube. Vortex briefly.
6. Aliquot 1.0 mL of each collected sample into separately labeled sterile Falcon tubes.
7. Add 25 μL of trypsin solution (1 $\mu\text{g}/\mu\text{L}$) to each aliquot. Vortex briefly. Incubate in a 37°C water bath for 4 h.
8. Quench the digestion by adding 15 μL of 10% TFA to each tube.
9. Store the digests at -70°C prior to liquid chromatography tandem mass spectrometry (LC-MS/MS) analysis.

Dilute 1 mg of same protein samples (control and 4 week thermally stressed) without ^{18}O incorporation to 1.0 mL with denaturing buffer. Perform trypsin digestion by following steps 2–9 in this Subheading 3.2 described above.

3.3. Liquid Chromatography Tandem Mass Spectrometry Analysis of Tryptic Peptides

1. Separate the tryptic peptides on a Jupiter C18 column using a capillary HPLC system with optimized HPLC system and running conditions (see Note 2).
2. Connect the effluent from the HPLC directly to the electrospray ionization source of LTQ Orbitrap mass spectrometer operating in a positive ion mode with the following parameters. The spray voltage is 4.5 kV, and the capillary temperature is 250°C. The mass spectrometer is operated in the data-dependent fashion to switch automatically between MS and MS/MS modes (see Note 3).
3. Acquire survey full scan MS spectra from m/z 300 to m/z 2,000 in the FT-Orbitrap with a resolution set for $R=60,000$ at m/z 400.
4. Fragment the five most intense ions in the linear ion trap using collision-induced dissociation (CID) at normalized collisional energy of 35% with an activation time of 30 ms and isolation width of 2.5 m/z units.
5. Enable the dynamic exclusion (DE) function to reduce data redundancy and allow low-intensity ions to be selected for data-dependent MS/MS scans. The dynamic exclusion parameters are as follows: a repeat duration of 30 s, an exclusion list size of 500, an exclusion duration of 90 s, a low exclusion mass width 0.76, a high exclusion mass width of 1.56, and a repeat count of 2.

3.4. Data Analysis

1. Submit all acquired LC-MS/MS data of digested protein samples without ^{18}O incorporation to a Mascot search engine.
2. Identify peptides using a Mascot algorithm (see Note 4) by comparing acquired mass spectral data with theoretical parent and fragment ions predicted for the recombinant protein sequence in the database.
3. Verification of the presence of isoAsp-containing peptides:
 - Generate extracted ion chromatograms (EIC) for the molecular ions (most abundant charge state) with a mass tolerance of ± 5 ppm. This should be done for each Asp-containing peptide separately. This will ensure that corresponding isoAsp peptides, if present, are clearly seen to elute earlier (see Note 5) (Fig. 2).
 - If peaks in addition to the native peptide are observed in extracted ion chromatogram, then compare the corresponding CID spectra to ensure that the new peak in EIC really corresponds to isoAsp-containing peptide. CID spectra of Asp and isoAsp-containing peptides should be similar (see Note 6).

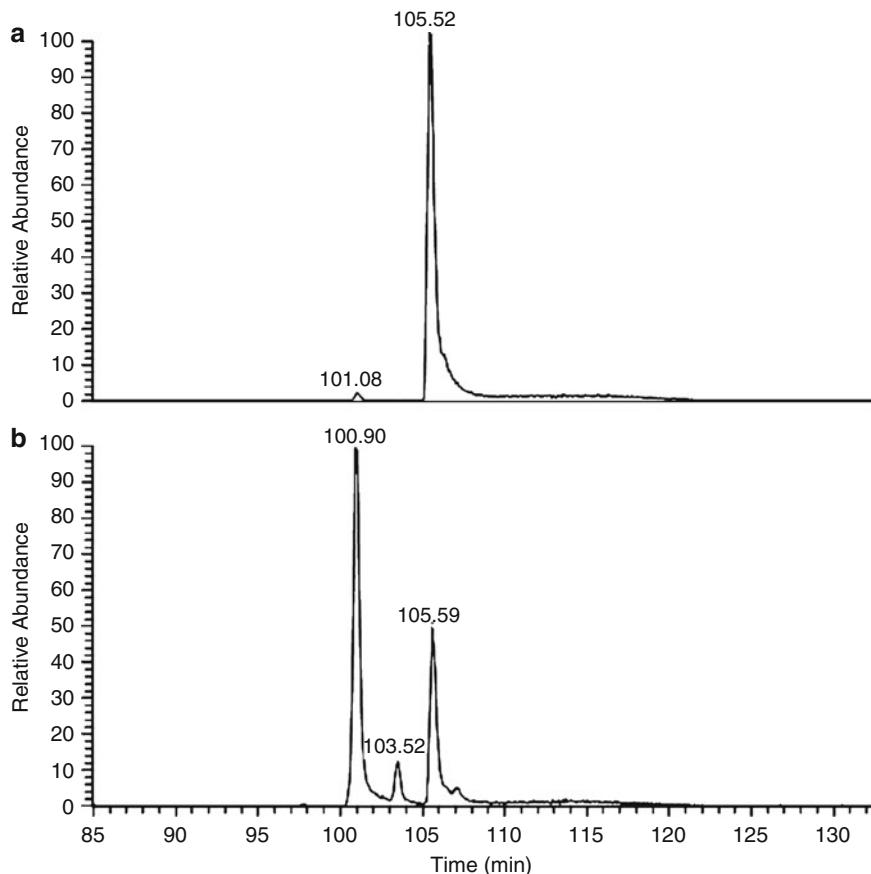


Fig. 2. Extracted ion chromatograms of m/z 1,415.6491 (corresponding to 2+ ion of peptide L2) from the LC-MS/MS analysis of the tryptic digests of (a) the control; (b) 4 week thermally stressed samples. The mass tolerance window is ± 5 ppm. Modified with permission from ref. 25.

4. Identifying the specific Asp involved in isomerization using the ^{18}O -labeled protein digest (Subheading 3.2):

- Using the retention time of the isoAsp-containing peptide as a guide, compare the molecular ion regions in the full scan spectra. A mass shift of +2 Da is to be observed (see Note 7) (Fig. 3).
- Compare the CID spectra of native and ^{18}O -incorporated peptides. The fragment ions that do not contain the isoAsp will not shift. Fragment ions containing the isoAsp will show a +2 Da shift (see Note 8) (Fig. 4).
- For any Asp-containing peptide, if the EIC with a mass tolerance of ± 5 ppm shows more than one peak (in addition to the native peptide), there is a possibility that multiple Asp residues in the same peptide are involved in isomerization or racemization. Comparison of the product ion

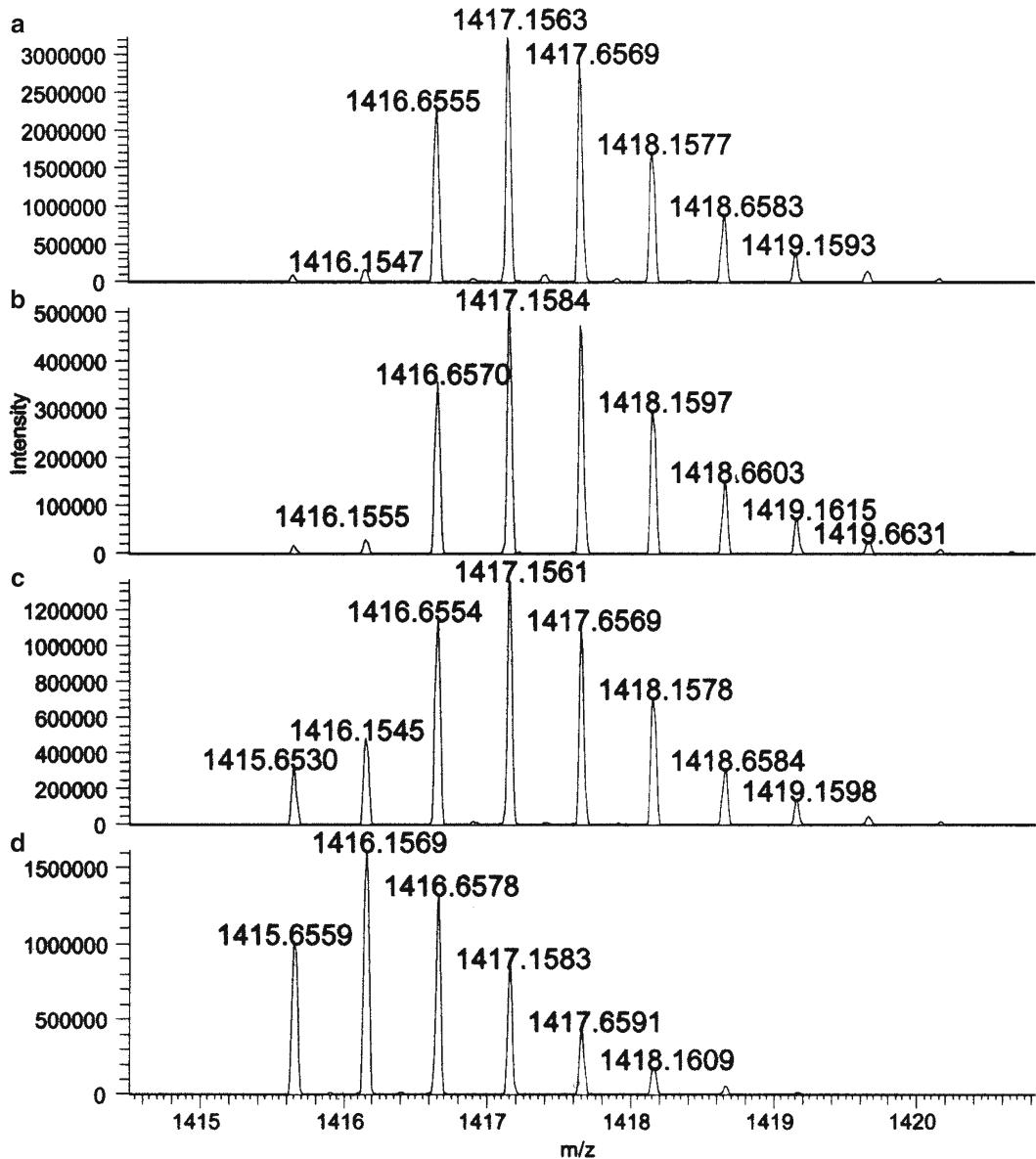


Fig. 3. Detailed mass spectra regions around the doubly charged ion of peptide isoL2 eluting at (a) 101 min; (b) 104 min; (c) peptide L2 eluting at 106 min after ^{18}O incorporation in the 4 week thermally stressed sample. (d) Mass spectrum of doubly charged ion of peptide L2 eluting at 106 min in the same sample after tryptic digestion without ^{18}O incorporation. Reproduced with permission from ref. 25.

spectra should indicate if multiple Asps are involved. If only one Asp residue is involved, some racemization may be taking place. This phenomenon can be verified by co-injecting synthetic standards (see Note 9 for the details about the model protein studied).

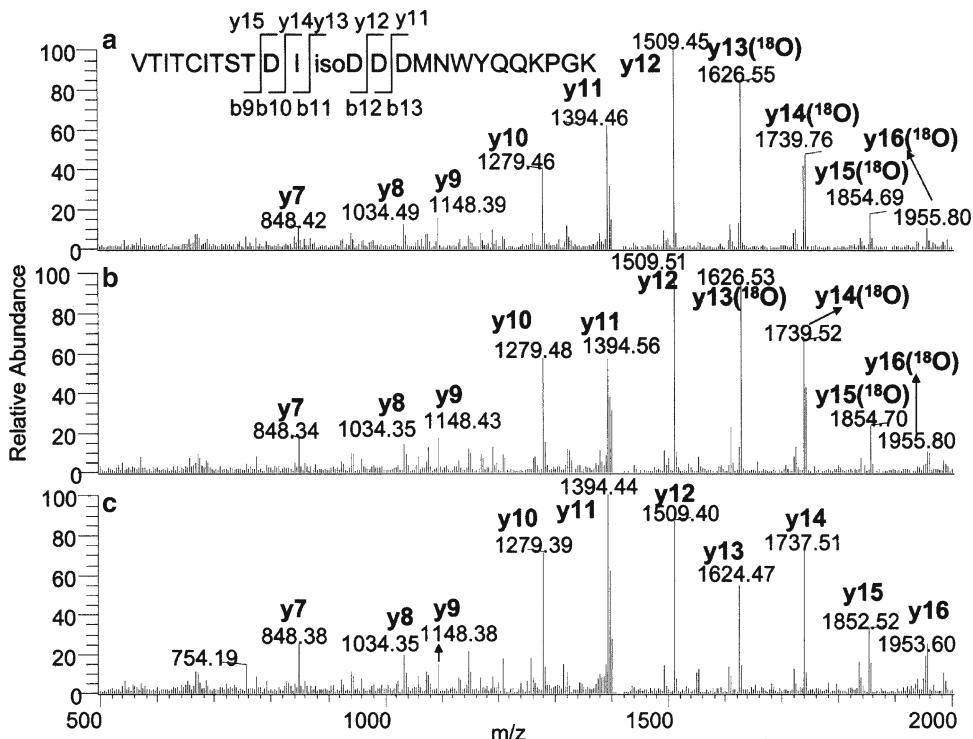


Fig. 4. Tandem mass spectra for the isoAsp peptide L2 VTITCITSTDIDDDMNWYQQKPGK from the 4 week thermally stressed sample. (a) With ^{18}O incorporation eluting at 101 min; (b) with ^{18}O incorporation eluting at 104 min; (c) without ^{18}O incorporation. Data were obtained by collision-induced dissociation of the $(\text{M} + 2\text{H})^{2+}$ precursor ion (m/z 1,416) in the linear ion trap. ^{18}O incorporation into isoAsp was revealed by the mass spectra that showed y_{13} -ions (containing isoAsp) in both (a) and (b) were 2 Da heavier than the corresponding y_{13} -ions in (c). The y_{12} -ions (not containing isoAsp) had the same m/z in all of (a), (b), and (c). Reproduced with permission from ref. 25.

4. Notes

1. Add small amount of sodium bicarbonate powder into ^{18}O labeling buffer using spatula and dissolve it completely. Pipette 2 μL of buffer onto pH paper (pH 7.5–9.5) to measure the buffer pH. Repeat this process until the pH is achieved at 8.2.
2. The HPLC conditions described here are specific for this molecule studied. The user may use different conditions (column, gradient, and solvent that are compatible with ESI-MS) optimized for separating isoAsp- and Asp-containing peptides. The HPLC column temperature is maintained at 55°C. The flow rate is 65 $\mu\text{L}/\text{min}$. 30 μL of protein digest is injected into the column for analysis. Solvent A is 0.1% TFA in water, and solvent B is 0.09% TFA, 90% acetonitrile in water. Before sample injection, equilibrate the column with 99.9% solvent A. An optimized 148-min step gradient is used as

follows (min/%B): 0/0.1, 3/0.1, 111/30, 148/43, 149/100, 154/100, 155/0.1. The total HPLC run time is 175 min. The UV detector is set at 214 nm. The effluent is diverted to waste for the first 2 min to keep the ESI source clean.

3. Mass spectrometric conditions described here are specific to Orbitrap with Ion Max electrospray source. Users may use different systems/conditions that can clearly provide high-resolution mass spectra in MS mode and good fragmentation efficiency for the peptide of interest.
4. The following parameters are used in a Mascot matching search: a peptide tolerance of 10 ppm, an MS/MS tolerance of 0.8 Da, a maximum missed cleavage of 2, carboxymethylation of cysteine as the fixed protein modification, methionine oxidation as the variable modification, fully tryptic as cleavage specificity for tryptic digestion. The probability-based ion score is set for ≥ 25 for a significant match.
5. Typically, a peptide with isoAsp residues would elute earlier than that with aspartyl residues in reversed-phase chromatography (4). As Asp/isoAsp residues have same mass, evidence for isomerization is derived from examining the EIC for every Asp-containing peptide. For example, Fig. 2 shows the EIC of a peptide VTITCITSTDIDDDMNWYQQKPGK (peptide L2) at m/z 1,415.6491(+2) with a mass tolerance of ± 5 ppm in the control and thermally stressed sample, respectively. For the thermally stressed sample, two new peaks are detected at 101 and 104 min in addition to the native peptide L2 eluting at 106 min. These two early eluting peaks are attributed to peptide L2 with one isoaspartic acid (isoL2).
6. The MS/MS spectra of these two new peaks are similar to that of peptide L2, therefore, the MS/MS data could not be used to identify which Asp residue was involved in isomerization. To overcome this difficulty, we incorporate ^{18}O into the *Suc* hydrolysis products at isoAsp and Asp residue in the protein level (see Subheading 3.1). The protein is then digested with trypsin in regular ^{16}O -containing buffer (see Subheading 3.2). The resulting tryptic peptides will have ^{18}O incorporation into the Asp residues that are involved in *Suc* formation, thus they will be 2 Da heavier than the ^{16}O -containing counterparts.
7. For example, Fig. 3 shows the mass spectra of the doubly charged peptide isoL2 and L2 eluting at 101, 104, and 106 min after ^{18}O incorporation in the thermally stressed sample. Compared to the mass spectrum of peptide L2 in the same sample without ^{18}O incorporation (Fig. 3d), the isotopic distribution of all these isomers show a shift to the heavy side, indicating that ^{18}O is successfully incorporated into peptide L2 and isoL2. The first two peaks (Fig. 3a, b) clearly show the mass

shift of 2 Da suggesting that they contain isoAsp. The peak eluting at 106 min in Fig. 3c is the combination of peptide L2 containing ^{16}O aspartyl residue and ^{18}O aspartyl residue resulting from the *Suc* hydrolysis in ^{18}O buffer. Since the conversion ratio of isoAsp and Asp during hydrolysis is expected to be about 3:1, the ^{18}O incorporation in Asp-containing species is much less than that of the isoAsp-containing peptide.

8. Figure 4 compares the tandem mass spectra of the isoL2 peptides (24 residues in length) eluting at 101 min (see Fig. 4a) and 104 min (see Fig. 4b) after ^{18}O incorporation with that of the L2 peptide without ^{18}O incorporation (Fig. 4c). The y -ions ($y_7, y_8, y_9, y_{10}, y_{11}, y_{12}$) have the same m/z value in spectra for all samples. At the same time the y -ions ($y_{13}, y_{14}, y_{15}, y_{16}$) containing residue 12 are 2 Da heavier than the corresponding y -ions of the peptides from the sample prepared in ^{16}O buffer. The 2-Da increase of y_{13} but not of y_{12} is evidence that ^{18}O is incorporated into residue 12 but not into other sites. Therefore, it is inferred that the isomerization site in both the isoL2 peptides (eluting at 101 and 104 min, respectively) in 4 week thermally stressed samples is at Asp-12.
9. The two isoL2 peptides, both showing residue 12 as isomerized but eluting at 101 and 104 min, are attributed to the simultaneous racemization of isoAsp-12. This is determined by the co-injection of synthesized peptide L2 variants: L-Asp-12, L-isoAsp-12, D-Asp-12, and D-isoAsp-12, with the digest of the 4-week thermally stressed sample, respectively. It is clearly observed that the peak eluting at 101 min corresponds to the peptide L-isoAsp-12, while the 104- peak matches the peptide D-isoAsp-12 (25).

Acknowledgments

The authors gratefully acknowledge Clifford Quan for synthesis of peptide variants and Benson Gikanga for preparing the thermally stressed samples. We would also like to thank Reed Harris, John Stults, Mary Cromwell, and Charles Morgan for their valuable discussion and critical review.

References

1. Wakankar AA, Borchardt RT (2006) Formulation considerations for proteins susceptible to asparagine deamidation and aspartate isomerization. *J Pharm Sci* 95:2321–2336
2. Vlasak J, Ionescu R (2008) Heterogeneity of monoclonal antibodies revealed by charge-sensitive methods. *Curr Pharm Biotechnol* 9:468–481
3. Vlasak J, Bussat MC, Wang S, Wagner-Rousset E, Schaefer M, Klinguer-Hamour C, Kirchmeier M, Corvaia N, Ionescu R, Beck A (2009) Identification and characterization of
- monoclonal antibodies revealed by charge-

- asparagine deamidation in the light chain CDR1 of a humanized IgG1 antibody. *Anal Biochem* 392:145–154
4. Geiger T, Clarke S (1987) Deamidation, isomerization, and racemization at asparaginyl and aspartyl residues in peptides. *J Biol Chem* 262:785–794
 5. Wakankar AA, Borchardt RT, Eigenbrot C, Shia S, Wang YJ, Shire SJ, Liu JL (2007) Aspartate isomerization in the complementarity-determining regions of two closely related monoclonal antibodies. *Biochemistry* 46: 1534–1544
 6. Chu GC, Chelius D, Xiao G, Khor HK, Coulibaly S, Bondarenko PV (2007) Accumulation of succinimide in a recombinant monoclonal antibody in mildly acidic buffers under elevated temperatures. *Pharm Res* 24: 1145–1156
 7. Xiao G, Bondarenko PV, Jacob J, Chu CG, Chelius D (2007) ^{18}O labeling method for identification and quantification of succinimide in proteins. *Anal Chem* 79:2714–2721
 8. Manning MC, Patel K, Borchardt RT (1989) Stability of protein pharmaceuticals. *Pharm Res* 6:903–918
 9. Clarke S, Stephenson RC, Lowenson JD (1992) Liability of asparagine and aspartic acid residues in proteins and peptides; spontaneous deamidation and isomerization reactions. In: Ahern TJ, Manning MC (eds) *Stability of protein pharmaceuticals: Part A, Chemical and physical pathways of protein degradation*. Plenum, New York, pp 1–29
 10. Liu DT (1992) Deamidation: a source of microheterogeneity in pharmaceutical proteins. *Trends Biotechnol* 10:364–369
 11. Cleland JL, Powell MF, Shire SJ (1993) The development of stable protein formulations: a close look at protein aggregation, deamidation, and oxidation. *Crit Rev Ther Drug Carrier Syst* 10:307–377
 12. Aswad DW (1995) Deamidation and isoaspartate formation in peptides and proteins. CRC, Boca Raton
 13. Powell MF (1996) A compendium and hydropathy/flexibility analysis of common reactive sites in proteins: reactivity at Asn, Asp, Gln and Met motifs in neutral pH solution. In: Pearlman R, Wang YJ (eds) *Formulation, characterization, and stability of protein drugs: case histories*. Plenum, New York, pp 1–140
 14. Cacia J, Keck R, Presta LG, Frenz J (1996) Isomerization of an aspartic acid residue in the complementarity-determining regions of a recombinant antibody to human IgE: identification and effect on binding affinity. *Biochemistry* 35:1897–1903
 15. Harris RJ, Kabakoff B, Macchi FD, Shen FJ, Kwong M, Andya JD, Shire SJ, Bjork N, Totpal K, Chen AB (2001) Identification of multiple sources of charge heterogeneity in a recombinant antibody. *J Chromatogr B* 752:233–245
 16. Doyle HA, Gee RJ, Mamula MJ (2003) A failure to repair self-proteins leads to T cell hyperproliferation and autoantibody production. *J Immunol* 171:2840–2847
 17. Doyle HA, Zhou J, Wolff MJ, Harvey BP, Roman RM, Gee RJ, Koski RA, Mamula MJ (2006) Isoaspartyl posttranslational modification triggers anti-tumor T and B lymphocyte immunity. *J Biol Chem* 281:32676–32683
 18. Yang ML, Doyle HA, Gee RJ, Lowenson JD, Clarke S, Lawson BR, Aswad DW, Mamula MJ (2006) Intracellular protein modification associated with altered T cell functions in autoimmunity. *J Immunol* 177:4541–4549
 19. Stephenson RC, Clarke S (1989) Succinimide formation from aspartyl and asparaginyl peptides as a model for the spontaneous degradation of proteins. *J Biol Chem* 264:6164–6170
 20. Oliyai C, Borchardt RT (1993) Chemical pathways of peptide degradation IV: pathways, kinetics, and mechanism of degradation of an aspartyl residue in a model hexapeptide. *Pharm Res* 10:95–102
 21. Radkiewicz JL, Zipse H, Clarke S, Houk KN (2001) Neighboring side chain effects on asparaginyl and aspartyl degradation: an ab initio study of the relationship between peptide conformation and backbone NH acidity. *J Am Chem Soc* 123:3499–3506
 22. Oliyai C, Borchardt RT (1994) Chemical pathways of peptide degradation. VI. Effect of the primary sequence on the pathways of degradation of aspartyl residues in model hexapeptides. *Pharm Res* 11:751–758
 23. Potter SM, Henzel WJ, Aswad DW (1993) In-vitro aging of calmodulin generates isoaspartate at multiple asn-gly and asp-gly sites in calcium-binding domain-ii, domain-iii, and domain-iv. *Protein Sci* 2:1648–1663
 24. Xiao G, Bondarenko PV (2008) Identification and quantification of degradations in the Asp-Asp motifs of a recombinant monoclonal antibody. *J Pharm Biomed Anal* 47:23–30
 25. Zhang J, Yip H, Katta V (2011) Identification of isomerization and racemization of aspartate in the Asp-Asp motifs of a therapeutic protein. *Anal Biochem* 410:234–243
 26. Donato AD, Ciardiello MA, de Nigris M, Piccoli R, Mazzarella L, D'Alessio G (1993) Selective deamidation of ribonuclease A. Isolation and characterization of the resulting

- isoaspartyl and aspartyl derivatives. *J Biol Chem* 268:4745–4751
27. Kwong MY, Harris RJ (1993) Identification of succinimide sites in proteins by N-terminal sequence analysis after alkaline hydroxylamine cleavage. *Protein Sci* 3:147–149
28. Zhang W, Czupryn JM, Boyle PT Jr, Amari J (2002) Characterization of asparagine deamidation and aspartate isomerization in recombinant human interleukin-11. *Pharm Res* 19: 1223–1231
29. Sadakane Y, Yamazaki T, Nakagomi K, Akizawa T, Fujii N, Tanimura T, Kaneda M, Hatanaka Y (2003) Quantification of the isomerization of Asp residue in recombinant human alpha A-crystallin by reversed-phase HPLC. *J Pharm Biomed Anal* 30:1825–1833
30. Johnson BA, Shirokawa JM, Hancock WS, Spellman MW, Basa LJ, Aswad DW (1989) Formation of isoaspartate at two distinct sites during in vitro aging of human growth hormone. *J Biol Chem* 264:14262–14271
31. Rehder DS, Chelius D, McAuley A, Dillon TM, Xiao G, Crouse-Zeineddini J, Vardanyan L, Perico N, Mukku V, Brems DN, Matsumura M, Bondarenko PV (2008) Isomerization of a single aspartyl residue of anti-epidermal growth factor receptor immunoglobulin $\gamma 2$ antibody highlights the role avidity plays in antibody activity. *Biochemistry* 47:2518–2530
32. Cournoyer JJ, Pittman JL, Ivlevaver AB, Fallows E, Waskell L, Costello CE, O'Connor PB (2005) Deamidation: differentiation of aspartyl from isoaspartyl products in peptides by electron capture dissociation. *Protein Sci* 14:452–463
33. Cournoyer JJ, Lin C, O'Connor PB (2006) Detecting deamidation products in proteins by electron capture dissociation. *Anal Chem* 78:1264–1271
34. Cournoyer JJ, Lin C, Bowman MJ, O'Connor PB (2007) Quantitating the relative abundance of isoaspartyl residues in deamidated proteins by electron capture dissociation. *J Am Soc Mass Spectrom* 18:48–56
35. Chan WYK, Chan TWD, O'Connor PB (2010) Electron transfer dissociation with supplemental activation to differentiate aspartic and isoaspartic residues in doubly charged peptide cations. *J Am Soc Mass Spectrom* 21:1012–1015

Chapter 24

Monitoring of Subvisible Particles in Therapeutic Proteins

Satish K. Singh and Maria R. Toler

Abstract

The unintended presence of particulate matter in injectable products is an indicator of the quality of the product. Subvisible particulates have historically been monitored through methods such as light obscuration and membrane microscopy, as outlined in the United States Pharmacopeia (USP) General Chapter <788> or the equivalent Ph Eur 2.9.19 and the Japanese Pharmacopeia General Chapter 20. These methods were designed to protect patients against the risk of capillary occlusion through the infusion of “foreign” particulate matter. With the development and commercialization of protein therapeutics, the application of these methods has to be adapted to the special requirements posed by such products. Apart from the “foreign” particulates, therapeutic protein products may also contain particulates that are inherent to the product, arising as a consequence of protein self-association or aggregation. The nature of these inherent “proteinaceous” particulates is generally different than the traditional “foreign” particulates. Proteinaceous particulates tend to be of an amorphous irregular morphology, soft, with a refractive index resulting in low contrast against an aqueous background, making them more difficult to detect and count compared to the “foreign” particulates. The growing realization of the importance of monitoring all (foreign as well as inherent) particulates in therapeutic protein products has led to a number of developments in this area including techniques of measurement. Here, we summarize a number of the procedures used for subvisible particle measurements, with new techniques as well as refinements to existing techniques.

Key words: Subvisible particulates, Parenterals, Injectables, Light obscuration, HIAC, Optical microscopy, Membrane microscopy, Flow imaging, Dynamic imaging, Micro-Flow imaging, Electrozone sensing, Coulter counter, Single particle tracking, NanoSight

1. Introduction

Subvisible particles are a critical quality attribute for pharmaceutical products that are administered by injection, i.e., parenterals or injectables. The ability to measure (detect and count) such particles is important during all stages of product lifecycle from development to commercialization. With protein therapeutics, the usual

category of “foreign” particulates has to be expanded to include particulates inherent to the product, arising from protein self-association or aggregation. The nature of these inherent “proteinaceous” particulates is generally different than the traditional “foreign” particulates. Proteinaceous particulates tend to be of an amorphous irregular morphology, soft, with a refractive index resulting in low contrast against an aqueous background, making them more difficult to detect and count compared to the “foreign” particulates. The growing realization of the importance of monitoring all (foreign as well as inherent) particulates in therapeutic protein products has led to a number of developments in this area as well as refinements in techniques of measurement (1, 2). This chapter covers some established techniques as well as some new techniques for these measurements, with particular emphasis on the measurement and detection of proteinaceous particles.

The term “subvisible” implies particulates that are not visible. However, since the visual detection of particles is probabilistic dependent upon size and number, a universal definition of what is visible versus subvisible is not available. However, for most discussions, 100 µm is taken as an upper limit for subvisible particles (3).

Since this chapter covers a number of techniques, we have provided a compare and contrast guide to the use of these methods in Table 1.

Some general aspects of subvisible particle measurement that are important for an analyst to be aware of are provided in Notes 1–3.

Apart from the methods covered in this chapter, there are methods based on laser diffraction and light scattering that can also be used for measuring particles. However, these methods do not provide a proper measure of particle numbers and have not been covered here.

1.1. Light Obscuration Particle Analysis

Among the many methods available for monitoring subvisible particulates in protein products, the most commonly accepted and applied method is based on the principle of Light Obscuration (LO). An example of the implementation of this principle is the Hach instrument. As the name suggests, the particles are counted based upon the obscuration of light as they pass in front of a detector, while the change in intensity is correlated to their size (illustrated in Fig. 1). This method is also referred to as Method 1 in the pharmacopeia.

Obscuration of light is a function of the difference in refractive index of the particle and the aqueous background, which provides the optical contrast required for the “shadow” to be cast. Particles that arise from the manufacturing environment or precipitates of buffer salts generally have good contrast and are easily counted. Similarly, silicone oil droplets (e.g., from stopper coatings or in

Table 1
Summary comparison of various methods to measure subvisible particles in protein therapeutics

Method	Advantages/benefits	Disadvantages/limitations
Light obscuration	Primary pharmacopeial method. Readily available instrumentation. Easy to use.	Low maximum count limit. Lower size range 2 µm. May not be sensitive to low contrast small proteinaceous particles, thus undercounting them.
Optical microscopy/ membrane microscopy	Pharmacopeial method.	Counting requires particles to have good contrast against membrane. Proteinaceous particles are likely to be optically transparent and significantly undercounted.
Flow image analysis	Relatively new technique in pharmaceutical industry. Good orthogonal technique for measurement and characterization. Provides ability to perform morphological analysis. Small measurement volume makes it useful for development work. Images can be saved for re-analysis.	Counts depend on image analysis algorithms. Small measurement volume can be a limitation for ensuring representative counting of sample.
Electrozone sensing/Coulter counter	Good resolution. Size range down to 0.4 µm. Dilution allows small sample volume to be used.	Requires sample to be in an electrolyte solution. Dilution, if required, into electrolyte solution may impact proteinaceous particles. Aperture must be changed to cover a wide particle size range.
Submicron particle tracking/ NanoSight	Size measurement down to 30 nm. Primarily for characterization.	Counts are calculated based on number of particles detected. Counting is not the primary measurement.

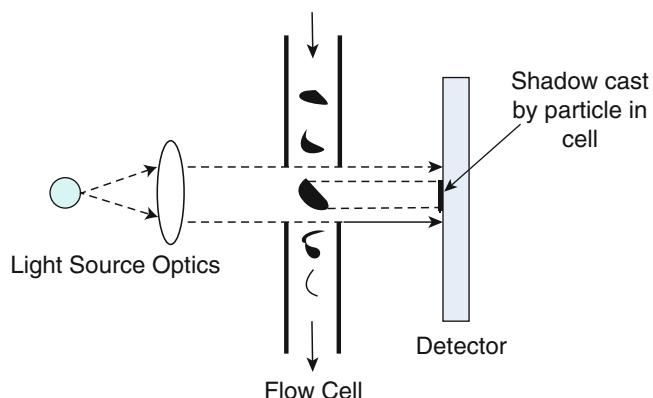


Fig. 1. Schematic representation of particle measurement by Light Obscuration.

prefilled syringes) have a refractive index that allows them to be counted. On the other hand, proteinaceous particles tend to have low contrast and are generally undercounted by this method, especially as the size gets smaller.

The commonly available LO instrument (HIAC) saturates at counts approx. 18,000 particles/mL. Saturation in one channel can cause nonlinear effects in other channels also. If such a situation arises, dilution of the sample will be required, in which case the precautions mentioned above must be considered to ensure that dilution does not itself alter the counts.

There has been an increasing interest in obtaining particle count information on particles smaller than the 10 μm size traditionally prescribed by the pharmacopeia (4–6). There are physical limitations that constraint the practically achievable lower limit. With the HIAC instrument, a lower detection of 2 μm is readily obtained. The procedure described includes the instrument setup to allow for the collection of particle count information down to 2 μm . If collecting only the 10 and 25 μm information, as outlined in USP, then the current USP SOP can be run in the HIAC instrument (Hach Company, Loveland, CO). If collection of data below 10 μm is desired, then the Run Counter manual analysis option must be used. The procedure below is written to allow for using the Run Counter option.

The current harmonized pharmacopeial methods require the use of a total sample volume of 25 mL. For small-volume injectables of less than 25 mL in volume, the contents of ten or more units may be combined in a clean container to obtain the requisite volume. The test solution may also be prepared by pooling the contents of a suitable number of test articles and diluting to 25 mL with a suitable diluent (e.g., particle-free water). However, the dilution scheme must be verified and qualified. For small-volume injectables greater than 25 mL in volume, units may be tested individually. For large-volume parenteral, single units are tested.

For therapeutic proteins, these volumes may not be amenable in all cases (e.g., high concentration and/or low volume products; during development, investigations, etc.). In these situations, a small-volume method can be adopted. The procedure described in this chapter is given for a 5 mL total sample volume with a 1 mL measurement volume.

1.2. Optical Microscopy/Membrane Microscopy

A second method for monitoring subvisible particles is the membrane microscopy method (also referred to as Method 2), as outlined in the pharmacopeias (4–6). This method is recommended when the primary method based on light obscuration (i.e., Method 1) is not feasible or as a confirmatory test if the sample does not pass the light obscuration test. This method is also applicable to small and large volume injectables. As with light obscuration testing, a clean environment is essential to successful testing.

The membrane microscopy method is well suited to measure and count particulates in parenteral products when they provide a good contrast against the membrane. Sizing is performed on a calibrated ocular micrometer called the Circular Diameter Graticule (see Fig. 2). Foreign particulates that arise from a manufacturing environment are therefore easily discerned. Silicone oil based particles are not counted on the membrane and thus enable a “correction” to be made to particle counts obtained by Light Obscuration Method 1. However, the ability of the membrane microscopy method to count proteinaceous particles is dubious

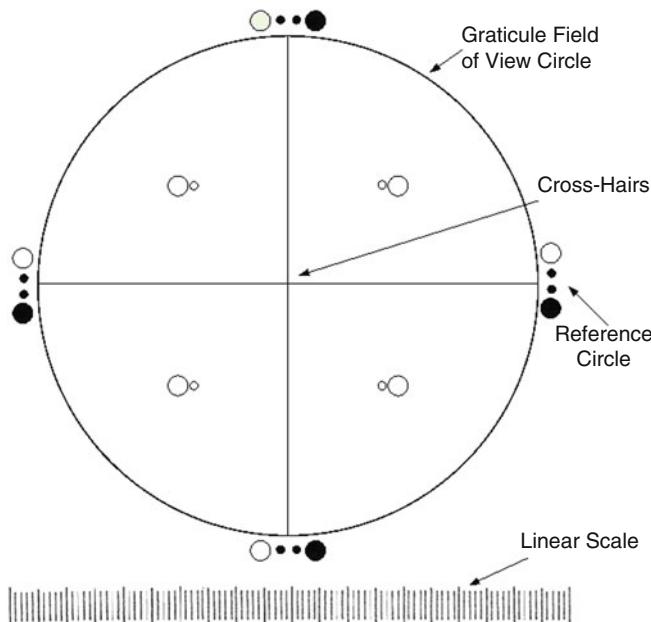


Fig. 2. The Circular Diameter Graticule used in particle measurement by Optical Microscopy/Membrane Microscopy.

since these particles are likely to be optically transparent or translucent and mechanically fragile. They can become deformed or even dehydrated on the membrane and can be difficult to detect and count. Membrane microscopy method will thus generally undercount proteinaceous particles and is thus a technique which though allowed for protein therapeutics, is not recommended.

Volume requirements per compendia are the same as for the Light Obscuration Method 1 given earlier. In principle, small-volume testing along the same lines as described earlier for Light Obscuration can also be performed with Method 2.

1.3. Flow Image Analysis

Flow (or Dynamic) imaging methods have lately been introduced for the counting of particles in parenterals, with particular emphasis on their application to protein therapeutics. The general principle of operation involves introducing a sample into the system allowing it to flow past a light source/detector (illustrated in Fig. 3). Digital images are captured by a high speed camera and analyzed, providing particle size and count information. Since this is an imaging technique, the particles counted can also be viewed. Various image analysis algorithms can be performed on the acquired images, including determination of particle size, count, shape factors and particle intensity measurements. Since the images are captured and stored, analysis can also be performed after the measurements are complete. Application of the analysis algorithms allows silicone oil and microbubbles to be differentiated from other particles based upon preset characteristics such as brightness and aspect ratio. The ability to create an image of a particle is also a function of the contrast presented by the particle against its background, along the lines of Light Obscuration. Furthermore, the smaller the (proteinaceous) particle, the closer is its refractive index to that of the surrounding medium. However, the imaging detector is purported to be more sensitive than the intensity detector in

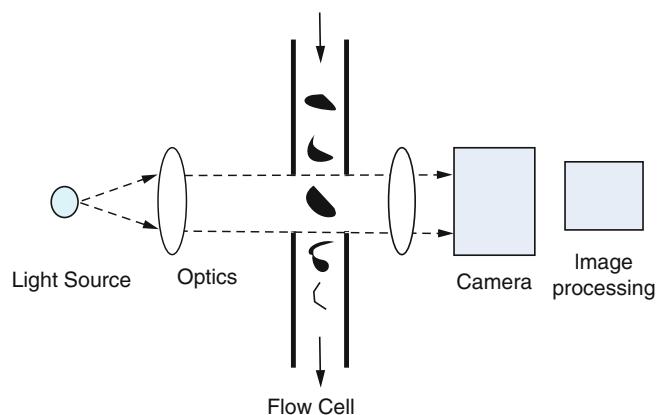


Fig. 3. Schematic representation of particle measurement by Flow Imaging.

LO, and thus generally detects a greater number of particles than LO especially in the size ranges below 10 µm (7).

Flow imaging analysis for measuring and counting subvisible particulates in injectables is not part of the compendial test methods. It is therefore currently utilized mainly as a developmental and (orthogonal) characterization tool. However, it is possible that a sponsor, after proper qualification and validation, could use this method from a regulatory perspective also. Some comparison with data from light obscuration would likely be expected.

The algorithms that examine the image and discern whether a region represents a particle and assess its characteristics are based upon an assessment of the gray-scale over the pixels occupied by the image. The smaller the particle, the fewer pixels it occupies and thus the lesser the information that can be gleaned from the image. It is generally accepted that significant loss of information occurs for sizes below 4 µm wherein only the presence or absence of a particle can be detected. The lower size limit for the technique is stated to be 1 µm but our experience suggests that 2 µm is the likely lower limit for good quantitation, especially for low contrast (proteinaceous) particles.

One example of this technology is the Brightwell Micro-Flow Imaging system (Brightwell Technologies, Ottawa, Ontario, Canada). This system has two models available, the DPA4100 and the DPA4200. The DPA4100 is an older model, and has the ability to size particles over a broad range (2–300 µm). The DPA4200 has a single magnification setting, and is optimized to analyze particles in the sub-visible size range (1–100 µm). The sample preparation considerations are the same regardless of the instrument model. This technique requires a fairly dilute solution of particles. If analyzing particle rich samples, it may require a dilution before analysis. Since the analysis is based on the acquired image, the particles must have some spatial resolution for the analysis to distinguish between particles. The maximum amount of particles that can be resolved and counted is approximately 850,000 counts/mL. Confirmation that dilution does not impact counts must be made, as discussed earlier. Other considerations about sample preparation such as degassing also apply.

Sample volumes down to 500 µL can be introduced into the instrument. A common approach for introducing small volume samples is using a 1 mL pipette tip. The tip will fit into the syringe port and analysis can proceed after it is seated in the port. The procedure provided in this chapter is based upon the DPA4200. Updated versions of the instruments are now available with an autosampler function.

1.4. Electrozone Sensing/Coulter Counter

The Coulter Counter Multisizer is a particle size and count instrument based on the Coulter Principle, also known as the Electrical Zone Sensing or Electrozone Sensing method. Particle size and count measurements are made by preparing suspensions of the sample of interest in a low concentration electrolyte. The particulate

suspension is made to flow through a small cylindrical tube with a defined opening (the aperture tube) that separates two electrodes. An electric current flows between the electrodes. As each particle passes through the aperture, it displaces its volume of the electrolyte and causes a transient change in the impedance across the aperture, leading to a voltage pulse. The amplitude of this pulse is directly proportional to the volume of the particle that produced it, with each pulse originating from a single particle. Since a known volume of suspension is drawn through the aperture, the number of pulses can be converted to yield a concentration of particles per unit volume of the test suspension (illustrated in Fig. 4). If a constant particle density is assumed, the pulse height is also proportional to the particle mass. Since the pulse height is determined by the volume of electrolyte displaced, porous particles and particles carrying a large fraction of enclosed electrolyte will be sensed as being smaller than their geometrical size.

Since the particles must pass through the sensing zone one at a time, the optimal concentration must be determined for each sample. In addition, the aperture tube has a defined opening, which will only allow a certain size range of particles (within 2–80% of its nominal diameter) to pass. A range of aperture sizes is available. If the sample particle size distribution is too broad, some particles may be excluded from analysis or possibly plug the aperture opening. In addition, the suspension must be present in a conducting media for the Coulter Principle to apply. These aspects should be considered when deciding on the particle size and count technique to employ for a given sample. The overall particle size measurement range is between 0.4 and 1,600 μm by combining data from several apertures. It is however important to remember that a

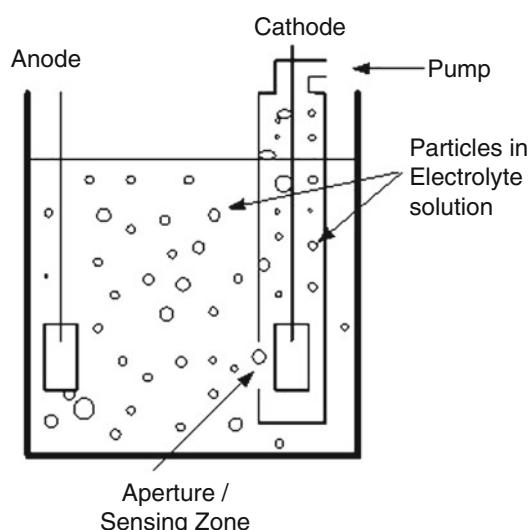


Fig. 4. Schematic representation of particle measurement by Electrozone Sensing/Coulter Counting.

particle count measurement is only as good as the dispersed state of the sample; therefore the analyst must ensure the sample preparation is properly dispersed and stable over the measurement time. Large particles may settle out before being analyzed.

Protein therapeutics are generally formulated in a buffer solution with a tonicity modifier such as NaCl or a disaccharide. The specific formulation buffer properties and experimental conditions may require modification of the ionic strength of the test sample by dilution or addition of salt. Any impact of this manipulation on the material to be measured is likely dependent on the protein and the nature of the aggregate constituting the proteinaceous particles. Foreign particles should in general not be impacted by this manipulation. Loose aggregate morphologies carrying a large fraction of enclosed electrolyte can also lead to difficulties in sizing by the Coulter principle since the effective drop in impedance will be smaller than that reflected by the geometric volume.

Additional guidance on the Coulter Principle can be found in the International Standard ISO 13319 “Determination of Particle Size Distributions-Electrical Sensing Zone Methods.” This technique is often used as a reference for qualifying other particle sizing instruments.

1.5. Submicron Particle Tracking Analysis/NanoSight

The NanoSight is a nanoparticle analysis instrument which functions by analyzing the motion of the particles. Particles are illuminated by a laser light as they move in solution by Brownian motion. Particle motions are tracked and recorded by a camera, as small points of (scattered) light, under a microscope objective (illustrated in Fig. 5). The path the particle takes over a suitable period of time (approximately 30 s) is tracked and the particle size profile is determined using the NanoSight analytical software program. Particles with overlapping paths or if out of focus, are eliminated from the analysis.

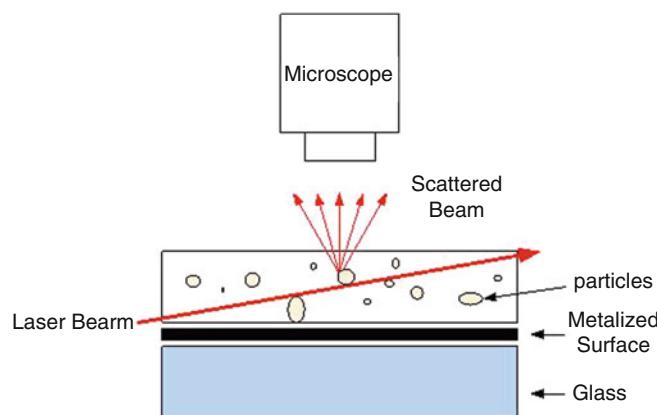


Fig. 5. Schematic representation of particle measurement by Submicron particle tracking/NanoSight.

Lower size limit of detection depends in part on the refractive index and scattering ability of the particle. For biological (proteinaceous) samples, the lower limit is approximately 30 nm. The upper size limit will depend upon whether the particle movement can be tracked accurately and will be problematic if the particle sediments during analysis. The stated limit is 1,000 nm.

Optimal sample concentration is 10^8 – 10^9 particles/mL, while the minimum concentration is 10^7 particles/mL and maximum concentration is 5×10^9 particles/mL. At these high concentrations, individual particles cannot be resolved accurately.

Accuracy of particle sizing is within 5% of the diameter (for standard spherical reference materials, such as polystyrene standards). Reproducibility of sizing is approximately 3% at optimal concentration (for standard spherical reference materials, such as polystyrene standards). The particle concentration is estimated from the number of particles detected and thus a particle size distribution can also be calculated. The accuracy of the concentration measurement depends on the number of particles detected (8).

The sample cell has a volume of 0.25 mL, with a depth of 0.5 mm. The dimension of the sample viewed is $100 \times 80 \times 5$ μL .

It should be noted that the NanoSight measures the diffusion coefficient (D_t) of a particle (this is the mean squared displacement) and uses the Stokes–Einstein equation, assuming a spherical particle, to assign a (hydrodynamic) radius (r_h) to that particle (see Eq. 1). This model applies to most spherical or near spherical particles. If the aspect ratio is >3 , the Stokes–Einstein equation will no longer be applicable.

$$D_t = K_B T / 6\pi n r_h \quad (1)$$

(Note: K_B is the Boltzmann's constant, T is temperature, and n is solvent viscosity).

Submicron particle tracking analysis should primarily be used as a developmental and (orthogonal) characterization tool.

2. Materials

2.1. Light Obscuration Analysis

1. Suitable calibrated instrument based on the principle of light blockage, e.g., HIAC/Royco Pacific Scientific Liquid Particle Counting System.
2. Sampler Model: 9703 or equivalent.
3. Syringe Size: 10 mL.
4. Sensor Model: HRLD 150, HRLD 400CE or equivalent.

5. Software: PharmSpec 2.0 or equivalent.
 6. Particle-free water ($0.22\text{ }\mu\text{m}$ filtered).
 7. Work and test area as described in USP <788> (HEPA hood with laminar air flow).
- 2.2. Optical Microscopy/Membrane Microscopy**
1. Membrane filters: 25 mm or 47 mm; gridded or nongridded; black or gray; of a suitable material (mixed cellulose ester commonly used); $1.0\text{ }\mu\text{m}$ or smaller pore size.
 2. Particle free water for rinsing and environmental control.
 3. Compound binocular microscope with two illuminators (episcopic and oblique) and a $10\times$ objective; $10\times$ eyepiece.
 4. A calibrated circular diameter graticule (see Fig. 2).
 5. Filter assembly consisting of a filter holder (preferably glass funnel), with a suitable vacuum source.
- 2.3. Flow Image Analysis**
1. 1-mL pipette tip.
 2. Particle-free water.
 3. Thermo Count-Cal latex microsphere count and size standards, $15\text{ }\mu\text{m}$ nominal diameter (or equivalent).
 4. Glass 10-mL syringe with luer-lock tip.
- 2.4. Electrozone Sensing/Coulter Counter**
1. Coulter Multisizer 3 Counter (or equivalent).
 2. Desired aperture tube.
 3. Sample beaker/cuvette, cleaned with particle free water (the ACCUVETTE by Coulter can be used for apertures up to $100\text{ }\mu\text{m}$. For larger apertures, the beaker should be used).
 4. Particle free water.
 5. Appropriate conducting diluent (such as Isoton).
 6. Particle size/count standards, e.g., Thermo polystyrene NIST traceable microspheres.
- 2.5. Submicron Particle Tracking Analysis**
1. Particle-free water for blank runs.
 2. Particle size/count standards, e.g., Thermo polystyrene NIST traceable nanoparticles.

3. Methods

3.1. Light Obscuration Analysis

Follow the general instructions for test article selection (see Note 1), test article preparation (see Note 2) and test environment (see Note 3).

1. Prepare the instrument as follows:

- Instrument is calibrated over the range 1.5–100 µm (single calibration) when using the HRLD 150 sensor. If using the HRLD 400 sensor, calibrate from 2–100 µm.
- The following sizes are suggested for monitoring, but can be modified per the user's requirements (note that the pharmacopeias require monitoring 10 and 25 µm): 2, 5, 8, 10, and 25 µm.
- Nominal flow rate: 25 mL/min.
- Syringe size: 10 mL.
- Measurement (sample) volume: 1.00 mL.
- Number of runs: 4.
- Dilution factor: 1.00.
- Tare Volume (mL): 0.2.
- Multi stroke tare (mL): 0.10.
- Check "discard first run."

2. Prepare all glassware, apparatus as per compendia using particle-free water.

3. Turn on instrument and allow to warm up for at least 15 min.

4. Verify instrument calibration as per compendia (4–6). (If possible, use Thermo Count Cal 15 µm latex standard). Standard counts should meet the vendor specifications for count.

5. Confirm an acceptable Environment Blank and System Blank as described in the compendia or local procedures.

- Environmental Blank: Suitability of the environment is determined by testing five samples of particle-free water each of same volume as the intended measurement volume (=1 mL). The system is considered not suitable if more than 1 particle/mL of size 10 µm or greater are present. The preparatory steps must be repeated until the environment, glassware and water are found suitable for the test.
- System Blank: Suitability of the sample handling procedure is assessed with the system blank which is particle-free water taken through the same procedure as the test sample. Measure particle counts in a sample of particle-free water of same volume as the intended measurement volume (=1 mL). The system is considered not suitable if more than 5 particles/mL of size 10 µm or greater are present. The system (environment, glassware and water) and the procedure must be repeated until a suitably low system blank is obtained.

6. Perform test according to compendia (4–6), and report the results for Small-Volume Injections or Large-Volume Injections,

depending on product volume. Sizes to be included are 2, 5, 8 µm in addition to the 10 and 25 µm. The collection of data on the 2, 5, and 8 µm sizes is recommended for protein therapeutics as additional information.

- Perform measurements in at least four samples of 1 mL each. Discard the results for the first aliquot and average the number of particles of the last three aliquots.
- 7. Use this average of cumulative particle counts to calculate the number of particles in the original test article.
- 8. Retain any remainder volume of the sample. Note that the pharmacopeias allow use of Method 2 (Membrane Microscopy) if failing results are obtained by Method 1 (Light Obscuration) (see Note 4).

3.2. Optical Microscopy/ Membrane Microscopy

Follow the general instructions for test article selection (see Note 1), test article preparation (see Note 2), and test environment (see Note 3).

1. Clean the filter membrane before use with a low pressure stream of particle free water. Rinse both sides of the filter
2. Confirm an acceptable Environment Blank as described in the compendia (4–6).
 - Environmental Blank: Analyze the blank filter after filtering a 50 mL volume of particle-free water. The system is considered not suitable if more than 20 particles of size 10 µm or greater are present or if more than 5 particles of size 25 µm or greater are present within the filtration area. The preparatory steps must be repeated until the environment, glassware and water are found suitable for the test.
3. Mix sample carefully by slowly inverting the sample container 20 times. (For protein therapeutics this may again lead to entrainment of bubbles, but bubbles are not a concern with this method).
4. Filter the test sample onto the clean filter membrane.
5. Place filter membrane in a cleaned petri dish or similar container and allow the membrane to dry.
6. Place the petri dish on the microscope and view the membrane under the microscope- mentally transform each particle into a circle and compare to the reference circles in the calibrated eyepiece.
7. Scan the entire membrane surface. A portion of the membrane can be analyzed, and the total particle counts by calculation are allowed.

8. Do not attempt to size or count amorphous, semiliquid, or otherwise morphologically indistinct particles that have the appearance of a stain or discoloration on the membrane filter.
9. Count the number of particles $\geq 10 \mu\text{m}$ and the number $\geq 25 \mu\text{m}$. The cumulative particle counts are reported (number of particles greater than or equal to the particle size of interest (see Note 4).

3.3. Flow Image Analysis

Follow the general instructions for test article selection (see Note 1), test article preparation (see Note 2) and test environment (see Note 3).

1. The following parameters are provided as an example of an analysis method (for Model DPA4200).
 - Set point = 3 (based on flow cell model, SP3 used for this example).
 - For a 1 mL total sample volume, use:
 - Total available volume: 0.90 mL.
 - Purge volume: 0.20 mL.
 - Analyzed volume: 0.60 mL.
 - It is recommended to select “Edge Particle Rejection” and “Fill Particles,” for consistency in results.
 - Set limits for the desired filters—For this example the following are monitored:
 - Equivalent Circular Diameter (ECD) (μm).
 - Circularity.
 - Intensity Mean.
 - Max Feret Diameter (μm).
 - Aspect Ratio.
 - Image Frames = Max frames set at 20 (this provides a representation of the images without generating an extremely large data file).
2. Add particle free water or preferred diluent to a clean 10-mL syringe and attach to the syringe port on the Brightwell.
3. Open the MVSS software and choose a method for analysis.
4. Rinse cell and perform the “optimize illumination” step to blank any background image noise. Optimization of illumination can be performed with sample matrix if desired.
5. Mix test sample well, without introducing any air bubbles (see Note 5).
6. Remove desired sample amount with syringe or pipette tip (depending on sample volume for testing).
7. Add sample to syringe port on the Brightwell.

8. Perform the “analyze sample” step.
9. A PDF file of the results is generated. Save the file if desired, or close the file. The data will be saved in the database. A summary page with all results is provided in the Sample Analysis Report, sorted by Equivalent Circular Diameters and Intensities.
10. An extended analysis of the images can be performed to obtain information for further characterization of the particles, e.g., separation of silicone oil, bubbles, particle morphology, etc. (see Note 6).

3.4. Electrozone Sensing/Coulter Counter

3.4.1. Instrument Setup

Follow the general instructions for test article selection (see Note 1), test article preparation (see Note 2) and test environment (see Note 3).

Begin by turning on the computer and ensure the analyzer is communicating. Allow the system to warm up approximately 15 min after powering up. Note that previously collected data can be reprocessed as long as the pulse data was collected.

1. The front of the Multisizer system has some features including:
 - Aperture Viewer-a small screen that allows visualization of the aperture opening.
 - Stirrer Rotation Control-sets the direction of stirring.
 - Stirrer Speed Control-sets the rotation speed as well as on/off functionality.
 - Sample Compartment: holds the sample beaker/cuvette, stirrer, and aperture tube. This is enclosed in a Faraday Cage to minimize electrical interference.
2. Ensure the waste container is in place and empty.
3. When installing a new aperture, it must be calibrated before the first use. The software contains a Wizard that will walk the user through installing and calibrating the aperture tube. The user manual also contains detailed information on calibration of the aperture. Calibration standards are available from the manufacturer. When replacing the aperture tube on the instrument, note that it is made of glass and is fragile. Never force the tube onto its connector.
4. After installing the aperture, align the optics using the three knobs on the side of the instrument (positions the tube up, down and focus depth).

3.4.2. Aperture Setup

1. The following items should be checked before analysis (see Note 7).
 - The correct aperture is installed.
 - The system is filled with electrolyte, and the aperture is immersed in electrolyte.

- The waste container is not full.
 - The appropriate analysis settings have been selected.
2. Place a beaker or cuvette with clean electrolyte on the sample platform (ensure the same electrolyte is in the container within the analyzer).
 3. Position the beaker to the appropriate height and adjust focus.
 4. Use the drop down menus in the application to select the electrolyte name (enter a name if it is not listed) and the dispersant name if one is used.
 5. If not already selected, choose “select new aperture tube.” This will initiate the Aperture Tube Wizard. In the Wizard, fill the system and adjust the metering pump.
 6. Set the current and gain. The auto-set function can be used (be sure to select whether you are analyzing cells or particle systems, as different current settings will be employed).
 7. Select ‘measure noise level’ in the drop down menu.

3.4.3. Sample Analysis

1. In the System menu, choose Change SOM to access the Standard Operating Method (SOM).
2. In the SOM menu choose the mode of operation (by time, volume, total counts, etc.). This will determine how the analysis will be performed and stopped. Also select the number of runs and ensure that the threshold setting is correct (the lower limit for data acquisition).
3. Ensure Current and Gain settings are correct (or set to Auto-set mode).
4. In the SOM menu, select to “flush aperture tube.” The instrument will automatically flush the aperture after every run.
5. Hit OK to exit the SOM menu.
6. To save all changes to the SOM, go to the Settings menu on the Main Menu Bar. Select Save Run Settings.
7. In the Settings menu, choose Enter Sample Info to enter all necessary sample information. If particle concentration values are needed, the following must be entered: Sample volume or Mass, Electrolyte Volume, Analytical Volume.
8. Analyze a background sample- use the same parameters as for the sample analysis. Choose Start in the status panel.
9. To enable a background subtraction for each sample run, choose Load Background Run from the Settings menu. Choose the appropriate background file and load.

10. Analyze the sample: Choose Preview in the status panel and Run. Make sure sample concentration is less than 10%. If concentration is appropriate, select Start to begin sample analysis.
11. It is advisable to clean the system every day after completion of analysis.

3.4.4. Cleaning the System

1. Disconnect the electrolyte reservoir from the instrument.
2. Choose System menu from the main menu. Select Drain System. After the drain is complete, connect a reservoir with cleaning agent (DI water, Coulter Clenz, or 2% bleach solution are recommended). Repeat the fill and drain procedures.
3. Leave the system stored in the cleaning agent. Before running a sample, the drain and fill procedures will be repeated with the appropriate electrolyte.

3.4.5. Data Analysis

1. Once the run is complete, the analysis file can be viewed and saved in many different formats.
2. The results can be previewed by Opening a file and in the dialog box choose Quick View (with the appropriate parameters such as by size, trend or pulses). A small viewing box in the lower left corner of the menu will display the data.
3. To overlay multiple files choose Overlay in the File menu. Choose the files to be overlaid in the dialog box. Files will be “added” to the bottom of the dialog box. When done, choose OK. A graph overlaying all the selected files will be displayed.
4. Several files can be averaged. This function allows the data from these files to be combined, channel contents averaged, sample statistics on each channel calculated and a new average distribution plotted. The Average function is found in the File menu.

3.5. Submicron Particle Tracking Analysis

Follow the general instructions for test article selection (see Note 1), test article preparation (see Note 2) and test environment (see Note 3).

1. Open the capture screen (choose basic or advanced mode). This procedure will refer to the basic mode.
2. Increase camera level to maximum.
3. Find the “thumbprint” flare spot. This image can be used to check for any unwanted vibration before analysis. Any movement suggests vibration.
4. Move to the optimal viewing region, where particles are seen next to the “thumbprint.”
5. Focus particles (should appear as smooth spheres or points of light).

6. Increase camera level to maximum, to check that all particles are being visualized.
7. Decrease camera level, ensuring that small particles are still imaged.
8. Check particle concentration visually (there should be around 20–100 particles in the field).
9. If particle concentration is not optimal, alter the sample concentration accordingly and repeat the steps above.
10. Set capture duration appropriate for a monodispersed sample.
11. Set capture duration based on polydispersity.
12. Record the video, entering the sample temperature if prompted.
13. Open the analysis screen and load the saved video if it's not already loaded. Choose basic or advanced mode- this procedure will refer to the basic mode.
 - Adjust the screen gain to see the dimmest particles in the image.
 - Set the detection threshold (a red cross should appear in each particle on the screen).
 - Use the blur function to reduce noise (false centers) found around particles.
 - Set the minimum expected particle size. If it is unknown, it is suggested to set it at 100 nm and adjust after rerunning the analysis as needed.
 - Click the process sequence button to run the analysis. The particle size distribution will be generated.

4. Notes

1. Some general critiera for test articles selection must be followed depending upon the objective of the measurement. An appropriately representative set of test articles (e.g., vials) must be used. Compendial testing (e.g., for quality control, release or formal stability) requires a minimum of 25 mL of solution per test. This applies especially to light obscuration and the optical/membrane microscopy methods. In our experience, this can however be scaled down to 5 mL with good results, especially during development stages. Note however that as for any counting process, the accuracy of the counts increases (and the variability decreases) as the numbers (and volume) counted increases. Thus, when particulate levels are low, sampling and counting errors will have a large effect on method variability. In this case, use of a small-volume method can be mis-representative of the actual product, especially if a multiplication factor is used to extrapolate the result to a per vial

basis (e.g., when the result for a 1-mL measurement volume is multiplied by 20 to obtain the counts for a 20 mL product). Subvisible particulate levels are also likely to be nonuniformly distributed between individual test articles. For the small particle-size ranges (generally below 10 µm), the counts are often high enough that a reasonable estimation of the standard deviation can be obtained. This becomes more difficult as particle counts decrease (e.g., for sizes above 10 µm), and often a larger test article set is required. The small volume method may also be applied for ophthalmic injectables which have to comply with USP <789>.

The other methods discussed in this chapter (Flow imaging analysis and Submicron particle analysis by particle tracking) are not compendial methods, and thus, the use of these methods has no predefined test volume. However, it is recommended that a standard protocol be established for the use of these techniques where possible, to enable results across samples and time to be compared.

2. Proper sample preparation is critical to obtaining reproducible and reliable results when measuring (counting) subvisible particles by any method. Test articles should be allowed to come to room temperature prior to pooling. Solutions should be transferred to a clean particle-free container if needed, taking care to not introduce any environmental contamination. In the case of lyophilized products, reconstitution must be performed per instructions with the appropriate diluent. The diluent should itself be examined for its contribution of particulates to the product.

In certain cases such as high-concentration products which may also have a high viscosity, direct injection of the product may not be feasible. In these cases, dilution of the product with an appropriate diluent (e.g., particle-free WFI) may be required. In such situations, the appropriateness of the dilution selected should be verified by testing a wide range of dilutions to ensure that a linear response is obtained in all the particle size ranges being monitored.

Degassing of test samples is critical to obtaining reproducible and representative counts. It is not recommended to degas protein solutions by sonication. Allowing the sample to stand or applying a gentle vacuum (e.g., ~75 Torr) for a suitable length of time (e.g., 10 min–2 h; (3)) will be more appropriate. Degassing time should be verified as high concentration solutions have a greater tendency to form and retain microbubbles, but can also begin to dry out, forming a surface film if degassed for too long. Since larger particles can settle during the degassing procedure, a gentle swirling motion should be utilized to re-entrain such particles after degassing, taking care to not generate new bubbles.

Table 2
Summary of pharmacopeial requirements for subvisible particulates in injectables

Attributes	USP 34 General Chapter <788>	Ph Eur 7.0 General Chapter 2.9.19	JP 16th edn. General Chapter 20
	Unit product volume ≤100 mL = Small volume parenterals (SVP) >100 mL = Large volume parenterals (LVP)	Unit product volume <100 mL = Small volume parenteral (SVP) ≥100 mL = Large volume parenteral (LVP)	Unit product volume <100 mL = Small volume parenteral (SVP) ≥100 mL = Large volume parenteral (LVP)
Specifications: Light obscuration (LO) (Preferred method)	SVP ≥10 µm: ≤6,000 counts/container ≥25 µm: ≤600 counts/container LVP ≥10 µm: ≤25 counts/mL ≥25 µm: ≤3 counts/mL	≥10 µm: ≤6,000 counts/container ≥25 µm: ≤600 counts/container ≥10 µm: ≤25 counts/mL ≥25 µm: ≤3 counts/mL	≥10 µm: ≤6,000 counts/container ≥25 µm: ≤600 counts/container ≥10 µm: ≤25 counts/mL ≥25 µm: ≤3 counts/mL
Specifications: Microscopy (M) (Second stage to LO or in case LO cannot be used)	SVP ≥10 µm: ≤3,000 counts/container ≥25 µm: ≤300 counts/container LVP ≥10 µm: ≤12 counts/mL ≥25 µm: ≤2 counts/mL	≥10 µm: ≤3,000 counts/container ≥25 µm: ≤300 counts/container ≥10 µm: ≤12 counts/mL ≥25 µm: ≤2 counts/mL	≥10 µm: ≤3,000 counts/container ≥25 µm: ≤300 counts/container ≥10 µm: ≤12 counts/mL ≥25 µm: ≤2 counts/mL
Intravenous injections	Limits apply	Limits apply	Limits apply
Injections solely for intramuscular (IM) or subcutaneous (SC) dosing	Limits do not apply per USP 33 <1>; USP PF35.3 proposal: Limits to apply for IM and SC products also	Limits apply. Higher limits may be appropriate	Not mentioned separately

Non IM/SC dispersed systems		M (No special procedure described)	M (No special procedure described)	M (No special procedure described) suspension particle $\leq 150 \mu\text{m}$ Emulsion drop $\leq 7 \mu\text{m}$
Non IM/SC solutions or powders for injection	With final filter before injection	Excluded from requirements per USP33 <1>; USP PF35.3 proposal: Parenteral products for which the labeling specifies the use of a final filter prior to administration are exempt from the requirements provided that scientific data are available to justify the exemption	Exempt from requirements, providing it has been demonstrated that the filter delivers a solution that complies	Not mentioned
Test protocol	High viscosity Other solutions	Dilution followed by LO LO alone or followed by M	Dilution followed by LO LO alone or followed by M	Dilution followed by LO LO alone or followed by M
	Statistically sound sampling plan	Required for $<25 \text{ mL}/\text{unit}$ For $\geq 25 \text{ mL}/\text{unit}$, 10 units acceptable	Required for $<25 \text{ mL}/\text{unit}$ For $\geq 25 \text{ mL}/\text{unit}$, 10 units acceptable	Required for $<25 \text{ mL}/\text{unit}$ For $\geq 25 \text{ mL}/\text{unit}$, 10 units acceptable
	$<25 \text{ mL}/\text{unit}$	Pool ≥ 10 units to obtain $>25 \text{ mL}$, Test $4\times \text{NLT } 5 \text{ mL}$ aliquots, discard first result	Pool ≥ 10 units, Test $4\times \text{NLT } 5 \text{ mL}$ aliquots, discard first result	Pool ≥ 10 units to obtain $>25 \text{ mL}$, Test $4\times \text{NLT } 5 \text{ mL}$ aliquots, discard first result
	$\geq 25 \text{ mL}/\text{unit}$	No pooling, Tested individually, Test $4\times \text{NLT } 5 \text{ mL}$ aliquots, discard first result	No pooling, Tested individually, Test $4\times \text{NLT } 5 \text{ mL}$ aliquots, discard first result	No pooling, Tested individually, Test $4\times \text{NLT } 5 \text{ mL}$ aliquots, discard first result

3. A basic requirement for accurate, reproducible and reliable particle counting is to keep the environment for testing as clean as possible. The use of particle free gloves, Tyvek sleeves, and Tyvek laboratory coats, while making all preparations and transfers in a laminar air flow hood are essential to keeping the extrinsic particle load to a minimum. Use of particle-free containers as well as particle-free water is important. Note that WFI is not necessarily particle-free and that filters must be flushed thoroughly prior to use since filters also shed particles.
4. The pharmacopeias have acceptance criteria for subvisible particles in injectables as measured by Light Obscuration or Membrane Microscopy. These are listed in Table 2.
5. The sample should be well mixed before introducing to the system. If a 10 mL glass syringe is used to introduce sample to the instrument, a small mixing unit is available that fits within the syringe barrel to help keep particles suspended.
6. Extended analysis of Flow Image Analysis data can be performed to further extract information from the images captured. Software filters can be created to distinguish between type of particles based on Morphology (aspect ratio), Brightness (intensity) and size, for sizes above 5 μm ECD. Silicone oil based particles tend to be spherical (high aspect ratio) and have a higher intensity than proteinaceous particles (see, e.g., Fig. 6). Similarly, bubbles tend to have a high intensity although

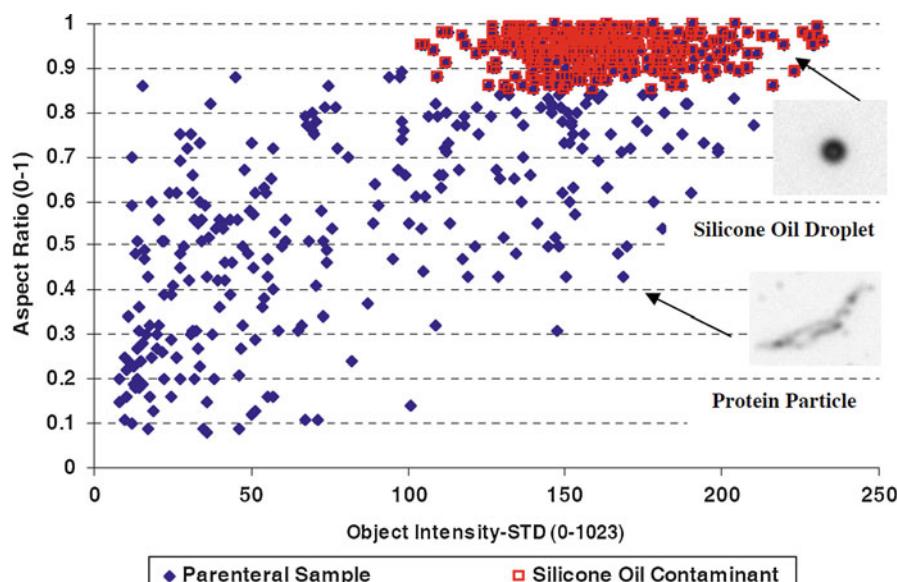


Fig. 6. Morphological filters can be applied to isolate silicone oil from proteinaceous aggregates in dynamic image analysis (7); with permission).

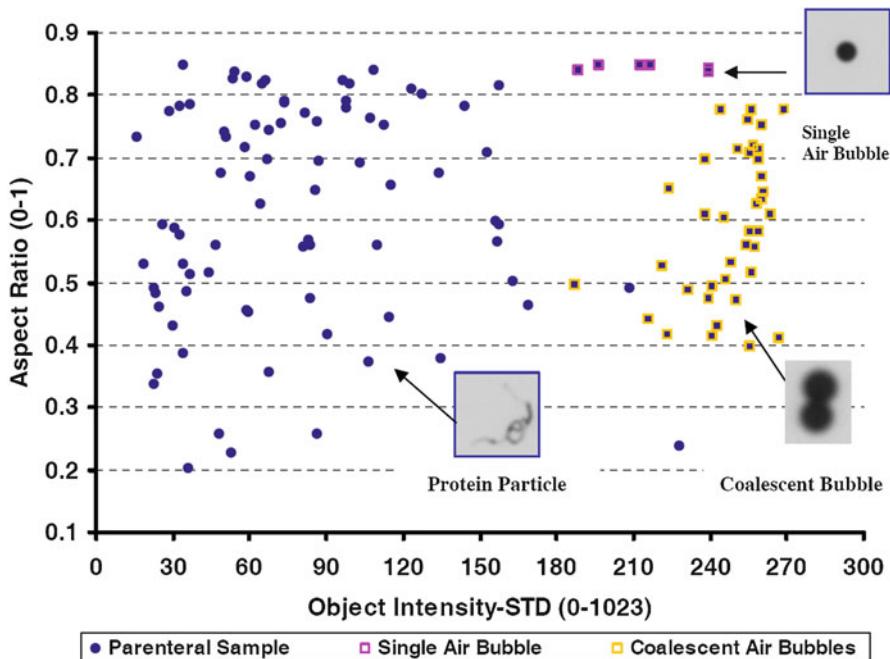


Fig. 7. Aspect ratio and brightness filters can be applied to isolate bubbles from proteinaceous aggregates in dynamic image analysis (7); with permission).

the aspect ratio can vary due to coalescence (see, e.g., Fig. 7). In all cases, software filters should be used with care in the context of a good understanding of the characteristics of the product being analyzed.

7. A specific aperture will be suited for a specific particle size range. It is good practice to run a calibration standard as a system suitability test before running any samples on any given day.

References

1. Carpenter JF, Randolph TW, Jiskoot W et al (2009) Overlooking subvisible particles in therapeutic protein products: gaps that may compromise product quality. *J Pharm Sci* 98(4): 1202–1205
2. Singh SK, Afonina A, Awwad M et al (2010) An industry perspective on the monitoring of subvisible particles as a quality attribute for protein therapeutics. *J Pharm Sci* 99(8): 3302–3321
3. Narhi LO, Jiang Y, Cao S et al (2009) A critical review of analytical methods for subvisible and visible particles. *Curr Pharm Biotechnol* 10(4): 373–381
4. US Pharmacopeia (2011) USP/NF General Chapter <788>. Particulate matter in injections. In: US Pharmacopeia, National Formulary, USP34/NF-29 (suppl. 1). Rockville, MD
5. European Pharmacopoeia (2011) 7th edn. (7.2). General Chapter 2.9.19. Particulate contamination: sub-visible particles. European Pharmacopoeia Commission, Council of Europe, European Department for the Quality of Medicines
6. Japanese Pharmacopoeia (2011) 16th edn. General Chapter 20. Foreign insoluble particulate matter test for injections. Society of Japanese Pharmacopoeia
7. Sharma DK, King D, Moore P et al (2007) Flow microscopy for particulate analysis in parenteral and pharmaceutical fluids. *Eur J Parenter Pharm Sci* 12(4):97–101
8. Filipe V, Hawe A, Jiskoot W (2010) Critical evaluation of Nanoparticle Tracking Analysis (NTA) by NanoSight for the measurement of nanoparticles and protein aggregates. *Pharm Res* 27(5):796–810

Chapter 25

Size-Exclusion Chromatography with Multi-angle Light Scattering for Elucidating Protein Aggregation Mechanisms

Erinc Sahin and Christopher J. Roberts

Abstract

In this chapter, application of size exclusion chromatography with inline multi-angle light scattering (SEC-MALS) to protein systems is reviewed, in particular for its use in elucidating mechanistic details of net-irreversible aggregation processes. After motivating why SEC-MALS or analogous techniques are natural choices to interrogate such aggregating systems, the individual techniques (SEC and MALS) are reviewed briefly, as needed for the context of the remainder of the chapter. Illustrative examples are provided to highlight when and how SEC-MALS can be applied to test mass-action kinetic models for protein aggregation. Limitations of the technique, as well as recommendations for troubleshooting and potential errors in data interpretation are also provided.

Key words: Light scattering, Protein aggregation, Chromatography, Aggregation models, Protein stability

1. Introduction

This chapter presents methodology for utilizing size-exclusion chromatography with multi-angle light scattering (SEC-MALS) to interrogate net-irreversible, often non-native, aggregation mechanisms for *in vitro* protein systems. The results from SEC-MALS provide a natural means for qualitative mechanistic analysis of protein aggregation kinetics, as well as quantitative analysis if one or more candidate (mathematical) models are available for regression against the data (1–4). Mechanistic mathematical models for aggregation kinetics are typically cast in the form of mass-action frameworks, or analogous population-balance models (5–9); brief illustrations for applying such models to SEC-MALS data are included as part of this chapter, including a relatively general and simple “model-free” approach.

In the present section of the chapter, relevant theoretical aspects of protein aggregation that motivate the use of SEC-MALS or analogous techniques are first reviewed. This includes relatively simple tests to determine whether SEC-MALS will provide advantages over analogous techniques, over SEC alone, or whether an alternative approach is more appropriate altogether. Unlike chapters that focus on techniques with relatively new protocols for sample preparation or instrument operation, this chapter focuses on a methodology for applying an established experimental technique (SEC-MALS) to a specific application—i.e., protein aggregation. The basic principles of SEC and MALS are therefore only briefly summarized in this section, and users are directed to consult vendor-specific documentation regarding operating procedures and/or sample preparation for particular instruments. Many of the approaches described herein are in principle applicable to any system for which SEC and MALS measurements can be done. They do not require inline MALS with SEC, although inline capabilities can be desirable from the perspective of user convenience.

Subheading 3 focuses first on the main steps and considerations when generating accurate SEC-MALS data for irreversibly aggregating protein systems; for concreteness, this assumes a non-native mechanism in which temperature can be used to effectively quench the kinetics of aggregation, but in principle it also applies to aggregation via covalent bond formation (e.g., disulfide crosslinking) if the kinetics are sufficiently slow to not occur during the SEC-MALS measurement(s). In the remainder of Subheading 3, two approaches are briefly reviewed for using SEC-MALS for analysis of aggregation kinetics and mechanisms. The first approach is essentially model-independent, within simple assumptions that hold for most systems, but it provides only qualitative information regarding the importance and semi-quantitative magnitude of different stages in the aggregation process. The second approach requires one to assume one or more different models and then test them by regression against the SEC-MALS data.

Subheading 4 is subdivided into trouble-shooting sections regarding SEC and SEC-MALS, as well as recommendations regarding model discrimination and the importance of additional techniques and a range of solvent conditions and protein initial concentration when studying aggregation mechanisms. Recommendations are also summarized regarding which types of aggregating systems are most appropriately studied with SEC-MALS or analogous approaches, and which systems may be more appropriate for study with alternative techniques.

1.1. Mass-Action Models Related to Experimental Kinetics and Utility of SEC-MALS

Mechanistic mass-action models postulate elementary kinetic steps—e.g., a two-state unfolding step from folded or native monomer (N) to unfolded monomer (U)—and the law of mass action indicates that the net rate for a given step is proportional to the concentration of each species that is involved in that step. When

translated to a mathematical description, this naturally results in one or more ordinary differential equations, wherein the physical quantities are number concentrations (e.g., mol/L) of different species, and there are (typically unknown) rate coefficients or equilibrium constants associated with different steps (10).

From a practical perspective, one ideally would like to directly track each species in the model experimentally, either as a function of time or some other measure of rates of production/consumption of each species. Given the intrinsic difficulties with spectroscopically distinguishing structural differences between different intermediates in protein aggregation pathways (11), it is often more tenable to track the mass concentrations and/or mass-weighted average molecular weight (M_w , more commonly referred to as weight-average molecular weight) as a function of time. That is, one tracks the change in the distribution of species—illustrations of this are provided later in this chapter, as SEC-MALS is a natural choice for monitoring aggregation in this way. SEC provides a means to separately quantify the mass concentration of monomer (c_m), and possibly small soluble aggregates such as dimers and trimers, as a function of sample incubation time. Inline MALS provides a means to monitor M_w for the same injection on SEC. Equivalently, the fraction monomer (m) in a given sample at a given time can be determined simply by dividing c_m by the initial concentration (c_0).

It is worth mentioning that alternative fractionation techniques, such as flow field-flow fractionation (F4) can also fill the role of SEC (12, 13). All the analysis later in this chapter can readily be adapted to an F4-MALS format; the examples below are given from the SEC-MALS perspective because of the added considerations from SEC that may not be as relevant for F4, and because SEC currently is still a main workhorse within many industrial, academic, and government analytical laboratories.

One can, in principle, propose many different mass-action models to try to quantitatively describe an aggregating system. It is beyond the scope of this chapter to enumerate such models, and many are available in published review articles (14, 15). However, from a qualitative perspective it is useful to distinguish between a number of commonly observed behaviors for aggregating protein and peptide systems. Table 1 summarizes different scenarios in terms of qualitative observables that allow one to assess whether SEC-MALS, SEC or MALS alone, or neither are appropriate for analysis of in vitro aggregation kinetics and mechanism(s).

Inspection of Table 1 shows that the main features of an aggregating system that dictate whether SEC-MALS or a subset of the techniques are appropriate are: (a) whether aggregation rates are effectively arrested by quenching via cold temperatures (e.g., ice-water bath or refrigerated conditions); (b) whether aggregates are easily reversible upon dilution and/or small changes in salt content or pH; (c) whether aggregates are soluble upon formation, and/or

Table 1
Overview of some key considerations when deciding whether SEC-MALS, or an alternative approach, is most appropriate for monitoring and quantifying aggregation kinetics

Technique/ approach	Aggregates remain soluble in time course of experiment	Mobile phase/column conditions			Detectable levels of reversible aggregates	Aggregation proceeds at refrig./ ambient
		No adsorption of monomer or aggregates	Adsorption of some aggregates	Not all aggregates are soluble		
SEC	+	+	+	+	X	X
SEC-MALS	+	+	X	X	X	X
F4-MALS	+	+	X ^a	X ^a	X ^a	X
MALS (ex situ)	+	n/a ^b	n/a ^b	n/a ^b	+ ^c	X
MALS (in situ)	+	n/a ^b	n/a ^b	n/a ^b	+	+

(+) and (x) indicate a given technique or approach is or is not compatible, respectively

^aThis issue can typically be avoided if running buffer conditions match sample conditions

^bNot an issue; however, not compatible if ex situ conditions (e.g., temperature) cause aggregates or monomer to adsorb to scattering cell and/or to precipitate

^cNot compatible if ex situ conditions (e.g., temperature) shift equilibria from in situ conditions

whether aggregates rapidly coalesce with each other upon changes in temperature and/or small changes in pH or salt content; (d) whether mobile phase conditions are identified so as to eliminate significant losses of aggregate (or monomer) via adsorption to the column.

The first feature is essential if one is to use SEC or any fractionation technique, due to sample preparation and practical limitations of needing aggregates to not be evolving irreversibly during the time course of the fractionation itself. The second feature is important for SEC, because one must often use mobile phase conditions that differ in pH and/or salt content from the actual sample conditions of interest (see also Subheadings 3 and 4). This may be less of a concern for some fractionation techniques, such as F4 (12, 13). The third feature is an often overlooked aspect that will be illustrated in more detail later in the chapter. In brief, aggregate solubility dictates whether it will be practical to obtain accurate values of aggregate molecular weight as a function of time. This is also the primary problem associated with feature (d), as material lost to the column therefore skews the apparent molecular weight, and possibly the value of m . Overall, Table 1 shows that SEC-MALS is most appropriate under conditions where all aggregates remain soluble and elute quantitatively from the column. Under these conditions, one can realistically monitor m and M_w as a

function of sample incubation time (t) using the methods described in Subheadings 3 and 4.

For use later in the chapter, it is useful to define two different but related average molecular weights. The first is the weight average molecular weight of the aggregates (M_w^{agg}). This is defined mathematically as

$$\frac{M_w^{\text{agg}}}{M_{\text{mon}}} = \frac{\sum_{j>1} j^2 a_j}{\sum_{j>1} j a_j} \quad (1a)$$

or in terms of quantities that are measureable using inline SEC-MALS

$$\frac{M_w^{\text{agg}}}{M_{\text{mon}}} = \frac{\sum_i^{(\text{agg})} c_i M_i}{\sum_i c_i} \quad (1b)$$

In Eq. (1a, 1b), M_{mon} is the monomer molecular weight, a_j is the number concentration of aggregates composed of j monomers, c_i is the mass concentration of protein within the i th small “slice” of eluent from the SEC column (typically $\sim 10 \mu\text{L}$ of eluent), and M_i is the molecular weight (from MALS) for that i th slice of eluent. The summation in Eq. (1a) is over all aggregate species (i.e., $j > 1$), provided all aggregate sizes remain soluble (cf. Table 1). The summation in Eq. (1b) is over all slices within the aggregate peaks. Equation (1b) holds as equivalent to Eq. (1a) if the assumptions in Table 1 hold regarding aggregate solubility and quantitative elution from the column.

If one cannot properly achieve baseline resolution between monomer and the aggregate peak(s), then the more appropriate quantity to monitor is the weight average molecular weight of the total protein population (M_w). By analogy to Eq. (1a, 1b), M_w is defined by

$$\frac{M_w}{M_{\text{mon}}} = \frac{\left(m + \sum_{j>1} j^2 a_j \right)}{\left(m + \sum_{j>1} j a_j \right)} \quad (2a)$$

or

$$\frac{M_w}{M_{\text{mon}}} = \frac{\sum_i^{(\text{tot})} c_i M_i}{\sum_i c_i} \quad (2b)$$

More details regarding when M_w or M_w^{agg} is more appropriate are provided in Subheadings 3 and 4. It is also notable that M_w is equivalent to that obtained from a batch static light scattering experiment, and thus one could in principle perform batch MALS instead of inline MALS if needed or preferred.

1.2. Size Exclusion Chromatography

In simple terms, SEC is based on the ability of a porous stationary phase to retain smaller sized solutes (in comparison to pore size of stationary phase) within its maze-like structure longer than the larger solutes, which pass only through larger pores, and/or between the stationary phase particles without penetrating into their pores. In a system with controlled flow rate, such differences in migration routes result in separation of solutes into distinct peaks: higher molecular weight (MW) species eluting faster than the lower MW solutes with longer retention times.

In classical SEC analysis, the sizes of the solutes can be estimated based on the comparison of their retention times to that of standards with known MWs. However, size determination via SEC is performed under four major assumptions:

1. The molecule of interest is assumed to have similar geometric compactness, compared to standards (typically globular proteins).
2. The stationary phase is assumed to be “inert” towards the components of the mixture to be separated. Any interactions (e.g., hydrophobic or van der Waals attractions, hydrogen bonding, electrostatics, etc.) between solutes and stationary phase will affect retention times and consequently the apparent molecular weights based on SEC.
3. The MW distribution deduced from column retention times is assumed to be the same as that in the sample prior to injection. However, in case of easily reversible oligomeric species with weak associations, this assumption is likely not valid. Since injection of solutes into a column through a flowing mobile phase is essentially diluting the solute, and/or altering the solvent environment unless the mobile phase is the same as the sample conditions, this can result in a difference between apparent monomer/oligomer distributions (via SEC) from that of the bulk sample unless care is taken for weakly (reversibly) aggregating systems (cf. Subheading 4).
4. Even the largest MW species in the sample is assumed to be capable of eluting within one column volume without the observation of a “filtering effect” from the SEC column. In a case where filtering of large MW species via the column occurs, the mass balance between monomeric and aggregated samples will be violated and the distribution obtained from SEC-MALS will not represent the actual MW distribution of the sample.

1.3. Multi-angle Laser Light Scattering

Static laser light scattering (SLS) is based on the time-averaged intensity of scattered light at a given angle (θ) between the incident beam and the detector. The scattered intensity is calibrated against a known reference standard, and the background from solvent is subtracted, to yield the excess Rayleigh ratio (R^{ex}) for a given (mass/vol) concentration of protein (c), with c measured independently (e.g., by UV or RI detection). If the average radius of gyration (R_g) of monomer and/or aggregate(s) within the scattering volume is small compared to the wavelength of laser light (λ), then R^{ex} is effectively independent of θ , and $R^{\text{ex}} = KcM_{w,\text{app}}$ in the limit of low c , with K defined to correct for optical and geometric factors, as well as the dependence of refractive index (n) on c at a given solvent composition. The apparent weight-average molecular weight ($M_{w,\text{app}}$) is historically assumed to be equal to the true M_w , although this is not strictly true if there are strong solvent–protein interactions (16, 17). However, it may be possible to correct for this if the same deviation occurs for both monomer and aggregates (see cf. Subheading 4, Note 11). This is a separate issue from assuming that the second osmotic virial coefficient (B_{22}) is negligible under these conditions (17).

When R_g is sufficiently large, such that λ/R_g is small (of order 10 or smaller) (18), R^{ex} has a detectable dependence on θ , and canonical analysis uses the Zimm expression to relate R^{ex} to M_w

$$\frac{Kc}{R^{\text{ex}}} = \left(\frac{1}{M_w} + 2B_{22}c \right) \left(1 + \frac{q^2 R_g^2}{3} \right) \quad (3)$$

with the magnitude of the scattering vector (q) defined (19) as $4\pi n\lambda^{-1}\sin(\theta/2)$. Strictly, R_g in Eq. (3) is the z -average value across all species in the scattering volume.

As a result of Eq. (3), extrapolating Kc/R^{ex} as a function of $\sin^2(\theta/2)$ gives $1/M_w$ as the intercept in the limit of low c . Note: the same issues with M_w vs. $M_{w,\text{app}}$ in principle exist for aggregates as they do for monomers. Thus, although it has not been considered systematically in the literature, it is expected that M_w in Eq. (3) and analogous expressions should be replaced with $M_{w,\text{app}}$ as discussed above (see also Subheading 4). In the limit of large R_g ($\sim\lambda$ or greater), a more complex dependence of R^{ex} on θ is expected unless one confines analysis to the Guinier region, where Eq. (3) is a good approximation (18).

Inline MALS instruments are composed of multiple detectors placed around a minimal-volume flow cell at various fixed angles with respect to the incident beam. Simultaneous measurement at multiple angles, and the small footprint of these instruments enable serial connection of MALS to size-based fractionation techniques such as SEC and F4, in combination with multiple concentration detectors (e.g. UV-VIS, refractive index). This allows one to

perform multi-angle SLS and apply Eq. (3) or its simpler, angle-independent counterpart, so as to obtain weight-average molecular weight values of the column eluent as a function of the elution time or retention time (t_R). Note: the average molecular weight can depend on t_R because different species elute at different times, and multiple species may co-elute at a given t_R . As noted above, R^{ex} is itself a time average. Thus, in practice, R^{ex} is obtained over a small but finite “slice” or range of elution times from the column (Δt_R). The value of Δt_R is intended to be sufficiently small that the M_w , c , and R_g can be treated as effectively constant during Δt_R , but large enough that one obtains a decent signal-to-noise ratio for R^{ex} for a given “slice” of eluent from the column.

2. Materials

Based on the principles of SEC and MALS, there are a number of key factors in utilizing SEC-MALS effectively. At a minimum, selection of the SEC conditions that prevent experimental bias is crucial for obtaining reliable analytical results. Table 2 provides some scenarios that may result in such potentially biased results and suggestions on how to diagnose the presence of these problems. This will also be revisited in Subheading 4. The primary materials and instrumentation that are needed for SEC-MALS are:

2.1. SEC Column

SEC columns are available from a variety of vendors, and are selected based on the putative MW range and/or hydrodynamic radius range of interest, based on column specifications from the vendor, and any limitations of column lifetime depending on mobile phase conditions (e.g., pH and organic content). It should be noted that certain column-HPLC combinations were reported to cause “column shedding”: an increase in the amount of particles released from the stationary phase itself, potentially causing interference in light scattering measurements at early retention times (typically earlier than excluded volume of the column). For detailed information on achieving a cleaner light scattering baseline via using compatible columns, HPLC systems and light scattering instruments, reader is advised to consult to the vendor of the specific light scattering instrument.

2.2. Mobile Phase Conditions

Mobile phase conditions can vary widely from method to method, but should be selected to minimize the interactions of protein (monomers and aggregates) with the column material, and assure quantitative recovery of all protein material from each injection; it is also desirable to have baseline resolution between monomer and small aggregates (e.g., dimers) if possible; the most common variables in adjusting mobile phase conditions are pH, buffer/salt

Table 2
Possible scenarios that can result in inaccurate representation of bulk sample in light scattering measurements and suggestions to diagnose such potentially skewed results

	Mobile/stationary phase in chromatography (diagnostic method given in italics)	Aggregation/storage conditions (diagnostic method given in italics)
Reversible/irreversible aggregates: SEC-MALS is suitable to study irreversible aggregates	Mobile phase should not alter oligomerization state of aggregates <i>Use of alternate mobile phases/flow rates/order of injection or comparison of overall M_w obtained from SEC-MALS to batch SLS</i>	Conditions at which aggregates are generated and stored should be selected so that aggregates are not reversible. <i>Injecting the same sample at different storage times after quenching</i>
Aggregate solubility: SEC-MALS is limited to the study of soluble aggregates	Mixing with mobile phase should not result in precipitation of component(s) of sample <i>Spiking aliquots of aggregate samples into mobile phase candidates to assess solubility changes</i>	Precipitation during generation or storage of aggregates needs to be visually assessed, noted and only soluble contents should be injected. <i>Careful visual observation of samples at every stage of sample handling</i>
Conservation of mass balance: Proper quantitative kinetic analysis of aggregation is dependent upon conservation of mass balance between monomer and aggregates in solution at different time points	Mobile/stationary phases should be selected/optimized to minimize protein-column interactions. Saturation injections with concentrated monomer + aggregate mixtures should be included in sample queues. <i>Changes in peak height/areas between multiple injections of same dilute protein samples (without saturation)</i>	NA

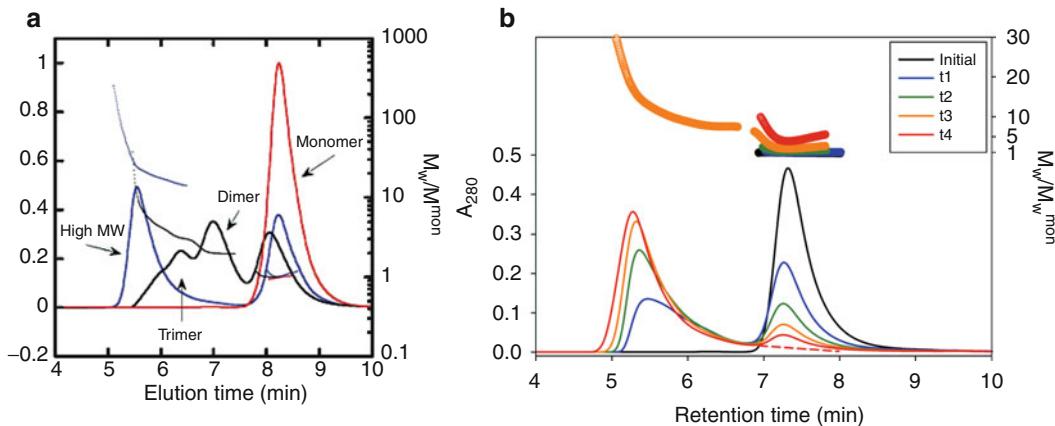


Fig. 1. Examples of information that can be obtained from SEC-MALS runs on thermally stressed/aggregated protein samples. Overlaid chromatograms (thin curves) represent SEC traces (left y-axis) while points represent 1-second slices for MALS data (right y-axis). M_w is the M_w value from MALS for an unheated, monomer sample. Panel (a) is from Brummitt et al. (2) for an IgG1 antibody across a range of acidic conditions, while panel (b) is for a different antibody at pH 4.5 and higher ionic strength conditions, but as a function of incubation time (labeled t1, t2, etc.) at elevated temperature.

concentration, and buffer/salt type; typical mobile phase flow rates are based on maximum column pressure drop balanced against achieving the most rapid separation so as to minimize peak dispersion.

2.3. HPLC and MALS Instruments

There are a variety of HPLC or FPLC instruments to select from. Simple isocratic elution conditions are typically all that is required. For MALS instruments there are also multiple options available commercially, depending on whether one is interested in multi-angle vs. single angle detection, as well as differences in price and software capabilities.

3. Methods

3.1. SEC-MALS for Samples Containing Oligomeric (Small) Aggregates

In the context of this chapter, “small” aggregates are those that are not fully excluded from the column, while high molecular weight (HMW) aggregates are fully excluded. Often, one can expect to reasonably separate dimer from monomer with near-baseline resolution. But it is often not possible to separate larger aggregates with baseline resolution. As a result, there is often overlap of species with different molecular weights within early retention-time peaks that may otherwise appear (by UV-VIS or RI) to behave as single species in a chromatogram. In these cases, one often observes MALS results for M_w as a function of retention time that are similar to what is shown in Fig. 1a (black curve). For example, there is essentially a constant value of M_w (i.e., a “flat” profile) across the middle of a monomer or dimer peak, coinciding with a peak maximum. When

there is overlap of a tail of an aggregate peak into a monomer peak, this often results in an upward-curving molecular weight profile for a monomer or dimer peak in SEC-MALS. This follows because the aggregates are a negligible fraction of the monomer (dimer) peak near the middle of the peak, but their stronger scattering power gives them a significant contribution at the leading edge and/or tail of the monomer (dimer) peak. This is illustrated by the behavior in Fig. 1a for the black and blue curves.

There also may be artifactual curving M_w values at the tails of a given peak (not shown). This behavior for the leading and trailing M_w profiles for peaks such as pure monomer or dimer are artifacts that are due, at least in part, to the low concentration of species in that portion of the chromatogram, and the low scattering power of the species since they are not large aggregates. As a result, the scattering intensity is not sufficiently large to be reliable; this can also result in jagged or “noisy” profiles of M_w vs. retention time. In addition, although the MALS instrument may return a fitted value for R_g , this value is not reliable unless it satisfies the considerations in Subheading 1 regarding when one is properly in the Guinier regime for R_g , relative to the wavelength of the laser.

3.2. SEC-MALS with HMW Aggregates

HMW aggregates often co-elute in the exclusion volume of the SEC column, and therefore exist as a mixed pool within a common (excluded) HMW peak (e.g., as in Fig. 1a or b). While it is sometimes tempting to think of each slice of such a peak as being homogeneous, this is not always the case (20, 21). As such, it is recommended that one instead uses only the average molecular weight across the entire peak (M_w^{agg}), or that across both aggregate and monomer peaks (M_w —sometimes denoted M_w^{tot}).

To determine whether M_w^{agg} can be reliably quantified, one must consider whether a significant mass of aggregates co-elutes with the monomer peak. Figure 1a (black or blue curves) shows an example where aggregates do co-elute significantly with monomer. As noted above, a characteristic observation in this case is that the M_w value from slices across the monomer peak are constant, and consistent with known monomer molecular weight for samples without aggregate present, but are significantly higher and have tails to higher values once the aggregate peak(s) is(are) present. This behavior can also happen with overlap of dimer or small oligomer peaks if the column is overloaded with the aggregate species (2). If such overlap occurs, then using M_w^{agg} from just the aggregate peak(s) can significantly over- or under-estimate the true M_w^{agg} value. This follows because there is a significant portion of the aggregate population that is “lumped” in with the monomer peak, and because the overlap with monomer peak is due to column overloading, not simply because the aggregates are similar in size to the monomer. In such cases, one should instead use M_w^{tot} obtained from integrating over both aggregate and monomer peaks (2, 4, 22). Additional considerations are included in Subheading 4.

3.3. Assessing Aggregation Mechanisms and Models Based on SEC-MALS Data

SEC-MALS can be used in at least two ways to test aggregation models. One is by using MALS to identify the molecular weight of different, well-defined SEC peaks such as monomer and dimer, and then simply tracking the concentration (peak area) using just SEC for those species as a function of incubation time. Unfortunately, in practice SEC is often only able to reliably quantify changes in monomer concentration, because dimer and higher-aggregate peaks convolute within the chromatogram(s). Alternative fractionation techniques may be able to better rectify small species (dimers, trimers, etc.), but ultimately high-molecular-weight (HMW) aggregates will be too large to realistically resolve into separate peaks for quantification by peak area or height. This leads naturally to the use of batch MALS or inline MALS to instead quantify M_w for the sample as whole (batch or inline mode), or for the sample after monomer and any other well-resolved peaks have been separated. In the latter case one obtains M_w^{agg} —a weight-averaged molecular weight of the HMW aggregates. If one uses the time dependence of m and M_w or M_w^{agg} , this is equivalent to a population-balance approach in which one monitors the first and second moments of the protein size distribution (3).

To proceed, we first consider a simple but potentially powerful “model-free” application of SEC-MALS to analyze aggregation pathways. This can be used alone, or also combined with the next approach if one has one or more specific mathematical mass-action models to test. Based on the methods and quantities defined in Subheadings 1 and 2, and earlier in this section, one first obtains m and M_w or M_w^{agg} as a function incubation time (t) at the solvent conditions and temperature of interest.

For each time point, there is a set of values for m and average molecular weight. If one has distinguishable aggregate peaks, one simply integrates them together. A plot of M_w^{agg} vs. $(1 - m)$, or M_w vs. $(1 - m)^2$ is then constructed, such as in Fig. 2a or b. Plotting in this way eliminates time as an explicit variable, and allows one to focus on qualitative aspects of the aggregation mechanism that may be common across a range of sample conditions and/or c_0 values (1, 2, 20), even though the relative time scales differ greatly for those different conditions. When plotting as in Fig. 2a or b, if the data are reasonably linear after the initial increase in molecular weight (e.g., solid curves) then aggregate growth is primarily via monomer addition; conversely if M_w^{agg} (or M_w) increases more rapidly than linearly in $1 - m$ (or $(1 - m)^2$)—i.e., an “upturn” in Fig. 2, then aggregate–aggregate coalescence contributes significantly to growth of soluble aggregates. Often, the latter behavior is also a precursor to macroscopic particle formation (1, 2, 4). Alternatively, the curves in Fig. 2 may be relatively flat, indicating that little aggregate growth occurs as new aggregates are formed (or “nucleated”).

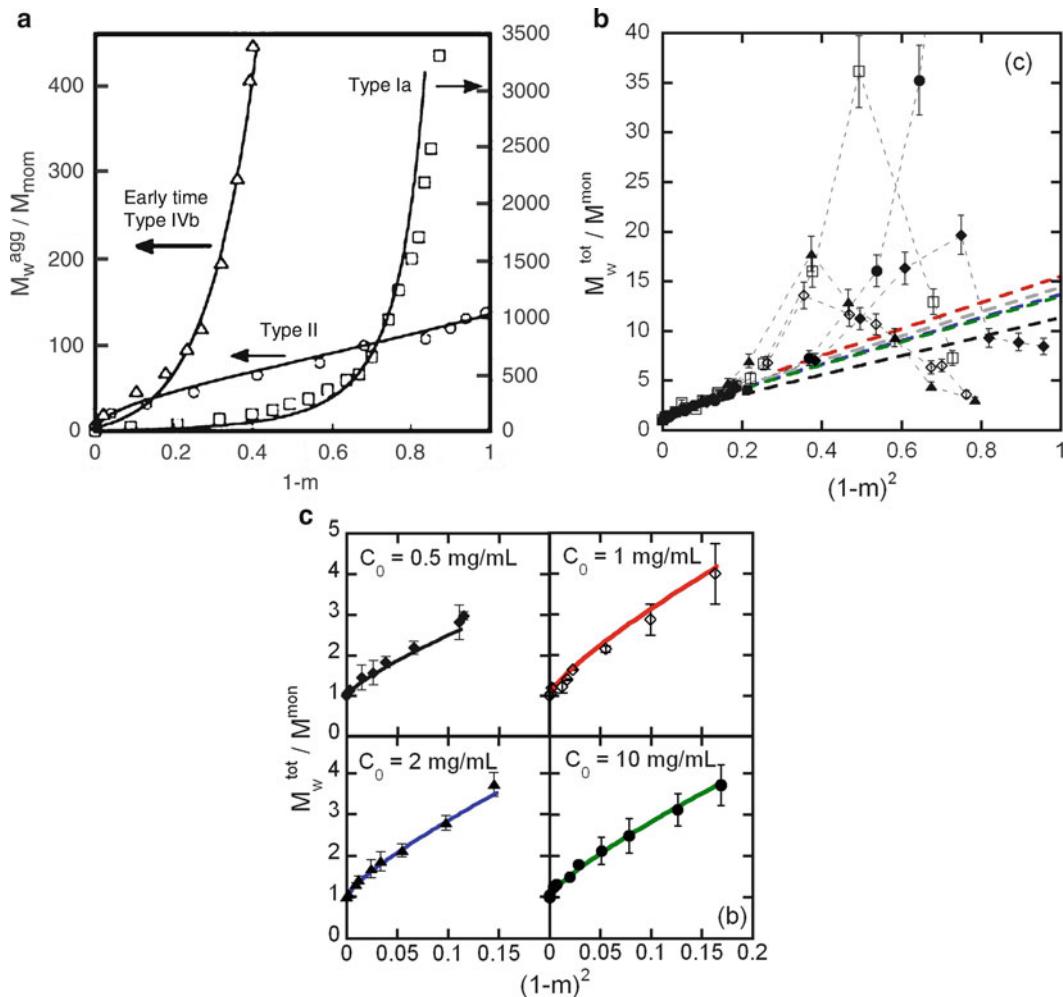


Fig. 2. Examples of qualitative and quantitative mechanistic analysis of SEC-MALS data for cases where aggregates are HMW and cleanly resolved from the monomer peak (panel (a), from Li et al. (20)), and for cases where dimer or HMW convolutes with the monomer peak (panels (b) and (c), from Brummitt et al. (2)). See text for additional details.

Physically, the above analysis holds because aggregate coalescence causes the average molecular weight to rise without a concomitant loss of monomer, while monomer addition causes M_w^{agg} to increase linearly with respect to loss of m (or increase of $1-m$). Of course, neither M_w nor M_w^{agg} will increase significantly if there is negligible growth. These behaviors can also be shown succinctly in mathematical terms by rewriting Eq. (1a) in an equivalently form as (3, 23)

$$\frac{M_w^{\text{agg}}}{M_{\text{mon}}} = \left(\frac{M_w}{M_n} \right)^{\text{agg}} \frac{M_n^{\text{agg}}}{M_{\text{mon}}} = \text{PD}^{\text{agg}} \frac{(1-m)}{\sum_{j>1} \alpha_j} \quad (4)$$

In Eq. (4), M_w^{agg} is the number-average aggregate molecular weight, PD^{agg} is the polydispersity of the aggregate size distribution, and the sum across a_j in the denominator is equal to the total number of aggregates (molar basis), divided by the initial protein concentration (molar basis). For systems in which aggregates do not grow rapidly by coalescence, PD^{agg} stays close to 1; if growth by monomer addition is rapid compared to initiation (or nucleation) of new aggregates, then the total number of aggregates quickly reaches a steady value (3, 23). As a result, M_w^{agg} is approximately linear in $1 - m$ after the short lead time for the summation in the denominator to reach a steady value. If aggregates do not grow significantly, then it is straightforward to show that $1 - m$ and that summation are proportional to each other (3), making M_w^{agg} effectively independent of $1 - m$ (i.e., essentially a flat line in Fig. 2).

If aggregate growth is via aggregate–aggregate coalescence, then the polydispersity will increase continuously, and the summation in the denominator will decrease continuously as aggregation proceeds. Thus M_w^{agg} will increase more greatly than linearly in $1 - m$. To see how to convert this analysis to M_w , one simply needs to use Eqs. (1a) and (2a), along with the identity that $1 + \sum_{j>1} j\alpha_j = 1$ to show that

$$\frac{M_w}{M_{\text{mon}}} = m + \sum_{j>1} j^2 \alpha_j = m + (1 - m) \frac{M_w^{\text{agg}}}{M_{\text{mon}}} \quad (5)$$

From Eqs. (4) and (5), it then follows that M_w (or M_w^{tot}) is plotted vs. $(1 - m)^2$ (cf. Fig. 2b) when seeking an analogous analysis to Fig. 2a for M_w^{agg} . For the examples in Fig. 2, it then follows easily that one can neglect aggregate–aggregate coalescence under circumstances such as the low-pH and low-ionic strength conditions for chymotrypsinogen (Fig. 2a, labeled Type II), or for the “early time” (i.e., low extents of monomer loss) portion of the aggregation process for the monoclonal antibody in Fig. 2b. This can greatly simplify any subsequent mathematical analysis and proposed mass-action models for those conditions. In contrast, one necessarily must consider coalescence mechanisms for the other cases or regimes shown in Fig. 2.

Finally, if one seeks to regress rate coefficients for a given mass-action model based on SEC-MALS data, the following illustration highlights some of the important features. As this chapter is not promoting any one particular model over another, the reader is referred to original sources for details of particular models. For conciseness, we only illustrate the procedure for data analysis here. The curves in Fig. 2a are based on a non-linear least-squares fit of M_w^{agg} vs. t , simultaneously with m vs. t (the data are simply plotted as M_w^{agg} vs. $1 - m$ for purposes noted above). The particular model

in that case included a term for nucleation, one for growth by monomer addition, and one for growth via coalescence. However, it was noted in subsequent analysis that alternative models for coalescence were statistically equivalent (3). This is an inherent limitation if one uses only SEC-MALS for analyzing aggregation kinetics, because distinguishing different coalescence mechanisms requires additional information, such as higher moments of the aggregate size distribution, or detailed morphological characterization (3, 24).

Figure 2c shows only the “early-time” regime from Fig. 2b, with different c_0 values in the different sub-panels. The curves are fits to an analogous model to that used for fitting in Fig. 2a. In the authors’ experience with globular and multi-domain proteins across a range of temperature, pH, and salt conditions (1, 2, 4, 20, 22, 25, 26), a majority of conditions that promote aggregation fall into one of three categories regarding kinetics monitored by SEC-MALS: (1) growth by condensation is negligible over multiple half lives of monomer loss; (2) growth by condensation is negligible if one considers only early-time regimes; (3) coalescence/condensation is the dominant mechanism of growth, and results in almost immediate formation of large, effectively insoluble aggregates once detectable monomer loss occurs. As highlighted above and elsewhere (1, 2, 20, 22), the first two scenarios allow one to quantitatively regress SEC-MALS data with models that provide quantitative determination of nucleation and growth time scales. Separating those time scales is not possible with monomer loss alone—i.e., via SEC or otherwise (5, 23, 25).

In scenario (3), it is arguable whether MALS data is much use at all if one is interested in quantifying the steps that consume monomer; in this case, aggregate growth has little contribution from monomer addition, and therefore the primary rate-limiting step(s) for monomer loss are at the nucleation stage(s) (5). This provides a more concrete example of the basis for the recommendations in Table 1 regarding when SEC-MALS is appropriate (e.g., scenarios (1) and (2) above), vs. when SEC alone may be sufficient—i.e., scenario (iii) if one is focused on the mechanism(s) or rate(s) of monomer loss.

4. Notes

This section provides a brief discussion and set of points to consider when developing and trouble-shooting SEC-MALS for characterization and analysis of protein aggregation mechanisms. At a minimum, Notes 1–3 should be first considered when developing the SEC-MALS method prior to performing mechanistic tests regarding aggregation. The later notes refer to issues or considerations

when determining if SEC-MALS is appropriate for use with a given aggregation phenomenon, and/or when analyzing the resulting SEC-MALS data.

1. Some simple tests to assure the mobile phase and/or stationary phase is not significantly altering the aggregation state of the system include: (1) visual inspection and/or turbidity measurements for sample injections that are spiked into mobile phase, to assure aggregates (and monomers) remain soluble in the mobile phase; (2) comparison of total integrated peak area(s) and overall M_w for samples injected with and without a column in place, to assure minimal losses to the column; (3) systematically altering the volume of the injection of the same sample, and/or the flow-rate on the column, to test whether aggregates can reverse or grow significantly during elution. Of course, comparison with orthogonal methods (e.g, F4 or AUC) is also recommended, acknowledging that such methods have their own limitations and/or logistical considerations.
2. In many cases, maintaining aggregate solubility and minimizing adsorption to the column will require one to work under mobile phase conditions that are significantly away from the pI of the protein of interest, and at intermediate ionic strength (often $\sim 10^2$ mM). Much higher ionic strength can jeopardize aggregate solubility, as electrostatic repulsions have been found to help minimize aggregate coalescence (1, 2, 27). In the authors' experience, purely aqueous mobile phases, along with organic or inorganic buffer salts and ionic strength modifiers are often sufficient if one is not working with highly hydrophobic entities such as membrane-associated proteins.
3. Carefully select the pore-size (or size distribution) and the packing material for the SEC column. As an example, SEC-MALS was successfully used to characterize aggregates of alpha-chymotrypsinogen A (aCgn) that were as large as hundreds of nanometers in length, and with $M_w^{agg} \sim 10^4$ kDa (20). However, unlike conventional wisdom regarding SEC, this required SEC columns with average pore sizes that were significantly smaller than the aggregates in question; otherwise the aggregates adsorbed effectively irreversibly to the column. Presumably, this occurred because such large aggregates could not extricate from the tortuous pore network once they had penetrated significantly into that network. As such, if such large yet soluble aggregates (sometimes termed sub-visible particles) are a significant mass fraction of the size distribution, then SEC-MALS may perform best under conditions where the aggregates are fully excluded from the intra-particle pore volume of the column (20). Additional discussion of HMW aggregates is included below.

4. Reversible vs. Irreversible Aggregation: Self-association of biomacromolecules via weak interactions is often concentration dependent; SEC (and thus SEC-MALS) may provide inaccurate results due to some intrinsic limitations of most commercial instruments: (1) they require relatively dilute protein samples (~10 g/L or less) due to detector limitations and column capacities; (2) samples face significant dilution and dispersion upon injection into the mobile phase and during migration in the column and detectors. As a result, it should not be surprising to have SEC analytical results that do not correctly represent the ratios (or even presence/absence in some cases) of multimers to monomers in a concentrated protein sample if the timescale of association/dissociation is faster than the timescale of the SEC experiment.
5. In some cases, oligomerization may not be reversible in the original sample buffer, but instead is induced by the mobile phase used for SEC. To prevent inaccurate representation of bulk sample in SEC, mobile phases should be selected so that oligomerization is not reversed. To diagnose such issues, one can refer to the guidelines in and Table 2 regarding SEC method selection; in addition, for SEC-MALS one can compare the M_w results across chromatograms from different SEC conditions to assure consistent results.
6. Further to Notes 4 and 5, compare M_w (i.e., M_w^{tot}) results from SEC-MALS to M_w values obtained from batch mode light scattering or analytical ultracentrifugation (AUC) measurements performed in the original sample conditions. This obviously must be done judiciously, as to perform such tests on every sample defeats the purpose of using inline SEC-MALS for greater automation and minimizing sample requirements compared to batch light scattering or AUC. Some of these concerns are mitigated if one is instead using F4-MALS, although F4 has its own potential pitfalls for aggregating systems (21). Finally, one must also consider that aggregates may continue to grow—for example via aggregate coalescence—during sample storage on autosamplers or otherwise, prior to injection on SEC-MALS. If this occurs rapidly compared to the time required for injecting the necessary samples, then in situ monitoring of the system may be required, such as by batch MALS alone (cf. Table 1 and discussion in Subheading 1).
7. Aggregate Solubility: Subheading 2 and Tables 1 and 2 highlight the need to assure aggregates remain soluble in the mobile phase—i.e., do not coalesce or otherwise precipitate. In addition, aggregate solubility can be temperature dependent, with lower solubility at refrigerated conditions compared to room temperature or accelerated conditions.

The authors have observed this phenomenon on multiple occasions (unpublished results), and it appears to often be overlooked by many workers.

8. Although the solubility of aggregates, as opposed to monomers, has not been systematically considered except in a few published cases (1), in the authors' experience two of the most important variables are pH and ionic strength of the mobile phase. Unfortunately, these parameters are also those that often must be adjusted to minimize protein interactions with SEC columns. At present, it remains an empirical exercise to optimize mobile phase conditions to balance all of these factors.
9. Subheading 3 and notes earlier in this section included tests for significant loss of protein mass due to adsorption, based primarily on assuring that total chromatogram area (over all peaks) is robust. It should not be overlooked that aggregates may become more easily adsorbed as they grow and evolve. Thus, monomer-only samples, or those with only small aggregates should not be the only ones used for SEC method development. In addition to the obvious "filtering" effects on very large aggregates/particles that is inherent to SEC, one should also consider that monomers from stressed samples may be structurally altered, and so may interact differently with a given stationary phase. Incorporation of "saturation injections" in SEC methods is a practice that the authors advocate in order to obtain better conservation of peak areas between samples with same total protein content (such as time points in a stability study). Saturation injections are of course common practice in chromatography, but it is often not appreciated that SEC with low levels of soluble aggregates may require many more saturation injections than is commonplace in other forms of analytical chromatography. Given the considerations above regarding differences in adsorption for monomers vs. aggregates, such saturation injections should contain a mixture of monomer and representative aggregates.
10. There are at least two caveats one must consider when using light scattering to obtain accurate values of molecular weight. These were alluded to in earlier sections, but are described in more detail here. The first issue is that M_w values are determined by the slope of R^{ex} vs. protein concentration (c) (see also Subheading 1). In SEC-MALS, one does not perform a series of dilutions of the same sample to obtain R^{ex} vs. c for each chromatographic slice. Instead, one uses the angular dependence and the Zimm equation (Subheading 1) if species are sufficiently large compared to the wavelength of the laser. Unfortunately, for many protein systems—large antibodies included—monomers and small oligomers (dimers, etc.) do

not have a significant angular dependence; as such, detection at multiple angles in MALS does not provide more than replicate measures of purely Rayleigh scattering for the sample or eluent “slice” for these species. Thus, the M_w value for a given slice from SEC-MALS for these species is essentially determined from a slope that uses just two concentration points—zero and the c corresponding to that detected in SEC. The net result is that accurate quantitative M_w^{agg} or M_w^{tot} values may be difficult to achieve when one is considering samples that have only small quantities of aggregates that are not large enough to have significant angular dependence to their scattering.

11. Further to Note 10, a second and much less commonly acknowledged issue is that, rigorously, it is not correct to assume that the slope of R^{ex} vs. c gives the absolute M_w . The original work by Stockmayer (16), as well as more recent and general treatments (17) show that in fact one only obtains an apparent molecular weight ($M_{w,\text{app}}$). The reason it is only apparent is that the slope of R^{ex} vs. c in the limit of low c has contributions from both M_w and interactions between the solvent and protein, and between any cosolutes or cosolvents and protein. To the best of the authors’ knowledge, the deviations due to these non-idealities have not been studied systematically. In lieu of a means to currently correct for these non-idealities, one can instead consider a reasonable hypothesis to be that the magnitude of the non-idealities will be similar in magnitude and sign, on a per-mass or per solvent-exposed surface area (SSA) basis, for a given protein and its aggregates if they are considered in the same solvent. Thus, a simple if only heuristic “fix” for this issue is to scale all M_w values obtained for aggregates in SEC-MALS by the value one obtains for a purely monomer sample (if one exists) under the same mobile phase conditions, and not to assume that M_w from a monomer sample should match the value obtained from, for example, amino acid analysis.
12. A common misconception when using SEC-MALS is that quantitative SEC-MALS analysis requires baseline-level resolution between peaks. Indeed, SEC-MALS can be used to provide absolute molecular weights for well-resolved peaks. However, it is important to highlight that valuable information can still be obtained in case of high polydispersity mixtures without clean separation, using the procedures noted in Subheading 3 regarding M_w^{tot} and M_w^{agg} (20).
13. Finally, it is important to highlight that researchers do not need to have access to an inline SEC-MALS instrument to study protein aggregation via SEC and light scattering. The same mechanistic understanding (using variants of the same

regression models) can be obtained using a combination of batch light scattering measurements on non-separated monomer-aggregate mixture (to obtain M_w^{tot}) and SEC chromatogram (to obtain monomer fraction) for the same sample. As was shown in Subheading 3, M_w^{agg} or M_w^{tot} (or M_w) give equivalent information if one has reliable values for monomer fraction (m). In this regard, one must be careful to not greatly overload columns with HMW aggregate, such that they significantly overlap with the monomer peak. Doing so artificially increases the area of the monomer peak, and then jeopardizes the accuracy of m values from peak area or peak height. This is an even greater problem when one uses UV absorbance to quantify monomer area, as HMW aggregates that overlap with monomer both absorb and scatter the incident UV light.

References

1. Li Y, Ogunnaike BA, Roberts CJ (2010) Multivariate approach to global protein aggregation behavior and kinetics: effects of pH, NaCl, and temperature for α -chymotrypsinogen A. *J Pharm Sci* 99:645–662
2. Brummitt RK, Nesta DP, Chang L, Kroetsch AM, Roberts CJ (2011) Nonnative aggregation of an IgG1 antibody in acidic conditions, part 2: nucleation and growth kinetics with competing growth mechanisms. *J Pharm Sci* 100:2104–2119
3. Li Y, Roberts CJ (2009) Lumry-Eyring nucleated-polymerization model of protein aggregation kinetics. 2. Competing growth via condensation and chain polymerization. *J Phys Chem B* 113:7020–7032
4. Sahin E, Grillo AO, Perkins MD, Roberts CJ (2010) Comparative effects of pH and ionic strength on protein–protein interactions, unfolding, and aggregation for IgG1 antibodies. *J Pharm Sci* 99:4830–4848
5. Roberts CJ (2007) Non-native protein aggregation kinetics. *Biotechnol Bioeng* 98:927–938
6. Lomakin A, Teplow DB, Kirschner DA, Benedek GB (1997) Kinetic theory of fibrillogenesis of amyloid β -protein. *Proc Natl Acad Sci U S A* 94:7942–7947
7. Goldstein RF, Stryer L (1986) Cooperative polymerization reactions. Analytical approximations, numerical examples, and experimental strategy. *Biophys J* 50:583–599
8. Powers ET, Powers DL (2006) The kinetics of nucleated polymerizations at high concentrations: amyloid fibril formation near and above the “supercritical concentration”. *Biophys J* 91:122–132
9. Lee C, Nayak A, Sethuraman A, Belfort G, McRae GJ (2007) A three-stage kinetic model of amyloid fibrillation. *Biophys J* 92: 3448–3458
10. Laidler KJ (1965) Chemical kinetics, 2nd edn. McGraw, New York, NY
11. Weiss WF, Young TM, Roberts CJ (2009) Principles, approaches, and challenges for predicting protein aggregation rates and shelf life. *J Pharm Sci* 98:1246–1277
12. Liu J, Andya JD, Shire SJ (2006) A critical review of analytical ultracentrifugation and field flow fractionation methods for measuring protein aggregation. *AAPS J* 8:E580–E589
13. Chuan YP, Fan YY, Lua L, Middelberg APJ (2008) Quantitative analysis of virus-like particle size and distribution by field-flow fractionation. *Biotechnol Bioeng* 99: 1425–1433
14. Roberts CJ (2003) Kinetics of irreversible protein aggregation: analysis of extended Lumry-Eyring models and implications for predicting protein shelf life. *J Phys Chem B* 107: 1194–1207
15. Morris AM, Watzky MA, Finke RG (2009) Protein aggregation kinetics, mechanism, and curve-fitting: a review of the literature. *Biochim Biophys Acta Proteins Proteomics* 1794: 375–397
16. Stockmayer WH (1950) Light scattering in multi-component systems. *J Chem Phys* 18:58–61
17. Blanco MA, Sahin E, Li Y, Roberts CJ (2011) Reexamining protein–protein and protein–solvent interactions from Kirkwood-Buff analysis of light scattering in multi-component solutions. *J Chem Phys* 134:225103/1–225103/12

18. Zemb T, Lindner P (2002) Neutron, X-rays and light scattering methods applied to soft condensed matter Rev Sub. Elsevier, North Holland
19. Zimm BH (1948) Apparatus and methods for measurement and interpretation of the angular variation of light scattering; preliminary results on polystyrene solutions. *J Chem Phys* 16: 1099–1116
20. Li Y, Weiss WF, Roberts CJ (2009) Characterization of high-molecular-weight non-native aggregates and aggregation kinetics by size exclusion chromatography with inline multi-angle laser light scattering. *J Pharm Sci* 98:3997–4016
21. Philo JS (2006) Is any measurement method optimal for all aggregate sizes and types? *AAPS J* 8:E564–E571
22. Sahin E, Jordan JL, Spatara ML, Naranjo A, Costanzo JA, Weiss WF, Robinson AS, Fernandez EJ, Roberts CJ (2011) Computational design and biophysical characterization of aggregation-resistant point mutations for γ D crystallin illustrate a balance of conformational stability and intrinsic aggregation propensity. *Biochemistry* 50:628–639
23. Andrews JM, Roberts CJ (2007) A Lumry-Eyring nucleated polymerization model of protein aggregation kinetics: 1. aggregation with pre-equilibrated unfolding. *J Phys Chem B* 111:7897–7913
24. Pallitto MM, Murphy RM (2001) A mathematical model of the kinetics of β -amyloid fibril growth from the denatured state. *Biophys J* 81:1805–1822
25. Weiss WF, Hodgdon TK, Kaler EW, Lenhoff AM, Roberts CJ (2007) Nonnative protein polymers: structure, morphology, and relation to nucleation and growth. *Biophys J* 93: 4392–4403
26. Andrews JM, Weiss WF, Roberts CJ (2008) Nucleation, growth, and activation energies for seeded and unseeded aggregation of α -chymotrypsinogen A. *Biochemistry* 47: 2397–2403
27. Krebs MRH, Domike KR, Cannon D, Donald AM (2008) Common motifs in protein self-assembly. *Faraday Discuss* 139:265

Chapter 26

Computational Methods to Predict Therapeutic Protein Aggregation

Patrick M. Buck, Sandeep Kumar, Xiaoling Wang, Neeraj J. Agrawal, Bernhardt L. Trout, and Satish K. Singh

Abstract

Protein based biotherapeutics have emerged as a successful class of pharmaceuticals. However, these macromolecules endure a variety of physicochemical degradations during manufacturing, shipping, and storage, which may adversely impact the drug product quality. Of these degradations, the irreversible self-association of therapeutic proteins to form aggregates is a major challenge in the formulation of these molecules. Tools to predict and mitigate protein aggregation are, therefore, of great interest to biopharmaceutical research and development. In this chapter, a number of such computational tools developed to understand and predict the various steps involved in protein aggregation are described. These tools can be grouped into three general classes: unfolding kinetics and native state thermal stability, colloidal stability, and sequence/structure based aggregation liabilities. Chapter sections introduce each class by discussing how these predictive tools provide insight into the molecular events leading to protein aggregation. The computational methods are then explained in detail along with their advantages and limitations.

Key words: Biotherapeutics, Aggregation-prone regions, Computational prediction, Monoclonal antibodies

1. Introduction

Protein based biotherapeutics undergo a series of processing steps including production, purification, refolding, freeze-thaw, shipping, drying, filling, filtration, and nebulization, all of which can impact the stability of the drug substance and product (1). Environmental factors such as high concentration, temperature and pH extremes, ionic strength, agitation, shear, air–water interface, and light can also impact stability (2). These stresses cause molecules to degrade by mechanisms such as oxidation, deamidation, fragmentation, denaturation, surface adsorption, and

aggregation (3). Of these degradations, aggregation is the most common and least understood one (4). From a pharmaceutical perspective, predicting and mitigating biotherapeutic aggregation has important consequences. Protein aggregates have the potential to elicit immune response and reduce target binding affinity (5, 6).

In this chapter, we define aggregation as an irreversible form of protein self-association. Aggregates are made up of oligomers that can range in size from a few monomers (dimers, trimers, etc.) to several hundreds of monomers. As aggregates continue to grow in size they eventually lose their solubility and fall out of solution as precipitates. Aggregates that lack any regular structure are amorphous, while highly ordered aggregates form fibril-like structures. Both morphologies can be described as beta-aggregates when they are enriched in beta-strands. Some aggregates bind staining dyes such as Thioflavin T (ThT) and Congo Red that indicate the formation of the cross-beta steric zipper motif (7). This structure is distinguished from regular beta-sheet structure by the distances between strands which form much tighter packing interactions (8, 9). On the other hand, not all aggregates that show enriched beta-sheet structure bind these staining dyes (10). Moreover, not all aggregates are enriched in beta-sheet structure. In this case, oligomerization may result from three-dimensional domain swapping (11–13).

A schematic overview of protein aggregation is shown in Fig. 1a. The process begins with native monomers in solution. These native monomers either unfold to an intermediate state (I) or a near-native state (N^*) that represents an aggregate competent conformation (2) (Fig. 1b). Unfolding to a completely denatured state (U) may not be necessary to initiate aggregation. There are several known examples of proteins aggregating under physiological conditions via a partially unfolded conformation that retains

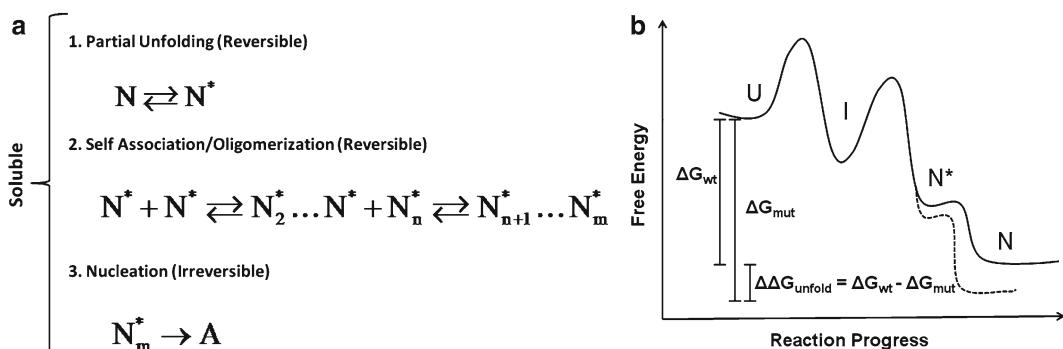


Fig. 1. (a) Mechanism of aggregation for partially unfolded proteins N^* . Later stage aggregation steps such as elongation and condensation are not shown. (b) Reaction scheme for unfolding to the aggregate competent state either N^* , or an intermediate, I. Monomers may not need to completely unfold to the denatured state, U, to initiate aggregation. The dashed line represents the free energy change from a stabilizing sequence mutation.

many features of the native state (14). Monomers remain in solution during the reversible process of unfolding to the aggregate competent state. For near native aggregation, self-association begins when N^* monomers interact to form soluble oligomers. Once the critical number of monomers within the oligomer (N_m^*) is reached, the monomers may bind tighter by undergoing a conformational change (or structural rearrangement) to form an “aggregation nucleus” which initiates irreversible aggregate formation (A). Subsequent polymerization can elongate the nucleus into a protofibril. Condensation of protofibrils can form higher ordered assemblies such as mature fibrils and plaques (15–17).

This chapter focuses on computational methods to predict and mitigate the initial stages of aggregate formation (partial unfolding/unfolding, oligomerization, and nucleation). These aggregation stages are more relevant to biopharmaceutical sciences from the stand point of target binding loss, potential immunogenicity, and safety issues (18, 19). Higher ordered aggregates such as fibrils and plaques are more important to the study of neurodegenerative diseases (20–22), nanomaterials (23), and genetic variation (24). Broadly speaking, the kinetics of protein aggregation depends on the rate-limiting step (25, 26). If unfolding happens slower than the “downstream” stages of aggregation, then the process is unfolding-limited. Similarly, if the observed aggregation rate depends on the rate of monomer self-association then the process is association-limited. Predicting protein aggregation will be most effective when the rate-limiting step is known.

Typically, the challenge of mitigating aggregate formation is met during formulation development studies. These studies seek to establish the dominant route of drug substance degradation and minimize it by optimizing the formulation conditions such as pH, buffer, salt, and excipients. This strategy addresses the *extrinsic* factors (solution conditions) that lead to aggregation and other physicochemical degradations. Currently, the availability of computational tools to predict aggregation based on extrinsic factors is limited. Their development requires comprehensive experimental data on a variety of molecules and conditions which is not publicly available. Biotherapeutic drug candidates can also be formulated *intrinsically* at the late discovery or early formulation development stages, where sequence changes are still allowed. Computational methods to predict the intrinsic factors associated with biotherapeutic aggregation formation are beginning to be developed (27). The success of these methods in mitigating aggregation depends in part on knowledge of the aggregation mechanism and strategizing on how to slow the various steps involved in aggregate formation. Modifying the intrinsic physicochemical liabilities could build quality via rational design and selection of candidates with an optimum mix of potency, developability, and safety. This endeavor aligns well with Quality by Design efforts (28, 29).

The remainder of this chapter consists of three sections. The first section addresses the kinetic and thermodynamic stabilities of the protein native state. We review methods to predict protein unfolding rates and stability (free energy) based on sequence and structure information. In the second section, methods for predicting protein self-association and colloidal stability are presented. In the third section, our focus turns to protein sequence and structural motifs which may be prone to aggregation by acting as “aggregation nuclei.” A synthesis of all these concepts is provided in a summary at the end. Thermodynamics of protein folding, unfolding and aggregation are major areas of study in computational biophysics. A comprehensive review of all the methods and studies is beyond the scope of this book chapter. Instead, we have tried to highlight a few select methods to relay the current state of the art. We would also like to make a distinction between protein stability and pharmaceutical stability. Pharmaceutical stability refers to the physicochemical degradations over the life span of the biotherapeutic drug substance or product. Protein stability on the other hand, involves the thermodynamic and kinetic considerations involved in the native, unfolded, and aggregation states. It is an important component of pharmaceutical stability.

2. Conformational Stability and the Partially Unfolded State

The first step towards irreversible aggregate formation is unfolding or a partial unfolding of the native monomer (Fig. 1a). Aggregation is an unfolding-limited (or conformational change) process when the rate of aggregation is dependent on the rate of unfolding. In this case, aggregation obeys first-order reaction kinetics and aggregate formation occurs at the rate of monomer loss. In experiments, a number of proteins have been found to follow first-order aggregation kinetics (2, 17, 30). In such cases, computational tools that accurately predict unfolding rates have the potential to rank order highly similar molecules under identical experimental conditions. Further, unfolding rate prediction tools can suggest ways to kinetically stabilize the native state by modifying the protein sequence.

2.1. Predicting Unfolding Rates

Although a lot of effort has gone into predicting the folding rates of proteins based on contact order (31), chain length (32, 33), long-range order (34, 35), and secondary structure content (36), several methods exist to predict unfolding rates of proteins as well (37–39). Here, one of the first methods developed by Gromiha et al. (37) is reviewed. Their model uses various amino acid properties to find correlations with a dataset of two and three state folding proteins. The authors use multiple regression to train the parameters of three different equations, one for each structural

class of protein (alpha, beta, mixed), to predict unfolding rates. Correlation coefficients of greater than 0.99 were obtained for all three structural classes/equations based on a leave one out or jack knife test. The authors also derive a single equation for all protein sequences for which the structure class of the protein is not known. The general equation is a combination of the functions derived for alpha, beta, and mixed structural classes. A correlation coefficient of greater than 0.9 was obtained for the general equation in the same test. Their model predicts unfolding rates based solely on the properties of the sequence. When the structure of the protein is known to be all-alpha the following equation was used,

$$\ln(k_u) = 109.4(\pm 0.12) \times N_s + 114.84(\pm 0.11) \times K^0 - 87.44(\pm 0.04)$$

where N_s is the number of surrounding residues, K^0 is the compressibility, and the plus minus represents the parameter fitting error. N_s and K^0 are averaged values for the sequence (P_{avg}) and computed by

$$P_{avg} = \frac{1}{N} \sum_{j=1}^N P(j)$$

where $P(j)$ is the property (N_s and K^0) value of the j th residue and the sum is over all residues in the chain (N). The property values are taken from a set of 49 diverse amino acid properties (physico-chemical, energetic, and conformational) which fall into various clusters analyzed by Tomii and Kanehisa (40). An explanation of these properties can be found in earlier works (41) and their values are available at http://www.cbrc.jp/~gromiha/fold_rate/prop_des.html. For the jack knife test, a dataset of 28 proteins was used to train the 16 parameters of the general equation. However, the authors report that 5 data points are necessary per adjustable parameter. Acknowledging potential overparameterization, a second test was performed on an independent dataset of 16 proteins. The structure class equations performed significantly better than the general equation in this second test.

Even with accurate unfolding prediction and knowledge that aggregation is unfolding-limited, there are a few caveats that need to be considered before using this tool to rapidly rank order biotherapeutic protein candidates. Although unfolding rates depend on experimental conditions, these equations cannot be used to predict unfolding rates for the same molecule under different conditions. Moreover, structure class predictors were trained using unfolding information from small two and three state folding proteins. It is therefore unclear how accurate predictions will be when applied to biotherapeutic proteins, especially antibodies that have large, multidomain structures. Furthermore, some proteins may only need to reach a partially folded aggregate competent state to

aggregate (14) which is distinct from the denatured state of the same molecule. Aside from these usage concerns, any attempt to slow unfolding by sequence modification should take care to avoid introducing aggregation-prone motifs (see Subheading 4) or increasing the self-association propensity.

2.2. Predicting Protein Thermodynamic Stability

Observed aggregation rates depend on the equilibrium distribution of native monomers. Simply stated, the number of monomers in solution that have reached the aggregate competent state always impacts the observed aggregation rate, regardless of the rate-limiting step. Therefore, the stability of the native state ΔG_{un} (Fig. 1b) should be assessed when attempting to rank order similar molecules under identical conditions. Furthermore, strategies to improve protein thermodynamic stability can utilize $\Delta\Delta G$ ($\Delta\Delta G_{\text{un}} = \Delta G_{\text{wt}} - \Delta G_{\text{mut}}$) prediction tools. These programs are quite accurate at predicting stability changes due to single and double point mutations (42–45).

This section reviews Eris, a stability prediction tool developed by Yin, Ding, and Dokholyan (45). Their model incorporates backbone flexibility to remove collisions that result from small side-chain mutations to large ones. These mutations are particularly difficult to predict, as scoring functions are often parameterized by alanine scanning experiments. Eris uses a physical force field that is a weighted sum of van der Waals forces, solvation energy, hydrogen bond interactions, and backbone-dependent statistical energies. The parameter weights were trained to reproduce the native sequences for 34 proteins by protein design (46). CHARMM19 force field parameters for united atoms (47) were used for the VDW terms (E_{attr} , E_{rep}). The solvation term (E_{solv}) is based on the method of Lazaridis and Karplus (48). Hydrogen bond terms are similar to those developed by Kortemme and Baker (49). All remaining energetic terms model backbone dihedral angle energies (50). The statistical nature of this equation gives an approximate ΔG .

$$\begin{aligned}\Delta G = & W_{\text{vdw_attr}} E_{\text{vdw_attr}} + W_{\text{vdw_rep}} E_{\text{vdw_rep}} + W_{\text{solv}} E_{\text{solv}} \\ & + W_{\text{bb_hbond}} E_{\text{bb_hbond}} + W_{\text{sc_hbond}} E_{\text{sc_hbond}} + W_{\text{bb_sc_hbond}} E_{\text{bb_sc_hbond}} \\ & + W_{\text{aa|j,\psi}} E_{\text{aa|j,\psi}} + W_{\text{rot|aa,j,\psi}} E_{\text{rot|aa,j,\psi}} - E_{\text{ref}}\end{aligned}$$

The authors tested their $\Delta\Delta G$ prediction tool (Eris-flexible backbone) on 595 mutants and found a correlation of 0.75 between predicted and experiment stability changes which is comparable to other available tools (45). However, Eris was able to predict stability changes for small to large side-chain mutations better than other tools because of its ability to relax structures. Figure 2 shows plots of predicted versus experimental free energy changes ($\Delta\Delta G$) upon mutations from small to large side-chains

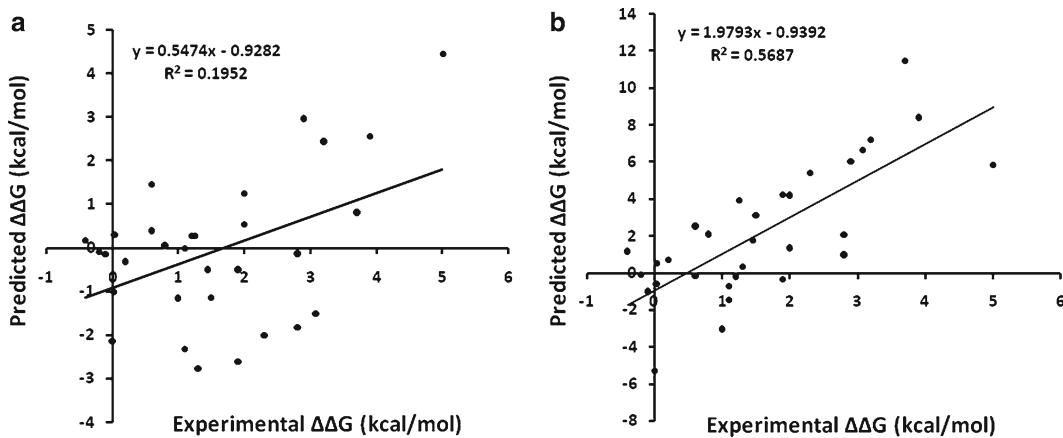


Fig. 2. Plots for predicted versus experimental stability changes ($\Delta\Delta G$) for small to large side-chain mutations using (a) Fold-X with a correlation coefficient of 0.44 and (b) Eris-flexible backbone with a correlation coefficient of 0.75. Correlation coefficients are taken from Yin et al. (45). Plots and fits were performed using data points provided in the supplementary material of Yin et al. (45).

for Fold-X and Eris. The data contained in the supplementary material accompanying the manuscript by Yin and coworkers was used for this purpose. The authors also report that prerelaxation improved their ability to predict stability changes for structures that were not solved at high resolution. For the NMR determined structure of alpha-spectrin domain R16, the correlation between predicted and experimental stability changes was 0.66 and 0.28 for surface and core mutants, respectively. When a prerelaxed structure of the same protein was used, the correlation for core mutants improved to 0.70. The overall correlation for all mutants improved from 0.09 to 0.69. The authors rationalized that imperfections in the structure model were compensated for by the prerelaxation step.

Predicting protein conformational stability changes upon mutation ($\Delta\Delta G$) presents an attractive approach to reducing aggregation by stabilizing the native state. On the other hand, recent studies on sequence variants of γ D-crystallin indicate that conformational stability enhancements may only decrease aggregation rates modestly (51, 52). Two sequence variants of γ D-crystallin were generated by different point mutations. One stabilized the molecule, as measured by urea denaturation, compared to the wild-type. The other, destabilized γ D-crystallin and, at the same time, disrupted an aggregation-prone region of the sequence. The aggregation rate for the destabilized variant was dramatically slowed compared to the wild-type sequence (52). In a separate study on the structural homodimer γ S-crystallin, which contains two Greek key domains, the authors created a variant with a single mutation in one of the structural domains. A second variant was created with a point mutation at the same sequence position of the

other structural domain. Both variants showed decreased thermal stability compared to the wild type (51). However, the less destabilized variant showed a greater propensity to aggregate than the more destabilized one (51). While mutations or environmental changes that destabilize the native structure generally increase the populations of all nonnative states, these cases seem to indicate that simply stabilizing the native state may not significantly decrease the rate of aggregation, particularly when aggregation is association or nucleation-limited. However, increases in stability should have a greater impact on observed rates when aggregation is unfolding-limited (53).

3. Protein Self-Association and Colloidal Stability

Self-assembly of protein monomers to form oligomeric species involves several weak attractive short range inter-molecular interactions including hydrophobic, electrostatic, van der Waals, and steric packing (2, 54). Self-association interactions are nonspecific and different from those involved in the formation of salt bridges and hydrogen bonds. Experimentally they can be quantified via a thermodynamic solution parameter called the osmotic second virial coefficient (B_{22}). Positive B_{22} values indicate protein–protein interactions are overall repulsive (colloidally stable). When protein–protein interactions are favored over protein–solution interactions, B_{22} values are negative, indicating overall attraction between individual protein molecules. The same interactions that impact colloidal stability are also expected to determine the morphology of protein crystals and precipitates during salting out (55). Based on this relationship, B_{22} values have also been used to predict protein solubility (55, 56) and crystallization conditions (57, 58).

Steric repulsions and nonelectrostatic attractions are weak and independent of pH and salt concentration. On the other hand, electrostatic interactions are strongly dependent on salt concentration and pH in the range of pK_a values near the isoelectric point. B_{22} values reflect this relationship. As the pH is lowered, B_{22} values become large and positive. Molecular surface charge increases and proteins experience repulsion. As solution pH moves closer to the pI , molecules lose their surface charge, attract one another, and become colloidally unstable. Similarly, with increasing ionic strength, protein surface charge is shielded and molecules experience attraction.

3.1. Predicting B_{22} Values

B_{22} values can be estimated using an electrostatic potential between two protein molecules in solution. Simplified estimates of B_{22} often treat proteins as uniform spheres with fixed charge (59). Although simplistic models can improve our understanding of

self-association, they may not be accurate near the molecular p*I*. The surface charge distribution can be nonuniform even when the net charge is zero. Anisotropic interactions can promote preferred orientations between proteins that are lost in simplified estimates. Using statistical mechanics, Long et al. (60) have derived a B_{22} estimate that models electrostatic interactions as a function of distance and orientation between proteins:

$$B_{22} = \frac{1}{2M_w\rho V} \int_{\Omega_1} \int_{\Omega_2} \int_0^\infty (1 - e^{-U_{\text{el}}/kT}) r_{12}^2 dr_{12} \Omega_1 \Omega_2$$

where ρ is the density of the protein, V is its volume, k is the Boltzmann constant, and r_{12} is the distance between molecules. The B_{22} estimate represents a Boltzmann average over all protein–protein configurations (Ω_1 , Ω_2). Orientations were sampled using a 5-axis Eulerian gimbal system. The potential of mean force captures electrostatic interactions between proteins (U_{el}):

$$U_{\text{el}} = \frac{e^2}{4\pi\epsilon_0\epsilon_r} \sum_i \sum_j \frac{q_i q_j}{r_{ij}}$$

where ϵ_0 is the permittivity of free space, ϵ_r is the dielectric constant, e is the charge on an electron, i and j are atom indices for proteins 1 and 2, q is the partial charge on a given atom, and r_{ij} is the distance between atoms i and j . To address the screening effects of salt, the authors modified their potential function based on DLVO (Derjaguin–Landau–Verwey–Overbeek) theory to avoid solving the Poisson Boltzmann equation for all configurations (not shown). A reduced protein representation was used to speed calculations.

To our knowledge, the model by Long et al. is one of the few attempts made at estimating B_{22} values to reproduce experimental information. However, their model was parameterized to recapitulate solubility trends for lysine mutants of Ribonuclease Sa, not the experimental B_{22} values. While B_{22} values are routinely measured during biotherapeutic formulation development projects, such data is not publically available. This limits the ability to properly parameterize computational methods to predict B_{22} values. On the other hand, modeling B_{22} values has been useful to understand changes in self-association behavior after varying ionic strength, pH, and the definition of colloidal interactions (54, 55, 59). For example, a B_{22} model developed by Wiess et al. was used to show that increasing the B_{22} value (by changing the pH) significantly impacted the self-association rate at low ionic strength. Self-association rates were less sensitive to B_{22} value increases at high ionic strength (54).

Even when B_{22} values are experimentally determined, their usefulness to predict aggregation propensities is still unclear.

Studies have shown that aggregation may relate better with conformational stability, especially when B_{22} values are insensitive to changes in environment conditions (61). Other studies have shown that B_{22} values correlate reasonably well with aggregation rates and mechanisms (10, 62). B_{22} estimates are perhaps most useful in predicting aggregation propensities when aggregation is association-limited (53). When evaluating similar proteins under the same condition, B_{22} values should always be used along with protein stability information.

3.2. Predicting pI Values

One way to reduce aggregation resulting from reversible self-association is to move the pH away from the pI by modifying the solution conditions. Another approach is to increase protein net charge by modifying the intrinsic factors that determine the molecular pI . The average pK_a of a sequence provides a simple estimate of the pI :

$$pI = \frac{1}{n} \sum_{i=1}^n pK_{a_i}$$

where, i represents all charged residues in the protein. Side-chain pK_a values are typically inferred from model compounds. More sophisticated methods to approximate residue $pK_a(s)$ (63) and $pI(s)$ (64) exist. Care should be taken to avoid introducing like charges in close proximity as this would destabilize the native structure. Wu et al. were able to modify the molecular pI of a monoclonal antibody by increasing the net charge (+2) for each Fab (fragment antigen-binding) domain (65). An isoelectric focusing gel confirmed a ~0.4 increase in molecular pI . The solubility for the modified molecule improved twofold over the wild type (measured by ultrafiltration). Further pI modification was constrained by sequence compatibility with the human germline.

4. Aggregation-Prone Regions

Irreversible aggregates are nucleated at specific sequence locations called aggregation-prone regions (APRs) (7, 66, 67). Once the “aggregation nucleus” is formed, it provides scaffolding for aggregates to elongate. APRs have unique amino acid residue compositions and sequence patterns. For example, amorphous beta-aggregates, with undefined structure and greater flexibility, tend to favor sequences composed of hydrophobic residues, while fibril-like aggregates may exclude residues with large side-chains to accommodate the need for tight packing interactions (68). On the other hand, the stability of the quasi-crystalline fibril structure may depend more upon sequence specific interactions (69). Fibrils from

the amyloidogenic peptide STVIIIE are extremely sensitive to residue substitutions at positions 3 and 4, while positions 1, 2, and 6 are much more tolerant (69). Furthermore, amorphous aggregates tend to be more sensitive to polar or charged residue substitutions at any position. Amino acids of this type can interfere with the hydrophobic interactions of amorphous aggregates (68). Which factor, sequence specificity or composition, is more important for predicting the aggregation propensity of a sequence, may, in part, depend on the solution conditions. Fibrillation can and often does occur at conditions far from the pI (where the protein is charged) (70). This suggests that specific side-chain interactions are more important for promoting fibrillation in this environment. As the pH moves closer to the pI , amorphous aggregation becomes more common (70). In this case, aggregation propensity may depend more upon sequence composition rather than specific side-chain interactions.

The development of computational tools to predict aggregation propensities was first attempted after point mutants of human Acylphosphatase exhibited different aggregation rates under identical solution conditions (71). Support for this endeavor also comes from knowledge that aggregates formed by different sequences can yield similar structural signatures (66, 72–75). Table 1 summarizes several sequence based computational tools to predict APRs in peptides and proteins. These tools fall into four general categories: (1) tools that use only sequence composition (71, 76–78), (2) tools that combine sequence composition with position specific patterns (79–81), (3) tools that utilize secondary structure prediction and conformational preferences (82, 83), and (4) tools that perform threading onto the cross- β structure (84–86). Some methods incorporate information on environmental conditions such as temperature, concentration, pH, and ionic strength (79, 81, 82). Outputs predict APR windows and propensities for individual residues to exist in APRs. Most tools were trained with sequence motifs that exhibit beta-aggregation behavior (81, 82). Others were trained for predicting amyloid formation only, as verified by a powder x-ray diffraction peak at ~4.6 Å (80). The reader is referred to excellent reviews published elsewhere for more information on these tools (4, 54, 87, 88). Here, we detail one method to show the overall strategy for sequence based APR prediction and validation.

4.1. Sequence Based Prediction

Data from aggregation studies on 50 sequence variants of Acylphosphatase (AcP) (7, 89) was used to derive one of the first computational tools to predict aggregation rate changes for proteins due to point mutations (71). Aggregation rate changes between wild-type and sequence variants of AcP were found to significantly correlate with differences in three sequence composition based terms: hydrophobicity, charge, and secondary structure

Table 1
Sequence based tools to predict aggregation-prone regions (APRs) in peptides and proteins

Tool	Category	Brief description
3D Profile (84)	Sequence threading	Molecular modeling method that evaluates compatibility of a sequence with the crystal structure of hexapeptide NNQQNY. The algorithm uses the Rosetta energy function and was validated with strongly predicted peptides known to form fibrils in experiment.
AGGRESCAN (76)	Sequence composition	Intracellular aggregation propensity for mutants of A β ₄₂ peptide (mutants were generated by point mutation in the sequence region 17–21). The algorithm was parameterized on A β ₄₂ peptide mutants and validated with an experimental data of 24 fibril forming polypeptides.
AMYLPRED (98)	Consensus prediction	Uses consensus among five different methods to predict APRs. Validated using experimental data on a set of amyloidogenic sequence stretches from 18 proteins.
Packing density (77)	Sequence composition	Number of neighboring residues is used to measure packing densities for individual amino acids. Average packing density for each amino acid was derived from protein crystal structures. Validated using 12 amyloid forming peptides and proteins.
PAGE (81)	Sequence composition and patterning	Aggregation propensity is calculated based on aromaticity, beta-strand propensity and charge. The algorithm was parameterized using peptides found in disease causing amyloidogenic proteins and validated with experimental data on a number of proteins known to cause amyloidogenic diseases.
PASTA (83)	Secondary structure and conformational preferences	Pair-wise interaction potentials for a pair of residues found facing each other in the cross-beta motif. The interaction potentials were determined from a dataset of 500 high resolution globular protein crystal structures. Validated using the TANGO testing set.
SALSA (78)	Sequence composition	Chou-Fasman beta-strand propensity. The algorithm was parameterized using a protein dataset and validated on alpha-synuclein, A β , and Tau.
TANGO (82)	Secondary structure and conformational preferences	Statistical mechanics based method. Takes into account physicochemical principles behind beta-sheet formation. The algorithm was parameterized using short aggregating and nonaggregating peptides and validated with an experimental dataset of 179 peptides.

(continued)

Table 1
(continued)

Tool	Category	Brief description
Waltz (80)	Sequence composition, patterning, and structure modeling	A combination of physicochemical properties of beta-sheet formation with position specific substitution matrices. The algorithm was parameterized using 200 aggregating and nonaggregating hexapeptides and validated with an experimental data of 120 hexapeptides plus the AmylHex database.
Zygggregator (79, 107)	Sequence composition and patterning	Relative propensities for folding and aggregation in a given sequence region. The algorithm was parameterized using short peptides and validated with a number of known amyloidogenic peptides.

propensity (see below for the calculation). To improve the strength of these correlations to predict aggregation rate changes, a summation of terms was used:

$$\ln(v_{\text{mut}} / v_{\text{wt}}) = A\Delta\text{Hydr.} + B(\Delta\Delta G_{\text{coil}-\alpha} + \Delta\Delta G_{\beta-\text{coil}}) + C\Delta\text{Charge}$$

where, ΔCharge represents the net charge difference between variant and wild-type sequences, $\Delta\text{Charge} = |\text{Charge}_{\text{mut}}| - |\text{Charge}_{\text{wt}}|$, at neutral pH. Similarly, the change in hydrophobicity was calculated as $\Delta\text{Hydro} = \text{Hydr}_{\text{wt}} - \text{Hydr}_{\text{mut}}$ using the Roseman scale (90). The propensity of the chain to convert from alpha-helix to a disordered conformation was calculated using $\Delta\Delta G_{\text{coil}-\alpha} = RT \ln(P_{\alpha}^{\text{wt}}/P_{\alpha}^{\text{mut}})$, where P_{α} values are determined using the AGADIR algorithm (91) and R is the gas constant in kJ. The propensity of the chain to convert from coil to beta-strand was calculated as $\Delta\Delta G_{\beta-\text{coil}} = 13.64 (P_{\beta}^{\text{wt}} - P_{\beta}^{\text{mut}})$, where P_{β} are normalized beta-sheet propensities taken from Street and Mayo (92). The constant parameters of the function, A , B , and C , are the slopes from linear regression fits between individual terms and aggregation rates changes for all 50 AcP variants.

The function was tested by predicting the aggregation rate changes for 27 mutations in a set of proteins and peptides sequences associated with neurodegenerative diseases. Aggregation rates for wild-type proteins and their variants were measured under conditions where proteins were unstructured. Linear regression on the predicted and experimental aggregation rates showed a significant correlation with an R -value of 0.85 and a p -value of 0.0001. Good agreement ($R=0.756$ and $P=0.0001$) was also observed for all AcP variants with mutations in two key sequence areas known to significantly perturb the observed aggregation rate for the wild

type. In a separate test, the equation derived from AcP variants correctly predicted aggregation rate changes for an independent dataset of sequence variants from other proteins. Given the generality of the equation to predict aggregation rate changes, the authors concluded that changes in sequence composition rather than specific side-chain interactions determine the aggregation rate for these proteins.

4.1.1. Commercial Biotherapeutics

Although sequence based APR prediction tools have been extensively validated against amyloidogenic proteins and sequences, their application to biotherapeutics is only beginning to emerge. Recently, Kumar and coworkers have combined two sequence based prediction tools TANGO (82) and PAGE (81) to show that experimentally validated APR disruption can be predicted (93) (see case studies below). Using multiple prediction tools helps avoid potential biases due to the training sets, parameterization and peculiarities of methods. Both TANGO and PAGE, developed by the Serrano and Caflisch groups respectively, use the amino acid sequence only and make predictions based on the physicochemical properties of amino acids.

TANGO is a statistical mechanics based algorithm to predict protein aggregation (82). Aggregation scores are based on the principles of beta-sheet formation. TANGO considers competing conformations such as the beta-turn, alpha-helix, beta-sheet, folded state, and assumes that nucleating beta-aggregation regions will be fully buried and satisfy their hydrogen-bond potential. A partition function is used to estimate the probability that protein segments will populate these conformational states. Environmental conditions such as stability, pH, protein concentration, ionic strength and concentration of denaturant TFE, are also factored into the aggregation propensity score. The tool was tested by predicting the aggregation propensities of 179 peptides, 67 of which were already known to aggregate experimentally. Peptide sequences were validated as aggregation-prone if their CD or NMR spectra exhibited concentration dependence or when binding to the reporter dye thioflavin T (ThT).

For a given amino acid sequence, PAGE computes aggregation propensities for each residue ($\ln \pi$) and the absolute aggregation rate ($\ln R$) by sliding a small window of 5–9 residues along the sequence (81). Aggregation scores are calculated using sequence aromaticity, beta-strand propensity, charge, solubility, average polar/nonpolar accessible surface area for each residue in a given window. PAGE was trained and tested against peptides from 16 known amyloidogenic proteins and peptides including alpha-synuclein, apolipoprotein, amyloid precursor protein, islet amyloid precursor protein, prion, Sup35, and tau.

In the work of Kumar and coworkers, PAGE scores ($\ln \pi$) were normalized to identify regions with statistically high aggregation propensities (93, 94). The Z -score of residue i is calculated as follows,

$$Z_i = \frac{\ln(\pi_i) - \langle \ln(\pi_i) \rangle}{\sigma(\ln(\pi))}$$

where $\langle \ln(\pi) \rangle$ is the average aggregation propensity of the sequence and $\sigma(\ln(\pi))$ is the standard deviation. Only regions with Z scores >1.96 were considered aggregation-prone when using PAGE. A similarly stringent cutoff of $\geq 10\%$ was used for TANGO scores (see Fig. 1 in (93)). Higher cutoff values ensure that predicted regions have a greater probability of being truly aggregation-prone. However, cutoffs were chosen arbitrarily and additional aggregation-prone motifs could have been identified with less stringent criteria (94).

Below, we present three case studies taken from literature where the experimental observations can be rationalized on the basis of TANGO/PAGE predictions.

4.1.2. Case Study 1

Bovine growth hormone (bGH) forms a folding intermediate that aggregates at concentrations of $10 \mu\text{M}$ or greater. The sequence region, 109–133 in bGH, appears to be directly involved in the aggregation process. It has been determined experimentally that human growth hormone (hGH), which differs from bGH at eight sequence positions in the sequence region 109–133, has a lower aggregation propensity than bGH under identical conditions. Lehrman et al. have designed a variant of bGH called 8H-bGH that contains residues 1–108, 134–191 of bGH and 109–133 of hGH (95). Under partially denaturing conditions, the observed aggregation rate was significantly slower for 8H-bGH as compared to bGH. Figure 3a shows the alignment of all three sequences.

Figure 3b and 3c compare TANGO and PAGE profiles of bGH, hGH and 8H-bGH. All three sequences have similar TANGO and PAGE profiles except at positions 109–133 where the sequence variation occurs. Consistent with experiments, both TANGO and PAGE indicate an APR in this region for bGH. As indicated by the absence of a peak in the TANGO profile, the APR (119-GILALM-124) is disrupted in 8H-bGH and is missing for hGH. Similarly, 8H-bGH shows lower $\ln \pi$ values for this APR in the PAGE profile. TANGO and PAGE also indicate additional APRs outside sequence positions 109–133 for all three proteins. These were not studied experimentally.

4.1.3. Case Study 2

Light chain amyloidosis (AL) is one of a number of protein conformational diseases, associated with somatic mutations in immunoglobulin light chains of type λ . Baden et al. have shown that the amyloidogenic immunoglobulin light chain (AL-09), isolated from

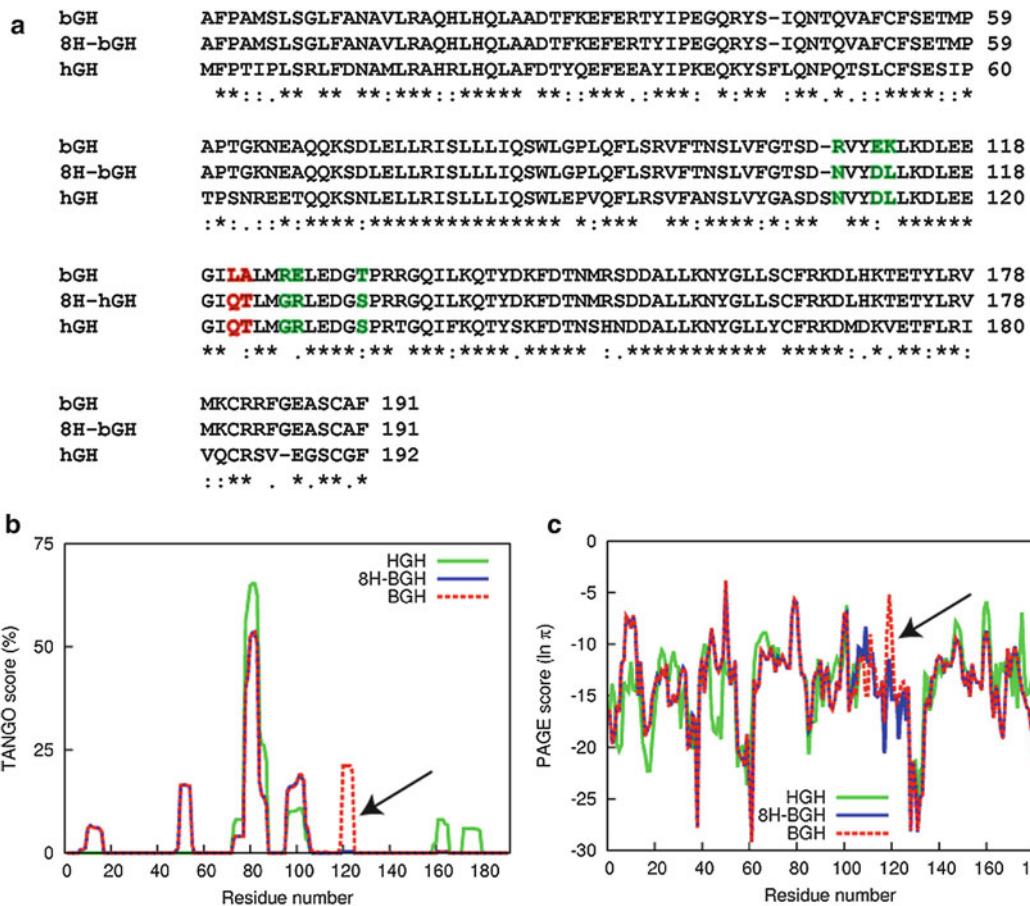


Fig. 3. (a) Sequence alignment for bGH, 8H-bGH, and hGH. Green and red sequence highlighting indicates the 8 residue differences between bGH and hGH. Sequence positions responsible for the APR disruption are highlighted red in the alignment. (b) TANGO profiles and (c) PAGE profiles for all three sequences. The APR of interest is indicated by a black arrow in (b) and (c).

an AL patient, differs from its germline light chain (κ IO18/O8) at seven sequence positions (96). The sequence alignment of AL-09 and κ IO18/O8 is shown in Fig. 4a. Three of these seven somatic mutations are located in the AL dimer interface. To study the aggregation for the AL protein, the authors performed systematic restorative mutations involving I34N, Q42K, and H87Y. From experiments, lag times were shown to increase in the following order: AL-09 < AL-09 I34N < κ IO18/O8.

The TANGO profile for AL-09 indicates a strong APR (32-YLIWY-36) peak containing the somatic mutation at position 34 (Fig. 4b). This APR is significantly weakened in κ IO18/O8. TANGO profiles did not show APRs around the other two somatic mutations located at positions 42 and 87. PAGE profiles (Fig. 4c) show an APR, 87-HCQQY-91 for AL-09 and 87-YCQQY-91 for κ IO18/O8. However, there were no significant differences between the PAGE profiles.

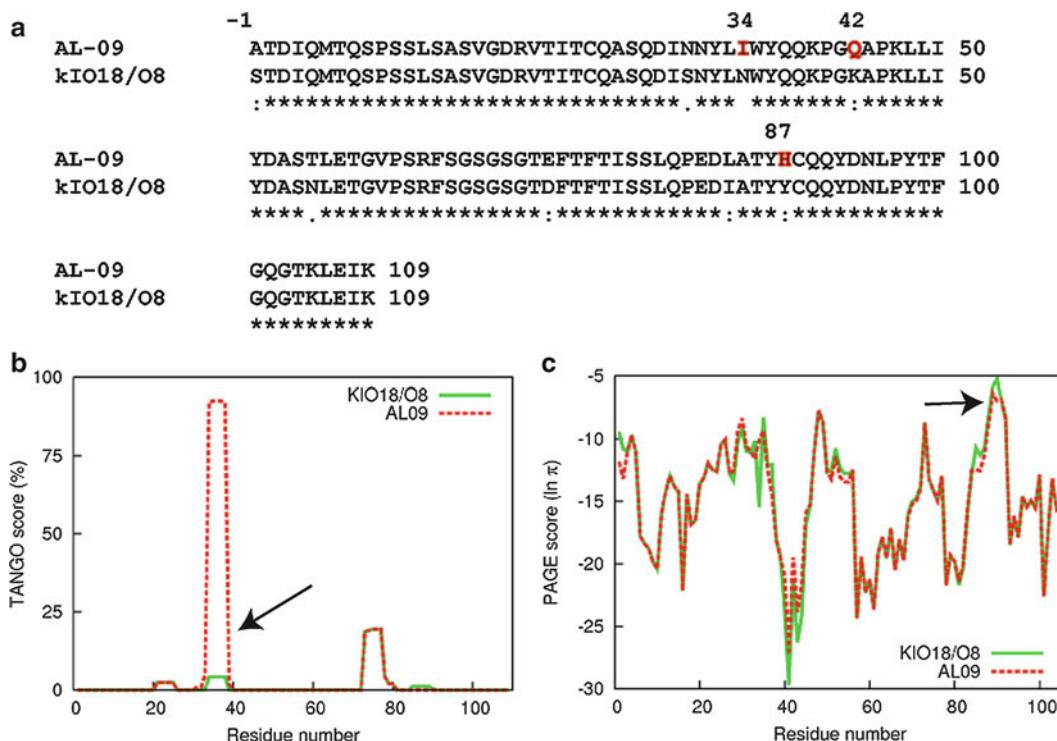


Fig. 4. (a) Sequence alignment for AL-09 and κΙΟ18/08. The three systematic restorative mutations located at the AL dimer interface (I34N, Q42K, and H87Y) are highlighted red. (b) TANGO profiles and (c) PAGE profiles for both sequences. Strongly predicted APRs are indicated by a black arrow. PAGE profiles for both sequences are similar near the APR.

4.1.4. Case Study 3

Constant regions of IgG heavy chains are predicted to contain several well conserved APRs by TANGO and PAGE. Many of them have been documented by Wang et al. (94). A highly conserved and strongly predicted APR, 302-VVSVLTVL-309, is found in the C_H2 domain. The same sequence motif was also detected as a structural APR (hydrophobic patch) by the Trout lab using SAP (described later in this chapter) (27). To disrupt this APR, a point mutation, L309K, was introduced into the sequence (Fig. 5a). The sequence change both reduced the aggregation propensity (as assessed by turbidity and HPLC assays) and improved the thermal stability (measured by DSC) (27). Comparing the TANGO profiles (Fig. 5b) for human IgG1 WT and variant, the mutation L309K disrupted an APR peak. In agreement, the PAGE profile also predicted a reduced aggregation propensity for the L309K variant (Fig. 5c).

4.1.5. Limitations for Sequence Based Tools

Sequence based approaches to APR prediction are fast and require minimal computer and human resources. Recently, many of these tools have become available due to the importance of neurodegenerative diseases. Given the accuracy of these tools to detect APR

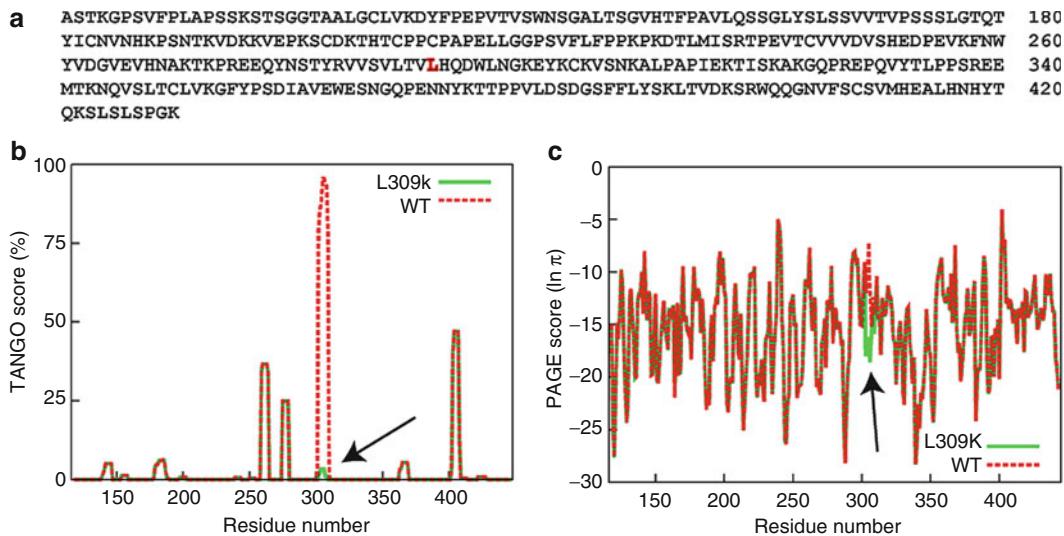


Fig. 5. (a) Heavy chain sequence for the constant region of a human IgG1 mAb. The mutated position for sequence variant L309K is highlighted in red. (b) TANGO profiles and (c) PAGE profiles for both sequences. The APR of interest is indicated by a black arrow in (b) and (c).

motifs in amyloidogenic proteins, they could also be used to screen biotherapeutic sequences during early formulation development. To improve the usefulness of sequence based tools for predicting biotherapeutic aggregation, more comprehensive data is needed to train these algorithms under different solutions conditions such as pH, ionic strength, and temperature. Often, sequence based tools identify several potential APRs, especially in large proteins such as monoclonal antibodies. However, the existence of multiple APR motifs in a candidate molecule is not expected to have an additive effect on the aggregation propensity. Some of these APRs will not participate in aggregation under a given set of conditions. Others may be buried in the core (97). Furthermore, many of these programs predict APR motifs based on peptide nucleation/elongation rates which may not correlate with protein aggregation rates that have the same sequence motifs. Therefore, the extent to which these tools are useful for predicting aggregation behavior in protein biotherapeutics remains to be seen.

Incorporating these tools into the R&D pipeline for biologics will require quantitative comparisons of their accuracies in standardized test sets. So far, such tests have been rarely performed. One of the first studies to provide comprehensive comparison data was done by the authors of Waltz (80). Their program outperformed several others on a dataset that included only x-ray verified fibril forming hexapeptides. However, large therapeutic proteins may be less likely to form aggregates such as fibrils or plaques (18).

Quantitative comparisons will need datasets that are relevant to biotherapeutic protein aggregation. Given the lack of standardization, the best way to detect APR motifs in sequences outside of the training sets for these tools is to seek consensus among different predictive tools. The server AMYLPRED runs five different programs and identifies APRs when there is consensus among two or more prediction methods (98). Comparisons issues aside, the accuracy of all sequence based tools to predict APR motifs can be improved by using molecular structure information.

4.2. Structure Based Prediction

Many proteins have aggregation-prone sequence motifs. In most cases, they are sequestered in the core (97). Excluding APRs from protein surfaces or disordered regions helps protect protein function by inhibiting aggregation (99). Recently, Hamada et al. have studied the ability of individual beta-strands to initiate amyloid-like fibril formation in beta-lactoglobulin (100). They found that sequence regions with high intrinsic aggregation propensities also need to undergo local unfolding events to initiate aggregate formation. Therefore, APRs will facilitate intermolecular interactions only when structural regions containing APRs are either surface exposed (14, 93, 101) or undergo transitions to become surface exposed. Mapping the predicted APRs of a sequence onto the corresponding three dimensional structure is one way to determine which predicted APRs can actually form molecular interactions (93). Beyond using structure information to filter sequence based prediction (102), new methods are incorporating structure and dynamic information to predict aggregation propensities for regions of the protein surface (103–105).

4.2.1. Spatial Aggregation Propensity

Spatial Aggregation Propensity (SAP) is an algorithm to predict protein structural regions that may participate in aggregation by identifying hydrophobic surface patches (27). Hydrophobic residues will form aggregation “hotspots” when they are clustered together on protein surfaces (Fig. 6). Hotspots can also change in size due to normal protein fluctuations under physiological conditions. SAP uses molecular dynamics to simulate these changes. Residue aggregation propensities are determined from SAP values for each atom, SAP_{atom} :

$$SAP_{atom\ i} = \sum_{\substack{\text{Simulation} \\ \text{average}}} \sum_{\substack{\text{Residues with} \\ \text{at least one side} \\ \text{chain atom within} \\ R \text{ from atom } i}} \left(\frac{\text{SAA of side chain} \\ \text{atoms within radius } R}{\text{SAA of side chain atoms} \\ \text{of fully exposed residue}} \times \text{Hydrophobicity}_i \right)$$

Hydrophobicity for each residue is taken from the scale of Black and Mould (106). The scale is normalized by positioning

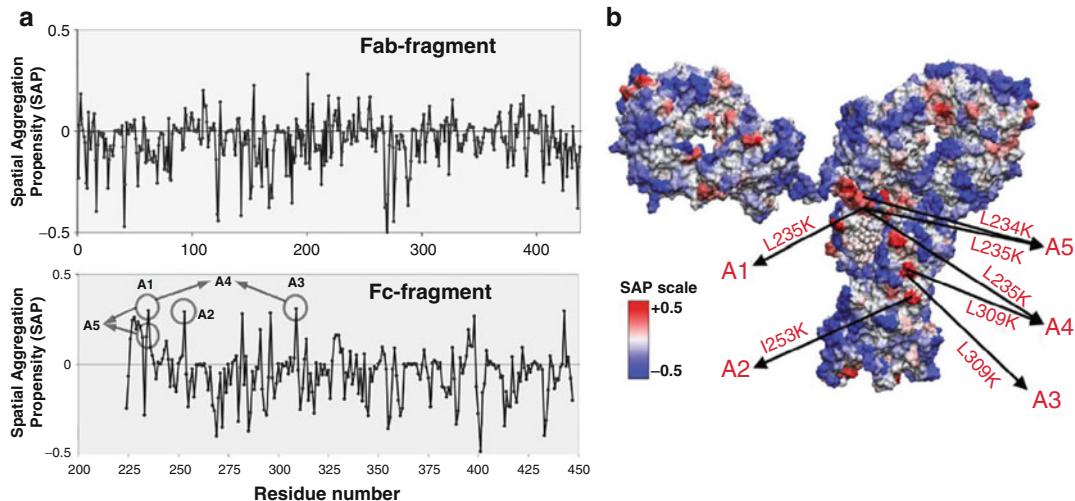


Fig. 6. (a) SAP values for an unidentified mAb. Values for the five sequence variants are *circled*. (b) mAb surface residues are *colored* according to their SAP values. Hydrophobic hotspots (positive SAP values) are colored *red*. Hydrophilic regions (negative SAP values) are colored *blue*. Color intensity is proportional to the magnitude of the SAP value. (Images taken from Chennamsetty et al. (27) and reprinted with permission).

glycine at a hydrophobicity of zero. Amino acids with greater hydrophobicity than glycine have positive values. Hydrophilic residues have negative values. The solvent accessible area (SAA) of side-chain atoms within a radius R is computed for each simulation frame. Residue SAA values for fully exposed side-chains were obtained by calculating the SAA of side-chain atoms for the middle residue of a fully extended tri-peptide, Ala-X-Ala. Residue aggregation propensities are taken as the average SAP_{atom} for all atoms of the residue.

Using SAP to understand the aggregation behavior of an unidentified mAb, the authors discovered three potential hotspots in the Fc region (27). To disrupt aggregation, five sequence variants were created, three with point mutations and two with double point mutations. Hydrophobic residues were mutated to hydrophilic ones. All five variants showed an increase in melting temperature for the C_H2 domain (measured by DSC). To test for differences in aggregation, wild-type and variant antibody samples were stressed at 65°C, up to 4 h, at protein concentrations of 150 mg/ml. SEC analysis indicated that monomer content increased from 91% (wild type) to 93–95% for the three point variants. A cloudy solution was observed for the wild type in turbidity assays. Solutions for all five variants remained clear. Using SAP on a second antibody, a hotspot was detected in a CDR loop. Mutating these residues improved the thermal stability but disrupted antigen binding.

Further studies on SAP have attempted to vary the parameters of the equation, investigate the value of the dynamic information, and experimented with different solvation models to simulate protein dynamics (104). Tests on changing the radius size in SAP calculations were performed to determine what SAP resolution best identified hydrophobic hotspots involved in aggregation. Using a cutoff of 5 Å, small hydrophobic surface patches were identified. Although small hydrophobic surface patches are not expected to play a major role in protein aggregation, disruption of these patches (by introducing a lysine residue) increased monomer retention 1–5% over wild type. Further evidence to support the role of small hydrophobic surface patches in influencing aggregation comes from double mutants of the same study which showed an enhanced stability and reduced aggregation over single mutants. Increasing the cutoff radius potentially identifies larger hotspots by forcing SAP to average over larger surface areas. However, since protein surfaces are a mix of hydrophobic and hydrophilic regions, this may result in lower sensitivity to smaller hydrophobic surface patches.

The authors of SAP have also analyzed the added value of performing molecular simulations to predict surface hotspots (104). Calculating SAP values using a static structure did not significantly change the ability of SAP to identify the same hydrophobic surface patches. However, predictions with simulation data took 200,000× longer to generate (104). The few hotspots missed were located in flexible regions such as loops and CDRs. Hotspots in regions of the structure that underwent only minor fluctuations were all captured. To reduce the CPU time, SAP values were also calculated using simulations performed with various implicit solvation models. Simulations with GBSW, the most accurate implicit solvation model, took approximately the same CPU time as explicit simulations. However, the hotspot prediction accuracy was reduced compared to explicit simulations. Using a low accuracy solvation model, EEF1, the CPU time could be significantly reduced, but large structural deviations during simulations increased the number of incorrectly predicted hotspots.

4.2.2. Other Structure Based Methods

Other studies have used structure information to improve sequence based prediction as well. In the work of Tartaglia et al., protection factors from H/D exchange were predicted using both protein structure and sequence (102). Aggregation propensities modified with predicted protection factors were able to remove some of the falsely predicted APRs identified by Zygggregator (79, 107). However, the structure modified algorithm did not predict any additional true positives. Using structure information in sequence based prediction is further supported by recent studies on protein–protein binding sites (105, 108). After modifying Zygggregator to identify surface hotspots,

Pechmann et al. discovered that protein–protein interaction sites often overlap with aggregation hotspots (105). Therefore, the same interactions which promote functional protein associations, including hydrophobicity and electrostatic charge, can lead to aggregation as well. Furthermore, surface hotspots tend to be regulated by disulfide bonds and salt bridges. These interactions form “structural gate keepers” that interfere with uncontrolled aggregation by stabilizing the native state.

5. Conclusions and Future Directions

In this book chapter, we have reviewed a number of computational methods that can be used to predict aggregation in biotherapeutics. These include computational methods developed to predict protein unfolding kinetics (37–39), native state thermodynamic stability (4–45), and colloidal stability (60). Structure and dynamics based methods to predict aggregation propensities were reviewed as well (8–86, 102, 103). Aggregation in protein biotherapeutics may have similar molecular level origins as proteins implicated in neurodegenerative diseases. Hence, the tools and methods developed to study disease causing protein aggregation can also be used to identify potential APRs in biotherapeutics (4, 93, 94). As with any experimental technique, all the above methods have advantages and limitations which have been pointed out. Our recommendation is to use these methods in combination with one another rather than relying on a single method.

While a number of computational tools have become available to study aggregation, only a few have been validated and shown to be of utility in predicting aggregation issues during biotherapeutic development (27, 94, 103). It is expected that more of such studies will become available in future. Validation issues aside, these computational tools potentially have a number of applications to improve in biotherapeutic development. One is to rank order molecules for their propensity to aggregate during early stages of formulation development. Another application is to mitigate aggregation without impacting target binding via rational structure based design. This second application can reduce costs associated with the development of novel and follow-on biologic drug products in two ways. Mitigating aggregation can increase protein expression yields, thus lowering production costs and improving pharmaceutical stability which may prevent the need for lyophilization and/or refrigeration. Furthermore, disrupting APRs that are in immune epitopes potentially improves safety of biotherapeutic drug products (5). Rational structure based biologic drug design is aligned with the FDA Quality by Design (QbD) initiative (28, 29).

Acknowledgments

The authors (P.B., S.K., X.W., and S.K.S.) thank Drs. Sandeep Neema, Kevin King, Russ Robbins, and Nick Warne for their interest and support. P.B. acknowledges a postdoctoral fellowship from Pfizer Inc.

References

- Wang W, Nema S, Teagarden D (2010) Protein aggregation-pathways and influencing factors. *Int J Pharm* 390:89–99
- Chi EY, Krishnan S, Randolph TW, Carpenter JF (2003) Physical stability of proteins in aqueous solution: mechanism and driving forces in nonnative protein aggregation. *Pharm Res* 20:1325–1336
- Manning MC, Chou DK, Murphy BM, Payne RW, Katayama DS (2010) Stability of protein pharmaceuticals: an update. *Pharm Res* 27: 544–575
- Kumar S, Wang X, Singh SK (2010) Identification and impact of aggregation-prone regions in proteins and therapeutic monoclonal antibodies, in aggregation of therapeutic proteins (eds W. Wang and C. J. Roberts), John Wiley & Sons, Inc., Hoboken, NJ, USA. doi: 10.1002/9780470769829.ch3
- Kumar S, Singh SK, Wang X, Rup B, Gill D (2011) Coupling of aggregation and immunogenicity in biotherapeutics: T- and B-cell immune epitopes may contain aggregation-prone regions. *Pharm Res* 28:949–961
- Rosenberg AS (2006) Effects of protein aggregates: an immunologic perspective. *AAPS J* 8:E501–E507
- Chiti F, Calamai M, Taddei N, Stefani M, Ramponi G, Dobson CM (2002) Studies of the aggregation of mutant proteins in vitro provide insights into the genetics of amyloid diseases. *Proc Natl Acad Sci USA* 99(Suppl 4): 16419–16426
- Nelson R, Sawaya MR, Balbirnie M, Madsen AO, Riek C, Grothe R, Eisenberg D (2005) Structure of the cross-beta spine of amyloid-like fibrils. *Nature* 435:773–778
- Sawaya MR, Sambashivan S, Nelson R, Ivanova MI, Sievers SA, Apostol MI, Thompson MJ, Balbirnie M, Wiltzius JJ, McFarlane HT, Madsen AO, Riek C, Eisenberg D (2007) Atomic structures of amyloid cross-beta spines reveal varied steric zippers. *Nature* 447:453–457
- Sahin E, Grillo AO, Perkins MD, Roberts CJ (2010) Comparative effects of pH and ionic strength on protein-protein interactions, unfolding, and aggregation for IgG1 antibodies. *J Pharm Sci* 99:4830–4848
- Domanska K, Vanderhaegen S, Srinivasan V, Pardon E, Dupeux F, Marquez JA, Giorgetti S, Stoppini M, Wyns L, Bellotti V, Steyaert J (2011) Atomic structure of a nanobody-trapped domain-swapped dimer of an amyloidogenic beta2-microglobulin variant. *Proc Natl Acad Sci USA* 108:1314–1319
- Liu C, Sawaya MR, Eisenberg D (2011) Beta-microglobulin forms three-dimensional domain-swapped amyloid fibrils with disulfide linkages. *Nat Struct Mol Biol* 18:49–55
- Sinha N, Tsai CJ, Nussinov R (2001) A proposed structural model for amyloid fibril elongation: domain swapping forms an interdigitating beta-structure polymer. *Protein Eng* 14:93–103
- Chiti F, Dobson CM (2009) Amyloid formation by globular proteins under native conditions. *Nat Chem Biol* 5:15–22
- Li Y, Roberts CJ (2009) Lumry-Eyring nucleated-polymerization model of protein aggregation kinetics. 2. Competing growth via condensation and chain polymerization. *J Phys Chem B* 113:7020–7032
- Andrews JM, Roberts CJ (2007) A Lumry-Eyring nucleated polymerization model of protein aggregation kinetics: 1. Aggregation with pre-equilibrated unfolding. *J Phys Chem B* 111:7897–7913
- Lumry R, Eyring H (1954) Conformation changes of proteins. *J Phys Chem* 58:110–120
- Ramshini H, Parrini C, Relini A, Zampagni M, Mannini B, Pesce A, Saboury AA, Nemat-Gorgani M, Chiti F (2011) Large proteins have a great tendency to aggregate but a low propensity to form amyloid fibrils. *PLoS One* 6:e16075
- Campioni S, Mannini B, Zampagni M, Pensalfini A, Parrini C, Evangelisti E, Relini A, Stefani M, Dobson CM, Cecchi C, Chiti F (2010) A causative link between the structure of aberrant protein oligomers and their toxicity. *Nat Chem Biol* 6:140–147

20. Ma B, Nussinov R (2006) Simulations as analytical tools to understand protein aggregation and predict amyloid conformation. *Curr Opin Chem Biol* 10:445–452
21. Thirumalai D, Klimov DK, Dima RI (2003) Emerging ideas on the molecular basis of protein and peptide aggregation. *Curr Opin Struct Biol* 13:146–159
22. Wu C, Shea JE (2011) Coarse-grained models for protein aggregation. *Curr Opin Struct Biol* 21:209–220
23. Berryman JT, Radford SE, Harris SA (2011) Systematic examination of polymorphism in amyloid fibrils by molecular-dynamics simulation. *Biophys J* 100:2234–2242
24. Jarosz DF, Lindquist S (2010) Hsp90 and environmental stress transform the adaptive value of natural genetic variation. *Science* 330:1820–1824
25. Roberts CJ (2006) Non-native protein aggregation: Pathways, kinetics, and shelf-life prediction, in *Misbehaving proteins: Protein (mis)folding, aggregation, and stability* (eds R.M. Murphy, A. Tsa), Springer, New York: pp. 17–46.
26. Roberts CJ (2007) Non-native protein aggregation kinetics. *Biotechnol Bioeng* 98:927–938
27. Chennamsetty N, Voynov V, Kayser V, Helk B, Trout BL (2009) Design of therapeutic proteins with enhanced stability. *Proc Natl Acad Sci USA* 106:11937–11942
28. Rathore AS (2011) Quality by design for biologics and biosimilars. *Pharm Technol* 35: 64–68
29. Rathore AS, Winkle H (2009) Quality by design for biopharmaceuticals. *Nat Biotechnol* 27:26–34
30. Kendrick BS, Cleland JL, Lam X, Nguyen T, Randolph TW, Manning MC, Carpenter JF (1998) Aggregation of recombinant human interferon gamma: kinetics and structural transitions. *J Pharm Sci* 87:1069–1076
31. Plaxco KW, Simons KT, Baker D (1998) Contact order, transition state placement and the refolding rates of single domain proteins. *J Mol Biol* 277:985–994
32. Galzitskaya OV, Garbuzynskiy SO, Ivankov DN, Finkelstein AV (2003) Chain length is the main determinant of the folding rate for proteins with three state folding kinetics. *Proteins* 51:162–166
33. Huang JT, Cheng JP, Chen H (2007) Secondary structure length as a determinant of folding rate of proteins with two- and three-state kinetics. *Proteins* 67:12–17
34. Gromiha MM, Selvaraj S (2001) Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: application of long-range order to folding rate prediction. *J Mol Biol* 310:27–32
35. Harihar B, Selvaraj S (2009) Refinement of the long-range order parameter in predicting folding rates of two-state proteins. *Biopolymers* 91:928–935
36. Gong H, Isom DG, Srinivasan R, Rose GD (2003) Local secondary structure content predicts folding rates for simple, two-state proteins. *J Mol Biol* 327:1149–1154
37. Gromiha MM, Selvaraj S, Thangakani AM (2006) A statistical method for predicting protein unfolding rates from amino acid sequence. *J Chem Inf Model* 46:1503–1508
38. Harihar B, Selvaraj S (2011) Application of long-range order to predict unfolding rates of two-state proteins. *Proteins* 79:880–887
39. Jung J, Lee J, Moon HT (2005) Topological determinants of protein unfolding rates. *Proteins* 58:389–395
40. Tomii K, Kanehisa M (1996) Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. *Protein Eng* 9:27–36
41. Gromiha MM, Oobatake M, Sarai A (1999) Important amino acid properties for enhanced thermostability from mesophilic to thermo-philic proteins. *Biophys Chem* 82:51–67
42. Capriotti E, Fariselli P, Casadio R (2005) I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res* 33: W306–W310
43. Guerois R, Nielsen JE, 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
44. Zhou H, 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
45. Yin S, Ding F, Dokholyan NV (2007) Eris: an automated estimator of protein stability. *Nat Methods* 4:466–467
46. Ding F, Dokholyan NV (2006) Emergence of protein fold families through rational design. *PLoS Comput Biol* 2:e85
47. Neria E, Fischer S, Karplus M (1996) Simulation of activation free energies in molecular systems. *J Chem Phys* 105:1902
48. Lazaridis T, Karplus M (1999) Effective energy function for proteins in solution. *Proteins* 35:133–152
49. Kortemme T, Morozov AV, Baker D (2003) An orientation-dependent hydrogen bonding

- potential improves prediction of specificity and structure for proteins and protein-protein complexes. *J Mol Biol* 326:1239–1259
50. Dunbrack RL Jr, Cohen FE (1997) Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci* 6:1661–1681
 51. Brubaker WD, Freites JA, Golchert KJ, Shapiro RA, Morikis V, Tobias DJ, Martin RW (2011) Separating instability from aggregation propensity in gammaS-crystallin variants. *Biophys J* 100:498–506
 52. Sahin E, Jordan JL, Spatara ML, Naranjo A, Costanzo JA, Weiss WF, Robinson AS, Fernandez EJ, Roberts CJ (2011) Computational design and biophysical characterization of aggregation-resistant point mutations for gammaD crystallin illustrate a balance of conformational stability and intrinsic aggregation propensity. *Biochemistry* 50: 628–639
 53. Chi EY, Krishnan S, Kendrick BS, Chang BS, Carpenter JF, Randolph TW (2003) Roles of conformational stability and colloidal stability in the aggregation of recombinant human granulocyte colony-stimulating factor. *Protein Sci* 12:903–913
 54. Weiss WF, Young TM, Roberts CJ (2009) Principles, approaches, and challenges for predicting protein aggregation rates and shelf life. *J Pharm Sci* 98:1246–1277
 55. Haas C, Drenth J, Wilson WW (1999) Relation between the solubility of proteins in aqueous solutions and the second virial coefficient of the solution. *J Phys Chem B* 103:2808–2811
 56. Allahyarov E, Löwen H, Hansen J, Louis A (2003) Nonmonotonic variation with salt concentration of the second virial coefficient in protein solutions. *Phys Rev E* 67:051404
 57. George A, Chiang Y, Guo B, Arabshahi A, Cai Z, Wilson WW (1997) Second virial coefficient as predictor in protein crystal growth. *Methods Enzymol* 276:100–110
 58. George A, Wilson WW (1994) Predicting protein crystallization from a dilute solution property. *Acta Crystallogr D Biol Crystallogr* 50:361–365
 59. Cheung JK, Truskett TM (2005) Coarse-grained strategy for modeling protein stability in concentrated solutions. *Biophys J* 89: 2372–2384
 60. Long WF, Labute P (2010) Calibrative approaches to protein solubility modeling of a mutant series using physicochemical descriptors. *J Comput Aided Mol Des* 24:907–916
 61. Bajaj H, Sharma VK, Badkar A, Zeng D, Nema S, Kalonia DS (2006) Protein structural conformation and not second virial coefficient relates to long-term irreversible aggregation of a monoclonal antibody and ovalbumin in solution. *Pharm Res* 23:1382–1394
 62. Li Y, Ogunnaike BA, Roberts CJ (2010) Multi-variate approach to global protein aggregation behavior and kinetics: effects of pH, NaCl, and temperature for alpha-chymotrypsinogen A. *J Pharm Sci* 99:645–662
 63. Li H, Robertson AD, Jensen JH (2005) Very fast empirical prediction and rationalization of protein pKa values. *Proteins* 61:704–721
 64. Sillero A, Ribeiro JM (1989) Isoelectric points of proteins: theoretical determination. *Anal Biochem* 179:319–325
 65. Wu SJ, Luo J, O’Neil KT, Kang J, Lacy ER, Canziani G, Baker A, Huang M, Tang QM, Raju TS, Jacobs SA, Teplyakov A, Gilliland GL, Feng Y (2010) Structure-based engineering of a monoclonal antibody for improved solubility. *Protein Eng Des Sel* 23:643–651
 66. Frare E, Mossuto MF, Polverino de Laureto P, Dumoulin M, Dobson CM, Fontana A (2006) Identification of the core structure of lysozyme amyloid fibrils by proteolysis. *J Mol Biol* 361:551–561
 67. Zibaei S, Jakes R, Fraser G, Serpell LC, Crowther RA, Goedert M (2007) Sequence determinants for amyloid fibrillogenesis of human alpha-synuclein. *J Mol Biol* 374: 454–464
 68. Rousseau F, Schymkowitz J, Serrano L (2006) Protein aggregation and amyloidosis: confusion of the kinds? *Curr Opin Struct Biol* 16:118–126
 69. Lopez de la Paz M, Serrano L (2004) Sequence determinants of amyloid fibril formation. *Proc Natl Acad Sci USA* 101: 87–92
 70. Krebs MR, Devlin GL, Donald AM (2007) Protein particulates: another generic form of protein aggregation? *Biophys J* 92: 1336–1342
 71. Chiti F, Stefani M, Taddei N, Ramponi G, Dobson CM (2003) Rationalization of the effects of mutations on peptide and protein aggregation rates. *Nature* 424:805–808
 72. Serpell LC, Sunde M, Blake CC (1997) The molecular basis of amyloidosis. *Cell Mol Life Sci* 53:871–887
 73. Sunde M, Serpell LC, Bartlam M, Fraser PE, Pepys MB, Blake CC (1997) Common core structure of amyloid fibrils by synchrotron X-ray diffraction. *J Mol Biol* 273:729–739
 74. Eakin CM, Berman AJ, Miranker AD (2006) A native to amyloidogenic transition regulated by a backbone trigger. *Nat Struct Mol Biol* 13:202–208

75. Serag AA, Altenbach C, Gingery M, Hubbell WL, Yeates TO (2002) Arrangement of sub-units and ordering of beta-strands in an amyloid sheet. *Nat Struct Biol* 9:734–739
76. Conchillo-Sole O, de Groot NS, Aviles FX, Vendrell J, Daura X, Ventura S (2007) AGGRESCAN: a server for the prediction and evaluation of “hot spots” of aggregation in polypeptides. *BMC Bioinformatics* 8:65
77. Galzitskaya OV, Garbuzynskiy SO, Lobanov MY (2006) Prediction of amyloidogenic and disordered regions in protein chains. *PLoS Comput Biol* 2:e177
78. Zibaei S, Makin OS, Goedert M, Serpell LC (2007) A simple algorithm locates beta-strands in the amyloid fibril core of alpha-synuclein, Abeta, and tau using the amino acid sequence alone. *Protein Sci* 16:906–918
79. DuBay KF, Pawar AP, Chiti F, Zurdo J, Dobson CM, Vendruscolo M (2004) Prediction of the absolute aggregation rates of amyloidogenic polypeptide chains. *J Mol Biol* 341:1317–1326
80. Maurer-Stroh S, Debulpaep M, Kuemmerer N, Lopez de la Paz M, Martins IC, Reumers J, Morris KL, Copland A, Serpell L, Serrano L, Schymkowitz JW, Rousseau F (2010) Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. *Nat Methods* 7:237–242
81. Tartaglia GG, Cavalli A, Pellarin R, Caflisch A (2005) Prediction of aggregation rate and aggregation-prone segments in polypeptide sequences. *Protein Sci* 14:2723–2734
82. Fernandez-Escamilla AM, Rousseau F, Schymkowitz J, Serrano L (2004) Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat Biotechnol* 22:1302–1306
83. Trovato A, Seno F, Tosatto SC (2007) The PASTA server for protein aggregation prediction. *Protein Eng Des Sel* 20:521–523
84. Thompson MJ, Sievers SA, Karanicolas J, Ivanova MI, Baker D, Eisenberg D (2006) The 3D profile method for identifying fibril-forming segments of proteins. *Proc Natl Acad Sci USA* 103:4074–4078
85. Yoon S, Welsh WJ (2004) Detecting hidden sequence propensity for amyloid fibril formation. *Protein Sci* 13:2149–2160
86. Zhang Z, Chen H, Lai L (2007) Identification of amyloid fibril-forming segments based on structure and residue-based statistical potential. *Bioinformatics* 23:2218–2225
87. Caflisch A (2006) Computational models for the prediction of polypeptide aggregation propensity. *Curr Opin Chem Biol* 10:437–444
88. Agrawal NJ, Kumar S, Wang X, Helk B, Singh SK, Trout BL (2011) Aggregation in protein-based biotherapeutics: computational studies and tools to identify aggregation-prone regions. *J Pharm Sci* 100:5081–5095
89. Chiti F, Taddei N, Baroni F, Capanni C, Stefanini M, Ramponi G, Dobson CM (2002) Kinetic partitioning of protein folding and aggregation. *Nat Struct Biol* 9:137–143
90. Roseman MA (1988) Hydrophilicity of polar amino acid side-chains is markedly reduced by flanking peptide bonds. *J Mol Biol* 200:513–522
91. Lacroix E, Viguera AR, Serrano L (1998) Elucidating the folding problem of alpha-helices: local motifs, long-range electrostatics, ionic-strength dependence and prediction of NMR parameters. *J Mol Biol* 284:173–191
92. Street AG, Mayo SL (1999) Intrinsic beta-sheet propensities result from van der Waals interactions between side chains and the local backbone. *Proc Natl Acad Sci USA* 96:9074–9076
93. Wang X, Singh SK, Kumar S (2010) Potential aggregation-prone regions in complementarity-determining regions of antibodies and their contribution towards antigen recognition: a computational analysis. *Pharm Res* 27:1512–1529
94. Wang X, Das TK, Singh SK, Kumar S (2009) Potential aggregation prone regions in biotherapeutics: a survey of commercial monoclonal antibodies. *MAbs* 1:254–267
95. Lehrman SR, Tuls JL, Havel HA, Haskell RJ, Putnam SD, Tomich CS (1991) Site-directed mutagenesis to probe protein folding: evidence that the formation and aggregation of a bovine growth hormone folding intermediate are dissociable processes. *Biochemistry* 30:5777–5784
96. Baden EM, Randles EG, Aboagye AK, Thompson JR, Ramirez-Alvarado M (2008) Structural insights into the role of mutations in amyloidogenesis. *J Biol Chem* 283:30950–30956
97. Tzotzos S, Doig AJ (2010) Amyloidogenic sequences in native protein structures. *Protein Sci* 19:327–348
98. Froussios KK, Iconomidou VA, Karletidi CM, Hamodrakas SJ (2009) Amyloidogenic determinants are usually not buried. *BMC Struct Biol* 9:44
99. Linding R, Schymkowitz J, Rousseau F, Diella F, Serrano L (2004) A comparative study of the relationship between protein structure and beta-aggregation in globular and intrinsically disordered proteins. *J Mol Biol* 342:345–353

100. Hamada D, Tanaka T, Tartaglia GG, Pawar A, Vendruscolo M, Kawamura M, Tamura A, Tanaka N, Dobson CM (2009) Competition between folding, native-state dimerisation and amyloid aggregation in beta-lactoglobulin. *J Mol Biol* 386:878–890
101. Routledge KE, Tartaglia GG, Platt GW, Vendruscolo M, Radford SE (2009) Competition between intramolecular and intermolecular interactions in an amyloid-forming protein. *J Mol Biol* 389:776–786
102. Tartaglia GG, Pawar AP, Campioni S, Dobson CM, Chiti F, Vendruscolo M (2008) Prediction of aggregation-prone regions in structured proteins. *J Mol Biol* 380:425–436
103. Chennamsetty N, Helk B, Voynov V, Kayser V, Trout BL (2009) Aggregation-prone motifs in human immunoglobulin G. *J Mol Biol* 391:404–413
104. Chennamsetty N, Voynov V, Kayser V, Helk B, Trout BL (2010) Prediction of aggregation-prone regions of therapeutic proteins. *J Phys Chem B* 114:6614–6624
105. Pechmann S, Levy ED, Tartaglia GG, Vendruscolo M (2009) Physicochemical principles that regulate the competition between functional and dysfunctional association of proteins. *Proc Natl Acad Sci USA* 106:10159–10164
106. Black SD, Mould DR (1991) Development of hydrophobicity parameters to analyze proteins which bear post- or cotranslational modifications. *Anal Biochem* 193:72–82
107. Pawar AP, Dubay KF, Zurdo J, Chiti F, Vendruscolo M, Dobson CM (2005) Prediction of “aggregation-prone” and “aggregation-susceptible” regions in proteins associated with neurodegenerative diseases. *J Mol Biol* 350:379–392
108. Chennamsetty N, Voynov V, Kayser V, Helk B, Trout BL (2011) Prediction of protein binding regions. *Proteins* 79:888–897

Chapter 27

Coarse-Grained Simulations of Protein Aggregation

Troy Cellmer and Nicolas L. Fawzi

Abstract

Protein aggregation is believed to be responsible for a number of human diseases and limits the yields of pharmaceutical proteins during production. Computer simulations can be used to develop novel experimentally testable hypotheses pertaining to aggregation. While all-atom simulations with explicit solvent are too computationally intensive to address the multitude of relevant time scales, coarse-grained models make it possible to observe the transition of monomers to an equilibrium containing aggregates. Here, we provide the reader with background information and a list of steps for setting up, performing, and analyzing computer simulations of aggregating coarse-grained (CG) proteins.

Key words: Computer simulations, Molecular dynamics, Protein aggregation, Amyloid fibril formation

1. Introduction

Protein aggregation is of great importance to the biomedical, biopharmaceutical, and nanotech industries. It is believed to be the causative agent in a number of diseases (1) including Alzheimer's and Parkinson's diseases, can limit the yields of pharmaceutical proteins (2), and decrease the efficacy of protein formulations (3). Furthermore, protein assemblages are being used as novel nanomaterials (4–6). This wide array of applications makes protein aggregation a very active area of scientific research.

Computer simulations are a viable tool for studying protein aggregation (7–11). Ultimately, one would like to simulate proteins and solvent with atomistic detail to observe the transition of monomers to an equilibrium state consisting of aggregates and monomers. Repeating this simulation a large number of times would provide the statistics necessary to make the most reliable experimental predictions. However, computational constraints prevent fully atomistic aggregation simulations from being a realistic approach. For example, the longest simulations for single proteins

now reach milliseconds in length (12). Adding more protein and solvent would significantly slow simulations, making it impossible to observe events that occur on longer time scales. Thus, one must coarse-grain the details of the protein and solvent to obtain a computationally tractable simulation.

The purpose of this article is to provide the reader with an outline for performing coarse-grained (CG) simulations of protein aggregation. The first step in this process is to choose the solvent and protein representation. The proper choice depends on a number of factors, including the problem being investigated. Thus to begin, we provide an abbreviated literature review to familiarize the reader with different approaches used to perform aggregation-related simulations. We briefly discuss different protein representations, the potential functions that set the interactions between the different species in the simulation, and different algorithms for moving the particles. We also highlight the variety of aggregation-related problems addressed by the different modeling approaches. Based on our collective simulation experience, we then list a series of milestones on the path to performing coarse-grained simulations of aggregating proteins. We expect these steps will provide a clear path forward for someone interested in performing such simulations.

The first challenge is to coarse-grain the protein and solvent but maintain enough of the essential physics to achieve a sufficiently accurate answer to the problem in question. Because of the time and length scales involved, solvent is usually not explicitly represented in aggregation simulations. Instead, the physical effects of the solvent on the protein are implicitly included in the protein potential function. The hydrophobic effect is usually accounted for through some form of two-body attraction between hydrophobic species, while thermal motions and viscosity effects may be added through the random force and friction terms in the Langevin equation of motion.

Due to computational constraints, most early simulation studies of aggregation relied on very simple protein representations, such as two-dimensional and three-dimensional lattice proteins (13–17) (Fig. 1). Each amino acid or collection of amino acids is represented by a bead and occupies a single lattice site that cannot be occupied by any other bead. Lattice models have been useful in examining protein folding and aggregation-related problems, such as sequence effects on aggregation (17, 18), the effects of dilution protocol on protein folding from the denatured state (16), and the origins of prion-like behavior (19). More recently, lattice models have been used to study fibril formation (20). The lattice greatly reduces the amount of space conformational the protein can sample significantly increases the accessible time scales. The weakness of this approach, of course, is that the lattice does not realistically represent the true geometry of proteins and makes it very difficult to develop models of real proteins.

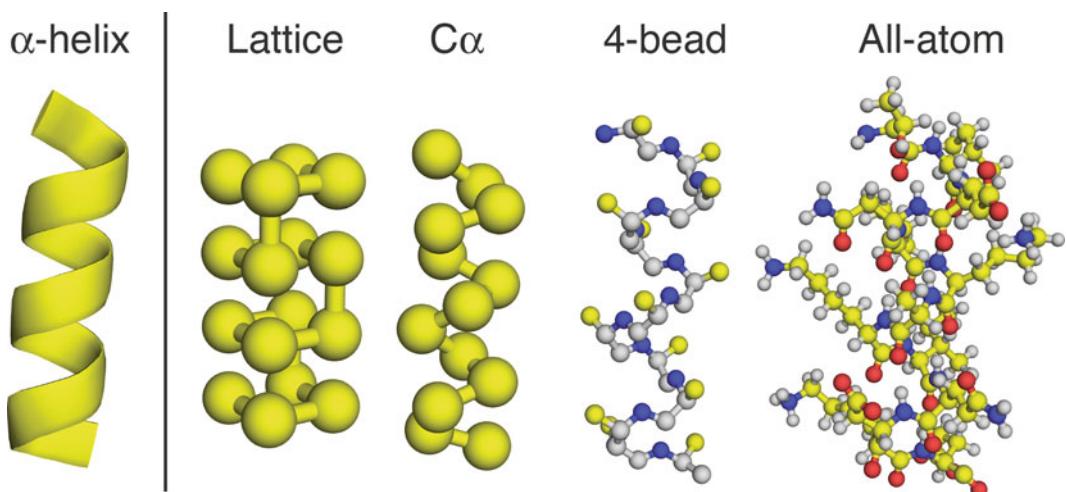
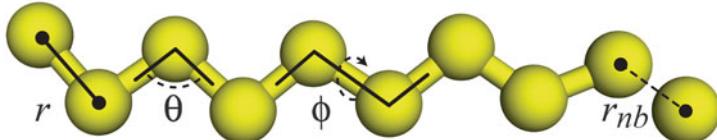


Fig. 1. An α -helix (*left*) as represented in models of increasing detail: 3D lattice, $C\alpha$ (off-lattice), 4-bead intermediate resolution, and all-atom.

Off-lattice models, on the other hand, can be readily adapted to mimic real proteins, thus affording the simulator the ability to make predictions regarding specific systems. Realistic models can be generated by starting with coordinates from structures determined by nuclear magnetic resonance (NMR) or X-ray crystallography. Tozzini provides a comprehensive review of coarse-graining approaches, thus we only discuss a few utilized in aggregation simulations (21). A widely-used and computationally friendly example is the $C\alpha$ model (Fig. 1), which centers a single bead on the α -carbon ($C\alpha$) sites of the protein. The potential functions that define the interactions of the $C\alpha$ model usually share a similar form (Fig. 2). There are bonded interactions that occur between beads that are local with respect to primary sequence. Bonded interactions are used to keep covalently bonded beads attached at their appropriate bond length, give the protein a realistic backbone configuration, and bias the chain conformations toward protein-like secondary structure, namely α -helices, β -sheets, and reverse turns. Nonbonded interactions usually occur between beads that are separated by three or more beads in primary sequence. Nonbonded interactions between beads that can attract typically take the form of a Lennard-Jones (LJ) potential. The attractive portion of the LJ potential is removed for bead pairs that have purely repulsive interactions. Fawzi and coauthors have added an anisotropic two-body interaction for the $C\alpha$ model meant to mimic the hydrogen bond between backbone atoms (22, 23). Other additions include long-range interactions (24) meant to account for electrostatics.

$C\alpha$ models have been used in a number of different aggregation studies. For example, Clark used $C\alpha$ models of proteins L and

Potentials for Cα Models with Molecular or Langevin Dynamics



$V(r)$ forces bonds to the optimum length (σ).
 r is the distance between beads.

$$V(r) = \sum_{bonds} k_r (r - \sigma)^2$$

$V(\theta)$ forces angles between three beads to realistic values (θ_0).

$$V(\theta) = \sum_{angles} k_\theta (\theta - \theta_0)^2$$

$V(\phi)$ controls dihedral angles and is used to bias four consecutive beads towards protein-like secondary structure.

$$V(\phi) = \sum_{dihedrals} [A(1 + \cos(\phi_0 + \phi)) + B(1 - \cos(\phi_0 + \phi)) + C(1 + \cos 3(\phi_0 + \phi)) + D(1 + \cos(\phi_0 + \phi + \frac{\pi}{4}))]$$

$V(r_{nb})$ sets the interactions between non-bonded beads. ϵ is the unit of energy.

$$V(r_{nb}) = \sum_{pairs} 4\alpha\epsilon \left[\left(\frac{\sigma}{r} \right)^{12} - \beta \left(\frac{\sigma}{r} \right)^6 \right]$$

Potentials for Four-Bead Models with DMD

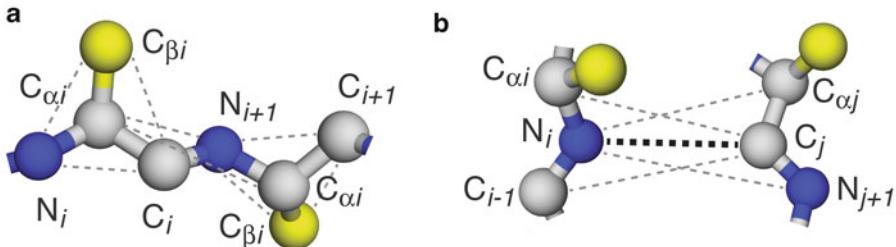


Fig. 2. Commonly used components of potential functions used in (a) Langevin dynamics simulations of C_α models and (b) DMD simulations of intermediate-resolution models.

G to predict that folding cooperativity plays an important role in aggregation propensity (25). Using a model of a β-barrel protein, Cellmer and coauthors studied the competition between folding and aggregation at high protein concentrations with a simulation methodology designed to match the one-step dilution approach used to refold proteins from *Escherichia coli* inclusion bodies (26). The authors found that aggregation was abated by the presence of structured intermediates that shielded aggregation-promoting beads from forming interprotein interactions. Fawzi and coauthors used a C_α model of the Alzheimer's disease associated A_β peptide to build an amyloid fibril (22) (Fig. 3). They systematically studied how fibril length affected fibril stability to arrive at a prediction for the number of A_β peptides in the critical nucleus, the transition

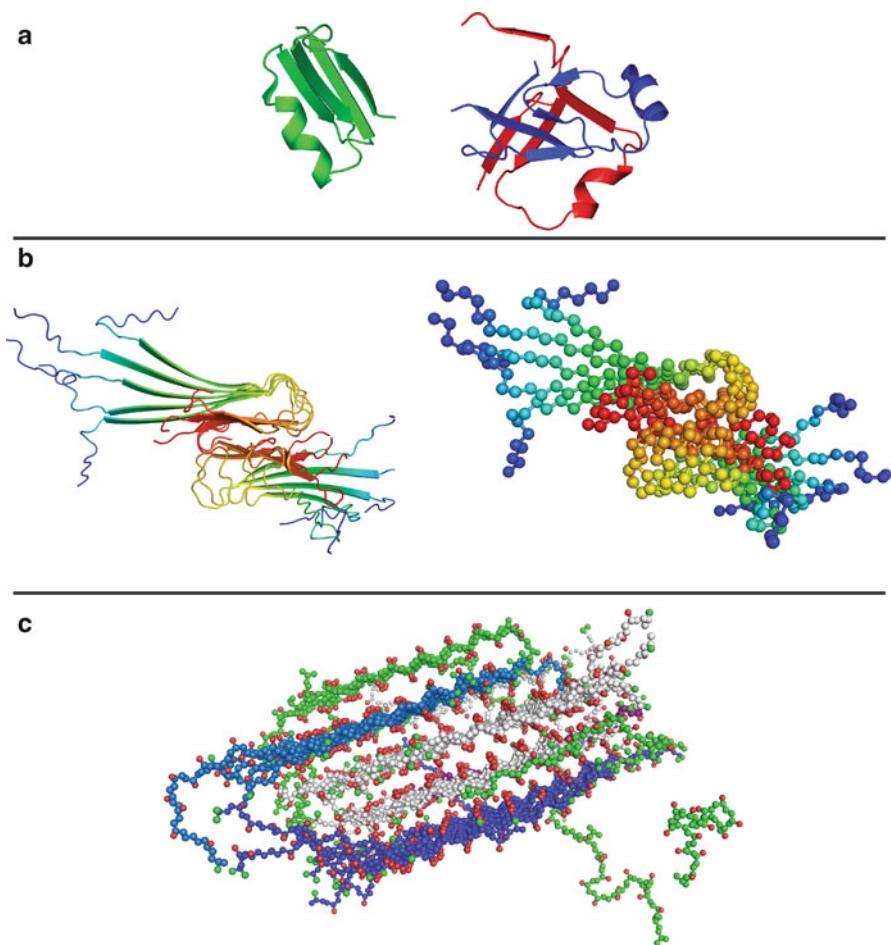


Fig. 3. Snapshots of aggregates from coarse-grained simulations. (a) Folded $C\alpha$ coarse-grained model of protein L (left) near an amorphous aggregate of two models of protein L stabilized by hydrophobic contacts in the beta-sheet region. (b) A $C\alpha$ coarse-grained model of amyloid- β fibril critical nucleus displayed as $C\alpha$ and cartoon representation. (c) Spontaneously formed, multilayered fibril of intermediate resolution models of polyalanine, transiently associating with nearby soluble peptides. Coordinates for figure (c) courtesy of Dr. Hung Nguyen.

state between disordered aggregates and the fibril. These studies provide examples of how one may tailor the simulation algorithm to answer a specific question regarding an aggregation mechanism.

In recent years, higher resolution models have become more prevalent. These models typically use four or more beads per amino acid, three representing the backbone (NH , $C\alpha$, CO) and one or more representing the side chains. Such models are referred to as intermediate-resolution (Fig. 1), as they represent a compromise between the $C\alpha$ and all-atom models. This protein representation makes it easier to incorporate backbone hydrogen bonds that often play an important role in aggregation. This is especially true in the case of amyloid fibril formation where interchain hydrogen bonds are a major stabilizing force for the aggregates. Furthermore, side

chains can be represented by multiple beads, thus making it easier to directly address questions pertaining to side-chain packing and point mutations that affect the size or character of the side chain.

The Hall group used an intermediate-resolution model called PRIME to study aggregation in biologically relevant systems (27–30). They used a four-bead model in simulations of polyalanine peptides, which showed the spontaneous formation of fibrils (27–29) (Fig. 3). The entire aggregation pathway was observed, allowing the authors to study the physics of monomer transitions, early aggregate formation, nucleus formation, and fibril elongation. The Hall group also studied the aggregation of polyglutamine peptides, which has been linked to Huntington's disease (30). In these simulations, the backbone was again represented by three beads and the side chain by four beads. They observed an enhanced ability to form aggregates at longer chain lengths, as observed in experiments. Most recently, the Hall and coauthors have extended PRIME to represent all 20 amino acids (31).

The Stanley group used an intermediate-resolution model proteins for simulations of the A β peptide, which has been implicated in the pathology of Alzheimer's Disease (32–35). Initial studies focused on the dimerization and oligomerization of A β 40 and A β 42 (33, 34). Work focusing on these aspects of the aggregation process is particularly important because of the toxicity and poor stability of early aggregates, the latter of which makes experimental investigation particularly challenging. Other studies examined the role of electrostatic interactions on A β oligomerization and found oligomer distributions that agree with experiments (35).

The Dokholyan group also uses intermediate-resolution models to study disease-related proteins (33, 36–38). Chen investigated differences in the aggregation behavior of oxidized and reduced β 2-microglobulin (β 2 M). β 2 M aggregates into amyloid fibrils and can cause amyloidosis in patients receiving long-term renal dialysis. The authors found that the Cys25-Cys80 disulfide bond mediates the aggregation pathway; oxidized β 2 M aggregates via domain swapping, while reduced β 2 M aggregates via parallel stacking of β -sheets (38). Ding investigated the conversion of an α -helical peptide into β -sheet topology (37), a transition observed in many aggregating systems and often associated with prion-like behavior. Other studies from this research group focused on simulations of aggregation-prone monomers in isolation (39). This approach may be preferable when the system of interest is too large or too complicated to study via aggregation simulations. By observing the conformational behavior of the monomer, it may be possible to develop hypotheses relevant to the basis for aggregation.

An important feature of the aforementioned intermediate-resolution simulations is the use of discontinuous molecular dynamics (DMD) (40, 41). Traditional molecular dynamics (MD)

uses Newton's equations of motion to move particles at very small, discrete time steps of the same size. DMD, on the other hand, moves forward in time by proceeding to the next event. Examples of events include particle collisions and local events such as bond, pseudobond, and torsion events, which are intended to maintain physically realistic bond lengths, molecular geometry, and dihedral propensity, respectively. The event-based nature of the simulation algorithm requires that the potentials be of the square well form (Fig. 2); therefore, nonbonded contacts are either formed or broken. This is in contrast to the traditional MD approach, in which the interaction potentials change gradually as a function of distance in a more realistic manner. However, the event based algorithm allows the simulations to proceed more rapidly and access longer time scales. Given the sacrifices made during the coarse-graining process, the additional physical sacrifice required for DMD simulations may represent a useful trade-off for the substantial increase in accessible time scales.

2. Materials

Computational requirements depend heavily on the problem being investigated, system details, simulation algorithm, and efficiency of the code. Nonetheless, we wish to note that a single workstation can provide enough computer power to perform simulations of fibril formation. For example, Hall (41) reports observing protofibril formation in ~160 h in DMD simulations of 96 intermediate resolution peptides carried out on a single AMD Athlon 2200+ workstation.

3. Methods

Simulations of CG aggregating proteins are usually tailored to answer a specific question, thus the precise set of steps to conduct and analyze these simulations will vary from system to system. Therefore in this tutorial, we simply provide a set of milestones that, generally speaking, will be required to perform and generate experimentally testable hypotheses from these simulations. We focus on the process of setting up, running, and analyzing simulations of CG models of particular proteins so that direct comparisons can be made to experiments.

3.1. Construct a Model of the Protein Structure and Sequence

1. Coarse-grain the protein representation.

The starting point for a simulation study of the aggregation of natively folded proteins is the conversion of an atomistic, three-dimensional structure of the protein into a CG representation.

Published structures determined by X-ray crystallography or NMR spectroscopy can be obtained from the RCSB Protein Data Bank (PDB, <http://www.pdb.org>). If no experimentally determined structure is available, a theoretical model structure derived from the structures of homologous proteins (known as a “homology model”) may make a good starting point for CG simulations. With the experimental structure or homology model in hand, a computer script specific to the chosen CG model replaces each amino acid with one or multiple “beads” or “centers” at the appropriate positions.

One may also be interested in simulating peptides that do not fold into stable structures. Examples include the A β peptide, whose aggregation has been implicated in Alzheimer’s disease and amylin, which forms aggregates thought to play a role in type 2 diabetes. If biophysical data concerning the average protein conformation at the individual residue level is available, this data may be chosen as a starting point to generate a CG model as described above for a folded protein. Other recent simulation work has focused on the aggregation of small peptides (24, 42–45). Coordinates of relevant conformations of the monomer or an aggregate from the PDB or previous simulation or model building approaches may also serve as useful starting structures. If no previous coordinates are available, a coordinate file of the initial model in an arbitrary, extended conformation can be constructed from only the amino-acid sequence using one of the community codes listed above and then converted to the CG representation using the same procedure as for an experimental structure.

2. Choose the potential function.

One must define the strength and character of the interactions between the beads representing each amino acid. Figure 2 highlights generic potential functions for C α and intermediate-resolution models. While the potentials that govern bonded interactions are not likely to deviate too far from those represented in Fig. 2, there may be significant variation in the interactions between nonbonded beads. Miyazawa and Jernigan (46), as well as Hall (31), have developed interaction potentials for all 20 amino acids. Also, hydrogen bonds typically play an important role in aggregation. Head Gordon and coauthors (22, 23) have developed a hydrogen bond potential for C α models, while Hall (47) and Ding (48) use hydrogen bond potentials in their simulations of intermediate resolution models. Finally, one may also choose to bias potentials based on a priori knowledge or to test a particular hypothesis. For example, if the aggregation is thought to be primarily characterized by

domain-swapping, where intraprotein native interactions are replaced with same interactions between separate proteins, one may use an interprotein, symmetric version of a Go potential biased for the formation of native structure (49).

3. Choose the algorithm for moving the particles.

There are a number of different algorithms for moving the particles. Smoothly varying, continuous potentials often used in C α models (Fig. 2) are appropriate for use with a number of discrete time-step methods of integrating Newton's equations of motion, including the leap-frog and velocity-Verlet integrators. In addition, integrating the Langevin equation, which adds a random force and drag terms to Newton's equations of motion, is a convenient way to implicitly incorporate the random forces, friction, and temperature bath effects of solvent. The random forces can be useful in alleviating kinetic trapping issues common with some CG models (T. Cellmer, unpublished observations). Conversely, discontinuous step function potentials are integrated using the DMD method where at each step, Newton's equations of motion are solved for the time of the next event or change in the interaction energy of any two particles. DMD is more computationally efficient than Langevin dynamics and is likely the best choice for large systems.

3.2. Characterize the CG model of the Single Protein

Once an initial structure and sequence has been generated, the characterization of the protein in isolation can be performed to evaluate the accuracy of the model and determine important parameters for the aggregation simulations. Unphysical behavior or behavior grossly inconsistent with experimental observations may require changes to the potential function. Head-Gordon and coworkers describe an iterative method for changing the CG sequence to match the experimentally observed folding and unfolding behavior (50). While performing an exhaustive analysis of the monomer requires effort, we believe it is necessary to minimize the chance that significant model artifacts of the may complicate the interpretation of aggregation simulations. Furthermore, a working model of the monomer may provide important information regarding aggregation behavior. For example, in many situations some degree of unfolding may be required for aggregation. Simulations of the monomer can help identify regions of the protein most likely to be unfolded.

Lastly, extensive characterization of the monomer may not be necessary for studying the aggregation of short disordered peptides. This approach also has the advantage that aggregation tends to proceed more quickly as the monomer does not have to undergo extensive structural changes to adopt an aggregation-competent state.

1. Obtain the lowest energy structure.

The first step in characterizing the folding of the monomer is to search for the lowest energy conformation (26, 51). This simply involves heating and slowly cooling the monomer a number of times to locate the structure that has the lowest energy. To ensure that an acceptable description of the native state is at least locally stable in the CG potential function, the energy-minimized structure can then be compared to the experimental structure by calculating the C α RMSD of the two structures and by inspecting the structures using a molecular visualization program.

2. Characterize the thermodynamics of monomer folding.

A thorough characterization of the thermodynamics of folding and conformational changes involves repeatedly heating and cooling the system, or examining several simulations in parallel at different temperatures in order to monitor the folding and unfolding behavior of the protein. This may include replica-exchange simulations (13), designed to enhance thermodynamic sampling, as well as the weighted histogram method (52) for calculating free energies. Many publications pertaining to protein folding offer excellent examples of characterizing the folding of CG proteins (51, 53, 54). The approach to characterizing the monomer is similar in situations where the protein does not fold into a stable native structure. If NMR data such as scalar coupling measurements that provide constraints on the backbone dihedral angles is available (55), one can compare the conformational preferences of the monomer in the simulation to the experimental data.

3.3. Prepare the System for Aggregation Simulations

1. Choose the number of proteins.

In general, representing as few proteins as possible will allow for the most efficient simulations. Many studies have been performed with just two proteins interacting in the simulation box (18, 49, 56). Unlike an *in vitro* experiment where macroscopic numbers of proteins potentially interact, the simulation will proceed most efficiently and will be easiest to analyze if the minimum number of representations are simulated for the maximum amount of time. If a higher order aggregation mechanism involving a trimer or larger species is of interest, the simulation should include at least this many representations and simulations with tens of proteins (or more) are an option. The effect of adding additional representations can be evaluated after initial simulations with fewer chains are completed.

2. Choose the size of the simulation box.

To mimic bulk solutions, simulations are generally conducted in a repeating box with periodic boundary conditions, where

proteins diffuse out one face of the box and reenter on the opposite face. The size of this box and number of proteins determine the effective concentration of the protein in the simulation, though it is common practice to set a concentration considerably higher than that in the experiment. From a practical standpoint, a high concentration will speed up the simulated aggregation process. From a theoretical standpoint, large initial distances separating proteins are usually not necessary since long-range interactions, i.e., charge attraction and repulsion, that decrease slowly as a function of increasing distance ($1/r$) can be faithfully represented by short-range potentials decreasing more rapidly ($1/r^4$ or $1/r^6$) in CG models due to dielectric screening and counter ion effects normally present in experiments. Therefore, as long as the proteins in the initial state are separated such that they are not in contact and have near-zero interprotein interaction energy, the effect of box size on aggregation kinetics should be due primarily to increased time necessary for diffusion over larger distances. To maximize the distance of initial separation in a cubic box for two proteins, the center of mass of one protein is set to the origin (coordinates 0, 0, 0) and the second protein to half the box length, l , along each dimension (coordinates $l/2, l/2, l/2$). For three proteins, the maximum separation is achieved when the proteins are placed at $(l/2, 0, 0)$, $(0, l/2, 0)$, and $(0, 0, l/2)$.

3.4. Run the Aggregation Simulations

The aggregation simulations can be launched once the starting structure incorporating multiple proteins has been constructed. A variety of possible simulation protocols are available, depending on the problem at hand. We briefly discuss two below.

1. Replica-exchange simulations.

If the primary goal is a rigorous thermodynamic characterization of the system, one may choose to run replica-exchange simulations. Although a single very long simulation will in principle sample and correctly weight all the available states of the system to calculate the thermodynamic averages of interest, the length of a simulation necessary to do so is impossible to predict *a priori* and may require years of computer time depending on the system and temperature of interest. In replica-exchange simulations, on the other hand, one launches simulations at a number of different temperatures and periodically attempts to swap replicas at different temperatures using a Metropolis criterion. This leads to improved sampling, though some effort is required to optimize the procedure. One needs to investigate the number of replicas, the temperatures of the replicas, and the frequency of the swapping criterion. Replica-exchange has been employed successfully in a number of different studies aimed at studying aggregation ([13, 56–59](#)).

2. Single-temperature simulations.

If aggregation kinetics is the primary concern, one may perform simulations at a single temperature of interest. Nonetheless, a single aggregation simulation, though it may show an example of a pathway to an aggregate, will not be sufficient to fully characterize the aggregation properties of the system. Small changes in starting structure, relative protein orientation and the use of Langevin dynamics (i.e., stochastic dynamics) can change the aggregation path and the outcome from simulation to simulation. Therefore, starting a large number of simulations with a different random number seed and with different (rotated) initial orientations of the proteins enables sampling of this ensemble of aggregation outcomes. The number of simulations required for sufficient statistics is difficult to predict a priori. One way to approach this problem is to divide the simulations into three sets and compare the resulting observables. If one obtains the same values within an acceptable variance from each set of simulations, it is a good indication that the observables of interest are near convergence.

3.5. Evaluating the Aggregation Simulations

Once an aggregation simulation is running, its progress can be monitored by calculating aggregation metrics from snapshots of the system evolving in time. Below we discuss several analysis-related tasks, the unifying theme being their relevance to experimental comparisons.

1. Monitoring aggregation.

Most metrics used to monitor aggregation utilize either the interprotein interaction energy or number of nonbonded contacts between chains, including hydrogen bonds. Contacts are formed when pairs of nonbonded beads come within some predetermined distance. The characteristics of the aggregated species, such as their size, can serve as a point of comparison with experiments (34). When equilibrium data is available, free-energy diagrams (14, 18) that reveal the presence of stable species and characteristics of the transition states that separate them can be constructed. An example of such a diagram can be seen in Fig. 4.

2. Determining the aggregation rates.

Rates can be determined by monitoring the fraction of aggregated proteins as a function of time. In most cases, the transition of a monomer to an aggregated state is cooperative in terms of an appropriate metric, making it straightforward to distinguish between the two states and calculate the percentage of aggregated material. As stated above, most metrics used to monitor aggregation involve the energy or number of contacts between proteins.

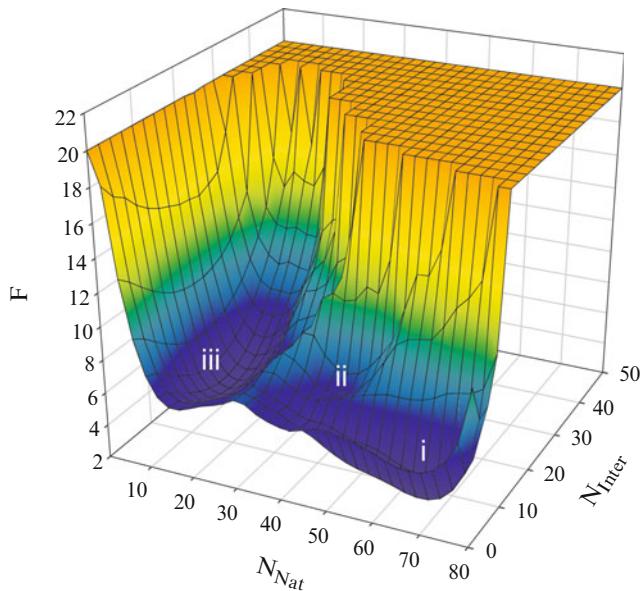


Fig. 4. Free-energy (F) as a function of the number of native contacts (N_{Nat}), i.e., intraprotein contacts found in the native state, and number of interprotein contacts (N_{Inter}). The plot was constructed from simulations of two lattice-model proteins. State (i) corresponds to folded proteins with a small number of interprotein contacts. States (ii) and (iii) correspond to misfolded, aggregated proteins with varying amounts of native contacts and interprotein contacts.

Due to the coarse-grained nature of the modeling, obtaining quantitative agreement with experiments is very difficult. Nonetheless, we believe that CG models are well equipped to offer valuable insight into the basis for qualitative differences in aggregation rates. For example, if experimental data is available on mutation effects, one could calculate rates from CG simulations on a model of the wild-type protein and a modified version of that model meant to mimic the mutant. Presuming the same qualitative effect on the rate is observed, one can then exploit the molecular-level viewpoint of the simulation to develop a hypothesis for the basis of the observed effect.

3. Determining the amino acids involved in aggregation.

CG simulations usually offer insight at the amino-acid level, thus one can readily determine the amino acids that play the most important role in various stages of the aggregation process. A generic approach for locating these amino acids involves clustering aggregates based on some structural metric or metrics, and then generating individual contact maps (11, 56) for these clusters (Fig. 5). The contact map simply shows the probability of observing a contact between each possible pair of amino acids. One can then propose mutations aimed at altering the stability of the aggregates, test them computationally, and suggest experiments based on the outcome.

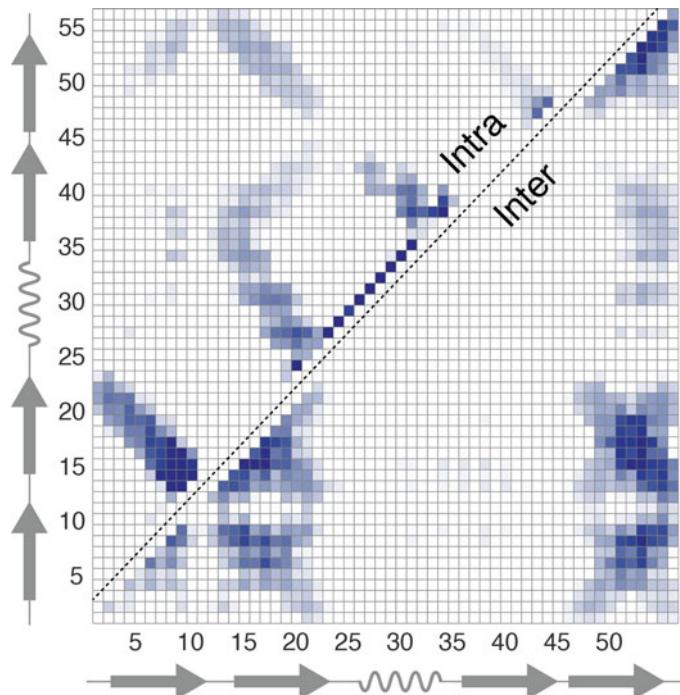


Fig. 5. Contact map from protein aggregation simulations of Cx model of protein L for the 60% of the simulations with the fastest times to form a stable aggregate, as in Fig. 3b (11). The contacts formed between one protein and itself when involved in an aggregate, the intrachain contact map, is plotted above the diagonal, while the contacts formed between different proteins that stabilize the aggregate, are plotted on and below the diagonal. The different secondary structure elements are plotted adjacent to the residue numbers. The plot demonstrates that the intraprotein, antiparallel β -hairpin involving residues 1 through 21 is maintained (*lower left, above diagonal*), while the β -hairpin involving residues 35–56 is largely disrupted (*upper right, above diagonal*), releasing residues 45–56 to participate in interprotein contacts (*right, below diagonal*).

4. Notes

Listed below are a series of software packages and their Web addresses that could be useful for one interested in performing coarse-grained simulations of protein aggregation.

1. *Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank (PDB)* (Web site: <http://www.rcsb.org>). The primary source of protein structural data where experimental structures, sequences, and other structural information are deposited. The atomic coordinates found in PDB files can be converted to coarse-grained representations.
2. *Modeller* (Web site: <http://salilab.org/modeller>). Modeller uses homology modeling to predict a three-dimensional structure for an amino-acid sequence when it is believed the structure is similar to a known structure from another sequence.

3. *Rosetta* [Web site: <http://www.rosettacommons.org>]. Using innovative algorithms combined with searching through structural fragments based on known protein structures, Rosetta predicts protein structure solely from its amino-acid sequence.
4. *AmberTools* [Web site: <http://ambermd.org>]. Amber is a molecular dynamics simulation and analysis package, available with an inexpensive academic or commercial license. The AmberTools free package offers many of the analysis and molecule building tools available in the full version of Amber. The LEaP module builds unstructured proteins or peptides in an initially extended conformation.
5. *NAMD* [Web site: <http://www.ks.uiuc.edu/Research/namd—NAMD>] is a freely available molecular simulation program ideal for running large systems with many molecules. Among many other fully atomistic force fields available, NAMD also offers an implementation of the MARTINI coarse-grained force field (<http://md.chem.rug.nl/cgmartini/>), which can be used for some protein aggregation problems. A description of this method, referred to as “residue based coarse-graining,” is available in the NAMD documentation: <http://www.ks.uiuc.edu/Research/CG/rbcg.html>.
6. *GROMACS* [Web site: <http://www.gromacs.org>]—**GROMACS** is another freely available program for running molecular dynamics simulations. It scales very well for large systems and parallel computing. Like NAMD, it can also be used to run simulations with the MARTINI coarse-grained force field. Several useful tutorials for setting up and running MARTINI simulations are found at: <http://md.chem.rug.nl/~marrink/MARTINI/Parameters.html>, including a script to convert a fully atomistic protein PDB file to a coarse-grained starting file (atom2cg_v2.1.awk).
7. *PyMOL* [Web site: <http://www.pymol.org/>]—**PyMOL** is an excellent computer program for molecular visualization. It integrates some analysis tools including distance measurements into an interactive environment. It is also useful for creating publication quality figures.
8. *UCSF CHIMERA* [Web site: <http://plato.cgl.ucsf.edu/chimera/>]—**UCSF Chimera** is another excellent visualization program. It creates animations, which can be particularly valuable for watching the progression of protein aggregation trajectories.
9. *Multiscale Modeling Tools for Structural Biology* [Web site: <http://mmtsb.org>]—**MMTSB** is a free set of tools for setting up and analyzing the results of molecular simulations.

Particularly useful for coarse-grained simulations of protein aggregation are the following scripts:

- convpdb.pl—standardizes PDB files before conversion to initial coarse-grained models.
- complete.pl—rebuilds a fully atomistic representation from a coarse-grained structure or trajectory of structures.
- countIntCont.pl—counts the number of contacts between two different chains, an excellent reaction coordinate for following protein aggregation.

References

1. Dobson CM (2003) Protein folding and misfolding. *Nature* 426:884–890
2. Clark ED (2001) Protein refolding for industrial processes. *Curr Opin Biotechnol* 12:202–207
3. den Engelsman J, Garidel P, Smulders R, Koll H, Smith B, Bassarab S, Seidl A, Hainzl O, Jiskoot W (2011) Strategies for the assessment of protein aggregates in pharmaceutical biotech product development. *Pharm Res* 28:920–933
4. Cherny I, Gazit E (2008) Amyloids: not only pathological agents but also ordered nanomaterials. *Angew Chem Int Ed* 47:4062–4069
5. Ulijn RV, Smith AM (2008) Designing peptide based nanomaterials. *Chem Soc Rev* 37:664–675
6. Zhang SG (2003) Fabrication of novel biomaterials through molecular self-assembly. *Nat Biotechnol* 21:1171–1178
7. Bratko D, Cellmer T, Prausnitz JM, Blanch HW (2007) Molecular simulation of protein aggregation. *Biotechnol Bioeng* 96:1–8
8. Cellmer T, Bratko D, Prausnitz JM, Blanch HW (2007) Protein aggregation in silico. *Trends Biotechnol* 25:254–261
9. Wu C, Shea JE (2011) Coarse-grained models for protein aggregation. *Curr Opin Struct Biol* 21:209–220
10. Ma BY, Nussinov R (2006) Simulations as analytical tools to understand protein aggregation and predict amyloid conformation. *Curr Opin Chem Biol* 10:445–452
11. Fawzi NL, Yap EH, Okabe Y, Kohlstedt KL, Brown SP, Head-Gordon T (2008) Contrasting disease and nondisease protein aggregation by molecular simulation. *Acc Chem Res* 41: 1037–1047
12. Piana S, Lindorff-Larsen K, Shaw DE (2011) How robust are protein folding simulations with respect to force field parameterization? *Biophys J* 100:L47–L49
13. Bratko D, Blanch HW (2003) Effect of secondary structure on protein aggregation: a replica exchange simulation study. *J Chem Phys* 118:5185–5194
14. Cellmer T, Bratko D, Prausnitz JM, Blanch H (2005) Protein-folding landscapes in multi-chain systems. *Proc Natl Acad Sci USA* 102: 11692–11697
15. Gupta P, Hall CK, Voegler A (1999) Computer simulation of the competition between protein folding and aggregation. *Fluid Phase Equilib* 160:87–93
16. Gupta P, Hall CK, Voegler AC (1998) Effect of denaturant and protein concentrations upon protein refolding and aggregation: A simple lattice model. *Protein Sci* 7:2642–2652
17. Istrail S, Schwartz R, King J (1999) Lattice simulations of aggregation funnels for protein folding. *J Comput Biol* 6:143–162
18. Bratko D, Cellmer T, Prausnitz JM, Blanch HW (2006) Effect of single-point sequence alterations on the aggregation propensity of a model protein. *J Am Chem Soc* 128:1683–1691
19. Harrison PM, Chan HS, Prusiner SB, Cohen FE (1999) Thermodynamics of model prions and its implications for the problem of prion protein folding. *J Mol Biol* 286:593–606
20. Li MS, Klimov DK, Straub JE, Thirumalai D (2008) Probing the mechanisms of fibril formation using lattice models. *J Chem Phys* 129: 175101
21. Tozzini V (2010) Minimalist models for proteins: a comparative analysis. *Q Rev Biophys* 43:333–371
22. Fawzi NL, Okabe Y, Yap EH, Head-Gordon T (2007) Determining the critical nucleus and mechanism of fibril elongation of the Alzheimer's A beta(1–40) peptide. *J Mol Biol* 365:535–550
23. Yap EH, Fawzi NL, Head-Gordon T (2008) A coarse-grained alpha-carbon protein model with anisotropic hydrogen-bonding. *Proteins* 70:626–638

24. Bellesia G, Shea JE (2007) Self-assembly of beta-sheet forming peptides into chiral fibrillar aggregates. *J Chem Phys* 126(24):245104
25. Clark LA (2005) Protein aggregation determinants from a simplified model: cooperative folders resist aggregation. *Protein Sci* 14: 653–662
26. Cellmer T, Bratko D, Prausnitz JM, Blanch H (2005) The competition between protein folding and aggregation: off-lattice minimalist model studies. *Biotechnol Bioeng* 89:78–87
27. Nguyen HD, Hall CK (2006) Spontaneous fibril formation by polyalanines; discontinuous molecular dynamics simulations. *J Am Chem Soc* 128:1890–1901
28. Nguyen HD, Hall CK (2005) Kinetics of fibril formation by polyalanine peptides. *J Biol Chem* 280:9074–9082
29. Nguyen HD, Hall CK (2004) Molecular dynamics simulations of spontaneous fibril formation by random-coil peptides. *Proc Natl Acad Sci USA* 101:16180–16185
30. Marchut AJ, Hall CK (2007) Effects of chain length on the aggregation of model polyglutamine peptides: molecular dynamics simulations. *Proteins* 66:96–109
31. Cheon M, Chang I, Hall CK (2010) Extending the PRIME model for protein aggregation to all 20 amino acids. *Proteins* 78:2950–2960
32. Urbanc B, Borreguero JM, Cruz L, Stanley HE (2006) Ab initio discrete molecular dynamics approach to protein folding and aggregation. *Methods Enzymol* 412:314–338
33. Urbanc B, Cruz L, Ding F, Sammond D, Khare S, Buldyrev SV, Stanley HE, Dokholyan NV (2004) Molecular dynamics simulation of amyloid beta dimer formation. *Biophys J* 87: 2310–2321
34. Urbanc B, Cruz L, Yun S, Buldyrev SV, Bitan G, Teplow DB, Stanley HE (2004) In silico study of amyloid beta-protein folding and oligomerization. *Proc Natl Acad Sci USA* 101: 17345–17350
35. Yun SJ, Urbanc B, Cruz L, Bitan G, Teplow DB, Stanley HE (2007) Role of electrostatic interactions in amyloid beta-protein (A beta) oligomer formation: a discrete molecular dynamics study. *Biophys J* 92:4064–4077
36. Ding F, Dokholyan NV (2008) Dynamical roles of metal ions and the disulfide bond in Cu, Zn superoxide dismutase folding and aggregation. *Proc Natl Acad Sci USA* 105: 19696–19701
37. Ding F, LaRocque JJ, Dokholyan NV (2005) Direct observation of protein folding, aggregation, and a prion-like conformational conversion. *J Biol Chem* 280:40235–40240
38. Chen YW, Dokholyan NV (2005) A single disulfide bond differentiates aggregation pathways of beta 2-microglobulin. *J Mol Biol* 354: 473–482
39. Khare SD, Ding F, Dokholyan NV (2003) Folding of Cu, Zn superoxide dismutase and familial amyotrophic lateral sclerosis. *J Mol Biol* 334:515–525
40. Sharma S, Ding F, Dokholyan NV (2008) Probing protein aggregation using discrete molecular dynamics. *Front Biosci* 13:4795–4807
41. Hall CK, Wagner VA (2006) Computational approaches to fibril structure and formation. *Methods Enzymol* 412:338–365
42. Bellesia G, Shea JE (2009) Diversity of kinetic pathways in amyloid fibril formation. *J Chem Phys* 131(11):111102
43. Bellesia G, Shea JE (2009) Effect of beta-sheet propensity on peptide aggregation. *J Chem Phys* 130(14):145103
44. Auer S, Dobson CM, Vendruscolo M (2007) Characterization of the nucleation barriers for protein aggregation and amyloid formation. *HFSP J* 1:137–146
45. Auer S, Dobson CM, Vendruscolo M, Maritan A (2008) Self-templated nucleation in peptide and protein aggregation. *Phys Rev Lett* 101(25):258101
46. Miyazawa S, Jernigan RL (1996) Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J Mol Biol* 256:623–644
47. Smith AV, Hall CK (2001) Alpha-helix formation: discontinuous molecular dynamics on an intermediate-resolution protein model. *Proteins* 44:344–360
48. Ding F, Borreguero JM, Buldyrey SV, Stanley HE, Dokholyan NV (2003) Mechanism for the alpha-helix to beta-hairpin transition. *Proteins* 53:220–228
49. Ding F, Dokholyan NV, Buldyrey SV, Stanley HE, Shakhnovich EI (2002) Molecular dynamics simulation of the SH3 domain aggregation suggests a generic amyloidogenesis mechanism. *J Mol Biol* 324:851–857
50. Brown S, Fawzi NJ, Head-Gordon T (2003) Coarse-grained sequences for protein folding and design. *Proc Natl Acad Sci USA* 100: 10712–10717
51. Sorenson JM, Head-Gordon T (2000) Matching simulation and experiment: a new simplified model for simulating protein folding. *J Comput Biol* 7:469–481

52. Kumar S, Bouzida D, Swendsen RH, Kollman PA, Rosenberg JM (1992) The weighted histogram analysis method for free-energy calculations on biomolecules. 1. The method. *J Comput Chem* 13:1011–1021
53. Sorenson JM, Head-Gordon T (2002) Protein engineering study of protein L by simulation. *J Comput Biol* 9:35–54
54. Guo ZY, Brooks CL (1997) Thermodynamics of protein folding: a statistical mechanical study of a small all-beta protein. *Biopolymers* 42: 745–757
55. Karplus M, Grant DM (1959) A criterion for orbital hybridization and charge distribution in chemical bonds. *Proc Natl Acad Sci USA* 45:1269–1273
56. Cellmer T, Bratko D, Prausnitz JM, Blanch H (2005) Thermodynamics of folding and association of lattice-model proteins. *J Chem Phys* 122(17):174908
57. Cecchini M, Rao F, Seeber M, Caflisch A (2004) Replica exchange molecular dynamics simulations of amyloid peptide aggregation. *J Chem Phys* 121:10748–10756
58. Takeda T, Klimov DK (2009) Side chain interactions can impede amyloid fibril growth: replica exchange simulations of a beta peptide mutant. *J Phys Chem B* 113: 11848–11857
59. Takeda T, Klimov DK (2009) Replica exchange simulations of the thermodynamics of a beta fibril growth. *Biophys J* 96:442–452

Chapter 28

Chitosan-Based Nanoparticles as Delivery Systems of Therapeutic Proteins

Pedro Fonte, José Carlos Andrade, Vítor Seabra, and Bruno Sarmento

Abstract

Therapeutic proteins represent a significant part of the new pharmaceuticals coming on the market every year and are now widely spread in therapy to treat or relief symptoms related to many metabolic and oncologic diseases. The parenteral route remains as a primary strategy for protein administration essentially due to its specific physicochemical properties. However, the research on alternative nonparenteral delivery routes continues. The high molecular weight (MW), hydrophilicity, and charged nature of therapeutically valued proteins render transport through membranes very difficult. In this regard, chitosan arises as a promising candidate for the development of protein-containing drug formulations, due to its exceptional biological properties. Chitosan-based delivery systems have been proposed as valid approaches to provide protective conditions to proteins from denaturation and loss of activity, during preparation and delivery, as well as during long-term storage of the prepared formulation.

In this chapter, one production method of a chitosan-based nanoparticle formulation is presented, as well as several characterization techniques to assess both nanoparticles and proteins characteristics and stability.

Key words: Chitosan, Chitosan-based delivery nanoparticles, Insulin, Nanoparticles, Nanoparticle characterization, Therapeutic proteins

1. Introduction

Therapeutic proteins are macromolecular drugs (>1,000 Da) manufactured by biotechnology, usually involving live organisms or their active compounds, that are becoming extensively useful for the management and cure of many diseases, particularly in the treatment of many oncologic and metabolic diseases. Actually, these drugs can provide a better approach to treat organic diseases due to the similarity to biologic molecules, comparatively to synthetic drug molecules, and their selectivity and ability to provide

effective and potent action, causing fewer side effects (1). These drugs represent a rapidly growing part of marketed drugs taking place alongside other established therapies (2). Thus, nowadays therapeutic proteins are used to relieve patients suffering from many conditions, including strokes, heart attack, cystic fibrosis (enzymes, blood factors), anemia (erythropoietin), hemophilia (blood clotting factors), diabetes (insulin), and cancers (monoclonal antibodies, interferons). Peptides are also being used to generate therapeutics for enhancing cellular uptake, drug targeting, and vaccination (2, 3). However, their size and physicochemical properties are challenging problems to overcome when developing adequate systems that allow its use, because of their inherent instability, poor pharmacokinetics, and potential toxicity (4).

Indeed, the bioavailability of proteins is reduced, mainly due to their high molecular weight (MW) and hydrophilic characteristics, which affect their ability to cross membranes. In addition, these molecules have limited chemical stability in vivo, undergoing proteolytic cleavage and degradation, leading to a fast removal from the bloodstream. Furthermore, the reactivity of amino acids can lead to potential hydrolysis or oxidation reactions, which are dependent on the conditions of production and storage of the formulations, such as temperature, ionic strength, pH, and others (1). Thus, under these conditions, proteins may suffer adsorption, aggregation, and denaturation, which limit their in vivo concentration. To overcome these stability problems, different excipients can be used in formulations, such as nonionic surfactants that reduce its aggregation, metal chelators, and enzyme inhibitors that reduce the activity of various proteolytic enzymes (2). Glycocholate, bacitracin, leupeptin, camostat mesilate, and chymostatin are some examples of enzyme inhibitors used to increase stability in protein-containing formulations (5, 6). It is possible to reduce the proteolytic capacity of exo- and endopeptidases, but maintaining the proteins stability at constant levels remains a complex task, since these molecules have several linkages along its sequence that can be broken by various proteases. Moreover, the degradation of only one linkage may be sufficient to undergo the loss of biological activity. Besides, the success of the enzyme inhibitors used in the laboratory still represents a challenge for widespread acceptance by clinicians and regulatory organizations. For instance, the use of enzyme inhibitors in long-term therapy remains questionable due to the possibility of disturbance of nutritive proteins digestion, and stimulation of protease secretion as a result of feedback regulation (7).

Regarding the administration of therapeutic proteins, parenteral injection remains the most used route of administration due to the instability and reduced permeability of proteins through biological membranes. However, this invasive route may lead to reduced patient compliance and consequently therapeutic failure

because of its inherent discomfort and pain. Moreover, the frequent need for qualified personnel to carry out the administration, the social stigma associated with the use of needles, and the requirements of sterile manufacture and storage led to the development of needle-free formulations for administration of therapeutic proteins (2, 8). Therefore, different routes are proposed for the administration of therapeutic proteins, including transdermal, ocular, nasal, oral, and others (2). Beside all of these noninvasive routes, the delivery of proteins remains a challenge that deserves further investigation.

Polymer-based delivery systems have been proposed as valid approaches to provide successful alternative to formulate therapeutic proteins. Particularly, chitosan-based delivery systems have been projected to grant protective and favorable conditions to such bio-pharmaceutical drugs and have been extensively tailored according to the required needs. Chitosan is obtained by partial deacetylation of chitin, the second most abundant natural polysaccharide to cellulose, and mainly found in crustaceans, mollusks, marine diatoms, insects, algae, fungi, and yeasts, representing a family of *N*-deacetylated chitins (6, 7). Both polymers comprise linear β -(1-4)-linked monosaccharides but the functional groups connected to the second carbon in the repeating units differ (9). Thus, chitin is a linear homopolymer composed of β -(1,4)-linked *N*-acetyl-glucosamine units (10), while chitosan is a linear copolymer polysaccharide consisting of β -(1-4)-linked 2-amino-2-deoxy-D-glucose (D-glucosamine) and 2-acetamido-2-deoxy-D-glucose (*N*-acetyl-D-glucosamine) units (11). Therefore, chitin is not useful in the development of drug delivery systems due to its insolubility in water and in most common organic solvents used in pharmaceutical technology (12). In contrast, chitosan is insoluble in water and organic solvents, but soluble in diluted aqueous acidic solution ($\text{pH} < 6.5$), which can convert the glucosamine units into a soluble protonated amine form (12). The positive charge of chitosan is useful since it enables interaction with poly-anions (13). Thus, the biodegradability and biological properties of chitosan are dependent on the relative proportions of *N*-acetyl-D-glucosamine and D-glucosamine residues, as well as on the MW (12). Typically, the term chitosan is used when the degree of deacetylation (DD) is above 70% and the term chitin is used when the extent of deacetylation is negligible or below 20% (8). Thus, chitosan polymers may present different degrees of deacetylation, MW, and viscosity which influence its properties. Generically, chitosan is characterized in terms of DD, in the range of 50–95%, and MW, which ranges from about 10 to 1,000 kDa (15). Both chitin and chitosan are structurally similar to hyaluronic acid, chondroitin sulfate, and heparin, which are biologically important mucopolysaccharides in all mammals. However, there are small

differences in the structure of chitin and chitosan, which affect also their properties when considering drug delivery (10).

Currently, chitosan is also receiving a great interest because of its properties regarding pharmaceutical and biomedical applications. Thus, chitosan is pointed out as being biodegradable, biocompatible, and mucoadhesive, and presents low toxicity and low antigenic potential (12). Furthermore, other chitosan properties have been reported, such as antimicrobial (14), antioxidant (15), antihypercholesterolemic (16), analgesic, and antitumoral (12) activity. The chitosan biological properties are exceptional and unique, promoting its use as a drug carrier such as in solutions, hydrogels, films, tablets, microparticles, and nanoparticles, particularly to enhance the permeability of mucosal barriers (17) and increase the effect in cell permeability (18). This polymer is also widely used as a supporting material for tissue engineering applications, cell culture, and nerve regeneration (19).

As referred above, the chitosan MW influences its biodegradability and consequently the drug release, as well its solubility. Thus, high MW chitosan, unlike low MW, is insoluble at physiological pH, which may restrict its *in vivo* use (20). The chitosan toxicological profile is also affected by MW, so cytotoxicity increases with higher MW chitosan (20). It is possible to change chitosan structure by altering its functional groups (NH_2 and OH), leading to the synthesis of derivates with distinct properties like increased mucoadhesion, solubility, and immunostimulation (8). Some examples of chitosan derivatives are the mono-*N*-carboxymethylchitosan (MCC) and quaternary derivatives such as *N*-trimethylchitosan chloride (TMC), dimethyl-ethylchitosan (DMEC), diethylmethylchitosan (DEMC), triethylchitosan (TEC), and *N*-(2-hydroxy)propyl-3-trimethylammoniumchitosan chloride (HTCC). Due to their positive charge, quaternary chitosan derivatives, unlike chitosan, are soluble in a wide range of pH, including neutral pH, allowing their use in various applications and different routes of administration (8).

Chitosan-based delivery systems can be useful in different routes of administration such as oral, nasal, ocular, parenteral and transdermal, particularly due to its utility in the transport of hydrophilic compounds such as peptides and proteins across epithelial barriers (12). The mechanism behind this permeation enhancing effect seems to be based on the positive charge of chitosan, which interacts with the cell membrane leading to a structural reorganization of tight junction-associated proteins (*zonula occludens* (ZO)) (21).

Nanoparticles have been widely studied in recent years as therapeutic proteins carriers with different degrees of effectiveness (22). These carrier systems can be obtained via different preparation protocols. Various *in vivo* and *in vitro* studies in different animal and cell models, showed the ability of chitosan-loaded nanoparticles

produced with water-soluble derivatives, to increase the oral availability of octreotide (23, 24), insulin, and buserelin (25, 26). Regarding insulin, the potential of chitosan nanoparticles to improve its systemic absorption, has been widely investigated. These nanoparticles showed to be a valuable carrier for the transport of insulin through the nasal mucosa, after nasal administration (10). Insulin can also be entrapped, with high efficiency, in different polyanion/chitosan nanoparticles systems (27). These complexes, seems to have good properties for oral protein delivery, with higher insulin association efficiency and retention in gastric simulated conditions (28). Furthermore, it was demonstrated that blood glucose levels of streptozocin-induced diabetic rats could be effectively controlled by insulin-loaded chitosan nanoparticles administration, and the hypoglycemic effect was prolonged for more than 24 h (29). Elcatonin-loaded chitosan-modified PLGA nanospheres administered by pulmonary route reduced the blood calcium levels to 80% of the initial calcium concentration and prolong elcatonin pharmacological effect (30), which can be attributed to the retention of nanospheres adhered to the lung tissue and bronchial mucus and to the sustained drug release properties at the adherence site. Moreover, the absorption of the drug can be enhanced by chitosan nanoparticles probably by the opening of the intercellular tight junction of the lung epithelium. Calcitonin-loaded chitosan nanoparticles showed a significant hypocalcemic effect when administered by pulmonary route, being a promising and safe carrier system for proteins delivery (31). Other chitosan-coated system loaded with salmon calcitonin showed an important ability to enhance intestinal uptake, as verified by a significant and long-lasting decrease in the blood calcemia levels in rats (32).

In this chapter, a protocol for the production of chitosan-based nanoparticles for therapeutic protein delivery, using insulin as model, is presented. Some methods for nanoparticles characterization are briefly discussed.

2. Materials

2.1. Reagents

1. Low G-content sodium alginate ($F_G=0.39$) (≈ 291 kDa) (Sigma-Aldrich, Oakville, Canada) (see Notes 1 and 2).
2. Medium G sodium alginate ($F_G=0.48$) (≈ 369 kDa) (Degussa, France) (see Notes 1 and 2).
3. High G sodium alginate Manugel DMB ($F_G=0.62$) (≈ 447 kDa) (ISP, Canada) (see Notes 1 and 2).
4. Oligochitosan (≈ 5 kDa), 85% deacetylated (Sigma-Aldrich, Oakville, Canada) (see Notes 3 and Note 4).

5. Low MW chitosan (\approx 50 kDa), 85% deacetylated (Sigma-Aldrich, Oakville, Canada) (see Notes 3 and 4).
6. Medium MW chitosan (\approx 500 kDa), 85% deacetylated (Natural Biopolymer, Raymond, WA, USA) (see Notes 3 and 4).
7. Human crystalline zinc-insulin, 7.0 mg per vial (Lilly Farma, Carnaxide, Portugal) (see Notes 5 and 6).
8. Calcium chloride and 18 mM calcium chloride solution.
9. Glacial acetic acid and 1% (v/v) acetic acid solution.
10. 0.1 M HCl solution.
11. 0.1 M NaOH solution.
12. Deionized water (house prepared).

2.2. Equipment

1. 50-ml, 100-ml, and 250-ml glass beakers.
2. 10-ml, 25-ml, and 150-ml graduated cylinders.
3. 25-ml screw-capped cylindrical bottles (7.0 cm height, 3.3 cm diameter).
4. 50-ml centrifuge tubes.
5. 5-ml semistoppered glass vials with slotted rubber closures.
6. Parafilm.
7. Millipore #2 paper filters (Millipore, Billerica, MA, USA).
8. 100–1,000 μ l semiautomatic pipette.
9. Magnetic stirrer hotplate.
10. Cylindrical magnetic stirrer bars (2.9 cm length, 0.6 cm section diameter).
11. Benchtop pH Meter.
12. Laboratory Refrigerator, 2–8°C.
13. Superspeed centrifuge, 20,000 $\times g$, 4°C.
14. Freeze dryer.
15. Varian HPLC with a Varian 9012 Gradient Solvent Delivery System and a Varian 9050 Variable Wavelength UV-VIS Detector (Varian, USA).
16. XTerra RP 18 column, 5 μ m particle size, 4.6 mm id \times 250 mm (Waters, USA).
17. LiChrospher 100 RP-18 guard column, 5 μ m particle size (Merck, Germany).
18. Malvern Zetasizer and Particle Analyzer 5000 (Malvern Instruments Worcestershire, UK).
19. DTS1060 disposable zeta cell (Malvern Instruments, Worcestershire, UK).
20. Scanning electron microscope.

21. Spectrum RXI FT-IR Spectrometer (Perkin Elmer, Massachusetts, USA).
22. Spectrum® v5.3.1 Software (Perkin Elmer, Massachusetts, USA).
23. Olis DSM 10 Spectrophotometer (Olis, Georgia, USA).

3. Methods

The alginate–chitosan nanoparticles are prepared by a two-step procedure based on the ionotropic gelation of polyanion with calcium chloride followed by polycationic cross-linking. Different types of chitosan and alginate can be used in this preparation (see Subheading 2.1).

3.1. Preparation of Stock Solutions

The indicated working concentrations of the alginate and chitosan solutions needed to prepare the nanoparticles (see Subheading 3.2), must be obtained from the following stock solutions by dilution with deionized water.

3.1.1. 2.0% (w/w) Alginate Solution, pH 4.9

1. Sprinkle 1 g of alginate slowly over 49 ml of deionized water under gentle magnetic stirring.
2. Cover the beaker with parafilm in order to avoid water evaporation and stir overnight.
3. Adjust the pH of alginate solution to 4.9 by using 0.1 M HCl solution.
4. Store at 4°C until use.

3.1.2. 1.0% (w/w) Chitosan Solution, pH 4.6

1. Sprinkle 0.5 g of chitosan slowly over 49.5 ml of 1% (v/v) acetic acid solution under gentle magnetic stirring.
2. Cover the beaker with parafilm in order to avoid water evaporation and stir overnight.
3. Adjust the pH of chitosan solution to 4.6 by using 0.1 M NaOH solution.
4. Filter by Millipore #2 paper filter and store at 4°C until use.

3.2. Preparation of Nanoparticles

The production protocol comprises the following steps (see Note 7):

1. Place 117.5 ml of the alginate solution (0.063%) into a beaker (see Note 8).
2. Add 7.0 mg of insulin (equivalent to 200 IU), achieving a final concentration of 0.005% (w/w), and mix under magnetic stirring (800 rpm) for 5 min (see Note 8).

3. Add dropwise 7.5 ml of 18 mM calcium chloride solution, under magnetic stirring (800 rpm) for 60 min to obtain an alginate pre-gel.
4. Add dropwise 25 ml of chitosan solution (0.07%) to the alginate pre-gel under magnetic stirring (800 rpm) for 90 min, giving a final alginate and chitosan concentration of 0.05% and 0.012% (w/w), respectively (alginate–chitosan mass ratio=4.3:1) (see Notes 8 and 9).
5. After chitosan addition, maintain the nanoparticles formed, with additional stirring (800 rpm) for 30 min (curing time).
6. Collect nanoparticles by centrifugation at $20,000 \times g$ for 45 min at 4°C.
7. Collect and reserve supernatant for insulin determination (see below for details).
8. Store the recovered nanoparticles at 2–8°C or freeze-dry (see Note 10) until further use.

3.3. Nanoparticles Characterization

3.3.1. Association Efficiency and Loading Capacity

Several techniques have been found useful for assessing drug-loaded nanocarrier properties, in particular those for protein and peptide delivery (33, 34). In this section, a brief list and discussion of critical characterization tests is provided.

The association efficiency (AE) is determined indirectly after nanoparticles separation from the aqueous medium containing nonassociated insulin. The amount of insulin associated with the particles is calculated by the difference between the total amount used to prepare the particles and the amount of insulin present in the aqueous phase after centrifugation.

The loading capacity (LC) is determined by the difference between the total insulin amount initially used to prepare the particles and the amount of residual unassociated insulin after particle separation as a percentage of total weight of nanoparticles.

The protocol to assess these parameters comprises the following steps:

1. Use the whole or diluted supernatant of the formulation, obtained after nanoparticle centrifugation (see Subheading 3.2).
2. Assess the amount of free insulin in supernatant by HPLC (see Note 11).
3. Calculate the total amount of associated protein by the following equation:

$$\text{AE\%} = \frac{\text{Total amount of insulin} - \text{Free insulin in supernatant}}{\text{Total amount of insulin}} \times 100$$

AE% is the association efficiency in percentage of insulin.

4. Calculate loading capacity percentage (LC%) as follows:
AE% is the association efficiency in percentage of insulin.

$$LC\% = \frac{\text{Total amount of insulin} - \text{Free insulin in supernatant}}{\text{Total weight of nanoparticles}} \times 100$$

The total weight of nanoparticles is assessed by freeze-drying an aliquot (approximately 50 mg of the hydrated pellet) of hydrated nanoparticles obtained after isolation.

3.3.2. Nanoparticles Size and Zeta Potential

1. Determine nanoparticles hydrodynamic radius and zeta potential by dynamic light scattering (DLS) using a Malvern Zetasizer and Particle Analyzer 5000 (Malvern Instruments, UK) (see Note 12).
2. Use whole or diluted nanoparticles samples in the determination (see Note 13).
3. To the nanoparticles size determination, measure samples in triplicate at 25°C with a detection angle of 90°.
4. To the determination of nanoparticles zeta potential, measure samples in triplicate at a scattering angle of 173° and 25°C, using previously activated and water-flushed DTS1060 disposable zeta cell.

3.3.3. Nanoparticles Morphology by Scanning Electronic Microscopy

1. Mount samples of nanoparticle dispersions on metal stubs using adhesive tape.
2. Nanoparticles must be gold coated under vacuum before observation.
3. Observe nanoparticles using a scanning electron microscope.
4. Figure 1 presents a Scanning Electronic Microscopy (SEM) microphotograph of Alginate–Chitosan nanoparticles obtained by the described protocol, evidencing spherical-like particles with smooth surfaces, typical of this type of systems (see Note 14).

3.3.4. In Vitro Release Profile of Insulin

The release profile of proteins from nanocarriers is assessed by simulating the natural gastrointestinal pH pathway, from the highly acidic pH of the stomach to the nearly neutral pH of the proximal intestine. Figure 2 provides an example of the release profile for alginate–chitosan nanoparticles containing insulin, as a model of peptidic drug. The following protocol allows obtaining biorelevant data concerning the oral administration of developed nanosystems:

1. Place 200 mg of nanoparticles collected after centrifugation into 25-ml screw-capped cylindrical bottles containing 20 ml hydrochloride acid (HCl) buffer at pH 1.2 (USP30-NF25) at 37°C ($\pm 1^\circ\text{C}$) (see Notes 15 and 16).
2. Incubate under magnetic stirring (100 rpm) for 4 h.

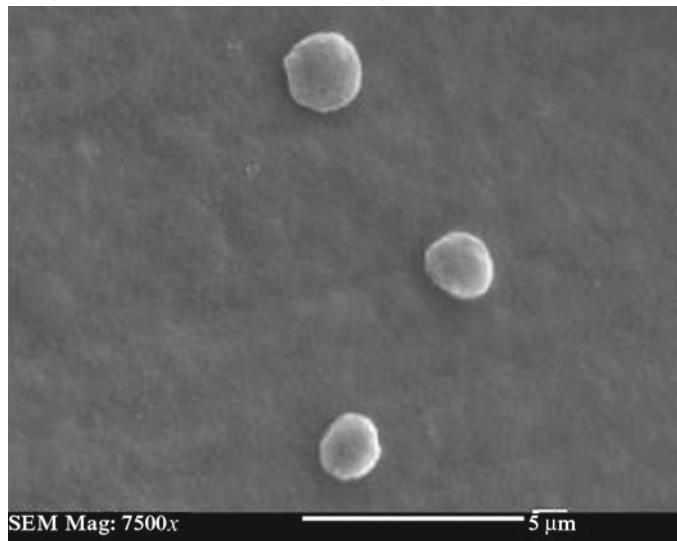


Fig. 1. SEM image of nanoparticles prepared with 0.05% (w/w) alginate and 0.01% (w/w) chitosan (reproduced from ref. 38 with permission from American Scientific Publishers).

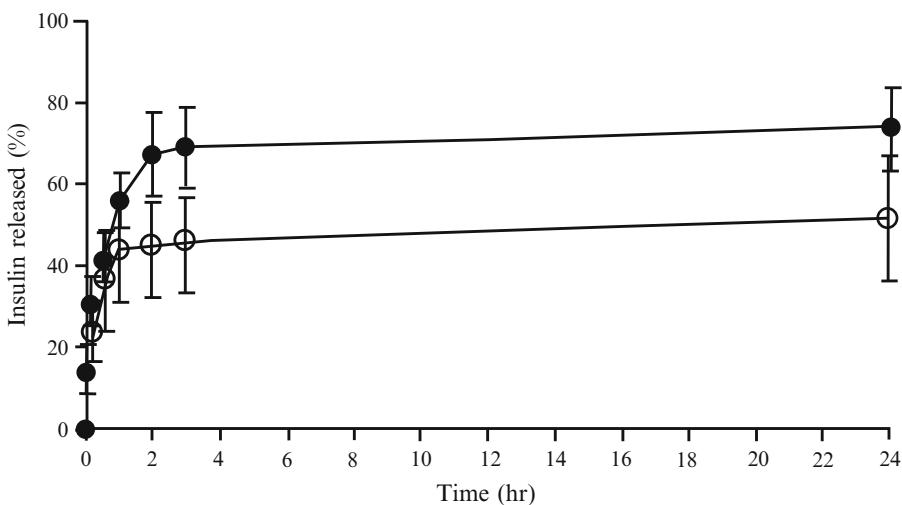


Fig. 2. Insulin release from insulin-loaded alginate–chitosan produced with alginate–chitosan mass ratio of 4.3:1 in gastric (*open symbols*) and intestinal (*close symbols*) pH simulated fluids at 37°C. The *vertical bars* represent standard derivations based on three replicates (reproduced from ref. 38 with permission from American Scientific Publishers).

3. At predetermined times, collect 0.4 ml samples and separate supernatant from nanoparticles by centrifugation ($20,000 \times g$ for 15 min, at 4°C).
4. Restore initial volume with 0.4 ml of fresh medium after each sample collection.
5. At 24 h, collect the total amount of medium and separate supernatant from nanoparticles by centrifugation ($20,000 \times g$ for 15 min, at 4°C).

6. Place 200 mg of nanoparticles collected after centrifugation into 25-ml screw-capped cylindrical bottles containing 20 ml phosphate buffer at pH 6.8 (USP30-NF25) (see Note 17).
7. Proceed with steps 2–5.
8. Repeat steps 1–7, two more times to have a good statistical correlation.
9. Determine drug content in the supernatant samples using HPLC (see Note 11).

3.3.5. Fourier Transform Infrared Spectroscopy

Information about proteins secondary structure can be provided by Fourier Transform Infrared (FTIR) spectroscopy (35). During the formulation process, the loss of the native protein structure can occur, leading to the loss of activity. Thus, the evaluation of protein arrangement and its preservation is crucial to ensure the biological activity. There are two types of proteins secondary structure: α -helices and β -sheets, which allow the amides to hydrogen-bond very efficiently with one another. The amide I absorption arise from the amide bonds that link the amino acids and is directly related to the backbone conformation with a major contribution from C=O stretching vibration and a minor contribution from the C–N stretching vibration. Studies with proteins of known structure have been used to systematically correlate the shape of the Amide I band, which occurs in the region 1,600–1,700 cm^{-1} , to the secondary structure content. Thus, the secondary structure content of proteins can be estimated by several numerical methods that increase the apparent resolution of the Amide I band. Curve fitting is the most widely used method for protein secondary structure quantification, mainly involving curve fitting of the amide I band (36). The basic principle of this procedure is to resolve the original protein spectrum into individual bands that fit the spectrum (35).

The main steps of the FTIR spectroscopy analysis for the assessment of insulin structural disposition are the following:

1. Collect the IR-spectra of encapsulated insulin using a Spectrum RXI FT-IR Spectrometer (see Note 18).
2. Collect also the IR-spectra of the unloaded nanoparticles and the water vapor spectra under identical conditions, and perform a double subtraction procedure of these last two spectra from the spectra collected in the previous step (see Note 18).
3. Obtain the second derivative spectra, using a seven-point Savitsky–Golay derivative function, and zap into the amide I region from 1,710 to 1,590 cm^{-1} (see Note 18).
4. Correct the baseline using a 3–4 point adjustment and area-normalize the spectra in that region (see Note 18).
5. Compare the spectra obtained in the previous step with an insulin standard, by area-overlapping both spectra. All samples

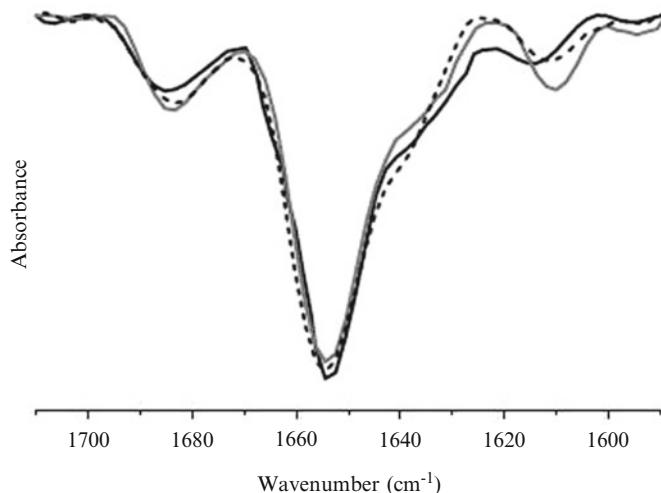


Fig. 3. Second-derivative FTIR spectrum of human insulin 10 mg/ml in 0.01 M HCl, pH 1.2 (black line), entrapped into fresh alginate–chitosan nanoparticles (dotted line), and entrapped into stored alginate–chitosan nanoparticles (gray line), (reproduced from ref. 39 with permission from Elsevier).

are run in triplicate and the data is presented as the average of three measurements (see Note 19).

Figure 3 provides an example of the second-derivative FTIR spectrum of an insulin standard and the protein entrapped into alginate–chitosan nanoparticles.

3.3.6. Circular Dichroism Spectroscopy

The circular dichroism (CD) spectroscopy is also used to examine the insulin structure and interactions between insulin and polymeric matrices after entrapment. The CD signals only arise where absorption of radiations occurs, so spectral bands are easily assigned to different structural features of a molecule, like a protein. Thus, in studies of proteins, the advantage of the CD is that complementary structural information can be obtained from a number of spectral regions. The most important chromophores, in proteins, include the peptide bond (absorption below 240 nm), aromatic amino acid side chains (absorption in the range 260–320 nm) and disulphide bonds (weak broad absorption bands centered around 260 nm) (37). Furthermore, if there is a close proximity between a number of chromophores of the same type, they can behave as a single absorbing unit, which will give rise to characteristic spectral features.

Induced CD signals can also arise from ligands that not having intrinsic chirality, they acquire chirality when bound in an asymmetric environment such as provided by a protein (37). Finally, the CD studies of proteins can give information essentially about the secondary structure composition of proteins, its tertiary

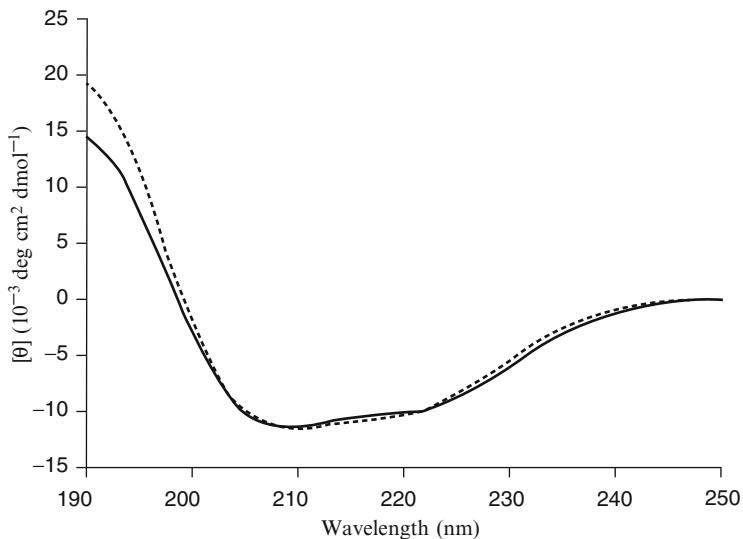


Fig. 4. Far-UV CD spectrum of human insulin in 0.01 M HCl solution (black line) and entrapped in alginate–chitosan nanoparticles prepared with alginate–chitosan mass ratio of 4.3:1 at pH 4.7 (dotted line), (reproduced from ref. 39 with permission from Elsevier).

structure fingerprint and, thus analyzing conformational changes in proteins.

The main steps of the CD spectroscopy analysis for the assessment of insulin structural disposition are the following:

1. Obtain the CD spectra at room temperature on the Olympos DSM 10 Spectrophotometer, using a 0.2 mg/ml insulin standard solution in 0.01 M HCl (see Note 20). Record the CD spectra, in the far-UV region, the CD in a 0.01 cm cell from 250 to 190 nm, using a step size of 0.5 nm, a bandwidth of 1.5 nm, and an averaging time of 5 s, with the lamp housing purged with nitrogen to remove oxygen.
2. Collect the spectra of the nanoparticles prepared (see Subheading 3.2) and subtract the spectra of the insulin-free nanoparticles (see Note 21). For all spectra, an average of five scans must be obtained.

Figure 4 provides an example of a CD spectrum of human insulin in solution and entrapped in alginate–chitosan nanoparticles. The molar ellipticity for insulin can be calculated as the CD signal \times MRW (115 Da, mean residual weight of each insulin residue) [insulin concentration (mg/ml) \times cell path length (0.1 mm)]. Finally, the insulin concentration can be determined by UV absorption at 276 nm using an extinction coefficient of 1.08 for 1.0 mg/ml.

4. Notes

1. Other suppliers of alginate may offer different MWs or even different types of alginate. Such variables may interfere on the final properties of nanoparticles and, thus, should be carefully evaluated.
2. The M/G alginate ratios are determined by NMR. The number average MW are obtained from the limiting viscosity number $[\mu] = 2.0 \times 10^{-5}$ MW using an Ubbelohde capillary viscometer (Canon Instruments, constants equals to $0.002564 \text{ mm}^2/\text{s}^2$), measuring the residence time of serially diluted alginate solutions containing 0.1 M sodium chloride on triplicate samples. The intrinsic viscosity is obtained by extrapolating the reduced viscosity to concentration zero.
3. Other suppliers of chitosan may offer different MWs and/or degrees of deacetylation. Such variables may interfere on the final properties of nanoparticles and thus should be carefully evaluated.
4. The presented MW and deacetylation degrees were provided by suppliers.
5. Although the described method is considered optimized for insulin from the supplier stated in the materials section, this protein is also available from other suppliers such as Sigma-Aldrich (Sintra, Portugal) and Novo Nordisk A/S (Denmark). Therefore, differences among products from different suppliers, namely insulin physical-chemical properties, other added excipients, and storage and handling conditions for optimal stability, should be taken into account and eventually justify minor adjustments to the described preparation procedure.
6. Insulin is used as a therapeutic protein model, so other proteins may be used, but the compatibility between the protein used and the carrier system must be known.
7. Before the preparation of nanoparticles, allow all solutions to attain room temperature ($20\text{--}25^\circ\text{C}$) if previously stored under refrigeration ($2\text{--}8^\circ\text{C}$).
8. These concentrations were deduced according to pre-gel viscometric investigations (38).
9. A colloidal dispersion at pH 4.7 is formed upon polycationic chitosan addition, being visible as the Tyndall effect.
10. Freeze-dry the obtained nanoparticles by dispensing the remaining pellet, with or without cryoprotectives (trehalose or sucrose corresponding to 10–30% of the total weight of nanoparticles), in 5-ml semistoppered glass vials with slotted rubber closures and add up to 2 ml of water, followed by freezing

during 24 h on the shelves of the lyophilization chamber at its minimum temperature (-85°C). Sublimation lasts 48 h at a vacuum pressure of 4×10^{-5} atm and without heating, being maintained at the condenser surface temperature of -60°C . Finally, glass vials are sealed under anhydrous conditions and stored until being rehydrated by using the same initial volume of water.

11. The amount of free insulin in supernatant is assessed by the HPLC system with the columns described (see Subheading 2.2.). The mobile phase is composed of acetonitrile (ACT) and 0.1% trifluoroacetic acid (TFA) aqueous solution operated in gradient mode at a flow rate of 1 ml/min. The protein identifications occur by UV detection at 214 nm. Change the gradient from 30:70 (ACT:TFA) to 40:60 in 5 min running following 5 min in isocratic 40:60 ratio. The method described is validated and linear in the range of 1–100 $\mu\text{g}/\text{ml}$ ($R^2=0.9996$) (38). Thus, you may run at least five standard samples in that range and your supernatant samples (diluted or not) and then interpolate the results to obtain the amount of insulin free in supernatant.
12. The determination of the hydrodynamic radius and zeta potential of nanoparticles can also be performed using similar equipments such as Zetasizer Nano ZS (Malvern Instruments Ltd, Worcestershire, UK) and ZetaPALS (Brookhaven Instruments Corp., Holtsville, NY, USA).
13. Sometimes, due to the solution viscosity or turbidity, it may be necessary to properly dilute the sample. Beware that the dilution used to measure the particle size, must be the same when measuring the zeta potential.
14. The nanoparticles must present unimodal size distribution in agreement with DLS analysis.
15. The in vitro release profile studies must be performed under sink conditions, maintaining a volume of dissolution media that is five to ten times greater than the volume at the saturation point of the drug contained in the drug delivery system being tested.
16. The temperature maintenance $37^{\circ}\text{C} (\pm 1^{\circ}\text{C})$, during the in vitro release profile studies is obtained using a thermostatic water bath.
17. Other buffers, such as acetate pH 4.5 or acetate pH 5.2, can be also used to mimic the physiological conditions encountered by the nanoparticles in the body, from the stomach to the proximal intestine.
18. This step is performed using the Spectrum[®] v5.3.1 Software, but similar and more recent software such as Grams/AI[®] 9.0 Software (Thermo Scientific, Thermo Scientific, Philadelphia,

PA, USA) and Horizon MB® Software (ABB, Zurich, Switzerland) can be also used.

19. This step can be performed using the same software of the previous note, or the data can be exported to other software such as OriginPro® 8 Software (OriginLab, Northampton, MA, USA) and Matlab® R2011a (MathWorks, Natick, MA, USA), that allow a more intuitive analysis of the spectra area-overlapping.
20. Normally, the CD spectrum of these solutions displays two minima at 209 and 222 nm (see Fig. 4), which is typical of predominant α -helix structure of proteins (39).
21. It is also possible to collect the spectra of the protein released in the buffers of the in vitro release profile assay, and analyze possible modifications of the protein structure. In this case, no subtraction process is needed because insulin is free in the buffer.

Acknowledgments

The authors acknowledge the financial support from Fundação para a Ciência e a Tecnologia, Portugal (PTDC/SAU-FCF/104492/2008).

References

1. Frokjær S, Otzen DE (2005) Protein drug stability: a formulation challenge. *Nat Rev Drug Discov* 4:298–306
2. Antosova Z, Mackova M, Kral V, Macek T (2009) Therapeutic application of peptides and proteins: parenteral forever? *Trends Biotechnol* 27:628–635
3. Lien S, Lowman HB (2003) Therapeutic peptides. *Trends Biotechnol* 21:556–562
4. Mahmood I (2008) Methods to determine pharmacokinetic profiles of therapeutic proteins. *Drug Discov Today Technol* 5:e65–e69
5. Cryan S-A (2005) Carrier-based strategies for targeting protein and peptide drugs to the lungs. *AAPS J* 7:E20–E41
6. Lee HJ (2002) Protein drug oral delivery: the recent progress. *Arch Pharm Res* 25:572–584
7. Morishita M, Peppas NA (2006) Is the oral route possible for peptide and protein drug delivery? *Drug Discov Today* 11:905–910
8. Amidi M, Mastrobattista E, Jiskoot W, Hennink WE (2010) Chitosan-based delivery systems for protein therapeutics and antigens. *Adv Drug Deliv Rev* 62:59–82
9. Shih C-M, Shieh Y-T, Twu Y-K (2009) Preparation and characterization of cellulose/chitosan blend films. *Carbohydr Polym* 78:169–174
10. Sinha VR, Singla AK, Wadhawan S, Kaushik R, Kumria R, Bansal K, Dhawan S (2004) Chitosan microspheres as a potential carrier for drugs. *Int J Pharm* 274:1–33
11. George M, Abraham TE (2006) Polyionic hydrocolloids for the intestinal delivery of protein drugs: alginate and chitosan—a review. *J Control Rel* 114:1–14
12. Rinaudo M (2006) Chitin and chitosan: properties and applications. *Prog Polym Sci* 31:603–632
13. Alonso MJ, Sánchez A (2003) The potential of chitosan in ocular drug delivery. *J Pharm Pharmacol* 55:1451–1463
14. Fernandez-Saiz P, Lagaron JM, Ocio MJ (2009) Optimization of the film-forming and storage conditions of chitosan as an antimicrobial agent. *J Agric Food Chem* 57:3298–3307
15. Tomida H, Fujii T, Furutani N, Michihara A, Yasufuku T, Akasaki K, Maruyama T, Otagiri M,

- Gebicki JM, Anraku M (2009) Antioxidant properties of some different molecular weight chitosans. *Carbohydr Polym* 344:1690–1696
16. Sashiwa H, Aiba SI (2004) Chemically modified chitin and chitosan as biomaterials. *Prog Polym Sci (Oxford)* 29:887–908
 17. Fonte P, Nogueira T, Gehm C, Ferreira D, Sarmento B (2011) Chitosan-coated solid lipid nanoparticles enhance the oral absorption of insulin. *Drug Deliv Translational Res* 1(4): 299–308
 18. Shi Y, Huang G (2009) Recent developments of biodegradable and biocompatible materials based micro/nanoparticles for delivering macromolecular therapeutics. *Crit Rev Ther Drug Carrier Syst* 26:29–84
 19. Zhao QS, Ji QX, Xing K, Li XY, Liu CS, Chen XG (2009) Preparation and characteristics of novel porous hydrogel films based on chitosan and glycerophosphate. *Carbohydr Polym* 76: 410–416
 20. Huang X, Du Y-Z, Yuan H, Hu F-Q (2009) Preparation and pharmacodynamics of low-molecular-weight chitosan nanoparticles containing insulin. *Carbohydr Polym* 76:368–373
 21. Andrade F, Antunes F, Nascimento AV, da Silva SB, das Neves J, Ferreira D, Sarmento B (2011) Chitosan formulations as carriers for therapeutic proteins. *Curr Drug Discov Technol* 8(3):157–172
 22. Gan Q, Wang T (2007) Chitosan nanoparticle as protein delivery carrier—systematic examination of fabrication conditions for efficient loading and release. *Colloids Surf B Biointerfaces* 59:24–34
 23. Thanou M, Verhoef JC, Marbach P, Junginger HE (2000) Intestinal absorption of octreotide: N-trimethyl chitosan chloride (TMC) ameliorates the permeability and absorption properties of the somatostatin analogue in vitro and in vivo. *J Pharm Sci* 89:951–957
 24. Thanou M, Verhoef JC, Verheijden JH, Junginger HE (2001) Intestinal absorption of octreotide using trimethyl chitosan chloride: studies in pigs. *Pharm Res* 18:823–828
 25. Van Der Merwe SM, Verhoef JC, Verheijden JHM, Kotzé AF, Junginger HE (2004) Trimethylated chitosan as polymeric absorption enhancer for improved peroral delivery of peptide drugs. *Eur J Pharm Biopharm* 58:225–235
 26. Qian F, Cui F, Ding J, Tang C, Yin C (2006) Chitosan graft copolymer nanoparticles for oral protein drug delivery: preparation and characterization. *Biomacromolecules* 7:2722–2727
 27. Sarmento B, Martins S, Ribeiro A, Veiga F, Neufeld R, Ferreira D (2006) Development and comparison of different nanoparticulate polyelectrolyte complexes as insulin carriers. *Int J Pept Res Ther* 12:131–138
 28. Sarmento B, Ribeiro A, Veiga F, Sampaio P, Neufeld R, Ferreira D (2007) Alginate/chitosan nanoparticles are effective for oral insulin delivery. *Pharm Res* 24:2198–2206
 29. Sarmento B, Ribeiro A, Veiga F, Ferreira D, Neufeld R (2007) Oral bioavailability of insulin contained in polysaccharide nanoparticles. *Biomacromolecules* 8:3054–3060
 30. Yamamoto H, Kuno Y, Sugimoto S, Takeuchi H, Kawashima Y (2005) Surface-modified PLGA nanosphere with chitosan improved pulmonary delivery of calcitonin by mucoadhesion and opening of the intercellular tight junctions. *J Control Release* 102:373–381
 31. Makhlof A, Werle M, Tozuka Y, Takeuchi H (2010) Nanoparticles of glycol chitosan and its thiolated derivative significantly improved the pulmonary delivery of calcitonin. *Int J Pharm* 397:92–95
 32. Prego C, García M, Torres D, Alonso MJ (2005) Transmucosal macromolecular drug delivery. *J Control Release* 101:151–162
 33. Hall JB, Dobrovolskaia MA, Patri AK, McNeil SE (2007) Characterization of nanoparticles for therapeutics. *Nanomedicine (Lond)* 2: 789–803
 34. Peltonen L, Hirvonen J (2008) Physicochemical characterization of nano- and microparticles. *Curr Nanosci* 4:101–107
 35. Van de Weert M, Hering JA, Haris PI (2005) Fourier transform infrared spectroscopy. In: Jiskoot W, Crommelin D (eds) *Methods for structural analysis of protein pharmaceuticals*. AAPS, Arlington, VA, pp 131–166
 36. Jørgensen L, Vermehren C, Bjerregaard S, Froekjaer S (2003) Secondary structure alterations in insulin and growth hormone water-in-oil emulsions. *Int J Pharm* 254:7–10
 37. Kelly SM, Jess TJ, Price NC (2005) How to study proteins by circular dichroism. *Biochim Biophys Acta (BBA)—Proteins Proteomics* 1751:119–139
 38. Sarmento B, Ribeiro AJ, Veiga F, Ferreira DC, Neufeld RJ (2007) Insulin-loaded nanoparticles are prepared by alginate ionotropic pre-gelation followed by chitosan polyelectrolyte complexation. *J Nanosci Nanotechnol* 7: 2833–2841
 39. Sarmento B, Ferreira D, Jørgensen L, van de Weert M (2007) Probing insulin's secondary structure after entrapment into alginate/chitosan nanoparticles. *Eur J Pharm Biopharm* 65: 10–17

Chapter 29

Challenges in the Development and Manufacturing of Antibody–Drug Conjugates

Laurent Ducry

Abstract

Advances in antibody–drug conjugates (ADCs) will permit sensitive discrimination between healthy and cancer cells. Promising clinical results generated much hope that this targeted prodrug therapy will offer more effective treatment options to patients. Manufacturing such highly potent biopharmaceuticals presents a series of unique challenges. Some specific skills required for the process development and production of ADCs are discussed. In addition to the accuracy and reliability needed to handle these potent and costly materials, coworker safety and equipment cleaning are of particular importance. The ideas and concepts shared in this article are based on the experience that Lonza has gained in the ADC field since 2004.

Key words: Antibody–drug conjugate, ADC, Bioconjugate, Conjugation, Cytotoxic, Process development, Manufacturing

1. Introduction

Antibody–drug conjugates (ADCs) are monoclonal antibodies (mAbs) linked to highly potent cytotoxic agents through a linker moiety ([1–6](#)). By combining the unique targeting of mAbs with the cell killing ability of cytotoxic drugs, ADCs allow sensitive discrimination between healthy and diseased tissues together with high cytotoxic activity. The concept of using a mAb to deliver a cytotoxic drug to cancer cells is over 50 years old ([7](#)), but the development of such a “magic bullet” has been paved with difficulties. It was not until the 1980s that non-immunogenic mAbs could be developed. Nowadays, the availability of humanized mAbs directed against antigens that are less frequently expressed on normal tissues makes them ideal targeting agents. In parallel, advances in conjugation technologies led by ImmunoGen

and Seattle Genetics have yielded linkers that are sufficiently stable in the bloodstream but nonetheless allow release of cytotoxic payloads inside of cancer cells (8). These efforts have resulted in a number of ADCs being investigated in clinical trials, with trastuzumab emtansine (T-DM1; Genentech/Roche) for treatment of HER2-positive metastatic breast cancer (9, 10), brentuximab vedotin (SGN-35; Seattle Genetics)* for CD30-positive Hodgkin's lymphoma (11, 12), and inotuzumab ozogamicin (CMC-544; Pfizer) for CD22-positive non-Hodgkin lymphoma (13, 14) as the most advanced examples. The promising Phase II results of these leading products should result in regulatory approval over the next 1–2 years, validating the ADC concept and creating the need for reliable commercial supply.

2. Process Development

More than for any other active pharmaceutical ingredient (API) class, the quality of the ADCs is controlled by the manufacturing process. Although the complexity of the analytical method as well as the overall effort to analyze ADCs are higher than those for most drugs, the insight gained is not as wide. This is the result of the high molecular weight of such biological drugs, and the fact that ADCs are heterogeneous mixture of compounds with varying number of drugs per antibody and multiple conjugation sites. In order to reproduce a heterogeneous subset of compounds, a well-developed process is essential. Meticulous process development requires experienced and well-trained workers, together with the suitable lab equipment (Fig. 1). Manufacturing plants and QC labs typically follow cGMP rules, whereas for R&D laboratories no regulation ensures that the accuracy of lab instruments is checked and documented. For ADCs even more than for any type of API, the implementation of “good scientific practice” rules for lab instruments such as thermometers, pipettes, pH electrodes, and balances is a logical requirement to generate reliable data. In early development stages, a Design of Experiments (DoE) approach is typically used to determine the impact of single variables as well as interactions of multiple variables. In order to be able to run many experiments with little material, tiny jacketed vessels which allow running conjugation experiments at milligram scale while accurately controlling the temperature proved very useful. This reaction model was shown to be predictive of reaction performance up to several hundred liters. For the purification process, however, gram scale runs are needed to assess the tangential flow filtration (TFF) conditions for removal of process-related contaminants.

*The FDA has granted accelerated approval of Adcetris™ (brentuximab vedotin) in August 2011.



Fig. 1. ADC process development lab (©Lonza Ltd).

Process development shall first ensure that the desired drug-to-antibody ratio (DAR) is achieved. This key quality attribute is generally controlled during the modification reaction (attachment of the linker to the mAb in the case of lysine conjugation) or during the interchain disulfide reduction (cysteine conjugation). Investigation and control of many process parameters are nonetheless needed in order to ensure process consistency. Mild conditions, low sheer forces, as well short process and hold times are also important in order to prevent degradation. The conjugation process may indeed destabilize the antibody, and addition of hydrophobic drugs on the mAb surface increases their susceptibility to aggregate formation (15). Selection of appropriate processing, storage, and handling conditions during manufacturing shall be based on the physicochemical stability of the intermediates as well the ADC.

3. Manufacturing of ADCs

Perhaps the greatest challenges in ADC manufacturing are design, construction, and operation of a biological manufacturing environment that allows safe manipulation of highly potent cytotoxic drugs. Since the conjugation process is performed in aqueous biological buffers which are capable of supporting growth of a wide range of microorganisms, aseptic conditions and clean utilities must be present to support manufacturing effort. Bioburden



Fig. 2. Isolator next to cGMP conjugation suite (©Lonza Ltd).

reduction steps can be designed in the process, for example through sterile filtration. However, efficient and cost-effective removal of endotoxins formed from cell membrane of Gram-negative bacteria and liberated during cell growth, cell division, or cell death is challenging (16). Since the presence of small amounts of endotoxins in injectables activates the immune system of the host causing side effects (like fever, endotoxin shock, tissue injury, and even death), the maximum level is set to five endotoxin units (EUs) per kilogram of body weight per hour. An aseptic manufacturing environment with area classification is thus needed to prevent bioburden growth and maintain a low endotoxin level. A biological manufacturing environment is typically designed to reduce microbial contamination, but not to handle highly potent cytotoxic compound with occupational exposure levels (OELs) in the 10^{-9} g/m³ range. A plant dedicated to ADC manufacturing (PCP for Potent Compound Production) was designed and built by Lonza in Visp, Switzerland. This new plant allows the safe handling of highly potent compounds in a biopharmaceutical environment. Emphasis was placed on engineering controls, more specifically on containment and/or capture of potential contaminants at the source:

- Handling of solid cytotoxic compounds in isolators (Fig. 2).
- Work with solutions containing cytotoxic material in closed vessels and material transfers via closed systems.
- Entry into the processing rooms via air locks (separated material and coworker flows).
- Maintenance of desired pressure relationships between the rooms; the entry air lock typically operates as a low-pressure sink to avoid contamination not only from the workroom to the outer area (e.g., cytotoxic contamination), but also from the outer area to the workroom (e.g., bioburden or cross-contamination).

- Air filtration over HEPA filters.
- Training programs and awareness programs in order to ensure proper use of the equipment and adherence to safety concepts.
- Personal protective equipment program to cover residual risks not eliminated by engineering controls, together with an emergency response program for the case of spill of potent compounds.

Before the first ADC project, the effectiveness of our safety measures was tested through a surrogate monitoring study. Having demonstrated that safety limits were not exceeded, the work with toxins whose OEL level is at 40 ng/m³ or below has been initiated. Regular hygiene monitoring is nonetheless needed to ensure that a sufficient safety standard is kept over time.

Equipment cleaning is another key activity when manufacturing highly potent cytotoxic drugs. Single-use technologies can be a convenient way to avoid or at least reduce equipment cleaning. Disposable systems are certainly very useful for preparing buffer solutions and many single-use small parts like filters and membranes are typically used. However, single-use materials are generally expensive and leachable and extractable data are needed, which can be critical for the part of the conjugation process where an organic co-solvent like *N,N*-dimethylacetamide or DMSO is used. Additionally, for the most critical operations, they hardly offer sufficient containment. Glass or stainless steel vessels are thus generally preferred for the conjugation reaction so as to offer optimal safety to the operators (Fig. 3). Dedicating vessels to a drug substance may be a valid strategy in early-phase programs in order to simplify equipment cleaning, but is not deemed economical at commercial scale where larger, automated vessels are normally used. An attractive compromise may be to dedicate equipment pieces to a particular class of toxin (e.g., Maytansine or Auristatin derivatives), considering that the most stringent cleaning requirements come from the cytotoxic drug rather than from the targeting agent.

In order to tackle the cleaning challenge as early as possible and to collect sufficient data and experience, Lonza has chosen to operate multipurpose suites for early-phase projects. However, it must be mentioned that if our ADC plants are designed as multipurpose plants, they are used solely for manufacturing bioconjugates and are dedicated to one therapeutic area, namely oncology. Nonetheless, an effective and validated cleaning process is needed. The facilities and equipment were designed with the ability to steam in place (SIP) and clean in place (CIP). For the proteinic part of the ADC, degradation by an aqueous caustic solution followed by rinse with purified water down to the lowest possible limits, namely those of water for injection (WFI), is applied (total



Fig. 3. 200 L stainless steel vessel (©Lonza Ltd).

organic carbon (TOC), conductivity, pH, endotoxin, and bioburden). An additional concern is obviously the ability to clean process equipment between manufacturing campaigns when different toxins are used in consecutive projects, where the toxin cannot be readily inactivated. In that case, an extra cleaning limit based on a Maximum Allowable Carryover (MAC) calculation is set for the cytotoxic drug between different ADC campaigns. This limit is typically very low, in the $\mu\text{g}/\text{L}$ (rinse) or $\mu\text{g}/\text{m}^2$ (Swab) range, and detection of residual cytotoxin by methods such as TOC and HPLC-based methods do not provide sufficient sensitivity. In order to support product changeover, validated ELISA or LC-MS/MS assays are typically needed.

4. Analytical Considerations

Analysis and characterization of ADCs are challenging due to their heterogeneous nature as well as large molecular size. Routine lot release requires a reliable control of the critical quality attributes. The following assays are most commonly used.

- The purity is generally analyzed by size-exclusion chromatography (SEC) followed by UV detection or possibly multi-angle light scattering (MALLS). The extent of fragmentation and, most importantly, aggregation is quite essential for biopharmaceuticals bearing relatively hydrophobic drugs. Sedimentation

velocity analytical ultracentrifugation (SV-AUC) and dynamic light scattering (DLS) are alternative analytical techniques for the determination of the protein aggregation.

- Subvisible particles can be analyzed by the light obscuration particle count test, and visible particles by flow cell microscopy or microscopy imaging analysis (FCM).
- Determination of the DAR by UV spectrometry when the mAb and payload have sufficiently different absorption profiles. In the case of cysteine conjugation, hydrophobic interaction chromatography (HIC) allows determination of species distribution in addition to average DAR. This method is, however, not applicable to ADCs obtained from lysine conjugation since the higher heterogeneity complicates the chromatographic separation.
- Charged-based separation assays such as capillary electrophoresis (CE), ion-exchange chromatography (IEC), isoelectric focusing gel electrophoresis (IEF), capillary isoelectric focusing (cIEF), and imaged capillary isoelectric focusing (icIEF) provide information on charge heterogeneity and thus on the drug distribution pattern.
- Sodium-dodecyl-sulfate-poly-acrylamide gel electrophoresis (SDS-PAGE) or capillary electrophoresis sodium-dodecyl-sulfate (CE-SDS) analyses of reduced and non-reduced ADCs under denaturing conditions allow molecular weight separation of protein fragments, indicating on which chain the payloads are located.
- Binding assays, like enzyme-linked immunosorbent assay (ELISA) and fluorescence-activated cell-sorting (FACS) assays, to assess antigen binding.
- Cell-based assays to measure target-dependent cytotoxicity.
- RP-HPLC is widely used to quantify residual, unconjugated drug and drug-related impurities. Alternatively, ELISA assays and MS-based approaches can be used for this purpose.
- Determination of process-related impurities from the conjugation reaction, such as residual co-solvent, by GC or HPLC techniques.
- Finally, endotoxin and bioburden testing are necessary to demonstrate that aseptic conditions have been maintained throughout the production chain. Microbial evaluation is necessary to assess if the lot can be safely used as an injectable.

For stability testing as well as ADC characterization, these tests can be complemented by additional analytical assays. Peptide mapping coupled with various detection techniques, MS analysis of drug distribution, and thermal analysis methods

such as differential scanning calorimetry (DSC) allow studying the effect of conjugation on the conformational stability of the mAb (14, 17, 18).

5. Conclusion

Regulatory approval of the ADCs currently in advanced clinical phases will most likely open a new era for cancer treatment, stimulating new ADC research programs and generating the need for commercial manufacturing. The successful development of a reproducible conjugation process and of the associated analytical assays, as well as technical transfer and process validation at scale, requires specific experience and assets. Both biopharmaceutical and chemical know-how is needed in order to manufacture ADCs. Currently, only a few companies are operating cGMP conjugation suites for that purpose. This is the result of the complex technical requirements, as well as the relatively small although promising ADC market.

References

- Wu AM, Senter PD (2005) Arming antibodies: prospects and challenges for immunoconjugates. *Nat Biotechnol* 23:1137–1146
- Ricart AD, Tolcher AW (2007) Technology insight: cytotoxic drug immunoconjugates for cancer therapy. *Nat Clin Pract Oncol* 4:245–255
- Carter PJ, Senter PD (2008) Antibody-drug conjugates for cancer therapy. *Cancer J* 14:154–169
- Kratz F, Müller IA, Ryppa C, Warnecke A (2008) Prodrug strategies in anticancer chemotherapy. *ChemMedChem* 3:20–53
- Senter PD (2009) Potent antibody drug conjugates for cancer therapy. *Curr Opin Chem Biol* 13:235–244
- Alley SC, Okeley NM, Senter PD (2010) Antibody-drug conjugates: targeted drug delivery for cancer. *Curr Opin Chem Biol* 14:529–537
- Mathé G, Loc TB, Bernhard J (1958) Effet sur la leucémie 1210 de la souris d'une combinaison par diazotation d'A-méthopérine et de γ-globulines de hamsters porteurs de cette leucémie par hétérogreffé. *C R Hebd Séances Acad Sci* 246(10):1626–1628
- Ducry L, Stump B (2010) Antibody-drug conjugates: linking cytotoxic payloads to monoclonal antibodies. *Bioconjug Chem* 21:5–13
- Lewis Phillips GD, Li G, Dugger DL, Crocker LM, Parsons KL, Mai E, Blättler WA, Lambert JM, Chari RVJ, Lutz RJ, Wong WLT, Jacobson FS, Koeppen H, Schwall RH, Kenkare-Mitra SR, Spencer SD, Sliwkowski MX (2008) Targeting HER2-positive breast cancer with Trastuzumab-DM1, an antibody-cytotoxic drug conjugate. *Cancer Res* 68:9280–9290
- Krop IE, Beeram M, Modi S, Jones SF, Holden SN, Yu W, Girish S, Tibbitts J, Yi J-H, Sliwkowski MX, Jacobson F, Lutzker SG, Burris HA (2010) Phase I study of Trastuzumab-DM1, an HER2 antibody-drug conjugate, given every 3 weeks to patients with HER2-positive metastatic breast cancer. *J Clin Oncol* 28:2698–2704
- Bartlett N, Forero-Torres A, Rosenblatt J, Fanale M, Horning SJ, Thompson S, Sievers EL, Kennedy DA (2009) Complete remissions with weekly dosing of SGN-35, a novel antibody-drug conjugate (ADC) targeting CD30, in phase I dose-escalation study in patients with relapsed or refractory Hodgkin lymphoma (HL) or systemic anaplastic large cell lymphoma (sALCL). *J Clin Oncol* 27:8500, Abstract
- Okeley NM, Miyamoto JB, Zhang X, Sanderson RJ, Benjamin DR, Sievers EL, Senter PD, Alley SC (2010) Intracellular activation of SGN-35, a potent anti-CD30

- antibody-drug conjugate. *Clin Cancer Res* 16:888–897
- 13. DiJoseph JF, Dougher MM, Kalyandrug LB, Armellino DC, Boghaert ER, Hamann PR, Moran JK, Damle NK (2006) Antitumor efficacy of a combination of CMC-544 (inotuzumab ozogamicin), a CD22-targeted cytotoxic immunoconjugate of calicheamicin, and rituximab against non-Hodgkin's B-cell lymphoma. *Clin Cancer Res* 12:242–249
 - 14. DiJoseph JF, Khandke K, Dougher MM, Evans DY, Armellino DC, Hamann PR, Damle NK (2008) CMC-544 (inotuzumab ozogamicin): a CD22-targeted immunoconjugate of calicheamicin. *Hematology Meeting Reports* 5:74–77
 - 15. Wakankar AA, Feeney MB, Rivera J, Chen Y, Kim M, Sharma VK, Wang YJ (2010) Physicochemical stability of the antibody-drug conjugate Trastuzumab-DM1: changes due to modification and conjugation processes. *Bioconjug Chem* 21:1588–1595
 - 16. Magalhães PO, Lopes AM, Mazzola PG, Rangel-Yagui C, Penna TCV, Pessoa A Jr (2007) Methods of endotoxin removal from biological preparations: a review. *J Pharm Sci* 10:388–404
 - 17. Wakankar A, Chen Y, Gokarn Y, Jacobson FS (2011) Analytical methods for physicochemical characterization of antibody drug conjugates. *mAbs* 3:164–175
 - 18. Stephan JP, Kozak KR, Wong WLT (2011) Challenges in developing bioanalytical assays for characterization of antibody–drug conjugates. *Bioanalysis* 3:677–700

INDEX

A

- ABD. *See* Albumin binding domain (ABD)
ABD molecular imaging 119
ADC. *See* Antibody-drug conjugate (ADC)
A* discrete search algorithm 128, 129, 131, 134, 139–141
Affibody molecules 103–122
Affinity
 ligands 104
 maturation 28, 44, 74, 122, 127, 128
Aggregation models 414
Aggregation-prone regions (APRs) 430, 434–446
Agrobacterium-mediated transformation 241
Albumin 1, 5, 15, 32, 76, 118, 121–122
Albumin binding domain (ABD) 118, 119, 121–122
Amyloid fibril 456–458
Antibody
 derivatives 157, 158, 171
 domain 13, 14, 21, 73–83, 89
 effector function 12–16
 engineering 73
 fragments 27, 63, 65, 86, 158
 framework 17, 33
 libraries 13, 27–40, 86
 synthetic 27–40
Antibody-drug conjugate (ADC) 11, 489–496
Antigen 7, 12, 14, 17, 20, 21, 27–31, 33, 39, 64, 65, 67, 71, 74, 78–82, 85, 86, 106, 148, 155, 204, 206, 207, 214, 265, 434, 444, 474, 489, 495
APRs. *See* Aggregation-prone regions (APRs)
Arabidopsis thaliana 239–262
Asp-Asp motifs 366
Automation 129, 211, 233, 235, 299–301, 316, 335, 340, 347, 419, 493

B

- Bibodies 157–175
Bioconjugate 493
Biodistribution 104–105, 110–111, 115, 117–119, 121
Bioinformatics 45, 57, 326, 333, 342, 466
Biomanufacturing 227–228
Biopharmaceuticals 11, 121–122, 277–278, 288, 326, 351–363, 427, 453, 492, 494, 496

- Biosimilar therapeutic proteins 22, 326
Biotherapeutics 227–236, 343, 425–427, 429, 433, 438–439, 442–443, 446
Bivalent 106

C

- Cancer
 breast 6, 18, 19, 490
 cells 11, 20, 105, 110, 489, 490
 colorectal 6, 7, 19
 diagnosis 105
 prostate 19
Capto™ adhere 316–318, 321, 322
Carbohydrates 86, 325, 327–335, 337, 343–347
Carbohydrate structure database (CCSD) 328, 330, 332, 335
CCSD. *See* Carbohydrate structure database (CCSD)
CDRs. *See* Complementarity determining regions (CDRs)
Cell line development 228–231, 234, 236
CH2 14, 74, 76, 85–101
Chelator 104, 108, 110–116, 118–121, 472
Chinese hamster ovary (CHO) cells 203, 205–214, 222–225, 230, 233, 236, 281
Chitosan 211, 471–486
Chitosan-based delivery nanoparticles 471–486
Chromatography 153–154, 161, 170, 171, 191, 255, 265–274, 279, 281, 286–289, 291, 295, 296, 298, 308–311, 317–319, 335, 340, 353, 366, 369, 370, 374, 403–422, 494
CID. *See* Collision-induced dissociation (CID)
Collision-induced dissociation (CID) 352, 353, 359, 363, 370, 371, 373
Complementarity determining regions (CDRs) 17, 28–30, 33–35, 39, 444, 445
Computational
 methods 425–446
 prediction 425–446
 protein design 128, 430
Computer simulations 453
Conjugation 11, 12, 15, 16, 23, 32, 76, 78, 79, 106, 107, 110, 114, 115, 118, 119, 121, 161, 172, 206–208, 325, 489, 490, 491, 493, 495
Coulter counter 385, 393–395
Cytotoxic 5, 7, 12, 19, 20, 146, 212, 489–495

D

- Database 240, 245, 251, 277–291, 297, 310, 324–347, 370, 393
 Dead-end elimination 127–143
 Diagnostic 5, 103–122, 127, 300–301, 410–411, 419
 Diethylmethylamine 352
 Diversity 21, 27–31, 33, 38–40, 45, 57, 63, 71, 74, 81–82, 127–143, 228, 295, 429
 Downstream purification 28, 194–195, 258, 286–288
 Drug
 circulation time 107, 128
 potency 11, 18, 23, 119, 128, 426–427, 471–472, 474–475, 489, 493
 protein-based 5, 16, 23, 127, 326, 425
 Dual-specific targeting 107, 118, 145–156, 472, 489
 Dual variable domain immunoglobulin (DVD-Ig) 145–156
 DVD-Ig. *See* Dual variable domain immunoglobulin (DVD-Ig)
 Dynamic imaging 384, 400–401

E

- Efficacy 5, 16–22, 44, 73, 81, 122, 146–148, 156, 326, 453
 Electrostatics 129, 131, 136–138, 140, 142, 408, 418, 432, 433, 446, 455, 458
 Electrozone sensing 381, 385, 386, 389, 393
 Endotoxin 265–274, 492, 494, 495
 Epitope 21, 86, 122, 446
Escherichia coli 28, 32, 33, 35–40, 49, 57, 62, 63, 68, 69, 71, 87, 150, 155, 159, 164, 166, 168, 169, 177–184, 187–199, 242, 246, 250, 257, 260, 266, 268

F

- FACS. *See* Fluorescence-activated cell sorting (FACS)
 Fast protein liquid chromatography (FPLC) 87, 91, 268–273, 412
 FDPB. *See* Finite-difference Poisson–Boltzmann (FDPB)
 Finite-difference Poisson–Boltzmann (FDPB) 129, 136–140, 142
 Fleximers 130, 131, 133, 141
 Flow imaging 381, 384–385, 389, 392, 397, 400
 Fluorescence-activated cell sorting (FACS) 74, 76, 79, 80, 82, 209, 495
 FPLC. *See* Fast protein liquid chromatography (FPLC)
 Framework 12, 17, 28, 29, 33, 278, 343, 403
 Free energy of folding 129, 130, 132, 137–139, 142, 143

G

- Genome 43, 45–46, 50–52, 204, 214, 230–232, 236, 240, 280–281, 327–328, 345–346
 Glycoinformatics 325–330, 341–347
 Glycoinformatic tools 328–329
 Glycomics 325, 327, 344–347
 Glycosylation 21, 86, 103, 155, 213, 227–228, 230, 258, 293–295, 315–316, 320–321, 326, 341

H

- Half-life 12, 14, 15, 19, 20, 105, 111, 117–119, 121, 122, 146, 351
 HEK293 cells 153, 159, 166, 203–212, 230
 HIAC instrument 382
 High pressure liquid chromatography (HPLC) 214, 295, 297, 298, 301–303, 305, 308, 309, 310, 318, 322, 327–330, 333, 335, 336, 339, 341, 342, 353, 354, 358, 361, 362, 366, 368, 370, 373, 374, 410, 412, 441, 476, 478, 481, 485, 494, 495
 High-throughput 13, 38, 69, 211, 233, 234, 293–312, 315–322, 330, 341–344
 analysis 321, 322
 purification 315–322
 HILIC. *See* Hydrophilic interaction liquid chromatography (HILIC)
 HPLC. *See* High pressure liquid chromatography (HPLC)
 Hydrophilic interaction liquid chromatography (HILIC) 295, 296, 302, 309, 310

I

- Imaging 104–119, 122, 234, 384–385, 397, 495
 Immune diseases 3, 17, 20, 146
 Immunogenicity 5, 12, 17–18, 21, 28, 122, 145, 147–148, 157, 230, 265, 326, 352, 427, 489
 Immunoglobulin 8–9, 11, 28, 74, 104–105, 116–117, 146, 306, 439
 Immunotherapy 20, 145–146
 Infection 6–7, 36, 40, 58, 256, 262
 Inhibitor 8, 10, 12, 19–20, 42–59, 63, 68, 191, 212, 233, 255, 472
 Injectables 379, 382–383, 385, 397–400, 492, 495
 Insertase 189
 In-source fragmentation (ISF) 352–353, 357, 359–360, 362
 Insulin 3–5, 8, 12, 16, 18, 160, 212, 241–242, 248, 250, 255, 257–259, 472, 475–486
 Isomerization 21, 337–338, 365–375

L

- Large-scale production 184, 221, 223, 225, 235, 239, 240
 LC/MS. *See* Liquid chromatography/mass spectrometry (LC/MS)
 Library
 antibody 13, 27–40, 86
 construction 14, 28, 30, 35, 36, 39, 40, 43–59, 62, 66, 74–76, 80, 86–88, 91, 92
 diversity 28, 29, 33, 38–40, 45, 57, 63, 71, 82
 peptide 28, 43–59
 phage 14, 36–38, 40, 44, 74, 96
 Light obscuration 380–385, 388–389, 391, 396, 398, 400, 495
 Light scattering 380, 403–422, 479, 494, 495
 Limited mutagenesis 91

- Linker 11, 119, 146–151, 158, 162, 164, 166, 167, 173, 174, 259, 357, 360, 489–491
- Lipofection 206, 208, 209, 214, 230
- Lipoglycan 265
- Lipopolysaccharides (LPS) 265, 266
- Liquid chromatography / mass spectrometry (LC/MS) 44, 99, 110, 155, 268, 305, 319, 329, 338, 344, 351–363, 365–375, 494
- Log reduction values 279
- LPS. *See* Lipopolysaccharides (LPS)
- M**
- MACS. *See* Magnetic-activated cell sorting (MACS)
- Magnetic-activated cell sorting (MACS) 74, 76, 78–79, 82
- Manufacturing 14, 38, 50, 52, 70, 77, 80, 95, 145, 146, 152, 171
- MazF 177–179, 182–184
- Melting temperture (Tm) 104, 117, 122, 444
- Membrane microscopy 290, 381, 383–384, 389, 391–392, 396, 400
- Membrane protein 63, 107, 178–179, 183, 186–200, 274, 287, 381
- Microflow imaging 385
- Molecular chaperone 189, 190
- Molecular dynamics 443, 458, 467
- Molecular modeling 129, 436
- Monoclonal antibodies (mAbs) 2, 6–7, 80, 85, 145–147, 157, 212, 221, 224, 278, 279, 281, 282, 284, 286, 289, 291, 317, 326, 416, 442, 472, 489
- Monoclonal antibody production 221, 224, 278, 289
- Multimodal strong anion exchange chromatography 318, 319
- Multivalent 158
- N**
- Nanoparticle characterization 475, 478–483
- Nanoparticles 387, 389, 471–486
- Nanosight 381, 387–388
- N-glycan analysis 293–313, 318
- N-glycosylation 316
- NMR protein structure. *See* Nuclear magnetic resonance protein structure
- Non-affinity protein purification 317–318
- Nuclear magnetic resonance (NMR) protein structure 455, 459–460
- O**
- ¹⁸O incorporation 369–375
- Oligosaccharides 265, 295, 331, 333, 337, 344, 345
- Optical microscopy 381, 383–384, 389, 391–392
- P**
- Parenterals 379, 382–384, 398, 399, 474
- PAT. *See* Process analytical technology (PAT)
- PCR. *See* Polymerase chain reaction (PCR)
- PEG. *See* Polyethylene glycol (PEG)
- PEGylation 118, 351–353, 357, 359, 361
- PEGylation site mapping 352
- Phage display 2, 27–29, 32, 33, 36–40, 44, 45, 50, 56, 74, 81, 86, 104, 121
- Phagemid 28, 29, 33, 38–40, 45, 47–49, 56, 86–90, 94, 96, 100
- Pharmacodynamics 351
- Pharmacokinetics 17, 122, 148, 156, 227, 472
- Phylomer 43–59
- Pichia pastoris* 157–175, 315–322
- Polyethylene glycol (PEG) 15, 32, 33, 38, 49, 56, 58, 351–353, 357, 359–363
- Polyethylenimine 149, 208
- Polymerase chain reaction (PCR) 5, 31, 50–54, 57–59, 62–71, 74–77, 81, 83, 87–89, 91–95, 97, 98, 149–153, 155, 158, 159, 163–168, 174, 193, 194, 196, 197, 240, 248, 251, 252, 256, 262, 279, 281, 321
- Post-column addition 352–354, 356, 357, 362
- Process analytical technology (PAT) 242, 294
- Process development 21, 228–230, 234, 240, 250, 278, 289, 490–491
- Protein administration 471
- administration 16, 472, 473
- aggregation 86, 155, 403–422, 425–446, 453–468, 494–495
- design 15, 127–143, 147, 430
- engineering 5, 15, 20, 21, 73, 86, 104, 105, 111, 117, 118, 121, 147, 148, 158, 187–199, 205, 206, 230, 239–240, 316, 326
- stability 428, 430–434, 472
- Purification 322
- Q**
- QbD. *See* Quality by design (QbD)
- Quality by design (QbD) 291, 294, 427, 446
- R**
- Rabbit reticulocytes 63, 66, 67, 70
- Radionuclides 5, 15, 105–110, 116, 119–121
- Radiotherapy 118–121
- Recombinant protein expression 239
- Ribosome display 61–71
- Robotics 299, 305, 306, 318, 322
- Rotamers 128–142
- S**
- Safety 5, 11, 16–18, 21, 22, 117, 122, 230, 270, 278, 288, 289, 291, 326, 427, 446, 493
- Scaffold protein 5, 15, 23, 63, 86, 104, 116
- Screening 27, 43–45, 56, 57, 61–71, 131, 142, 155, 164, 196, 221, 228, 231, 234–236, 248, 250, 255, 316, 317, 321, 345, 396, 433, 463

- SEC. *See* Size exclusion chromatography (SEC)
- Side effects 15, 16, 18, 118, 190, 472, 492
- Signal recognition particle (SRP) 188–192, 197, 199
- Single-chain antibody 63, 65, 213, 257
- Single particle tracking 379
- Single protein production (SPP) system 177–180, 183, 184
- Site-specific modification 104
- Size exclusion chromatography (SEC) 171, 255, 403–422, 444, 494
- Solubility 14, 122, 147, 199, 258, 267–269, 271, 285, 290, 352, 369, 406, 407, 411, 418–420, 426, 432–434, 438, 474
- SPP system. *See* Single protein production (SPP) system
- SRP. *See* Signal recognition particle (SRP)
- Stable expression 158, 166, 211, 227–236
- Stable transfection pool 221–225
- Subvisible particulates 379–401, 495
- Succinimide 365, 367, 369
- Synthetic antibodies 27–40
- Synthetic libraries 21, 27–40
- T**
- Tandem mass spectrometry 353, 365–375
- Targeted therapy 20, 105, 117, 119
- Therapeutic protein 1–23, 61–71, 239–262, 277–291, 379–401, 425–446, 471–486
- Toxins 2, 4, 7, 9–11, 104, 118, 493, 494
- Transgenic seed 241
- Transient expression 150, 151, 153, 155, 159–160, 166, 203–215
- Transient transfection 203–205, 208, 211, 213–215, 221
- Tribodies 158, 161–172, 174
- Triethylamine (TEA) 352
- Trigger factor 189, 190
- Triton X-114 266, 267
- Tumor targeting 106, 113, 114, 117–119
- U**
- Uptake 28, 107, 109–116, 119, 205, 208, 210, 305, 472, 475
- V**
- Vaccines 5, 19–20, 63, 278, 326
- Variable domain 28, 30, 86, 146–148, 150–151, 155
- VH 14, 85–101, 146, 150, 151, 164, 174, 175
- Viral clearance 277–291
- Virus
- filtration 278, 279
 - inactivation 278–281, 284, 285, 289, 290
 - removal 278–284, 286–289, 291
 - safety 230, 278, 288, 291
 - validation 278
- W**
- WAX HPL. *See* Weak anion exchange (WAX) HPLC
- Weak anion exchange (WAX) HPLC 297, 298, 302–304
- X**
- Xenograft 109–111, 113, 115, 117, 119–121
- Y**
- Yeast display 44, 73–83
- Yeast expression 158, 161, 166–169, 174
- YidC 188–192, 197
- Z**
- Z-domain 104, 105, 112, 116, 117