Use of Toxicological Information in Drug Design

Edwin J. Matthews, R. Daniel Benz, and Joseph F. Contrera

U.S. Food and Drug Administration (FDA), Center for Drug Evaluation and Research,
5600 Fishers Lane, Rockville, Maryland 20857. Tel 301-827-5180, Fax 301-827-3787.
matthewse@cder.fda.gov; contrerajf@cder.fda.gov

Abstract

This paper is an extension of the keynote address and another talk at the Symposium on the Use of Toxiciological Information in Drug Design. The symposium was organized by American Chemical Society's Chemical Information Division at the 220th National Meeting of the American Chemical Society in Washington, DC, August 20-24, 2000. We outline an approach for meeting the scientific information needs of the U.S. Food and Drug Administration (FDA). Ready access to scientific information is critical to support safety-related regulatory decisions and is especially valuable in situations where available experimental information from in vivo/in vitro studies are inadequate or unavailable. This approach also has applications for lead selection in drug discovery.

A pilot electronic toxicology/safety knowledge base and computational toxicology initiative is underway in the FDA Center for Drug Evaluation and Research (CDER) that may be a prototype for an FDA knowledge base. The objectives of this effort are: (i) to strengthen and broaden the scientific basis of regulatory decisions, (ii) to provide the Agency with an electronic scientific institutional memory, (iii) to create a scientific resource for regulatory and applied research, and (iv) to establish an internal Web-based support service that can provide decision support information for regulators that will facilitate the review process and improve consistency and uniformity. An essential component of this scientific knowledge base is the creation of a comprehensive electronic inventory of CDER-regulated substances that permit identification of clusters of substances having similar chemical, pharmacological or toxicological activities, and molecular structure/substructures. Furthermore, the inventory acts as a pointer and link to other databases and critical non-clinical and clinical pharmacology/toxicology studies and reviews in FDA archives. Clusters of related substances are identified through the use of: (i) an extensive index of alternative names for each substance, (ii) a molecular structure key field consisting of a rudimentary or core structure represented as an ISIS™.mol-file, (iii) global search terms (molecular group, chemical class, clinical indication, or pharmacologic activity), and (iv) molecular clustering using structure/sub-structure similarity indices.

The information contained in a toxicology knowledge database has limited value unless means are available to extract information, identify relationships, and create and test hypotheses. One such means is computational toxicology, also called in silico toxicology, ComTox, or e-TOX. Computational toxicology is the application of computer technology and information processing (informatics) to analyze, model, and estimate chemical toxicity based upon structure activity relationships (SAR). A computational toxicology software package, MCASE, has been evaluated and successfully improved by CDER through the incorporation of data from FDA archives and concomitant alterations of the logic used in the interpretation of the results to reflect the data analysis and hazard identification practices and priorities of the Center. Our modifications and uses of the MCASE program are discussed in detail.

Abbreviations

ADME (absorption, distribution, metabolism, excretion, and bioavailability); ADR (FDA/CDER adverse drug reaction database); AI (artificial intelligence software program); ARES (FDA/CDER adverse event reporting system database); CAS (Chemical Abstract Service registry); CDER (Center for Drug Evaluation and Research, the Center); CDER:SI (CDER substance inventory); CFSAN (Center for Food Safety and Applied Nutrition); CIR (Cosmetic Ingredient

Journal of Molecular Graphics and Modelling
© 2000 by Elsevier Science Inc.
655 Avenue of the Americas, New York, NY 10010
1093-3263/00/$–see front matter

News and Views

INDEX

News and Views Editor
Shauna Farr-Jones, Ph.D.
915 Cole St. P.M. Box 375
San Francisco, CA 94117-4315
shauna_farrjones@yahoo.com

ELSEVIER

Review); CPD (L. Gold Carcinogenicity Potency Database); CRADA (cooperative research and development agreement); e-ADME (in silico ADME); EPA (U.S. Environmental Protection Agency); ES (human expert system); e-TOX (in silico toxicology); FDA (The U.S. Food and Drug Administration; the Agency); FDAMA (FDA Modernization Act); FCS (food contact substance); GLP (Good Laboratory Practices); GRAS (generally recognized as safe); IAG (government interagency agreement); ICH (International Conference on Harmonization); IND (investigational new drug application); IUPAC (International Union of Pure and Applied Chemistry); MCASE (multiple computer automated structure evaluation program); MCASE-ES (multiple computer automated structure evaluation-expert system); MRD (maximum recommended therapeutic dose); MTA (material transfer agreement); MTD (maximum tolerated dose); MW (molecular weight); NCI (National Cancer Institute); NIDA (National Institute of Drug Abuse); NIEHS (National Institute of Environmental Health Science); NIH (National Institutes of Health); NOEL (no-effect level); OCAC (FDA/CFSAN/Office of Colors and Cosmetics); NTP (National Toxicology Program); OPA (FDA/CFSAN/Office of Premarket Approval); OTR (FDA/CDER Office of Testing and Research); NDA (new drug application); PharmTox (pharmacology and toxicology review); PMN (Premanufacture Notification Program); QSAR (quantitative structure activity relationships); QSBR (quantitative structure bioactivity relationships); RRAS (Regulatory Research and Analysis Staff); SAR (structure activity relationships); SDF (ISIS structure data ".sdf file"); SI (ISIS similarity index); SOPs (standard operating procedures); and TSCA (Toxics Substances Control Act).

Introduction

The FDA is a science and information based agency. A major goal of the FDA and FDA's Center for Drug Evaluation and Research (CDER) is to strengthen the scientific basis of regulatory decisions and at the same time find effective methods for improving and expediting the regulatory review process while using fewer resources.

This article outlines the progress of a pilot program of the Regulatory Research and Analysis Staff (RRAS) of FDA/CDER's Office of Testing and Research (OTR) to develop the Center's electronic toxicology database and computational toxicology (e-TOX and e-ADME) capability. The objectives of this program are to provide access to a resource of scientific information and information applications for regulatory decision support and applied research. A CDER Intranet-based computational toxicology consultant service was created to facilitate the application of e-TOX and e-ADME decision support information to FDA regulatory and research scientists to strengthen the scientific basis of regulatory decisions.

Computational toxicology, also called in silico toxicology, ComTox, or e-TOX, is the application of computer technology and information processing (informatics) to analyze, model, and estimate chemical toxicity based upon structure activity relationships (SAR).[1-4] These SARs can be developed for either: (i) individual/discrete toxicological endpoints (e.g., carcinogenicity in rats, mutagenicity in Salmonella typhimurium, etc.) or (ii) estimating an effective dose related to a toxicological endpoint (e.g., maximum tolerated dose in a two-year carcinogenicity study, LD50 in an acute toxicity test, etc.). The toxicological data sets analyzed for SAR must be compiled using data generated in standardized in vivo/in vitro assays, and the data must be subjected to uniform assay evaluation criteria. The relationships between chemical structure and toxicity can be described either qualitatively (e.g., toxic/non-toxic, active/inactive, etc.), or quantitatively (QSAR) in terms of toxicological (or biological) potency (e.g., potent/weak, 80/30 CASE units, etc.).

Commercial e-TOX software programs can be divided into two different approaches for estimating chemical toxicity: (i) human expert/rule-based methods and (ii) statistical/correlative methods. The expert/rule systems (e.g., DEREK [LHASA, UK] and ONCOLOGIC [LogiChem, Inc.]) establish SAR based upon prior knowledge of the activities of single chemicals or groups of chemicals with similar structures (i.e., congeneric chemicals), and consideration of known (or hypothetical) chemical and biological mechanisms of action. The ONCOLOGIC and DEREK programs have two major differences. DEREK contains rules for predicting chemical toxicity for a panel of toxicological endpoints and can estimate chemical metabolites (METEOR); ONCOLOGIC only contains rules for estimating carcinogenic potential. In contrast, statistical/correlative systems (e.g., TOPKAT and MCASE) rely upon specialized algorithms to establish SAR or QSAR based on existing data from large and heterogeneous groups of chemicals (i.e., noncongeneric chemicals). MCASE and TOPKAT have a large number of differences, including: operational differences (utility operating system platform, data input formats, data import/export functions, batch mode capability, metabolite prediction, etc.) and theoretical differences (method for development of the QSAR equations; fragment theory, modulators, expert system Wizard, artificial intelligence, etc.). The predictive performance of all of these systems and programs is limited by: (i) the size of the toxicological endpoint data set in the knowledge database, (ii) the degree to which the diverse chemical structure universe is represented, (iii) the ability of the program's operating system logic to duplicate/reflect FDA toxicity report evaluation standards, and (iv) the measure to which our understanding of chemical toxicity mechanisms of action are reflected in the program's output.

Until now, the FDA has not routinely used commercially available e-TOX programs in part because they did not meet the needs of the Agency. In limited testing in our laboratory carried out in the mid-1990s, commercially available e-TOX programs were evaluated and were found to exhibit unreliable predictions of toxicity, poor coverage for FDA-regulated substances, and inflexible and often incorrect logic with respect to FDA hazard identification practices. In addition, to our knowledge, no attempts had been made to directly model the effects of chemicals in humans using human study data. For regulatory use, e-TOX programs should: (i) provide reliable estimates of potential chemical toxicity for a wide range of toxicological endpoints, (ii) exhibit comprehensive coverage/representation for the diversity of FDA-regulated substances, (iii) possess flexible logic that can reflect differences/changes in regulatory strategy (e.g., hazard/risk/benefit scenarios, variations in weight of evidence schemata, etc.), and (iv) use data from clinical studies to model specific toxicological and pharmacological effects of substances in humans.

In order for e-TOX programs to be considered for regulatory uses by the FDA, data from FDA files must be incorporated into the software and program logic must be adjusted to reflect FDA toxicology evaluation policies, strategies, and priorities. However, non-clinical and clinical study data for FDA-regulated substances are not readily available for a variety of reasons including proprietary issues. Most studies submitted to the agency are contained in large volumes of paper files that make data retrieval and extraction difficult, although electronic submissions will improve this situation in the near future. Retrieval of information from regulatory reviews to meet the needs of Agency scientists, or Freedom of Information (FOI) requests from the public, currently is a lengthy process of physically retrieving the documents, and/or screening of microfiche, when available. The substances in FDA files are coded and identified by Agency admission/tracking/accounting numbers (e.g., FDA Center login number/code, investigational new drug application (IND) numbers, etc.) and by the industrial trade names/codes. The identity of these substances is often not linked to common chemical names, or to Chemical Abstract Service (CAS) registry numbers. Substances with identical active ingredient except for different salts/esters/formulations generally have entirely separate and unconnected administrative/tracking histories within the Agency. Thus, it is a very difficult task to compile toxicological and pharmacological data from FDA archives for a single substance or a cluster of structurally related (congeneric) substances.

Toxicology Knowledge Database Program

The CDER Substance Inventory

OTR/RRAS is constructing a comprehensive electronic inventory of the Agency's regulated substances to permit identification of clusters of substances having similar regulatory status, pharmacological/toxicological activities, and molecular structures/substructures.[5] This inventory is a critical first step in the development of a Center and Agency toxicology knowledge database. It serves as both a pointer to relevant non-clinical and clinical pharmacology/toxicology reviews and studies in Agency electronic and paper archives and as a powerful search engine to identify substances with related structure and/or function. For pragmatic considerations related to the cost in time and resources, it was necessary to restrict this substance inventory's capabilities to these two functions at this time.

OTR/RRAS has elected to create ORACLE tables containing critical fields of information in the substance inventory and to use molecular structure as the key field to retrieve information quickly and efficiently. The inventory also has been designed to identify clusters of related substances based upon the use of a large index of alternative substance names, a molecular structure key field, global search terms, and molecular clusters using structure/sub-structure similarity indices.

Name Index: The inventory has a name index that includes a comprehensive list of alternative identifiers for substances that includes: (i) common/generic names; (ii) synonyms and industrial trade names/codes; (iii) IUPAC names; (iv) Agency regulatory/tracking codes and numbers (e.g., Investigational New Drug Application [IND], New Drug Application [NDA], and National Drug Code [NDC], Bureau of Drug [BD], and Center for Food Safety and Applied Nutrition (CFSAN) food additive document number, etc.); and (v) internationally recognized code numbers (e.g., Chemical Abstract Service [CAS] registry numbers, National Cancer Institute [NCI/NSC] chemotherapeutic registry numbers, etc.).

Global Search Terms: The inventory also links each substance to three categories of global search term categories, including: substance molecular group (organic, polymer, protein, salt, etc.), regulatory function (therapeutic clinical indication, pharmaceutical aid, food additive, etc.), and pharmacologic activity (enzyme antagonist, etc.).

Key Field: The key field in the inventory is the molecular structure of the substance represented as the ISIS ".mol-files" and graphic (".skc") files prepared using the ISIS/DRAW software program (MDL Information Systems, Inc.). The individual substance molecular structure data sets are linked together as ISIS structure-data-files (".sdf files"). In order to maximize the clustering of structurally similar chemicals, we employ the .mol-files for organic chemicals without consideration of salts, stereo chemistry, atomic charges, etc.

Similarity Inde: OTR/RRAS uses two different software programs to identify structurally similar substances in the substance inventory. ISIS/HOST is used to identify congeneric clusters of substances based upon similarity index estimates for the entire molecular structure. MCASE is used to identify clusters of substances that share submolecular fragments that are highly correlated with toxicological activities (see below).

CDER Toxicology Databases

Databases are essential for identifying and summarizing the results of large numbers of studies and organizing information into a concise usable form. Databases are also valuable for identifying information gaps that can be used to set research priorities. Toxicology databases can facilitate regulatory review and improve regulatory consistency by making available background information on chemically or pharmacologically related compounds. There are six major types of toxicology studies for most pharmaceutical and food products:

1. Genotoxicity
2. Acute toxicity
3. Chronic toxicity
4. Reproductive toxicity
5. Developmental toxicity
6. Carcinogenicity

Due partly to similar requirements of regulatory agencies, the design of toxicity studies is relatively consistent, which facilitates data extraction and database development. In the case of pharmaceuticals, studies initiated after the 1978 Good Laboratory Practices (GLP) rules represent a good source of quality toxicology studies. The ultimate goal of the CDER toxicology database effort is to create databases for all of these categories of studies.

The CDER rodent carcinogenicity database was the first CDER toxicology database.[6-7] This database was initially created to meet the needs and priorities of the International Conference on Harmonization (ICH) of Technical Requirements for the Registration of Pharmaceuticals for Human Use. This database supported a redefinition of dose selection criteria, carcinogenicity test methods, and study evaluation procedures by the ICH. The ICH guidances defined multiple acceptable endpoints for dose selection, acceptable test methods, and an integrative "weight of evidence" approach to evaluating the relevance of study results. The value of CDER toxicology and carcinogenicity databases has been demonstrated by the major role they played in the formulation of five ICH safety guidances for human pharmaceuticals that have now been adopted and implemented world wide and are available on the FDA/CDER website (http://www.fda.gov/cder/guidance).

ICH S1A: The Need for Long Term Rodent Carcinogenicity Studies of Pharmaceuticals

ICH S1B: Testing for the Carcinogenic Potential of Pharmaceuticals

ICH S1C: Dose Selection for Carcinogenicity Studies of Pharmaceuticals

ICH S1CR: Use of Limit Dose in Dose Selection for Carcinogenicity Studies

ICH S4; S4B: Duration of Chronic Toxicity Testing in Animals

OTR/RRAS is continuing to expand and update the rodent carcinogenicity database and to construct additional electronic toxicology databases for the Center. For practical reasons of cost in time and resources, and the complexity of the process of data extraction and entry, it was necessary to limit the number of the database fields. Our efforts were focused on identifying a minimum number of data fields necessary to adequately cross-reference information in the Agency's/Center's toxicology/pharmacology files and to construct database modules for e-TOX programs (see below). We also want to identify an optimal field structure to link toxicology/pharmacology data from non-clinical and clinical studies. These databases are being constructed in ORACLE tables and linked to the Center's substance inventory and the ISIS/HOST search engine.

FDA Toxicology Databases; Creating an FDA Knowledge Database

The FDA is a science- and information-based organization and a repository of scientific data on a vast array of chemicals and pharmaceuticals. The toxicology and clinical data for pharmaceuticals in FDA files is an extremely valuable scientific resource. With the major advances in computer and information technology, this resource can be more effectively used to improve the scientific basis of regulatory decisions and product development while still protecting proprietary information. Better means need to be developed to link separate data sources and extract meaningful knowledge from large data sets. A knowledge database is the combination of databases and computational methods to discover meaningful relationships to better manage information. A pilot project is underway to create such a system by combining existing FDA scientific databases and information resources (see Figure 1).

As part of the FDA toxicology knowledge database pilot project, we are developing procedures to establish links from the CDER substance inventory and ISIS/HOST search engine to non-OTR/RRAS toxicology/pharmacology databases within the Agency. As a feasibility experiment we are now planning to link the Priority-based Assessment of Food Additives (PAFA) toxicology database that is maintained by FDA's CFSAN to the CDER inventory and toxicology databases. The PAFA database was selected because it contains important toxicological study data for many substances that are regulated by the Agency as food additives, excipients, and inert ingredients in pharmaceutical preparations, and as ingredients in cosmetics. To establish this linkage, an alternative name index, a molecular structure key field, global search terms for direct food additives and food contact substances was obtained and is being added to the CDER inventory. These data will be added as an ISIS/HOST ".sdf file" for direct food additives and food contact substances, and this file will be resident in the inventory and searchable using ISIS/HOST. If the toxicological and administrative data in the PAFA database were converted from its current M-204 software platform to ORACLE tables, these toxicology data could also be subject to ISIS/ BASE searches.

Intranet/Internet

The CDER substance inventory and toxicology databases have been set up on an Intranet Web site that is currently only accessible to Center scientists. The use of the ISIS/HOST search engine is now restricted based upon the current license agreement with MDL Information Systems, Inc. For this substance inventory and/or toxicology databases to become publicly available, these data will be redacted, and a non-proprietary version of the information will be prepared.

Computational Toxicology Program

The US Environmental Protection Agency (EPA) has a long history of relying upon SAR/QSAR to screen new chemicals for potential adverse health effects under the Premanufacture Notification (PMN) program requirement of the Toxics Substances Control Act (TSCA).[8] In conjunction with estimated production and human exposure data, the SAR/QSAR decision support information for the EPA. PMN substances may be used to limit chemical production, to require further testing, or to justify taking no regulatory action. In contrast, the FDA has only recently begun to con-

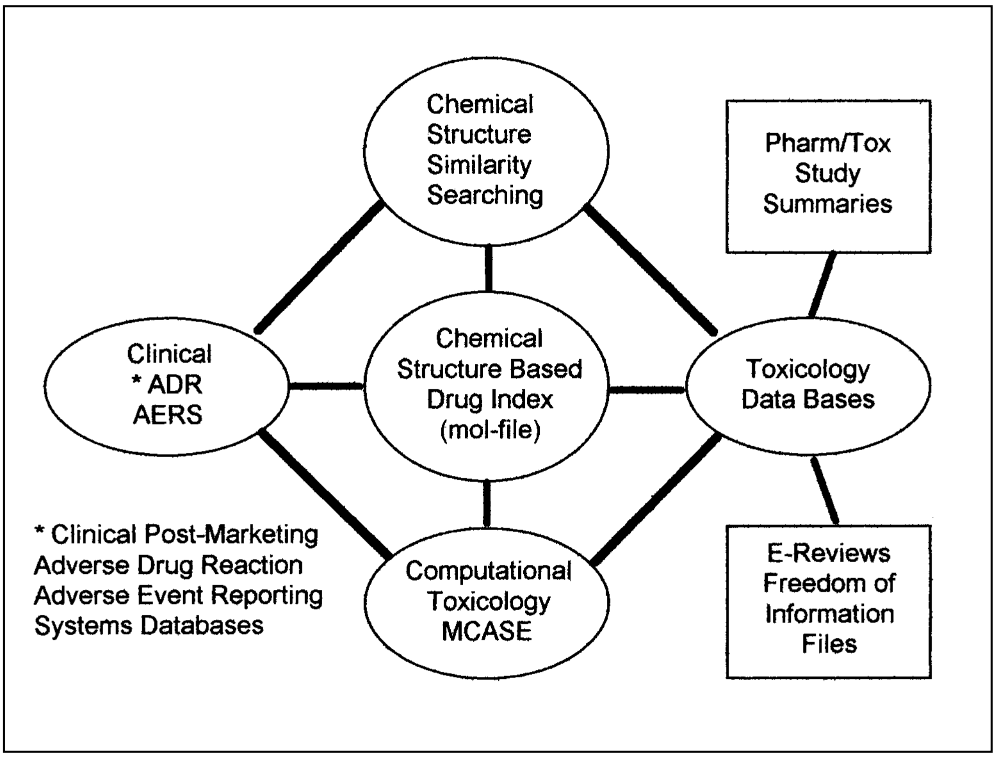
Chemical Structure Similarity Searching
Pharm/Tox Study Summaries
Clinical * ADR AERS
Chemical Structure Based Drug Index (mol-file)
Toxicology Data Bases
* Clinical Post-Marketing Adverse Drug Reaction Adverse Event Reporting Systems Databases
Computational Toxicology MCASE
E-Reviews Freedom of Information Files

Figure 1. FDA-CDER Knowledge Base for Decision Support and Discovery

sider SAR/QSAR decision support information for regulatory decisions.

OTR/RRAS began a computational toxicology program in 1994 with the mission of testing and evaluating e-TOX software programs for decision support applications in situations in which toxicology data are unavailable or limited.[9,10] When none of the then available SAR programs met the needs of the Agency, we investigated the deficiencies of the software and procedures for modifying existing programs to improve their predictive performance. Within the Center, primary regulatory application for these programs was anticipated to be the estimation of the potential toxicological activities of contaminants and degredants in pharmaceutical preparations. The following summarizes our experience with one of these programs.

Development of MCASE-ES

The experimental methods we used to develop the new MCASE-ES software program to estimate the trans-gender and trans-species carcinogenicity of chemicals in rodents have been described in detail.[9-12] Modules for other toxicological endpoints have since been developed using similar methods. CASE, CASETOX and MCASE were developed by Dr. Gilles Klopman at Case Western Reserve University beginning in about 1985. Our investigations use both the CASETOX and MCASE programs, database modules constructed by Dr. Herbert Rosenkranz at the University of Pittsburg, and database modules constructed by OTR/RRAS under a CRADA agreement between FDA/CDER/OTR and Multicase, Inc.

Briefly, the MCASE program reduces the SMILEs codes of organic chemicals to all possible 2 to 10 consecutive atom molecular fragments. The program then compares the fragments of active and inactive molecules, and through molecular subtraction identifies those fragments that are only associated with active/positive molecules (MCASE biophores/structural alerts). The MCASE program then identifies QSAR attributes and/or molecular fragments that are modulators, i.e., chemical molecular structure parameters that correlate with enhanced or diminished activity of chemicals that share a common structural alert (e.g., activating fragments, inactivating fragments, deactivating fragments, log-p, graph index, etc.). The combination of these data is used to develop a quantitative estimate of the potential toxicity of a test chemical in animals. The MCASE and MCASE-ES programs are a complete QSAR toolkit that permits the investigator to modify and construct new database modules, adjust the biologic activity/potency of training set chemicals, and investigate procedures to improve the predictive performance of the program.

We have discovered a number of experimental parameters that have improved the predictive performance of the MCASE program, and the cumulative changes have resulted in the development of a new operating system we call MCASE-ES.[9-12] The following summarizes the differences between the two programs using rodent carcinogenicity modules and divides the difference in terms of: (i) the data used to construct the database modules and (ii) the decision logic used to predict the activities of test chemicals:

MCASE-ES Database Module Improvements

1. Data Source: MCASE-ES uses rodent carcinogenicity studies obtained from the National Institute of Environmental Health Science (NIEHS), National Toxicology Program (NTP) and NCI; FDA/CDER archives; the L. Gold Carcinogenic Potency Database (CPD); and the published literature. In contrast, MCASE only used NIEHS/NTP studies.
2. Proprietary Data: The majority of the rodent carcinogenicity studies included in MCASE-ES modules are non-proprietary. However, a technical procedure was developed to identify structural alerts for the proprietary chemicals, if present, and to delete the identity and chemical structure of the proprietary chemicals from the com-

mercial version of the program. This procedure was developed so that the publicly available MCASE-ES software programs and those used by the FDA would identify the same molecular fragments correlated with carcinogenicity in rodents and would produce the exact same estimates of activity.

3. Module Type: A total of four different rodent carcinogenicity database modules were constructed for the MCASE-ES program in order to detect sex and species-specific carcinogenicity response (i.e., separate modules for male and female rats or mice). The MCASE program only has modules for rats, mice, or rodents.
4. Module Size: The average number of individual chemicals in the MCASE and MCASE-ES modules were ca. 300 and ca. 1,050, respectively.
5. Molecular Universe: There was a large difference in the number of 2-10 atom fragments obtained from the 300 chemical and 1,050 chemical database modules. The MCASE-ES module contains roughly 500,000 different fragments compared to ca. 40,000 fragments in the MCASE 300 chemical module. The 300 chemicals were largely aliphatic and aromatic chemicals of low molecular weight tested by the NIEHS/NTP; conversely, the 1,050 chemicals contained a large number of complex heterocyclic therapeutic molecules with relatively high molecular weights.

MCASE-ES Improvements in Software System Logic

1. Quantification of Carcinogenic Potency: The MCASE-ES program uses a detailed scale to distinguish chemicals having different carcinogenic potency in rodents, including: (i) carcinogens inducing tumors at multiple sites and/or two or more cells (assigned 30-79 CASE units), (ii) carcinogens inducing tumors at single sites in one cell or chemicals with equivocal findings (assigned 20-29 CASE units), and (iii) non-carcinogens (assigned 10-19 CASE units). In contrast, the original program did not use such a carcinogenic potency scale and, unlike MCASE-ES, it classified many chemicals with single site/single cell or equivocal findings as carcinogenic. This scheme is based upon the Tennant hypothesis for stratification of carcinogenicity bioassay results to reflect relative human hazard.[13]
2. Quantification of Structural Alert Frequency: The MCASE-ES program quantifies the carcinogenic activity of chemicals in terms of the number of chemicals in the training data set that share a MCASE structure alert for carcinogenicity. An alert identified in one or more database modules in five or more (≥5) training set chemicals is considered highly biologically significant, in three to five (3-5) chemicals is possibly significant, and in 1-2 chemicals is evaluated as not significant. In contrast, the MCASE program evaluates all alerts identified in one or more modules, in one or more training set chemicals, as biologically significant.
3. Quantification of Potency of Structural Alerts: We have empirically determined for several different toxicological endpoints that structural alerts with high total CASE unit activity (i.e., >150 total units) are most likely to be biologically significant and yield accurate estimates of toxicity. The total CASE units is the product of the average CASE unit activity of the structural alert times the number of chemicals that share the alert in one or more toxicology endpoint-related database modules. In contrast, alerts with less than 100 (<100) total units are not biologically significant and predictions based upon these alerts are unreliable. We evaluate alerts with 100 to 150 total CASE units as possibly biologically significant. In contrast, the MCASE program does not use a total CASE unit evaluation procedure.
4. Quantification of Biological Significance: We have determined empirically that positive predictions based upon an increasing number of database modules that are closely related in terms of the toxicological endpoint (e.g., rats and mice for carcinogenicity in rodents) are invariably associated with increased specificity, predictive value, and lower false positives, as well as lower sensitivity and higher false negatives. These studies showed that positive predictions from two or more modules for related toxicological endpoints typically yield predictive values and specificity of over 85 percent (>85%), and three or more modules 92 to 100 percent. In contrast, the predictive value and specificity of predictions based upon a single module for a toxicological endpoint are far less reliable (75% or less).
5. Artificial Intelligence/Wizard Functionality: The most recent versions of the MCASE-ES program (>3.40) links the results of the MCASE program to an artificial intelligence (AI) logic modeled after the C4.5 AI software. The resulting procedure links information on structural alerts detected in two or more database modules that are toxicologically related, and bases its prediction of activity based upon the cumulative data. This “Wizard” capability automatically duplicates many of the assay evaluation criteria rules developed by OTR/RRAS to evaluate the program’s data.

The relative sensitivity, specificity, and predictive value of the MCASE-ES program are adjustable, investigator-defined experimental parameters.[15] A positive prediction for a test chemical can be made based on: (i) one or more related toxicological endpoint database modules (ii) changes in the criteria for quantification of biological potency and structural alert frequency. Thus, the MCASE-ES program estimates can be adjusted for risk management to try to rule out all potential hazards, or for risk identification to try to identify specific potential hazards. For risk management, the investigators inflate the relative sensitivity of the program in an attempt to detect the largest number of active/positive chemicals at a given toxicological endpoint. The effect of optimizing this program for risk management is that it would exhibit higher sensitivity and lower false negatives, but have low specificity, low predictive value, and many false positives. In contrast, for risk identification the investigator optimizes the specificity/predictive value of the program in an attempt to use it to screen chemicals and identify with high confidence those chemicals with potential toxicity, and to study possible mechanism of toxicity induction. The consequence of optimizing e-TOX programs for hazard identification is that, although they exhibit high specificity, high predictive value, and few false positives, they have higher false negatives and lower sensitivity.

Our own investigations with the MCASE-ES program have demonstrated that this program is best used as a method for hazard identification. We deliberately set the program to achieve high predictive value, high specificity, and low false positives. Others have reported the same conclusion.[1,15] The benefits of this decision include:

1. False Negatives are Correctable: Although false negatives are increased when the MCASE-ES program was optimized for high specificity and predictive value, we have shown that the percentage of false negative predictions is correctable. We have observed that enhancement of our database modules with additional training set data invariably increases the number of alerts, the biological significance of individual alerts, and/or the profile of modulating factors for individual alerts. The combination of these factors invariably results in enhanced coverage and sensitivity of the program.
2. Predictions Reflect Mechanisms: When the MCASE-ES program is optimized for high specificity/predictive value, the structural alerts often identify molecular fragments that are experimentally correlated with given toxicological endpoints. Thus, the program's estimates are often readily defensible with studies from the knowledge database, and these predictions often reflect known mechanisms of action of chemicals. Taken together, the MCASE-ES estimates are readily defensible and implemented as decision support information for risk identification and regulatory decisions.
3. New Insights: In some cases, the MCASE-ES program, optimized for high specificity/predictive value, identifies structural alerts that are not obviously linked to the knowledge database. In this situation, the program may be providing interesting new insights and discovery that reflect additional toxicological endpoint structural alerts of concern and alternative mechanisms of action not previously considered.
4. Lead Chemical Selection: A MCASE-ES program optimized for high specificity/predictive value is ideally suited for screening possible lead chemicals for development. The program operates in batch mode and has a low probability of killing promising lead chemicals with false positive estimates of toxicity. Furthermore, the program can be used to suggest modifications of lead chemicals to create molecular structures that are less likely to be toxic, or to improve upon the efficacy of the therapeutic.
5. Alternative for in vitro/in vivo testing: The program can also be used as a legitimate and cost effective alternative for in vitro/in vivo testing for screening and hazard identification. The program can be used in batch mode to identify large numbers of chemicals with a high probability of being toxic/active for a battery of short-term tests, and thereby reduce the number of substances that might have to be tested in vitro/in vivo.

In contrast, when the MCASE-ES program optimized for hazard management by increasing its relative sensitivity, we discovered the program had a number of undesirable performance characteristics.

1. False positives are not correctable: Under these conditions, the program has a high frequency of false positives, low specificity and low predictive value.
2. Predictions do not reflect mechanisms: The program often elicited whimsical estimates of chemical toxicity. Although the addition of new data to the database modules improved the coverage of the program for test chemicals, it had no effect on the program's predictive performance. Thus, when the MCASE program was optimized for high sensitivity, the underlying logic of the operating system was flawed.
3. New Insights: Estimates of toxicity had a high probability of being controversial and hard to defend. These predictions were difficult to support using information from the Agency's knowledge database, and they did not reflect known mechanisms of action of chemical-induced toxicity. Likewise, because the estimates from high sensitivity MCASE programs were often flawed, these estimates can not be used to provide insights to unknown mechanisms.
4. Lead Chemical selection and alternative for in vitro/in vivo testing: Since it is desirable to exclude as few chemicals as possible with high potential efficacy as pharmaceuticals during the preliminary screening of lead chemicals, high frequency of false positive estimates is undesirable. Thus, a MCASE-ES program optimized for risk management would not be a useful tool for lead chemical selection, or as an alternative for in vitro/in vivo studies.

The overwhelming majority of FDA-regulated substances are relatively simple organic chemicals eligible for MCASE-ES analyses (75 to 5000 MW). Nevertheless, the software program should not be used for some substances. The program cannot be used for non-organic chemicals (salts and metals), organometallics, and polymers (fibers, polysaccharides, proteins, etc.). Nevertheless, the program could be used to evaluate subunits of polymers or organic chemicals (e.g., monomers, dimers, etc.) up to ca. 5000 MW. Furthermore, although organometallics cannot be evaluated, it is possible to substitute for the heavy metal/salt atom an acceptable atom (e.g., phosphorus, carbon, nitrogen, sulfur), or to remove the heavy metal/salt atom and test the remaining portion of the organometallic.

The MCASE-ES program should not be used to analyze the activities of certain organic molecules, including: (i) very small organic molecules (1-7 atoms, excluding hydrogen), (ii) molecules containing two or more unknown fragments, and (iii) molecules over ca. 5000 MW. Furthermore, the toxicological activities of mixtures of chemicals should not be used as data in the training database modules because the program has no means of determining which component of the mixture had the highest (most toxic) activity. However, MCASE-ES can be a very effective tool for risk identification of the potential hazard of individual components of a complex mixture.

Additional MCASE-ES e-TOX Endpoints

Based upon our initial success in using the MCASE-ES program to estimate the potential carcinogenicity of organic chemicals in rodents, OTR/RRAS has begun a number of separate investigations to develop additional MCASE-ES programs to estimate potential toxicities for chemicals at a variety of different non-clinical and clinical endpoints.[12,16] The endpoints have been chosen based upon: (i) the current needs and priorities of the Center and (ii) an objective of constructing a complete battery of software programs to correspond with the non-clinical tests recommended by the FDA.

Currently, OTR/RRAS has developed and/or is testing non-clinical database modules for estimating: (i) carcinogenicity in rodents (rats, mice), (ii) teratogenicity in mammals (rats, mice, rabbits), and (iii) maximum tolerated dose in rodents (MTD, rats, mice). In addition, we intend to compile databases for: (i) behavioral toxicity (Segment III studies, rats), (ii) reproductive toxicity (Segment I studies, rats), and (iii) the recommended battery of short-term genetic toxicity tests (mutagenicity in S. typhimurium and mouse lymphoma [L5178Y] cells, clastogenicity in CHO cells and the mouse micronucleus test, and cell transformation in BALB/c-3T3 and SHE cells). The data for these modules has been extracted from FDA archival files and publicly available records.

OTR/RRAS has also developed and/or is testing clinical database modules for estimating: (i) liver toxicity in adults, (ii) immunotoxicity in adults, (iii) maximum recommended therapeutic dose (MRD) in adults, and (iv) no-effect-level (NOEL) in adults.[16] In addition, we intend to compile databases for neurotoxicity and other organ and organ system toxicities in adults. The data for the organ and organ system toxicities has been extracted from FDA/CDER's Adverse Drug Reaction (ADR, ca. 1960-1997) and Adverse Event Reporting System (AERS, 1997-present) databases; the data for the MRD and NOEL modules were obtained from labelling and results of clinical trials. The details of all of these investigations will be the subject of additional publications.

e-ADME

In addition, OTR/RRAS is interested in expanding our service to include e-ADME consultant service capability because we feel these data are very important in understanding, investigating, and estimating the toxicological/pharmacological effects of therapeutics and other FDA regulated substances in humans. The objective of this service will be to provide quantitative estimates of a test chemical and structurally related congeneric chemicals, absorption, distribution, metabolism, excretion, and bioavailability (i.e., e-ADME). The structurally related chemicals will be identified using ISIS/HOST software as the standard; the e-ADME estimates will be obtained using QSBR,[17] MCASE-ES, and other appropriate software programs.

Other e-TOX Programs

In situations in which MCASE-ES programs are unavailable, not validated, and/or not recommended based upon the test substance, the FDA is using a variety of other e-TOX software programs to provide decision support information. For example, ONCOLOGIC is an important human expert program that provides estimates of carcinogenicity in rodents for a variety of substances not eligible for the MCASE-ES program (e.g., salts, metals, polymers, organometallics, etc.).[18] Likewise, DEREK has important human logic rules for skin sensitization/irritation that are useful for FDA-regulated substances with dermal applications (e.g., therapeutics and cosmetics). Finally, TOPKAT has been employed by FDA/CFSAN/OCAC[19,20] and the Cosmetic Ingredient Review (CIR) to prioritize potential toxicities of cosmetic ingredients.[21]

e-TOX Service

OTR/RRAS began a computational toxicology consulting service for the Center several years ago, and later expanded the service to the entire Agency. This service has the responsibility of using tested and approved e-TOX software to provide decision support information in situations in which toxicology data are unavailable or limited, and it has several objectives:

1. Consistent Decision Support Information: e-TOX and e-ADME software programs are for use by experts. They are SAR/QSAR toolkits that offer a plethora of data analyses options and an assortment of possible interpretations. It is the responsibility of FDA SAR/QSAR experts to determine the optimal experimental conditions for use of these software programs and the optimal data evaluation criteria to be used. Once these criteria have been determined, appropriate standard operating procedures (SOPs) can be formulated and used. This process has been employed by OTR/RRAS to furnish consistent decision support information for Agency regulatory decisions. Our objectives are to: (i) establish expert groups within the Agency that would use standardized study and data evaluation criteria, (ii) publish our observations and findings, and (iii) ultimately set Agency guidances for use of the e-TOX and e-ADME software programs posted on Center/Agency Web sites.
2. Minimize Cost/Limit Requirements: There is considerable equipment (computer software and hardware) expense and personnel time incurred in establishing and maintaining the OTR/RRAS centralized client support service. These costs include: (i) the purchase of e-TOX and e-ADME software programs, (ii) maintenance of the program annual license agreements, (iii) personnel time for training to maintain technical competence, (iv) research time to develop new software/databases to meet the needs of the Center/Agency, (v) time to prepare/maintain CDER Intranet/FDA Internet Web pages containing on-screen request forms, (vi) time to respond to requests for decision support information, (vii) time to upgrade existing e-TOX and e-ADME software database modules with new FDA study data, (viii) time to decode proprietary data for newly marketed substances, and (ix) time to prepare presentations, publications, and guid-

ances. By establishing one centralized SAR/QSAR service, a considerable cost of duplicating this capability and the considerable costs incurred in training large numbers of staff to use a battery of new software packages can be avoided.

3. Easy Access/Easy to Use/Rapid Response: By making the computational toxicology, biology, and chemistry consultants service data available via the CDER-intranet, the service makes important information readily available to all Center scientists for both regulatory and research decisions. The CDER Intranet also facilitates the use of simple/uniform on-screen request forms and a rapid response (two to three weeks) to requests for information. The service only requires that the chemical name/code and its chemical structure be submitted in any of a variety of acceptable formats.

The ultimate goal of the OTR/RRAS computational toxicology program and consultant service is that when a IND is submitted to the Center, its molecular structure (ISIS ".mol-file") will be entered into the Agency substance inventory and included in the inventory's ISIS ".sdf-files." OTR/RRAS will perform an ISIS structure similarity search of all Agency and Center ".sdf files" and will identify a cluster of congeneric chemicals with high ISIS similarity indices (SIs). This high SI/congeneric chemical cluster will be submitted for e-TOX and e-ADME evaluation using a battery of validated SAR software programs, and a profile of potential toxicological/biological/chemical activities will be estimated and registered in ORACLE tables in the Center/Agency databases linked to the Web. In addition, a brief OTR/RRAS consultant report of these data will be submitted to the medical officer and reviewers assigned to evaluate the IND.

Current and Potential e-TOX and e-ADME Applications within the FDA

The FDA is evaluating e-TOX as a source of supplemental supporting information, or to provide additional decision support information, in situations in which the results of in vivo/in vitro non-clinical studies is either inadequate or unavailable. The FDA does not consider either e-TOX or e-ADME data as substitutes for any in vitro/in vivo studies.

1. FDA/CDER (Pre-Market): (i) Evaluate potential hazards of contaminants, degradents, or excipients in new or generic drug products, and (ii) propose to supply additional decision support information for the entry of women of child bearing potential into phase I clinical trials when animal reproductive toxicity studies have not yet been completed.
2. FDA/CFSAN (Pre-Market): (i) Evaluate potential hazards of food contact substances (FCS). (FDA's CFSAN, Office of Premarket Approval [OPA] is using e-TOX data to provide decision support information for their evaluation of FCS [food contact substances, also called indirect food additives]. The FDA Modernization Act [FDAMA] of 1997 requires that FDA/CFSAN/OPA review FCS within a 120-day review period.)
3. FDA/CDER-NIH/NIDA (Pre-Market): The NIDA Drug Discovery Program for Medications Development for Addiction Treatment. The FDA/CDER and the National Institute of Health (NIH), National Institute for Drug Abuse (NIDA) are entering into a Government Interagency Agreement (IAG) through which FDA/OTR/RRAS will provide e-TOX and e-bioavailability data as decision support information for potential lead chemicals being evaluated by the NIDA Drug Discovery Program. In addition, NIDA will provide funding to support and expand the OTR/RRAS e-TOX capability.
4. FDA/CDER (Post-Market): FDA/CDER maintains electronic databases for the adverse effects of therapeutics noted in patients, the ADR and AERS databases. FDA/CDER/OTR and FDA/CDER/OB/QMRS are actively engaged in developing and evaluating analytical methods and models for evaluating the potential hazards of marketed therapeutics in humans utilizing these databases.
5. FDA/CFSAN (Post-Market): Evaluate potential hazard(s) of active ingredients of cosmetics. (FDA/CFSAN's OCAC and the Expert Panel of the CIR use e-TOX data to prioritize the active ingredients of cosmetic preparations. These data are particularly important because cosmetic ingredients are not subject to premarket approval, and limited toxicology data are usually available for these substances.)
6. FDA/CFSAN/FDA/CDER (Post-Market): Evaluate potential hazards of individual components of complex mixtures. Many complex mixtures are not subject to premarket approval or are allowed because they are GRAS. These substances include, but are not limited to, dietary and nutritional supplements, flavors, herbs, herbal medicines, etc. e-TOX and e-ADME capabilities may be useful to screen, prioritize, and identify potential hazards among the individual components of complex mixture substances of regulatory concern.

Discussion

The need for reliable e-TOX and e-ADME software, and the prerequisite e-databases and informatics for new product development and regulatory applications, has never been greater. Combinatorial chemistry and high throughput screening have increased the pace of drug discovery and development. These technical advances are now placing an increased burden on current animal based in vivo/in vitro toxicology screening methods that have become a rate-limiting factor in the lead selection process. The application of human genomics will likely increase the rate of discovery of new therapeutic compounds and will further increase the demands on toxicology testing. In addition, a vocal public sector is demanding that animal testing be reduced, replaced, and/or refined. Simultaneously, influential voices from industry regularly exhort Congress to lessen their regulatory burden and "streamline" the regulatory process (e.g., FDAMA, 1997). Nearly everyone would like the FDA to perform its regulatory mission faster, with fewer resources, but without diminishing or jeopardizing the safety and efficacy of FDA-regulated substances. Although the pharmaceutical and chemical industries have successfully developed high throughput technology and in vitro assays to process hundreds of thousands of chemicals daily, toxicological testing and screening have remained relatively slow and costly.

Fortunately, advances in computer technology have eliminated many of the technical impediments for the development of large databases and reliable e-TOX and e-ADME software. Modern computer hardware and software are currently available to process, manage, and mine immense data sets efficiently. Likewise, the development of these tools is no longer limited by theoretical principles and understandings. Recent research has demonstrated that certain e-TOX programs really do work, and the critical principles for obtaining reliable estimates of chemical toxicity have been discovered, defined, and characterized. These programs require: (i) robust training sets of data, (ii) a balance and adequate representation of compounds that are active and inactive (toxic/non-toxic), (iii) high quality data generated using consistent protocols and evaluated with consistent assay evaluation criteria, and (iv) program logic that reflects the nuances of regulatory policy.

The major remaining obstacle for systematic development and implementation of reliable e-TOX and e-ADME software, and the prerequisite e-databases and informatics, is the absence of readily available, comprehensive, electronic databases of toxicological/pharmacological data. The most robust sources of toxicological/pharmacological data remain inaccessible for modelling/mining in government and industry archives. The consummate need is to find a mechanism/procedure by which scientists can have access to both government and industry toxicology and at the same time preserve the proprietary status of these data, where required. These combined data are necessary to establish fundamental elements/attributes/understanding of chemical toxicity.

OTR/RRAS has successfully developed one possible procedure for sharing information from our proprietary studies, while simultaneously preserving the proprietary identity of the substance. We have developed procedures under our CRADA with Multicase, Inc., to use the MCASE-ES program to identify a combined list of MCASE-ES structural alerts and modifiers for a database module containing both proprietary and nonproprietary chemicals. These data are then used by the CASETOX program to make estimates of toxicity for test chemicals without divulging the name or the chemical structure of any proprietary chemical in the training data set. The identity of the proprietary chemicals are literally coded and removed from the database module after the composite pool of structural alerts and modifiers have been identified. This procedure has been reviewed by Agency legal experts and the software is now publicly available through Multicase, Inc. Thus, it is theoretically possible and feasible for the MCASE-ES program to identify a global library of structural alerts for any toxicological/pharmacological endpoint using all government, industry, and publicly available data. The CASETOX program could then be used for hazard identification screening and prioritization based upon the world's collective data set. The means of accomplishing global synthesis of data requires only an independent broker to review/evaluate the data and resources to fund their scientific evaluation.

Acknowledgments

We thank our partner and collaborator, Dr. Gilles Klopman, for his valuable technical comments, advice, and support.

Disclaimer

This is not an official U.S. Food and Drug Administration guidance or policy statement. No official support or endorsement by the U.S. Food and Drug Administration is intended or should be inferred.

References

1. Richard, A. Commercial toxicology prediction systems: a regulatory perspective. Toxicol. Lett. 1998, 102/103:611-616.
2. Richard, A. Application of artificial intelligence and computer-based methods to predicting chemical toxicity. The Knowledge Engineering Review 1999 14(4):307-317.
3. McKinney, J.D., Richard, A., Waller, C., Newman, M.C., and Gerberick, F. The practice of structure activity relationships (SAR) in toxicology. Toxicol. Sci. 2000, 56:8-17.
4. Richard, A.M., Bruce, R., and Greenberg, M. Structure activity relationship methods applied to chemical toxicity: making better use of what we have. Proc. World Congress of Alternatives to Testing, Bologna, Italy, August 28-Sept. 1, 2000.
5. Contrera, J.F. and Matthews, E.J. Recent advances in computational toxicology and regulatory applications. FDA Science Forum, Feb. 14-16, 2000.
6. Contrera, J.F., Jacobs, A.C., Prasanna, H.R., Mehta, M., Schmidt, W.J., and DeGeorge, J.J. A systemic exposure-based alternative to the maximum tolerated dose for carcinogenicity studies of human therapeutics. J. Am. Coll. Toxicol. 1995, 14(1):1-10.
7. Contrera, J.F., Jacobs, A.C., and DeGeorge, J.J., Chen, C.C., Choudary, J.B., DeFelice, A.F., Fairweather, W.R., Farrelly, J.G., Fitzgerald, G.G., Goheer, A.M., Jordan, A.W., Kelly, R.E., Lin, D., Lin, K.K., Meyers, L.L., Osterberg, R.E., Prasanna, H.R., Resnick, C.A., Sheevers, H.V., and Sun, J. Carcinogenicity testing and the evaluation of regulatory requirements for pharmaceuticals. Reg. Toxicol. Pharm. 1997, 25:130-145.
8. Wagner, P.M. Nabholz, J.V., and Kent, R.J. The new chemicals process at the Environmental Protection Agency: Structure-activity relationships for hazard identification and risk assessment. Toxicol. Lett. 1995, 79:67-73.
9. Matthews, E.J. Regulatory application of computational toxicology, now and in the future. FDA Science Forum, Feb. 14-16, 2000.
10. Contrera, J.F. and Matthews, E.J. Use of FDA databases, informatics and computational toxicology to estimate potential toxicity. Optimizing Lead Selection and Early Attrition. The Center for Business Intelligence, Philadelphia, PA, Sept. 27-28, 1999.
11. Matthews, E.J. and Contrera, J.F. A new highly specific method for predicting the carcinogenic potential of pharmaceuticals in rodents using enhanced MCASE QSAR-ES software. Reg. Toxicol. Pharmacol. 1998, 28:242-264.
12. Matthews, E.J. The development of new FDA-Multicase-SAR-Expert system software for estimating chemical toxicity and dose response. Predictive toxicology: pre-clinical testing, profiling tools and in silico models for successful early lead selection. Boston, MA, June 26-27, 2000.
13. Tennant, R. Stratification of carcinogenicity bioassay results to reflect relative human hazard. Mutat. Res. 1993, 286(1):111-118.
14. Cooper, J.A., Saracci, R., and Cole, P. Describing the validity

of carcinogen screening tests. Brit. J. Cancer 1979, 39:87-89.

15. Richard, A. M, and Woo, Y.T.A CASE-SAR analysis of polycyclic aromatic hydrocarbon carcinogenicity. Mutat. Res. 1990, 242(4):285-303.
16. Matthews, E.J. Assessment of the No-Effect-Level (NOEL) of chemicals in humans using pharmaceutical clinical trial data and MULTICASE software. QSAR2000, Ninth International Workshop on Quantitative Structure Activity Relationships in Environmental Sciences. Bourgas, Bulgaria, Sept. 16-20, 2000.
17. Andrews, C.W., Bennett, L., and Yu, Y.X. Predicting human oral bioavailability of a compound, development of a novel quantitative structure-bioavailability relationship. Pharm. Res. 2000, submitted for publication.
18. Cheeseman, M.A., Machuga, E.J., and Bailey, A.B. A tiered approach to threshold of regulation. Food Chem. Toxicol. 1999, 37(4):387-412.
19. Milstein, S.R., Yourick, J.J., Bronaugh, R.L., Wamer, W.G., Lowther, D.K., Meyers, M.B., Scher, A.L., and Bell, S.J. Safety assessment of cosmetic ingredients and color additives, evaluation of a quantitative structure-toxicity relationship approach for predicting toxicity. Society of Toxicology Abstracts, 1999.
20. Yourick, J.J., Milstein, S.R., Wamer, W.G., Lowther, D.K., Meyers, M.B., Scher, A.L., Bell, S.J., and Bronaugh, R.L. Integrating the QSAR paradigm into the safety assessment of cosmetic ingredients and color additives. FDA Science Forum, 1996.
21. Bergfeld, W.F. and Andersen, F.A. The Cosmetic Ingredient Review. In Cosmetic Regulation in a Competitive Environment. Norman F. Estrin and James M. Akerson, (ed.). Marcel Dekker, Inc., 2000.

CAREER COLUMN

New Industrial Training Program for Interns and Postdoctoral Fellows

Allen B. Richon and Merry Ambos

Network Science Corp., 412 Carolina Blvd., Isle of Palms, SC 29451. editors@netsci.org

Abstract

Pharmaceutical industry research managers consistently have difficulty finding job candidates trained to address industrial research projects. The current situation is even worse than usual because of high demand for research scientists, particularly those trained in solid phase synthesis, computational chemistry, and informatics. Network Science Corp., Chemical Computing Group, and pharmaceutical research groups are creating a website to assist students in gaining experience as industrial interns or postdoctoral fellows.

Introduction

On March 2, as a part of the 2000 Charleston Conference, we met with research managers from several pharmaceutical companies to discuss difficulties associated with locating entry level Ph.D.s who have an understanding of the requirements of industrial research and are thus productive in the early stages of their careers. During this discussion, Dr. Catherine Peishoff, Associate Director of Physical and Structural Chemistry at SmithKline Beecham, highlighted the problem when she commented that “both the variety of computational techniques and their use within industry have expanded so rapidly in recent years that the gap between academic training and industry needs is growing each year. On-the-job industry training has thus become more demanding making it increasingly difficult to provide the grounding necessary for new candidates to be successfully integrated into the company in a timely fashion.”

Unlike many other scientific areas, the group agreed that there is no well-defined academic training for industrial computer aided molecular design (CAMD). Dr. Christine Humblet, Senior Director of Biomolecular Structure and Drug Design at Pfizer, concluded, “computational chemistry is a field that builds upon a broad variety of disciplines, combining a palette of scientific components dictated by the industrial business. Within the pharmaceutical industry, an appropriately trained CAMD scientist is required to understand not only the multitude of computational techniques underlying computer-based methods but also the components of drug discovery and development and the business forces which drive the current industry. Since CAMD scientists work at the interface between chemistry, pharmacology, and the computer, they also are required to effectively communicate with other members of the research team. There is currently no formal academic curriculum that addresses such variety of skills.”

CAMD is constantly evolving; a technique learned in the past may no longer be sufficient to address today’s questions. This is reflected in the current impact of the emergence of combinatorial chemistry, high-throughput screening, and genomics. As shown in the Table 1, computational tools and molecular design methods have evolved to meet the requirements for faster design and more thorough analysis which are a result of these and other changes in compound discovery.

The Program

The group identified the use of internship programs and post-doctoral fellowships as a possible method to educate graduate students in the skills that industrial research requires. While several of the participants at the meeting stated that their companies offer these programs, many members of the group highlighted the fact that they would like to expand the use of interns and post-docs to promote their programs and to find qualified students. The primary obstacles to this program, they felt, were a lack of time to undertake a full recruiting effort and a lack of resources to support the search for qualified students for short-term assignments. Network Science, a registered 501(c)3 nonprofit organization, will address these tasks.

We are actively promoting education programs for drug discovery research. Companies in the pharmaceutical industry and Chemical Computing Group will support our efforts by contributing to the creation and maintenance of the program's website. "We are looking forward to expanding our support of educational initiatives for computational chemistry," stated Bill Hayden, Vice President Sales and Marketing of Chemical Computing Group. "This program complements the CCG Excellence Award by creating a single reference site that will link graduate students interested in augmenting their traditional academic training with projects in industry. Disseminating these applied programs more widely will assist young scientists with making a more rapid transition from graduate schools to technical careers."

As a part of the program, Network Science, with research managers at sponsoring organizations, will create job descriptions for positions within research, post these positions on the NetSci Website (http://www. netsci.org), and assist in filling these positions with students from the leading universities. The program will permit research organizations to get an advanced look at the students who will become candidates for employment. The resources created will be available only to the organizations that sponsor this effort. Specifically, for each of the disciplines in discovery, we will:

- Contact research managers to identify specific requirements and the academic research groups that traditionally place students within their groups.
- Contact industrial liaisons at the leading universities to advertise and promote the use of the student intern network.
- Create and manage a Web-based, password-protected database that contains descriptions of positions available within the industry and resumes of students who are interested in the program.
- Notify faculty members identified by the industry of the new opportunities available for training their students.
- Implement a Web-based advertising campaign to promote the program.
- Assist sponsoring organizations by prescreening resumes submitted by students against available openings and forwarding these resumes to designated contacts.
- Work with Human Resources to coordinate screening and interviews for candidates.

The program will grow as more companies join the initiative. Students who wish to participate in the program can submit their resumes at www.netsci.org/Resources/Initiative/resume.html or contact editors@netsci.org for more information.

MEETING REPORTS

The XIX International Conference on Magnetic Resonance in Biological Systems

Florence, Italy
August 20-25, 2000

Miriam Gochin

University of the Pacific School of Dentistry, Dept. of Microbiology, 2155 Webster Street, San Francisco, CA 94115 USA.
miriam@picasso.nmr.ucsf.edu

Over 1,000 participants from around the world attended the XIX International Conference on Magnetic Resonance in Biological Systems held in Florence, Italy.

The meeting was divided into morning and evening plenary lectures, with six parallel sessions during the day covering all aspects of NMR and EPR. There was a daily lecture on "The Way We Will Be" by internationally regarded magnetic resonance experts, who reflected on the way in which their field would impact the future. Session lectures covered the scope of magnetic resonance applications including protein and DNA/RNA structural studies, NMR spectroscopic technique development, solid state and membrane

Table 1. History of Computational Methods

Time Frame	Research Approaches	Computational Approaches
1960-1980	Random Screening	• QSAR - Free Wilson • Molecular modelling and conformational analyses
1980-1990	Rational Design	• Experimental structures for ligands or targets • Pharmacophore analysis
1990-1995	Structure-Based Design	• Molecular docking, de novo design, GA, neural network, cheminformatics, fingerprinting
1995-present	Combinatorial Chemistry High Throughput Screening Functional Genomics	• Molecular Diversity Analysis • Library Design • Protein folding, sequence alignments, fold recognition

proteins, EPR and ENDOR, computation and dynamics and imaging and in vivo spectroscopy. Nobel Laureate Richard Ernst, who had a rich collection of slides detailing the development of NMR spectroscopy in Europe and the United States over several decades, opened the meeting.

Two overriding themes emerged from the meeting. One was the role that NMR could play in structural genomics, a hot topic given the recent fanfare over genome sequencing. The other was the use of orientational ordering in NMR for obtaining structural constraints other than NOE's.

Structural Genomics

A special round table discussion on the role of NMR in structural genomics was convened. Participants included Shigeyuki Yokoyama, (RIKEN, Yokohama, Japan), Gaetano Montellione (Rutgers University, NJ) who has spearheaded automated NMR analysis method, Hartmut Oschinat (Research Institute for Molecular Pharmacolgy, Berlin, Germany), Ivano Bertini (University of Florence, Italy) conference chair, Cheryl Arrowsmith (University of Toronto, Toronto, Canada) and Kurt Wüthrich (ETH, Zurich, Switzerland). These speakers outlined and encouraged the use of large NMR Spectroscopic Centers where powerful instrumentation and a critical mass of scientists would come together to solve protein structures en masse. Only by instituting such centers, it was argued, could NMR hope to seriously compete with X-ray crystallography in finding a niche in the highly lucrative post-genome processing business. The session was then opened to questions and discussion from the audience. One detractor accused the panel of promoting its own agenda, since no scientists were included who did not head a large, centralized NMR facility. Other audience members questioned the cost of such centers, and how they would compete for the same pot of federal funding now distributed to smaller laboratories, how the formation of such centers would dispense power to a few and stifle innovation in favor of large NMR factories. Yet others questioned the underlying ability of NMR to compete in any way with X-ray crystallography, which can churn out structures of much larger molecules than NMR is capable of doing, in a fraction of the time. It was generally agreed that NMR spectroscopists should not sacrifice high-resolution quality structures in favor of high-speed low-resolution ones. This discussion will be taken up again at the International Conference on Structural Genomics to be held in Japan later this year.

Orientational Ordering

Orientational ordering in NMR is achieved by the use of an orienting solvent, such as a dilute liquid crystal that orients upon application of a magnetic field, or by the attachment of paramagnetic metal ions that impart anisotropic magnetic susceptibility to the molecule. Orientational constraints that can be obtained include paramagnetic shifts, cross-correlated relaxation measurements, and residual dipolar coupling. In particular, the use of residual dipolar coupling to improve structural definition and domain orientations was prevalent in talks and posters throughout the meeting. Residual dipolar coupling, for example, between an amide ^{15}N and its attached proton, is directly related to the angle that the ^{15}N-^{1}H vector makes with the principal axis of the orientational tensor in a partially ordered system, such as a liquid crystalline matrix. Having a set of these vector orientations for all ^{15}N-^{1}H bonds in a protein, for example, provides the spectroscopist with information on the global orientation of protein domains relative to each other. One of the principal developers of this technique, Ad Bax, (National Institutes of Health, Bethesda, MD) was pleased with the rapid proliferation of the method. He said that new types of ordering solvents could address previous problems such as the interaction of certain proteins with bicellar solutions developed to date or intractable line broadening that occurs in certain cases. Dr. Bax presented one of "The Way We Will Be" lectures, in which he outlined the possibility of using comparative sequence and structure analysis to derive starting models for the structure of newly identified proteins. Residual dipolar couplings could then be used to verify the theoretical model as well as to define the relative orientations of predicted helix and sheet domains. No complex and time-consuming NOE analysis would be needed at this stage, although these would be important for defining the high-resolution structure of potential ligand sites.

Imaging

Kamil Ugurbil (University of Minnesota, Minneapolis, MN) showed exceptional functional MRI images of neuronal activity in the cat brain. Blood vessels contain deoxy-hemoglobin which is paramagnetic and causes susceptibility inhomogeniety which can be detected by NMR methods. Increased blood flow and increased oxygen consumption is associated with neuronal activity and is accompanied by changes in local inhomogeneity as well as the blood T2. Particularly nice contrast and functionally lit-up images were obtained at higher fields (7T or 9.4T).

Simulations

Wilfred van Gunsteren (ETH, Zurich, Switzerland) discussed computer simulation of biomolecules and showed that he was able to accurately simulate the structure of small-folded peptides. He also delineated the importance of the role of solvent and of accurately defining the unfolded state in solving the folding problem.

Solid State NMR

The theme of structural and functional genomics was the subject of a presentation by Stanley Opella (University of Pennsylvania, Philadelphia, PA). Methods are being developed in his lab for the study of protein structure in the solid state. These techniques could be applied to the 30 percent of the genome that codes for membrane proteins.

Drug Discovery

The use of NMR in drug discovery was relatively underrepresented at the meeting, with the exception of Steve Fesik (Abbott Laboratories, Abbott Park, IL) and Luciano Mueller (Bristol-Myers Squibb, Princeton, NJ). Dr. Fesik discussed new directions in his Structure Activity Relationships (SAR) by NMR technique, including the use of NMR in the design of directed libraries and high-throughput structure determination. Dr. Fesik outlined his Linked Fragment approach whereby compounds that bind to adjacent sites are chemically linked to engineer a more potent ligand, and his Merged Fragment approach where compounds that bind to overlapping sites can be used as starting points for designing a more optimal binder. He gave an example of the development of an antiviral agent to human papilloma virus (HPV), by targeting the HPV-E2 protein. He also talked about the problem of many drugs binding non-specifically to serum albumin, leading to much larger dosage requirements. For example, diflunisol is 99 percent albumin bound. His group is looking at domain III of albumin as a target in developing drugs that can be given in conjunction with drugs like diflunisol, thus reducing dosage requirements.

Luciano Mueller stressed the importance of protein threading methods at Bristol-Myers Squibb for matching Conserved Essential Genes (CEGs) against known structures and trying to establish biological function. He expanded that, while NMR might be used to characterize whether proteins are folded or to give secondary structures based on rapidly assessed chemical shift information, hence verifying the threading results, X-ray crystallography would be used more typically for 3D structure determination. However, the crucial role of NMR then lay in its ability to do high throughput screening of a small molecule library for binding to a particular CEG product. He also suggested reverse screening, whereby a compound that might have been found, say to have antimicrobial properties, could be screened against a library of CEG products in order to determine its potential target.

One came away from this meeting with a sense that the role of NMR is not solely in static structure determination, but that it is most likely to be an important player in assaying targets to newly discovered proteins in the genome project. In all, the meeting closed with a general sense of satisfaction from the participants as to the high level of science and discussion and the inspirational effect of being in Florence, crucible for arts and sciences in Renaissance Europe.

Links

Conference Website
http://www.cerm.unifi.it/XIXicmrbs/icmrbs_p1.html

Center for Advanced Biotechnology and Medicine, Rutgers, NJ
http://www-nmr.cabm.rutgers.edu/

The Institute of Physical and Chemical Research
http://www.riken.go.jp/engn/index.html

Center for Magnetic Resonance Research, University of Minnesota
http://www.cmrr.drad.umn.edu/

Emerging Technology Symposium

Washington, DC
August 22, 2000

Donald B. Boyd

Department of Chemistry, Indiana University-Purdue University at Indianapolis (IUPUI), Indianapolis, IN, USA.
boyd@chem.iupui.edu

The first ever Emerging Technology Symposium was held at the American Chemical Society (ACS) National Meeting was organized under the auspices of the ACS Division of Computers in Chemistry (COMP). The purpose of the symposium was to stimulate advances in computational chemistry. The winning speaker was Amiram Goldblum (Hebrew University of Jerusalem, Israel) for work by himself and former student, Meir Glick.

At the Fall ACS meeting, seven speakers competed for a $1,000 prize sponsored by Schrödinger, Inc. The award was announced on the Computational Chemistry List at the beginning of this year and seven speakers were selected from the many submitted:

- Dr. Melissa L. Plount Price (Yale University)
- Mr. Matthew Randolph Lee (University of California, San Francisco)
- Prof. Amiram Goldblum (Hebrew University of Jerusalem)
- Prof. Randy J. Zauhar (University of the Sciences in Philadelphia)
- Mr. Shiang-Tai Lin (University of Delaware)
- Dr. Thomas F. Hendrickson (Agouron Pharmaceuticals)
- Dr. Joao M. Aires-de-Sousa (New University of Lisbon, Portugal)

Dr. Melissa L. Plount Price's paper, "Origin of binding selectivity for celecoxib analogs with COX-1 and COX-2 from combined docking and Monte Carlo simulations," was co-authored with Prof. William L. Jorgensen (Yale University). The paper "Getting 1.4 Å C-alpha RMSD structure predictions on two small proteins with molecular mechanics" was co-authored by Matthew Lee and Prof. Peter A. Kollman (University of California, San Francisco). The paper of Prof. Goldblum (Hebrew University of Jerusalem) and Meir Glick was entitled "A novel stochastic algorithm for structure predictions in proteins and for biomolecular interactions." Prof. Randy J. Zauhar (University of the Sciences in Philadelphia) co-authored his paper with William J. Welsh on "Application of the "shape signatures" approach to ligand- and receptor-based drug design." The fifth talk was entitled "A novel approach to improve group contribution predictions based on modern computational chemistry" by Shiang-Tai Lin (University of Delaware) and Prof. Stanley I. Sandler. Dr. Thomas F. Hendrickson (Agouron Pharmaceuticals of Pfizer Global Research), Fora Chan, Seigfried Reich, and Theodore O. Johnson worked on the paper "Design and evaluation of combinatorial libraries using protein crystal structures:

Methods and applications to drug discovery." The last talk, entitled "New representations of molecular chirality: Application to the prediction of enantiomeric selectivity in chromatography and chemical reactions" was by Joao M. Aires-de-Sousa (The New University of Lisbon, Portugal) and Prof. Johann Gasteiger.

Talks were evaluated on whether the research would have a large impact on the future of computational chemistry. Other considerations were: How widely applicable is the methodology? How likely is the methodology to be used by others? How original is the idea? Was the method explained in sufficient detail? Was the utility of the method adequately proved? Was the presentation clear and understandable? Did the speaker keep within the allotted time?

The judges were:

- Prof. Curt Breneman (Rensselaer Polytechnic Institute and Treasurer of COMP)
- Dr. George R. Famini (U.S. Army Edgewood RD&E Center and a past Chair of COMP)
- Dr. Charles H. Reynolds (R.W. Johnson Pharmaceutical Research Institute and a past Chair of COMP)
- Dr. Peter S. Shenkin (Schrödinger, Inc.)
- Dr. David C. Spellmeyer (DuPont Pharmaceuticals Research Laboratories and Chair-Elect of COMP)
- Dr. Terry R. Stouch (Bristol-Myers Squibb Pharmaceutical Research Institute)

The winning speaker, Amiram Goldblum, was presented with the award by this author and organizer of the symposium. Dr. Meir Glick is now a postdoctoral fellow in the lab of Prof. Graham Richards (Oxford University, UK). Prof. Goldblum summarized their work as follows: "We developed a novel stochastic approach to large combinatorial problems, which detects the global minimum as well as any number of solutions that are close to it. The method is based on a random choice of values for each of the system's variables and an evaluation of the full, randomly picked, system's configuration by some cost function (energy, distance, etc.). About 1,000 such configurations are probed, and variable values may then be evicted if they consistently contribute to 'bad' values for the full system. Iterations of this process lead to a manageable number of combinations for the system, which are all evaluated in an exhaustive calculation. Each problem requires some adaptation of this general algorithm. Success depends on the correct choice of cost function, but may be due, in part, to the probing of all values for each of the variables. Those values are either evicted or remain for the exhaustive phase. It has been shown already to be successful in positioning polar protons in crystal structures of proteins (Glick and Goldblum, Proteins 2000, 38:273-287), in rotamer positioning on a given native backbone, and in predicting large loop structures (n = 4-16) in proteins."

Prof. Goldblum commented, "This was my first ever participation and presentation at an ACS meeting. I knew beforehand that we had made some breakthrough, but did not fully appreciate it myself before the ACS and the gratifying winning of the prize. Naturally, it is not the financial value of the prize, but the recognition of some achievement, and the energizing effect that it made immediately on our efforts to continue probing deeper into the various aspects of this research. The exposure at the ACS meeting, due to the prize, has already helped us in making many scientific contacts that are necessary for such a continuation. I would like to take this opportunity to thank the organizers of the Emerging Technologies Symposium and the COMP Division of the ACS for an extremely interesting meeting in which I learned a lot and had the opportunity to interact and exchange ideas with so many excellent researchers, as well as to listen to very interesting presentations."

PHOTO CREDIT: ELIZABETH WILSON

Figure 1. Speakers at the 2000 Emerging Technologies Symposium. Front row (from left): Joao M. Aires-de-Sousa, Shiang-Tai Lin, Matthew R. Lee, and Melissa L. Plount Price. Back row (from left): Thomas F. Hendrickson, Randy J. Zauhar, Amiram Goldblum (winner), and Donald B. Boyd (organizer).

PHOTO CREDIT: JOANNE BOYD

Figure 2. Judges at the 2000 Emerging Technologies Symposium. Front row (from left): Terry R. Stouch, Amiram Goldblum (winner), Donald B. Boyd (organizer), and David C. Spellmeyer. Back row (from left): Charles H. Reynolds, George R. Famini, Curt Breneman, and Peter S. Shenkin.

Six runners-up were presented with a volume of "Reviews in Computational Chemistry."

The COMP division appreciates the generous sponsorship of Schrödinger, Inc. and plans to make the Emerging Technologies Symposium an annual event at the Fall ACS meetings.

Links

ACS Division of Computers in Chemistry
http://membership.acs.org/C/COMP

Schrödinger, Inc.
www.schrodinger.com

Ohio Supercomputer Center Computational Chemistry List
www.ccl.net/chemistry/

Amiram Goldblum
http://info.md.huji.ac.il/depts/pharmchem/gold.html

PUBLICATIONS

ISI Rankings of Chemistry Journals

Donald B. Boyd

Department of Chemistry, Indiana University-Purdue University at Indianapolis Indianapolis, IN, USA. boyd@chem.iupui.edu

Ratings of scientific Journals by the Institute of Scientific Information (ISI, Philadelphia, PA, www.isinet.com) have been released for 1999. The principal metric used by ISI in comparing serial publications is the “impact factor,” which is the total number of citations to a given journal and the total number of source items published in that journal for a given period. The time window is two years, so the new ratings count the number of citations in 1999 to articles published in 1997 and 1998. A high impact factor indicates that scientists are relatively frequently referring to articles in a given publication. Librarians use ISI data in deciding on journals subscriptions, as library budgets are limited, and ISI rankings frequently influence which journals are retained in a library’s collection.

The area of molecular modelling and computational chemistry is a high profile area, served by about 15 journals and book series. The top ranked serials in this field are shown in Table 1. One of the oldest journals in the field, Springer’s Theoretica Chimica Acta was renamed Theoretical Chemistry Accounts in 1997, but it is still listed under both names by ISI. By way of comparison, other journals that contain a relatively large number of papers reporting molecular modelling are shown in Table 2. The more general journals, Angewandte Chemie International Edition in English and the Journal of the American Chemical Society have very high impact factors.

Table 1. ISI Rankings of Serials in Computational Chemistry

Publication	Impact Factor
Journal Molecular Graphics & Modelling	4.206
Theoretica Chimica Acta	3.286
Journal Computational Chemistry	3.052
Theoretical Chemistry Accounts	2.827
Reviews Computational Chemistry	2.789
Journal Computer Aided Molecular Design	2.500
Journal Chemical Information Computer Science	2.066
Journal Chemometrics	1.821
Quantitative Structure Activity Relationships	1.803
Journal Biomolecular Structure & Dynamics	1.407
International Journal Quantum Chemistry	1.318
Journal Molecular Modeling	1.267
THEOCHEM	1.140
Molecular Simulations	0.896
Journal Molecular Structure	0.868
Journal Mathematical Chemistry	0.646

Table 2. ISI Rankings of Selected Other Serials

Publication	Impact Factor
Ang. Chem. International Ed. Engl.	7.996
Journal American Chemical Society	5.537
Journal Molecular Biology	5.501
Biophysical Journal	4.580
Biochemistry	4.493
Protein Science	4.457
Journal Medicinal Chemistry	4.079
Proteins	3.580
Journal Organic Chemistry	3.440
Journal Chemical Physics	3.289
Journal Physical Chemistry B	3.265
Protein Engineering	3.209
Journal Physical Chemistry A	2.695
Biopolymers	2.331

MEETINGS ABSTRACTS

Molecular Graphics and Modelling Society Meeting on Structural Genomics

Robinson College, Cambridge, UK
September 20-22, 2000

Speaker Abstract Titles

Poster Abstract Titles

Speaker Abstracts

Structural Genomics: A New Role of Structural Biology for Functional Genomics

Sung-Hou Kim, Department of Chemistry and Lawrence Berkeley National Laboratory, University of California, Berkeley, CA 94720 USA

Analysis of several genomic sequences indicates that no known functions can be inferred to a significant fraction of the genes. To infer functions for the products of these genes additional information, beyond sequences, is needed. Since the molecular (biochemical and biophysical) function of a gene product is tightly coupled to its three-dimensional structure, finding the structure or its folding pattern may provide an important insight into the molecular function of the gene product. That, in turn, may help in understanding its cellular function (genetic and physiological function: networks of many molecular functions) as well. We have started testing the premise that the structure infers molecular function of a protein with unknown function. Using the gene products of a hyperthermophile, Methanococcus jannaschii, we have tested the premise. The results of the test will be reviewed for three "hypothetical" proteins, where neither their functions nor the structures are known, and one protein for which its cellular function was inferred but molecular functions is not known.

Protein Families of Unknown Structure in Pfam

Alex Bateman[1] and Erik Sonnhammer,[2] [1]Sanger Centre, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SA, UK; [2]Centre for Genomics Research, Karolinska Institutet S-171 77 Stockholm, Sweden. Tel: +46-8-7286395, FAX: +46-8-337983. erik.sonnhammer @cgr.ki.se

The Pfam database contains a comprehensive collection of protein domain families defined by sequence homology. All Pfam families are linked to the PDB database by direct sequence comparison. It is thus relatively simple to use Pfam as a resource for creating a list of families of unknown 3D structure.[1] I will discuss some of the issues with creating such a list, such as completeness, the differences one might expect from domain family definitions in Pfam compared to structure-based definitions, and how to annotate the list for prioritizing targets.

Reference

1. Elofsson, A. and Sonnhammer, E.L. 1999. A comparison of sequence and structure protein domain families as a basis for structural genomics. Bioinformatics 15:480-500.

A Structural Genomics Pilot Project Based on the Genome of *Escherichia coli*

M. Cygler, A. Matte, Y. Li, J. Schrag, J. Sivaraman, C. Smith, V. Sauvé, R. Larocque, and S. Raymond, Biotechnology Research Institute, National Research Council of Canada, 6100 Royalmount Ave., Montréal, Quebéc, Canada H4P 2R2 and Montréal Joint Centre for Structural Biology, Montréal, Quebéc, Canada

Several structural genomics pilot projects are now underway world wide, with the eventual promise of high-throughput protein structure determination. An essential task within such projects is the development of effective methods to express, purify, and crystallize large numbers of proteins efficiently. We have initiated a pilot scale project based on gene targets selected from the genome of E. coli. Thirty-six genes have been selected for cloning initially and most of these were successfully over-expressed as soluble proteins. Target genes have been cloned as N-terminal fusions with either glutathione-S-transferase or $(His)_6$ tags. Initial affinity purification is achieved using glutathione Sepharose, or Ni-NTA resins, followed by thrombin cleavage to remove the fusion tag. Further purification using conventional FPLC ion exchange or gel filtration chromatography is then performed. Using this approach, over 20 proteins have so far been purified to apparent homogeneity as assessed by SDS-PAGE. Purified proteins are further characterized for homogeneity and suitable solution properties using a combination of dynamic light scattering, electrophoretic methods, mass spectrometry, and limited proteolysis. Purified protein samples are screened for initial crystallization conditions using a sparse-matrix approach. To follow the progress of various genes and to store all relevant experimental data required development of a specialized database to store these data. Web-based software for displaying and searching the information from this local database and for cross-referencing with other genomics databases has been developed.

To date, some crystals have been obtained for more than half of the purified proteins. Of these, diffraction quality crystals were obtained for five proteins. Using SeMet substituted proteins we have at this time determined the structures of three of these proteins. Another group has reported the structure of the fourth protein in the meantime. We will present the statistics related to various steps of the process and summarize the current results.

A Structural Genomics Pre-pilot Project: Study of 20 Yeast's ORFs

Sophie Quevillon-Cheruel,[1] Sylvie Auxillien,[1] Michel Desmadril,[1] Philippe Minard,[1] Joël Janin,[2] [1]Laboratoire de Modélization et d'Ingénierie des Protéines - Orsay, France; [2]Laboratoire d'Enzymologie et Biochimie Structurales — CNRS — Gif-sur-Yvette; France Bernard LABEDAN - Robert AUFRERE - Gilles HENCKES, Institut de Génétique et Microbiologie - Orsay, France

Among the about 6,200 ORFs of the yeast genome, we have selected 282 unique proteins and 202 families of paralog proteins with size ranging from 100 to 500 residues. These sequences were also selected in such a way that they correspond to cytoplasmic proteins. In this first list of proteins, we have chosen 20 ORFs to develop an efficient and cheap method for their cloning, overexpression in E. coli, purification, and crystallization. This method will be further applied to the other selected ORFs.

The 20 ORFs were amplified by PCR using S. cerevisiae genome as a template. Various constructions have been tested, by adding a 6His Tag either in C-terminus or N-terminus of the protein. The PCR products were then cloned into either pET9 or pET29 vector.

The proteins were overexpressed in B834(DE3)pLysS methionine auxotrophe strain, after induction by IPTG and using methionine. The optimized expression conditions and the solubility of the ORFs were tested. Under the selected conditions of expression induction, no significant difference was observed for the two vectors used: 1/3 of the ORFs are well expressed and soluble, 1/3 are expressed but not soluble, and 1/3 are not expressed. The localization of the His-Tag has no effect, neither on the level of expression, nor on the solubility of tested proteins.

The soluble proteins were purified by affinity chromatography (NiNTA) and gel filtration. The purity and integrity of the proteins were tested by SDS-PAGE and mass spectrometry. The conditions of crystallization for these proteins are in the course of development.

The Berlin-based "Protein Structure Factory" Project

U. Heinemann,[1,2] K.P. Hofmann,[3] G. Illing,[4] C. Lang,[5] C. Maurer,[6] H. Oschkinat,[7,2] W. Sanger[2] and M. Schroedter,[8] [1]Forschungsgruppe Kristallographie, Max-Delbrück-Centrum für Molekulare Medizin, Robert-Rössle-Str. 10, D-13125 Berlin, Germany; [2]Institut für Chemie, Freie Universität, Takustr. 6, D-14195 Berlin, Germany; [3]Institut für Medizinische Physik und Biophysik, Klinikum Charité der Humboldt-Universität, Ziegelstr. 5-9, D-10098 Berlin, Germany; [4]BMBF-Leitprojekt Proteinstrukturfabrik, Heubnerweg 6, D-14059 Berlin, Germany; [5]Fachgebiet Mikrobiologie und Genetik, Technische Universität, Gustav-Meyer-Allee 25, D-13335 Berlin, Germany; [6]Ressourcenzentrum im DHGP, Heubnerweg 6, D-14059 Berlin, Germany; [7]Forschungsinstitut für Molekulare Pharmakologie, Alfred-Kowalke-Str. 4, D-10315 Berlin, Germany; [8]Alpha Bioverfahrenstechnik GmbH, Im Biotechnologiepark, D-14943 Luckenwalde, Germany. Heinemann@MDC-Berlin.de

Structural genomics aims at the determination of the 3D structure of all proteins.[1] This international initiative follows as a logical consequence from the various genomic sequencing projects and can be seen as a subspecialty of the emerging science of functional genomics. The basic idea behind structural genomics is to determine by X-ray crystallography or NMR spectroscopy protein structures representing all protein families present in the biosphere and thereby allowing the homology modelling of virtually every protein structure. In order to finish this project within a reasonable time, methods for high-throughput structure analysis have to be developed. We hope that a comprehensive set of representative protein structures will have an important impact on biology and greatly accelerate rational drug development.

The Berlin-based Protein Structure Factory (PSF) contributes to the international structural genomics initiative. The PSF[2] is distinguished from other projects in this field by emphasis on technology development, especially concerning protein production, sample preparation, and data acquisition, use of X-ray crystallography and NMR spectroscopy, focus on human proteins, and targeting of predicted novel folds and potential drug targets.

References

1. www.structuralgenomics.org/main.html
2. U. Heinemann, J. Frevert, K.-P. Hofmann, G. Illing, H. Oschkinat, W. Saenger, and R. Zettl. In Genomics and Proteomics, S. Suhai (ed.). Kluwer Academic/Plenum Publishers, New York, p. 179-189, 2000.

High Throughput Expression for Structural Analysis: Pitfalls and Prospects

Owen Jenkins, Gene Expression Sciences, SmithKline Beecham Pharmaceuticals Research & Development, New Frontiers Science Park (North), 3rd Avenue, Harlow, Essex, CM19 5AW, UK

Current methods to express and purify proteins for crystallography are routinely performed in relatively low throughput mode, protein by protein. The absolute goal of producing crystallographic grade material on a limited number of idiosyncratic target proteins allows for time-consuming, reiterative approaches to expression, i.e., re-working of constructs, change of expression systems, novel purification and refolding protocols until ultimately, successful crystallization is achieved. To transform this process into a truly high throughput mode, for example 1,000 proteins per year, requires a complete rethinking of the philosophy and methodology of expression. This presentation will attempt to address the issues and feasibility of truly high throughput expression and purification using current technologies and those that may be available in the near future.

Crystallization for Structural Genomics: What We Have and What is Missing

Naomi E. Chayen, Biological Structure and Function Section, Division of Biomedical Sciences, Imperial College School of Medicine, London SW7 2AZ, UK

The subject of protein crystallization has gained a new strategic relevance in the next phase of the genome project in which X-ray crystallography will play a major role. The ability to express, purify, and crystallize large numbers of proteins will determine the success of structural genomics yet, even in cases where expression and purification are well under way, one often gets stuck at the stage of attempting to produce high quality crystals. Automation is crucial for crystallization (as well as for the other phases of structural genomics) since screening of numerous potential conditions is the first step in the search for crystals. Major effort and resources are currently being invested into automatic generation of high throughput crystallization trials. However, in spite of the ability to generate numerous trials and the manpower involved, so far only a small percentage of the proteins produced have led to structure determinations.

Some proteins will surely crystallize during the initial screening but many others are likely to yield microcrystals or low-ordered crystals. This is not surprising because the conversion of such crystals into useful ones requires intellectual input and individualized optimization techniques.

Dispensing crystallization trials automatically, especially for screening, is no longer a major problem. However, a number of problems need to be solved. For example: the large amount of manual preparation required prior to the actual dispensing, the issue of cleaning hundreds of syringes, and the viewing, follow-up, and analysis of the results. These stages can be and will be automated but the most important part — the optimization of the crystallization conditions for difficult cases — is more difficult to automate. It is only good methodology that has in the past solved difficult intractable crystallization problems, yet the issue of improving crystallization methods has been somewhat neglected in the rush to automate everything.

This presentation will describe simple optimization methods that have resulted in successful crystallization of proteins that could not be crystallized otherwise. These techniques have not yet been adapted as high throughput techniques, but they have the potential to become so. In combination with automated screening, the development of crystal optimization methods will equip the genome project to deal with its awesome task.

Structural Genomics at Structural GenomiX

Tom Peat, Structural GenomiX, 10505 Roselle Street, San Diego, CA 92121, USA. tom@stromix.com

Structural GenomiX was founded to capitalize on the value of protein structure information in drug and compound discovery. The company is developing a high-throughput platform to support the determination of hundreds of novel protein structures via X-ray crystallography. This platform integrates advances in genomics, X-ray crystallography, and bioinformatics. The company's approach to structure determination is genomic: families of target genes from a range of organisms are input into the platform, raising the odds of success and generating valuable information about family relationships. The company uses X-ray crystallography techniques to ensure fast structure solution, including the routine incorporation of selenium into proteins, the use of a third-generation synchrotron (the Advanced Photon Source in Illinois), and the use of MAD phasing. Bioinformatics tools enable target selection based on real-time information, and annotation that adds value for customers engaged in drug and compound discovery.

Structural GenomiX plans to make its protein structures available to pharmaceutical, biotechnology, agricultural, and other industrial customers through subscriptions to an annotated database and in strategic alliances. The speed and quantity of structures generated will enable customers to access protein structure earlier in the compound discovery process, changing target selection and bringing structure-based drug design into more universal applications.

Our process of target selection to structure determination is presented along with an overview of the current process as compared to what is being constructed for higher throughput in the near future. Recent progress towards scale up and automation is presented along with results in terms of structures completed to date.

Protein Modules — Targets for Structural Genomics and Beyond

Iain D Campbell, Department of Biochemistry, University of Oxford, South Parks Rd, Oxford, OX1 3QU, UK

A protein module can be defined as a domain with a contiguous sequence that appears repeatedly in diverse proteins. Expanding module databases, identified by multiple sequence alignments and other data, are arising from sequencing projects.[2-4] It is probable that about 4,000 families will match components of nearly all the proteins in the various genomes. Protein modules thus appear to be ideal targets for a co-ordinated program in structural genomics. We have been determining module structures, using NMR, for a number of years,[2] and current progress will be described. Since many module structures are already known, one can also consider how knowledge about module structure might be exploited in functional genomics. Illustrations of how this can be done will be given from on-going studies of proteins from the extracellular matrix and connective tissue.[4-8] Features where NMR has advantages, such as studies of ligand binding, module assembly, and module dynamics, will be emphasized.

References

1. Baron, M., Norman, D.G., and Campbell, I.D. 1991. Protein modules. TIBS 16:13-17.
2. Campbell, I.D. 1998. Modular architecture of cell-surface receptors. Immunol. Rev. 163:11-18.
3. http://smart.embl-heidelberg.de/: www.sanger.ac.uk/Software/Pfam/
4. Campbell, I.D. and Downing, A.K. 1998. NMR of modular proteins. Nature Struct. Biol. 5:496-499.
5. Bocquier A.A. et al. 1999. Solution structure of a pair of modules from the gelatin-binding domain of fibronectin. Structure 7:1451-1460.
6. Penkett, C.J. et al. 2000. Identification of residues involved in the interaction of S. aureus fibronectin-binding protein with the 4F15F1 module pair of human fibronectin, using heteronuclear NMR. Biochemistry 39:2887-2893.
7. Smith, S.P. et al. 2000. Interface characteristation of the type II module pair from fibronectin. Biochemistry 39:8374-8381.
8. Wilkins, M.B. et al. 2000. Drosophila dumpy is a gigantic extracellular protein required to maintain tension at epidermal cuticle attachment sites. Current Biol. 10:559-567.

From X-ray Maps to Protein Function

Tom J. Oldfield, Molecular Simulations Inc., University of York, UK. tom@ysbl.york.ac.uk

Recent developments in recombinant DNA techniques, crystallization protocols, X-ray data collection techniques and devices, and computing have led to a substantial increase in the speed and number of protein structure determinations in modern crystallographic laboratories. However, there remains a number of key stages in the crystallographic process that limit the rate of structure determination. One of these is fitting electron density maps, either in the initial stages of tracing a chain to a new map, or in the manual rebuilding during refinement. It is also apparent that solving a protein structure is not enough. It is necessary that an automated structure and functional classification is required to complete the process of protein structure determination.

The electron density applications already available within QUANTA represent novel and effective tools for speeding up all aspects of map interpretation. The various modules (X-AUTOFIT, X-LIGAND, X-SOLVATE, X-BUILD, and X-POWERFIT) have been developed in close collaboration with the large number of crystallographers working on projects in the Protein Group at York. These tools provide the crystallographer with automated CA-tracing, automated sequence assignment, automated model building, automated validation, automated ligand fitting, automated water fitting, structure classification, and functional classification all in a single program.

High Throughput X-ray Crystallography for Drug Discovery

Harren Jhoti, Astex Technology, UK

Astex is developing proprietary discovery platforms that will perform high throughput X-ray crystallography (HTX) to image target proteins at an unprecedented rate. A key component of HTX is a powerful new Internet-based software technology called AutoSolve which is able to determine crystal structures of protein/ligand complexes in a rapid and automated manner. Examples will be provided that highlight the performance of AutoSolve.

Underlying Methodology for High-throughput Structure Determination

Victor S. Lamzin[1] and Anastassis Perrakis,[2] [1]European Molecular Biology Laboratory (EMBL), Hamburg Outstation, c/o DESY, Notkestrasse 85, 22603 Hamburg, Germany; [2]European Molecular Biology Laboratory (EMBL), Grenoble Outstation, c/o ILL, B.P. 156, 6 Rue Jules Horowitz, 38043 Grenoble CEDEX 9, France

The vast majority of the macromolecular three-dimensional structures are nowadays determined by X-ray crystallography and it is foreseen that this technique will play a key role and will be further explored for the needs of structural genomics projects. Currently, rapid structure production is impeded by the time requirements to carry out a crystallographic experiment that may take anything from hours to years. Developments in crystallographic methodology and availability of tools that would allow determination of the macromolecular structures in a real high-throughput manner have now become one of the central goals.

There is an acute need for the re-examination of the whole process of X-ray structure determination. The data collection, phasing, model building, refinement, and validation are much more tied together than was generally believed to be, and should be considered as a single entity. The responsibility for construction of reliable macromolecular models will naturally shift from investigators to the developers of the underlying methodology.

The vast majority of the data will be recorded at synchrotron sources. The provision of on-site computational facilities directly linked with data collection will be the first step towards automation. Advances in molecular biology and availability of tuneable radiation, which enabled the success of the MAD/SAD technology, as well as the rapidly growing database of macromolecular structures are essentially re-defining the concept of the crystallographic phase problem and the emphasis should now be moved to obtaining higher quality X-ray data and faster and more reliable structure determination.

Several major challenges and major bottlenecks for high-throughput X-ray structure determination are the inspection of electron density maps, the construction of macromolecular models, and their refinement. The ARP/wARP suite is being developed to address these problems and may already be in a position to promote progress in automating the steps of deriving a complete structural model. Given the X-ray data extending to a resolution of 2.3Å or higher, the time required for building a protein structure can be shortened to a few CPU hours on inexpensive workstations.

Structures, Function, Weak Interactions, and NMR

Hartmut Oschkinat, NMR-supported Structural Biology, Forschungsinstitut fuer Molekulare Pharmakologie, Alfred-Kowalke-Str. 4, 10315 Berlin, Germany

Protein domains make up the structural code of life that has now gained special attention in structural genomics. However, it is difficult to read it in terms of protein function, because individual protein folds may be used for a variety of different biological tasks. In this context, an attempt to judge activities of signalling domains is given with the examples of WW, PDZ, and EVH1 domains, based on a combination of NMR and peptide library experiments. A strong component of structural genomics is the development of high-throughput technology for structure determination. Attempts to automate the NMR process within the Berlin project are outlined.

Structure-based Functional Genomics

Gaetano T. Montelione, Stephen Anderson, Daphne Palacios, Bonnie Dixon, Kristin Gunsalus, Yuanpeng Huang, Hunter Moseley, Daniel Monleon, Rajan Paranji, Parag Sahasrabudhe, G.V.T. Swapna, Roberto Tejero, Rong Xiao, and Deyou Zheng, Center for Advanced Biotechnology Medicine, Rutgers University, Piscataway, NJ 08854 USA

Genome sequencing projects have already determined nearly complete genome sequences of several organisms, including human. The products of these genes are widely recognized as the next generation of therapeutics and targets for the development of pharmaceuticals. While identification of these genes is proceeding quickly, elucidation of their three-dimensional (3D) structures and biochemical functions lags far behind. In some cases, knowledge of 3D structures of proteins can provide important insights into evolutionary relationships that are not easily recognized by sequence alignment comparisons. Thus, structure determination by NMR or X-ray crystallography can sometimes provide key information regarding protein fold class, locations and clustering of conserved residues, and surface electrostatic field distributions that connect a protein sequence with potential biochemical functions. The resulting limited set of putative biochemical functions can then be tested by appropriate biochemical assays. We are developing technologies that will significantly accelerate the process of protein structure determination by X-ray crystallography and NMR. These include bioinformatics methods for parsing novel genes into domain encoding regions, high-level "multiplexed" protein expression systems, database structures for keeping track of reagents and project data, and NMR pulse sequences, data collection methods, and expert-system software for automated analysis of protein resonance assignments and 3D structures. The goal of this work is to develop a high-throughput process for structural analysis of novel gene products on a genomic scale and to apply this in the analysis of novel gene products identified in the genome sequencing projects.

References

1. Montelione, G.T. and Anderson, S. 1999. Structural genomics: keystone for a human proteome project. Nature Struct. Biol. 6:11-12.
2. Moseley, H.N.B. amd Montelione, G.T. 1999. Automated analysis of NMR assignments and structures for proteins. Curr. Opin. Struct. Biol. 9:635 - 641.

RIKEN Structural Genomics Projects

Shigeyuki Yokoyama, Genomic Sciences Center, RIKEN Yokohama Institute, 1-7-22 Suehiro-cho, Tsurumi, Yokohama 230-0045, Japan

The RIKEN Institute has started the Structural Genomics Initiative to establish the relationship between structures and functions of proteins encoded by prokaryote and eukaryote genomes. The RIKEN Structural Genomics Initiative includes the Structurome Project (Leader, S. Kuramitsu), which is a structural genomics pilot project to determine crystal structures of as many proteins as possible from an extremely thermophilic eubacterium, Thermus thermophilus HB8, at RIKEN Harima Institute at SPring-8. The Protein Folds Project at Genomic Sciences Center (GSC), RIKEN Yokohama Institute, is to analyze structures and functions of mouse/human and plant proteins expressed from the full-length cDNAs collected and sequenced by other groups at the GSC. For this purpose, six 800-MHx and ten 600-MHz instruments have been installed. In addition to two RIKEN beam lines at SPring-8, construction of new high-throughput beam lines is planned. The cell-free protein synthesis is the major method for protein sample preparation with stable-isotope labelling for NMR and selenomethionine substitution for crystallography. A bioinformatics group headed by Y. Matsuo is also involved in the target selection and systematic analyses of the protein structures and functions.

Structural Proteomics in Prokaryote Systems

Aled M. Edwards,[1,2] Akil Dharamsi,[2] Masoud Vedadi,[2] Dinesh Christendat,[1] Adelinda Yee,[1] Cheryl Arrowsmith,[1,2] [1]Department of Medical Biophysics, Ontario Cancer Institute, University of Toronto, Toronto, Ontario, Canada; [2]Integrative Proteomics, Suite

520, 100 College St., Toronto, Ontario, Canada

Understanding the biology of an organism will require a range of genomics and proteomics approaches. Structural proteomics is an important part of the general strategy. Our academic and commercial arms have been combining structural proteomics with proteome-wide protein-protein interaction studies and data mining approaches to try to understand the biology of an archaeon and a bacterial pathogen. Highlights from these projects will be discussed.

The European Macromolecular Structure Database (EMSD) and Structural Genomics

K. Henrick, J. Ionides, P. Keller, J. Irwin, S. Velankar, and G.J. Barton, EMBL-European Bioinformatics Institute, Genome Campus, Hinxton, Cambs CB10 1SD, UK. Tel: +44 1223 494414, Fax: +44 1223 494496. geoff@ebi.ac.uk, http://barton.ebi.ac.uk

Approximately 250 new structures per month are deposited to the PDB (Protein Data Bank) collection, of these, 20% are deposited to and processed by the EMSD. Projects in structural genomics promise dramatically to increase the number of new structures deposited, many of which will be for proteins of unknown function. If these data are to be useful both to structural biologists and to the wider biological community, then it is essential that the data are saved in the public archives in a complete and accurate form and that the data are organized to allow complex questions to be answered with minimal effort. In this talk, I outline the work in progress at EBI that will allow fast and accurate deposition, sophisticated searches, and detailed cross-referencing with other EBI databases such as TrEMBL/SWISS-PROT and EnsEMBL genome annotation.

Structural Genomics: Building a Structural Foundation for Biology

Jean-Denis Pedelacq, Los Alamos National Laboratory, Los Alamos, NM 87545; The Consortium for Structural Genomics: Thomas Alber, James Berger, University of California, Berkeley; Edward N. Baker, University of Auckland; Joel Berendzen, Min Park, Tom Terwilliger, Geoffrey Waldo, Los Alamos National Laboratory; James Bowie, David Eisenberg, Juli Feigon, Jeanne Perry, Todd Yeates, UCLA; Axel Brunger, Paul Adams, Lawrence Berkeley National Laboratory; William Jacobs, Albert Einstein College of Medicine; Bernhard Rupp, Lawrence Livermore National Laboratory; James Sacchettini, Texas A&M University; Se Won Suh, Seoul National University; Manfred Weiss, Institute of Molecular Biology, Jena; Matthias Wilmans, Paul Tucker, Emke Pohl, EMBL-Hamburg; William Wood, University of Colorado, Boulder; Shigeyuki Yokoyama, RIKEN

The high-throughput determination and analysis of protein structures across whole genomes is one of the most exciting challenges in life science. The genome projects are changing biology by providing the opportunity to improve our understanding of cells, in particular, and of life, in general. Los Alamos is part of an effort to plan and promote the field of structural genomics. Participants in the 14-institution Consortium have carried out a pilot structural genomics project based on proteins from the hyperthermophile Pyrobaculum aerophilum, and are beginning a larger project to determine and analyze structures of functionally important proteins from Mycobacterium tuberculosis. The lessons learned in this pilot project will be discussed.

No Fold Recognition Method is Always Best! Results from Studies of Different Fold Recognition Methods

Arne Elofsson, Stockholm Bioinformatics Center, Stockholm University, 106 91 Stockholm, Sweden

Here we report results from two recent studies of different fold recognition methods. In the first study, we have performed the first large (10,000 pairs) test of alignment quality using several different alignment methods (local, global, profile alignment, hmmer, sam.t98, clustalW, sspsi) (Elofsson, 2000 submitted). We show that both evolutionary information and predicted secondary structure information improves the alignment quality. The best alignments are obtained from a method that combines a sequence profile obtained from psiblast with predicted secondary structures. In the second study, we present a novel, continuous approach aimed at the large-scale assessment of the performance of available fold-recognition servers (Bujnicki et al., 2000, submitted). Six popular servers were investigated: PDB-Blast, FFAS, T98-lib, GenTHREADER, 3D-PSSM, and INBGU. The assessment was carried out using as prediction targets a large number of selected protein structures released during October 1999 to April 2000. Overall, the servers were able to produce structurally similar models for one-half of the targets, but significantly, accurate sequence-structure alignments were produced for only one-third of the targets. We further classified the targets into two sets: “easy” and “hard.” We found that all servers were able to find the correct answer for the vast majority of the easy targets when a structurally similar fold was present in the server’s fold libraries. However, among the hard targets — where standard methods such as PSI-BLAST fail — we found that the most sensitive fold-recognition servers were able to produce similar models for only 40% of the cases, half of which having a significantly accurate sequence-structure alignment. Unfortunately, the increased sensitivity of the fold-recognition servers over standard methods came with the cost of low specificity.

Probably the most interesting observation from these studies is that no single method produces the best results (fold recognition or alignment). For instance, we show that almost twice as many good models can be created using any method compared with the best method for fold-related pairs and that each server had a number of cases with a correct assignment, where the assignments of all the other servers were wrong. This emphasizes the benefits of considering more than one method in difficult prediction tasks and implies that it would be possible to improve fold recognition performance significantly if a combination of several methods could be done without losing specificity.

In conclusion, we would like to encourage all protein structure predictors to take advantage of the variety of methods available. We have used novel methods to measure the quality of a model generated from a fold recognition method. We will also discuss the advantages of using these novel methods for measuring fold recognition capacity (Siew et al., 2000, in press; Cristobal et al., 2000, manuscript in preparation).

Structural Genomics in the Context of Other Genome Research

Richard Durbin, Sanger Centre, Wellcome Trust Genome Campus, Cambridge CB10 1SA, UK. rd@sanger.ac.uk

Structural genomics projects are getting going now around ten years after the large-scale sequencing programs started, at a time when many other systematic functional genomics approaches are also being started. I will review some of the history of the development of genomic sequencing that might be relevant to large scale structure data collection, and consider the context of other genomic-scale efforts, exploring how they might influence the practice and prioritization of structural genomics. Finally, the primary output of all these large- scale projects is data that must be managed in databases and accessed computationally. I will discuss how I see the requirements for future informatics resources developing to make these data available to research biologists in a maximally useful form.

Analysis of Gene Expression: Bridging the Gap from Sequence to Function

Tom Freeman, Sanger Centre, Hinxton, UK

The potential to correlate the genetic makeup of an organism to its biological function is moving into a new era. This is primarily being driven by the acquisition of the sequence of all the genes by the large-scale cDNA and whole genome sequencing programs. However even now, with approximately ninety percent of sequencing of the human genome completed, our knowledge of the transcriptome is still in its infancy. There remain great discrepancies in the estimates of the total number of mammalian genes, and the expression pattern of most genes and the function of the proteins they encode is largely unknown.

Knowledge of where and when a gene is expressed can provide valuable insights into the function of the encoded protein. If the expression of a gene can be shown to be restricted to a given tissue or cell type, then the protein's function is highly likely to contribute to the specific physiology of that system. Knowing a gene's expression pattern and comparison to that of others, also allows for the association of the function of one gene with that of another, as genes involved in the same pathway or protein complex, often exhibit highly similar expression profiles. Finally, expression profiling can now provide unparalleled insights into molecular mechanisms regulating biological systems. As the regulation of transcription is one of the primary controls of biochemical function, monitoring of the transcriptome during a change in the functional status of the system can, without any prior knowledge or hypothesis, reveal which genes may be regulating or underlying these changes. I will outline some of the approaches to the analysis of gene expression that we have been using and discuss their utility in revealing new insights into gene function and the biology of complex systems.

Functional Genomics Using 3D Structures: from ORFANs to Unknowns

Chantal Abergel, Vincent Monchois, Christian Cambillau,[1] J.-M. Claverie, Information Structurale et Génétiques et [1]Architecture et Fonction des Macromolécules Biologiques, 31 Chemin Joseph Aiguier, 13402 Marseille Cedex 20, France. http://igs-server.cnrs-mrs.fr, http://afmb.cnrs-mrs.fr

Newly sequenced microbial genomes routinely reveal up to 50% of genes without significant similarity to previously characterized genes and thus without any functional attribute. Our hope is that an approach combining large-scale bioinformatics and 3D structure determinations can shed some light on the function (and pharmaceutical relevance) of unknown "anonymous" genes. An overview of our projects and methods, as well as preliminary results on a pilot study of E. coli ORFAN genes, will be presented.

References

1. Abergel, C., Bouveret, E., Claverie, J.-M., Brown, K., Rigal, A., Lazdunski, C., and Benedetti, H. 1999. Structure of the Escherichia coli TolB protein determined by MAD at 1.95Å resolution. Structure 7:1291-1300.
2. Alimi, J.-Ph., Poirot, O., Lopez, F., and Claverie, J.-M. 2000. Reverse transcriptase-polymerase chain reaction validation of 25 "orphan."
3. Genes from Escherichia coli K-12 MG1655. Genome Res. 10:959-966.
4. Cambillau, C. and Claverie, J.-M. 2000. Structural and genomic correlates of hyperthermostability. J. Biol. Chem. (in press).
5. Ogata, H., Audic, S., Barbe, V., Artiguenave, F., Fournier, P.-E., Raoult, D., and Claverie, J.-M. 2000. Selfish DNA in protein coding genes of Rickettsia. Science (in press).

$(\beta/\alpha)_8$ Barrel Enzymes and the Evolution of Function and Pathways

Richard R. Copley, EMBL, Meyerhofstrasse 1, 69012 Heidelberg, Germany

We provide statistically reliable sequence evidence indicating that at least 12 of 23 scop $(\beta/\alpha)_8$ (TIM) barrel superfamilies share a common origin. This includes all but one of the known and predicted TIM barrels found in central metabolism. The statistical evidence is complemented by an examination of the details of molecular function and protein structure, with certain structural locations favoring catalytic residues even though the nature of their molecular function may change. The combined analysis of sequence, structure, and function also enables us to propose a phylogeny of TIM barrels. Based on these data, we are able to examine differing theories of pathway and enzyme evolution by mapping known TIM barrel folds to the pathways of central metabolism. The results favor widespread recruitment of enzymes between pathways, rather than a "backwards evolution" model, and support the idea that modern proteins may have arisen from common ancestors that bound key metabolites.

Practical Limits of Function Prediction

Damien Devos and Alfonzo Valencia, Protein Design Group, CNB-CSIC, Cantoblanco, Madrid E-28049, Spain. http://montblanc.cnb.uam.es/

The widening gap between sequences and functions has lead to the practice of assigning a potential function to an uncharacterized protein based on sequence similarity with other proteins of experimentally investigated function. Even if the reliability of those homology-based functional assignments is not well characterized, it represents common practice in whole genomes functional assignments. We propose here a systematic approach to the study of the margins of error in homology-based functional prediction by analyzing the conservation of the functional annotations in a large set of structural alignments. In particular, we analyze five aspects of protein function, commonly used in genome annotation, namely:

- PDB header line
- Enzymatic function classification: DE code, the standard definition of the chemical nature of the enzymatic function
- Functional annotations in the form of keywords, describing the biochemical function such as the interactions with compounds, cofactors, substrates, regulators, and other cellular components
- Classes of cellular function, capturing the main types of cellular activities in which proteins participate, e.g., carbon compound metabolism or DNA biosynthesis
- Conservation of the type of amino acid in the binding site, related with the binding activity of the protein, and in many cases, the specificity of binding different substrates and cofactors.

The screening of the full range of sequence functional similarities allows us to present an initial picture of the relation between sequence and functional similarity, and in particular, to derive a theoretical error rate for homology-based functional assignments.[1] With those data, we estimate the theoretical error rates of predicted functions in different genomes. Indeed, it is particularly interesting to think of the consequences of this study for whole genome annotations carried out by automatic systems[2] and to compare the expected level of error with the different values published by different groups of expert annotators.[3-5]

References

1. Devos, D. and Valencia, A. Proteins 41:98-107.
2. Andrade, M.A. et al. 1999. Bioinformatics 15:391-412.
3. Brenner, S. 1999. Trends Genet. 15:132-133.
4. Galperin, M.Y. and Koonin, E.V. 1998. In Silico Biol. 1:0007.
5. Ouzounis, C.A. et al. 1996. Mol. Microbiol. 20:985-900.

Further Developments Towards Reliable Genome-scale Fold Recognition

David T. Jones[1] and Caroline Hadley,[2] [1]Institute for Cancer Genetics and Pharmacogenomics, Department of Biological Sciences, Brunel University, Uxbridge, Middlesex, UK; [2]Department of Biological Sciences, University of Warwick, Coventry, UK

Protein fold recognition by threading has proven to be a very effective means for predicting protein tertiary structure from sequence, as witnessed by the number of successful threading predictions made in the various CASP prediction experiments.[1] Despite this success, the better fold recognition methods still often employ some degree of human expert intervention, which is clearly impractical if these methods are going to be applied to the annotation of uncharacterized genome sequences.

We have already described a method for identifying distant homologs to known 3D structures[2] using a combination of traditional sequence profile alignments, a set of potentials similar to those used for full optimal sequence threading, and a neural network based expert system. This very quick approach to fold recognition, whilst not being capable of recognizing analogous fold relationships, is very successful in reliably recognizing homologous fold similarities. Recently we have been exploring further developments of this method to extend its range both towards more distant evolutionary relationships (using "structure-function fingerprints") and towards analogous fold relationships using a new version of our threading program (THREADER 3) and a recently developed post-processing step, again involving neural networks. Preliminary results from both these new approaches will be discussed.

References

1. Jones, D.T., Tress, M., Bryson, K., and Hadley, C. 1999. Successful recognition of protein folds using threading methods biased by sequence similarity and predicted secondary structure. Proteins S3:104-111.
2. Jones, D.T. 1999. GenTHREADER: An efficient and reliable protein fold recognition method for genomic sequences. J. Mol. Biol. 287:797-815.

Exploiting Protein Structure in Genome Annotation

Michael Sternberg,[1] Patrick Aloy,[1,2] Francesc Xavier Aviles,[2] Paul Bates,[1] Lawrence Kelley,[1] Robert MacCallum,[1] Arne Mueller,[1] Enrique Querol,[2] and Mansoor Saqi,[1] [1]Biomolecular Modelling Laboratory, Imperial Cancer Research Fund, 44 Lincoln's Inn Fields, London WC2A 3PX, UK. www.bmm.icnet.uk; [2]Institut de Biologia Fonamental and Departament de Bioquimica, Universitat Autonoma de Barcelona, Bellaterra 08193. Barcelona, Spain. m.sternberg@icrf.icnet.uk

A strategy to annotate the structure and function of protein coding regions in genomes will be described. We have completed an initial characterization using standard programs such as PSIBLAST. Our plans are to use a method for fold recognition (3D-PSSM)[1] to identify remote homologies. Three-dimensional models for proteins will be constructed using our program 3D-JIGSAW.[2] Both of these programs can be used via web servers (www.bmm.icnet.uk). A strategy will be outlined to facilitate the interpretation of protein function from structure. To begin to include a high level view of protein function into structure-based genome annotation, we have analyzed the relationship between the conformation of a proteins and its assignment to metabolic pathways.

References

1. Kelley, L. et al. 2000. J. Mol. Biol. 299:501-522.
2. Bates, P. and Sternberg, M. 1999. Proteins S3:47-54.

The Evolution and Structural Anatomy of the Small Molecule Metabolic Pathways in *Escherichia coli*

Sarah A. Teichmann,[1] Stuart C.G. Rison,[2] Janet M. Thornton,[1,2] Monica Riley,[3] and Cyrus Chothia,[4] [1]Department of Biochemistry and Molecular Biology, University College London, Darwin Building, Gower Street, London WC1E 6BT, UK; [2]Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX, UK; [3]Josephine Bay Paul Centre for Comparative Molecular Biology and Evolution, 7 MBL St., Woods Hole, MA 02543-1015 USA; [4]MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB, UK

For the first time, sufficient sequence, structure, and functional data is available for a thorough examination of all the small molecule metabolic pathways of an organism in terms of the protein families of the enzymes. With information on the domain structure and evolutionary relationships of over three-quarters of the gene products in E. coli metabolic pathways, we can determine the extent to which domains are duplicated within and across pathways and are combined to form multi-domain enzymes. We have examined which functional features are conserved in families of homologs and thus shed light on the evolution of pathways and enzymes.

What Can Structure Tell Us About Bioinformatics?

Peter J. Artymiuk, University of Sheffield, UK

Bioinformatics will be of immense value in guiding the formulation of strategy for structural genomics initiatives and for the attribution of putative functions to the structures of proteins of unknown function. However, our present bioinformatics tools are far from perfect and also must be refined in the light of new structures.

Evolution of Function in Protein Superfamilies, from a Structural Perspective: Implications for Genome Annotation

Annabel E. Todd,[1] Christine A. Orengo,[1] and Janet M. Thornton,[1,2] [1]Department of Biochemistry and Molecular Biology, University College London, Darwin Building, Gower Street, London WC1E 6BT, UK; [2]Department of Crystallography, Birkbeck College, Malet Street, London, WC1E 7HX, UK

The recent growth in protein sequence and structural databases has revealed the functional diversity of many protein superfamilies. An understanding of how such diversity has evolved through sequence and structural changes is essential for the accurate functional annotation of the large number of uncharacterized gene products identified in genome sequencing projects. Given the large number of genes in the human genome, but a comparatively small number of folds, extensive combination, mixing, and modulation of existing folds has occurred during evolution to generate the multitude of functions necessary to sustain life. With the first working draft of the human genome complete, and the sequencing of other multi-cellular organisms underway, a grasp of these evolutionary processes is required if we are to benefit from this wealth of data.

We have analyzed how functional changes are implemented by modulation of sequence and structure with reference to 31 diverse enzyme superfamilies, and thus provide an overview of the mechanisms by which functional diversity has evolved. This has involved extensive reading of the literature combined with analyses of our own. Functional variation occurs mostly in more distantly related proteins (<40%) and the structural data have been essential for understanding the molecular basis of observed functional differences. A large number of variations and peculiarities are observed, at the atomic level through to gross structural rearrangements. Using selected examples, we present the structural and functional attributes that are conserved within some superfamilies and those that differ, and what bearing, if any, these similarities and changes have on protein function. The implications these observations have on structural genomics projects will be discussed.

New Developments Concerning the Swiss-PdbViewer Sequence to Structure Workbench

Nicolas Guex, Torsten Schwede, Alexander Diemand and Manuel Peitsch, GlaxoWellcome Experimental Research S.A. 16, chemin des Aulx, CH-1228 Plan-les-Ouates, Geneva, Switzerland

Initially, Swiss-PdbViewer (SPDBV; www.expasy.ch/spdbv/) was developed as a protein viewer, running only on the Macintosh platform. Over time, it evolved toward a cross-platform program with the same functionality and interface on MacOS, Windows, IRIX, and Linux.

As the program provides an interface to SWISS-MODEL as well as a large set of features (from basic display and measurement tools to computation of molecular surfaces, electrostatic potential, force field energy, 3D structure superposition, rotamer scanning, loop building, etc.), it has been widely adopted for teaching and routine work. However, one main limitation was the absence of scripting language. Thus, so far, it was not possible to automate tasks with SPDBV.

Four options have been considered to overcome this:

Release the Source Code

Although it would seem appealing to some users, it would automatically exclude several possible users who would not want to deal with C code and compilers. Moreover, the UNIX versions need to be linked with a commercial library, while dialogs and menus are taken directly from the Mac. Thus, the setup of a complete development environment is relatively complex. The risk that options are added in a nonsynchronous way among platforms is also present.

Releasing a Library and Providing an API:

Same problems as for the previous option.

Providing Macros Connected to the User Interface, Allowing to "Record" and "Play Back" Commands:

This option would be limited to the features already in place and would virtually require freezing the user interface to its current state, possibly limiting future extensions.

Provide a Complete Scripting Language

This option would let users automate repetitive tasks, and provide a way of adding new commands.

The fourth option was retained, and a complete interpreted language inspired from C and Perl syntax was developed using flex and yacc. The language supports variables, arrays, conditional branching, loops, access to external files, and to some extent, subroutines. Several key features of the program are already accessible via scripting, in a "natural language" way, and more will be added in the future. We think that this option will allow more people to contribute than option 1 or 2 (as no intimate knowledge of internal data structures or how to compile and link large projects is required). It should also be more appealing to developers than option 3, as it is more powerful and permits adding commands and functions to the program.

We hope to promote script sharing through publication in the SPDBV mailing list in the first place, and by maintaining a web database of scripts with description of their function and proper author credits if there is a growing interest.

Poster Abstracts

Determining Function from Structure: Application to a *Pyrobaculum aerophilum* Protein

Jean-Denis Pédelacq,[1] Elaine C. Liong,[1] Beom-Seop Rho,[1] Chang Y. Kim,[1] Kevin C. Menes,[1] Trevin Zyla,[1] Lisa Cornelius,[1] Min S. Park,[1] Sorel Fitz-Gibbon,[2] Jeffrey H. Miller,[2] Joel Berendzen,[1] and Tom Terwilliger,[1] [1]Los Alamos National Laboratory, Los Alamos, NM 87545 USA; [2]University of California, Los Angeles, CA 90095 USA

With the increasing number of genomes sequenced, one of the most challenging hurdles is to decipher the information encoded by these protein sequences. Sequence homology searches do not always provide all of the answers, as some proteins may not have retained sequence homology throughout evolution. On the contrary, the function of a protein is closely linked to its three-dimensional structure. Therefore, structural determination of the libraries of protein structures is a matter of high priority to overcome the obstacles involved. Los Alamos National Laboratory is part of a multi-laboratory effort to plan and promote the field of structural genomics. Along with researchers at University of California, Los Angeles, University of California, Berkeley, Lawrence Livermore National Laboratory, Pacific Northwest National Laboratory, Lawrence Berkeley National Laboratory, Caltech, and the University of Auckland we have carried out a structural genomics pilot project on the hyperthermophile, Pyrobaculum aerophilum.

Over the past two years, 130 proteins from P. aerophilum have been produced by members of the consortium. We have recently purified and crystallized one of these proteins for which no known function was available. The crystal of this 243 amino acid residue protein belongs to the orthorhombic space group C2 with cell parameters a = 161.8Å, b = 48.6Å, c = 60.1Å, and two molecules in the asymmetric unit. A MAD experiment was performed on a single crystal at 100 K using the NSLS beamline X8C (Brookhaven, NY). Complete highly redundant data were collected to a resolution of 2.1 Å. All selenium sites were found using the program SOLVE and the initial electron density map was of good quality. The experimental phases were improved by subsequent cycles of solvent flattening in the new RESOLVE program. Determining the function of this unknown protein was conducted using structural homology searches.

Directed Structural Genomics

James H. Naismith, Centre for Biomolecular Sciences, BMS Laboratories, The North Haugh, The University, St. Andrews, Fife Scotland, KY16 9ST UK, Tel: +44 1334-463792, Fax: +44 1334-467229. http://speedy.st-and.ac.uk/

Mycobacteria survive by assembling a complex coat. This coat is linked to the peptidoglycan by sugar molecules. We are determining the enzymes involved in coat assembly and synthesis. Thus far, we have determined five new enzyme structures by MAD techniques, shedding light on the chemistry of this process. The genes for capsule, cell wall, and LPS biosynthesis are grouped together on the bacterial chromosome, presenting an attractive target for a focused program in structural genomics. These proteins are all potential therapeutic targets and a structural understanding will aid research in this area in addition to assisting annotation of function.

Structure-based Functional Classification of Proteins for Structural Genomics

Michael A. Kennedy, John R. Cort, Aled M. Edwards, Cheryl H. Arrowsmith, Macromolecular Structure and Dynamics, Environmental Molecular Science Laboratory, Pacific Northwest National Laboratory, EMSL 2569 K8-98, Richland, WA 99352. Tel: 509-372-2168, Fax: 509-376-2303. ma_kennedy@pnl.gov

As part of a pilot study in structural genomics, we have used NMR to determine the solution-state structure of several hypothetical or uncharacterized proteins from various organisms. In each case, functional annotation of the protein prior to structural characterization was not possible from sequence analysis or comparison to sequences with known structure and function. We show several examples where knowledge of the protein structure immediately narrowed the related fold classes to one or two possibilities. We find that analyses of the functions represented for related fold classes frequently leads to hypotheses about the function of each protein that are easily testable using NMR methods. In one case, a MTH538 from Methanobacterium thermoautotrophicum (M. therm), a singleton in sequence space, was found to have a fold similar to flavodoxin and CheY. NMR mapping experiments indicated that MTH538 lacked flavin binding activity, ruling out a function related to a flavodoxin; however, it was found to have weak Mg2^{+} binding affinity. Collectively, sequence and structural analyses together with NMR mapping results indicated that MTH538 might represent a phosphorylation-independent receiver domain in a two-component signal transduction system. In another example, the structure of MTH1175 from M. therm was determined and its fold was found be similar to only one existing fold class according to CATH and SCOP, the RNaseH family. However, the fold of MTH1175 differs sufficiently from the RNaseH family that it may be designated as a novel fold. Unlike MTH538, MTH1175 is a member of a conserved family of proteins primarily of archaeal origin (COG1433), so the analysis of the structure and function of the protein can be carried out in the context of, and related to, the other members of the protein family from other organisms. In a third example, YciH from E. coli was found to represent a new protein superfamily similar to the ferredoxin-like topology (CATH) or DcoH-like fold (SCOP). Consequently, YciH now falls under a new fold classification in SCOP, the eIF1-like fold. Since YciH falls into a very common fold class, represented by more than 20 related superfamilies in SCOP and CATH, the YciH example illustrates how functional classification can be limited in highly populated fold spaces. However, it offers an opportunity to explore ancestral relationships between related folds including (i) convergent evolution, resulting in proteins with common structures but unrelated amino acid sequences or (ii) divergent evolution, resulting in expanded function for proteins originating from a common protein ancestor. In many cases, the putative active site in proteins occurs in highly dynamic regions that are well suited for characterization by NMR methods and that might not be as easily characterized by X-ray crystallographic methods. Because NMR provides a capability for rapid NMR mapping of ligand binding sites together with the ability to characterize dynamic regions of proteins, NMR should be considered a critical technology for post-structural genomics studies. The potential role that NMR can play in bridging the gap between structural and functional genomics and its relationship to proteomics will be discussed.

GeneAtlas — An Automatic High-throughput Pipeline for Structure Prediction and Function Assignment for Genomic Sequences

Lisa Yan, Zhan-Yang Zhu, Azat Badredinov, David Kitson, Krzysztof Olszewski, and David Edwards, Molecular Simulations, Inc., 9685 Scranton Road, San Diego, CA 92121 USA. lly@msi.com

With the vast amount of protein sequences determined from the genomic sequencing project, there is an emerging need to use a high-throughput method to predict structures and assign functions of the protein sequences. GeneAtlas is an automated, high throughput pipeline for the prediction of protein structure and function using sequence similarity detection, homology modelling, and fold recognition methods. It uses PSI-BLAST and SeqFold to search for homologous structures from PDB database and MODELER to build 3D models for the sequences based on the template structure. The quality of the 3D model is validated using Profile-3D/verify score. The accepted model gives the correct fold and indicates the possible function of the genome sequence from the known template. Furthermore, protein 3D structure contains much richer information for its function than sequence alone. Functional annotations based on the 3D model using a suite of methods give further details of the protein function that are crucial for target discovery, protein engineering, and inhibitor design. Using a “virtual” genome, a subset of PDB structures from SCOP database that consists of protein structures of less than 40% sequence identity, as a benchmark, we demonstrate that GeneAtlas detects additional functional relationships by building 3D models for genomic sequences in comparison with the widely used sequence searching method, PSI-BLAST. The method was applied to 22 publicly available genomes, including C. elegans, D. melanogaster, S. cerevisiae, H. sapiens, A. thaliana, etc. The modelling results of a small genome, M. genitalium, and the comparison with PSI-BLAST and Hidden Markov Model (HMMer/pfam) on function assignment will be discussed.

Toward the Crystal Structure of a Lectin-like Natural Killer Cell Activator Receptor Bound to its MHC Class I Ligand

Susana Cristóbal, Ylva Lindqvist, and Gunter Schneider, Molecular Structural Biology Group, Medical Biochemistry and Biophysics, Karolinska Institute, Tomtevodavägen 6 171 77 Stockholm, Sweden

Natural killer (NK) cell function is regulated by NK receptors that interact with MHC class I molecules on target cells. The murine NK receptor Ly-49D activates NK cells activity by interacting with H-2Dd through its C-type-lectin-like NK receptor domain. The activating Ly-49D receptor and the inhibitory Ly-49A receptor mediate opposing effects on NK cell cytotoxicity. The missing self hypothesis predicts that all NK cells express at least one inhibitory receptor for a self MHC I antigen, allowing NK cells to delete cells with lost or altered self MHC expression. Thus, the very existence of NK activating receptors specific for MHC class I remains somewhat perplexing, and the exact physiological role of these receptors has not been clarified. We are trying to co-crystallizate activating receptor and ligand to elucidate those questions by providing basis for an analysis of the interaction in the activating function. We have already overexpressed and purified MHC class I (H-2Dd) and we are trying to obtain recombinant Ly-49D in a soluble form using a novel overexpression approach. In prokaryotes, the recently characterized TAT pathway can transport folded substrates across the inner membrane, and signal peptides specific for this pathway bears a twin arginine motif. I have demonstrated that a folded large cytoplasmic domain of a non-TAT protein can be translocated by this machinery by fusion to a TAT signal sequence (Cristóbal et al., 1999) and we are using this strategy to overexpress and improve the solubility of the Ly-49D receptor.

Prediction of the 3D Structure for Proteins with Tandem Repeat Sequences

Andrey V. Kajava, Center for Molecular Modelling, CIT, National Institutes of Health, Bldg. 12A, 12 South Drive MSC 5626, Bethesda, MD 20892 USA

The genome sequencing projects have revealed a considerable number of protein sequences with tandem arrays of 10- to 30-residue repeats. Despite the established functional importance of many such proteins, only a few of their 3D structures are known. The lack of the structural information is explained by the fact that large molecular weight and elongated shape of these molecules hamper X-ray and NMR studies. On the other hand, inspection of the known structures suggests that structural prediction of such proteins can be more reliable than prediction of aperiodic globular proteins. The prediction can be facilitated by assuming repetitive spatial arrangements within the tandem repeat sequence and by more reliable distinguishing of structurally important residue positions in the repeats. Prediction and modelling of several proteins with 10- to 30-residue repeats, such as leucine-rich repeat proteins, human involucrin, and bacterial surface-associated adhesins are described. The modeled structures incorporate constraints from electron microscopy, circular dichroism, and other indirect structural experiments. The approach used for prediction and modelling of these proteins can be applied to other proteins containing internal repeats and will lead to a valuable tool of structural bioinformatics.

Solution Structure of the Tyrosyl-tRNA Synthetase C-terminal Domain: A Novel Type of Anticodon Binding Module

A. Pintar,[1] A. Prochnicka-Chalufour,[1] V. Guez,[2] C. Castagne,[1] H. Bedouelle,[2] and M. Delepierre,[1] [1]Laboratoire de RMN (CNRS URA 2185); [2]Unite de Biochimie Cellulaire, Institut Pasteur, 28 rue du Dr. Roux, 75724 Paris Cedex 15, France

Tyrosyl-tRNA synthetase (TyrRS) is a homodimeric protein that catalyzes both the activation of the amino acid through its reaction with ATP and the transfer of the aminoacyl-adenylate to the tRNA(tyr). In Bacillus stearothermophilus, each subunit of TyrRS comprises two structural domains, an N-terminal domain (residues 1-319), whose crystal structure is known, and a C-terminal domain (residues 320-419) that appears disordered in the crystals. The binding site of one tRNA-Tyr molecule encompasses both subunits of the TyrRS dimer. The folding state of the C-terminal fragment was characterized in solution by biophysical techniques and compared with those of full-length TyrRS. We present here the 3-dimensional structure of a recombinant protein TyrRS(de4) corresponding to the C-terminal domain of TyrRS solved by heteronuclear NMR spectroscopy. We suggest that the disorder observed in the crystal structure is due to a flexible linker between the N- and C-terminal regions. The TyrRS(de4) structure exhibits a novel fold among the anticodon binding domains of aminoacyl-tRNA synthetases, around two thirds of that appear to be shared with the ribosomal protein S4 and a heat shock protein HS. The common topology involves two α helices packed against an antiparallel ß sheet. Of six basic residues identified by site directed mutagenesis as essential for tRNA binding, four are clustered in this domain and are likely to interact with the anticodon arm of tRNA.

The Defining Characteristics of Immunoglobulin-like Proteins

A.E. Kister and I.M. Gelfand, Department of Mathematics, Rutgers University, 110 Frelinghuysen Rd., Piscataway, NJ, USA, Tel: 732-445-3478, Fax: 732-445-5530. akister@math.rutgers.edu

The main goal of this work is the analysis of the general relation between sequence and structure of immunoglobulins, i.e., analysis of sequence features that are consistent with a structure. Our recent investigations show that distantly related proteins (with no significant homology) that share one type of immunoglobulin fold also share a small set of residues at the same positions. This set of residues constitutes the defining characteristics for the immunoglobulin fold. Residues at each key position are chemically related and play approximately the same structural role in all proteins (residue-residue contacts, surface exposure, and other features, across all proteins, as well as almost identical coordinates in the system of coordinates unified for the protein family or fold. The result of the energy calculations in the threading test show that these residues have the decisive role in structure stability and they are sufficient for Ig fold recognition.

Knowledge of the distinguishing characteristics allows one to compare sequences with a low similarity (<20%) and, hence, assign these sequences to a proper protein fold by using several key residues only. In fact, it is not necessary to know all or almost all residues in a sequence as required for other traditional tools such as BLAST, FASTA, and HMM. Based on this analysis, a new method of protein classification was developed. The basic idea behind this method is that residues at key positions are taken into account only, all other residues are out of consideration. We will present the results of the classification using the defining characteristics for the different superfamilies of the immunoglobulin folds.

Accurate Formula for P-values of Gapped Local Sequence and Profile Alignments

Richard Mott, University of Oxford Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK. www.well.ox.ac.uk/~rmott/ariadne.html

We present a simple general approximation for the distribution of gapped local alignment scores, suitable for assessing significance of comparisons between two protein sequences or a sequence and a profile. The approximation takes account of the scoring scheme (i.e., gap penalty and substitution matrix or profile), sequence composition, and length. Use of this formula means it is unnecessary to fit an extreme-value distribution to simulations or to the results of databank searches. The method is based on our theoretical ideas.[1,2] Extensive simulation studies show that score-thresholds produced by the method are accurate to within ±5%, 95 percent of the time.

References

1. Mott, R. and Tribe, R. 1999. J. Comp. Biol. 6:91-112.
2. Mott, R. 2000. J. Mol. Biol. 300:649-659.

NMR Structure Determination and Ligand Screening of Hypothetical Proteins from *Haemophilus Influenzae*

Lisa Parsons, Nicklas Bonander, Edward Eisenstein, and John Orban, Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute, 9600 Gudelsky Drive, Rockville, Maryland 20850 USA

As the number of genomes sequenced continues to increase, a recurring observation is that approximately 20 to 40% of the open reading frames in these genomes have no known function.[1] These hypothetical proteins have no sequence homology with proteins of known function and typically have homologs in other organisms from bacteria to eukaryotes. Consequently, any information that can be obtained on these proteins will be important in understanding their role in a wide range of biological systems and will fill a large gap in knowing what is required for the viability of a free-living organism. The goal of this project is to obtain structures for soluble proteins in this hypothetical category and to narrow down potential biochemical functions that can then be assayed by other methods.[2] Examples from our current structural work on a number of Haemophilus influenzae proteins will be discussed together with results from small molecule ligand screening using NM Rmethods.

References

1. Fleischmann, R.D. et al. 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 269:496-512.
2. Eisenstein, E., Gilliland, G.L., Herzberg, O., Moult, J., Orban, J., Poljak, R.J., Banerjei, L., Richardson, D., and Howard, A.J. 2000. Biological function made crystal clear — annotation of hypothetical proteins via structural genomics. Curr. Opin. Biotechnol. 11:25-30.

Docking Large Proteins Using Spherical Polar Fourier Correlations

Russell S. Hamilton,[1] David W. Ritchie,[1,2] Graham J.L. Kemp,[1]
[1] Department of Computing Science, University of Aberdeen, King's College, Aberdeen, AB24 3UE, UK; [2] Department of Molecular and Cell Biology, University of Aberdeen, Polwarth Building, Foresterhill, Aberdeen, AB25 2ZD, UK

If we are to relate gene and protein structure to biological function, then it is necessary to develop good computational models of how large biomolecules might interact. There exist several algorithms for predicting how small ligands "dock" to proteins. However, these methods often require prior knowledge of the ligand binding site. In larger protein-protein complexes (where the interfacial surface area may range from about 800 to 1600Å^2), efficient and accurate surface matching algorithms are required to identify feasible docking orientations. The most successful current macromolecular docking algorithms are usually based either on (i) geometric hashing,[1] or (ii) fast Fourier transform (FFT) techniques.[2] However, none of the existing methods is well-suited to docking very large complexes such as the antibody Fab-hemagglutinin complex.[4] The geometric hashing algorithms are prone to a combinatorial increase in the number of features that need to be compared as the size of the molecules increase: none of the submissions to CASP2 used this method. The best single solution submitted used an FFT approach,[3] but the haemagglutinin moiety had to be broken into several fragments and large (low resolution) grids had to be used for the problem to fit into main memory.

To address many of the limitations of the grid-based FFT approaches, we recently developed a new Fourier-like algorithm based on spherical polar Fourier correlations.[5,6] By itself, our spherical polar approach is also unsuitable for such problems because our radial functions fall off rapidly beyond about 30Å from the chosen origin, hence molecular shapes larger than this are represented poorly. However, it is not necessary to rely on a single origin. We have developed an automatic method of generating multiple coordinate "centers," each of which may be used to capture an accurate representation of a local surface region. We can then perform high resolution angular docking searches over each surface patch. Taken together, these angular searches correspond to a full rigid-body search over the entire molecular surface. Selecting a suitable set of projection centers must be done with care to ensure full coverage of the molecular surface whilst minimizing the amount of computation that must be performed. This new macromolecular docking algorithm is fully automated and good docking predictions can now be obtained for very large complexes.

References

1. Fischer, D., Lin, S.L., Wolfson, H.L., and Nussinov, R. 1995. A geometry based suite of molecular docking processes. J. Mol. Biol. 248:459-477.
2. Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A.A., and Aflalo, C. 1992. Molecular surface recognition: Determination of geometric fit between proteins and their ligands by correlation techniques. Proc. Natl. Acad. Sci. USA 89:2195-2199.
3. Vasker, I. 1997. Evaulation of GRAMM low-resolution docking methodology on the Hemagglutinin-Antibody Complex. Proteins Struct. Funct. Genet. S1:226-230.
4. Dixon, J.S. 1997. Evaluation of the CASP2 Docking Section, Proteins. Struct. Funct. Genet. S1:198-204.
5. Ritchie, D.W. and Kemp, G.J.L. 2000. Protein docking using spherical polar Fourier correlations. Proteins Struct. Funct. Genet. 39:178-194.
6. www.biochem.abdn.ac.uk/hex/

The BALSAMIC (Basic Active and Ligand binding SurfAce Matching through Image Comparison) Project

Steven J. Pickering,[1] Andrew J. Bulpitt,[2] Nick D. Efford,[2] Nicola D. Gold,[1] and David R. Westhead,[1] [1]School of Biochemistry and Molecular Biology; [2]School of Computer Studies, University of Leeds, Leeds, W. Yorks, LS2 9JT. westhead@bmb.leeds.ac.uk

The relationship between biochemical function and protein structure is not straightforward. Sometimes function can be deduced directly from sequence and/or fold, but there are many examples of proteins with similar sequences and different functions, and examples of protein folds that support many unrelated functions. Even when proteins have related functions, small differences, for instance in enzyme specificity, are often only apparent when detailed structural information is available. Structural genomics initiatives are expected to generate structures for large numbers of proteins, the majority of which will be uncharacterized from a functional point of view. The aim of these initiatives is to provide structural information as a route to the determination of gene function. This will require the development of new bioinformatic tools able to predict functional characteristics from protein structure.

Many important biochemical processes, for example, molecular interaction and recognition or catalysis, occur on or near protein surfaces. Surfaces with similar shape, chemical, and physical properties are likely to perform similar functions. If a protein has an unknown function, a route to function prediction is therefore to locate similar surfaces in other proteins of better characterized function, and transfer the information. This process operates in direct analogy to the practice of predicting function from sequence by seeking to find related sequences of known function, for instance in a BLAST search.

Surface matching is a difficult computational problem, much studied in the field of computer vision. We report early progress in adapting methods from this field to the problem of protein surface matching. A multi-resolution approach views the surface in terms of differential properties, the curvedness and shape index, describing the degree of curvature and nature of the local surface shape (convex, concave, saddle), respectively. The surface matching algorithm uses a tree-structured search space where algorithmic efficiency is achieved through early pruning of branches corresponding to matches judged to be unreasonable by the above criteria. Preliminary results indicate that the method will be both efficient and useful. Future work will adapt the algorithms to include physico-chemical properties of the surface, with the aims of producing more reasonable matches from a biochemical point of view, and further increases in efficiency by increasing pruning of the search tree.

A database of surfaces for comparison will be generated and the algorithms will be used for similarity searches of this database. Other search algorithms will be implemented, including searches for user defined spatial arrangements of key functional residues. Ultimately these services will be made available on the Internet, and for analysis of structural genomics data.

Ablation of Cyclins in *Xenopus* Oocytes by Antisense Oligonucleotides Selected by Hybridization to Scanning Arrays

M. Sohail, H. Hochegger, A. Klotbucher, R. Guellec, T. Hunt, and E.M. Southern, University of Oxford, Department of Biochemistry, South Parks Road, OX1 3QU, UK

Arrays of antisense oligonucleotides corresponding to the first 120 nucleotides each of the cyclins B1, B4, and B5 were fabricated on the surface of aminated polypropylene. The arrays were hybridized with the appropriate radiolabelled transcript to assess the ability of the immobilized oligonucleotides to form heteroduplexes with their targets. Oligonucleotides that produced strong heteroduplex yield, as well as those that showed little annealing, were assayed for their effect on translation of endogenous cyclin mRNAs in Xenopus egg extracts and their ability to promote cleavage of cyclin mRNAs in oocytes by RNase H. Excellent correlation was found between the antisense potency and the affinities of the oligonucleotides for cyclin transcripts as measured by the arrays, despite the complexity of the cellular environment.

Homology Modelling of Hyperthermophilic Phosphoglycerate Kinases

Gina Crowhurst and Jennifer Littlechild, Schools of Chemistry and Biological Sciences, University of Exeter, Stocker Rd., EX4 4QD, UK

The hyperthermophilic archaeon Sulfolobus solfataricus lives at temperatures of up to 87ºC whilst Pyrococcus woesei survives temperatures in excess of 100ºC. The methods utilized in stabilizing intracellular proteins have been studied using homology modelling techniques using phosphoglycerate kinase (PGK) as an example. The glycolytic enzyme PGK is a well studied enzyme with over 100 primary sequences currently available. The enzyme is normally a monomer; however, S. solfataricus PGK is tetrameric whilst P. woesei PGK is dimeric. Homology models of the S. solfataricus, P. woesei, Methanococcus bryantii, and Haloarcula vallismortis PGKs have been generated and compared to the existing X-ray structures of the enzyme from Bacillus stearothermophilus, Thermotoga maritima, and Saccharomyces cerevisiae. These models have provided insights into the mechanisms of stabilization employed by hyperthermophiles and the substrate and cofactor binding sites of archaeal PGKs. The modelled S. solfataricus PGK structure also reveals potential areas for hydrophobic subunit interaction. Examination of the archaeal PGK protein sequences has confirmed that the essential catalytic residues have been conserved. A possible gene duplication event evident in the S. solfataricus PGK has also been observed.

β-Glucosyltranferase: Substrate Binding and Metal Site

S. Moréra,[1] L. Larivière,[1] W. Rüger,[2] P. Freemont,[3] [1]LEBS, UPR 9063 CNRS. Bât. 34 , 91198-Gif-sur-Yvette, France; [2]Arbeitsgruppe Molekulare Genetik, Ruhr Universität, Bochum, Germany; [3]MSFL, ICRF, 44 Lincoln's Inn Field, London WC2A 3PX, UK

β-Glucosyltransferase (BGT) is a DNA-modifying enzyme encoded by bacteriophage T4 that catalyses the transfer of glucose from uridine diphosphoglucose (UDPG) to 5-hydroxymethylcytosine (HMC) in double-stranded DNA. The glucosylation of T4 phage DNA is part of a phage DNA protection system aimed at host nucleases. We previously reported the complete BGT co-crystal structure in the presence of UDPG[1] where the glucose is missing due to BGT cleavage. This BGT structure has provided us with a basis for detailed modelling of DNA bound to BGT. Furthermore, using the structural similarity between the catalytic core of glycogen phosphorylase and BGT, we have been able to model the position of the missing glucose moiety from UDPG.

We now report two BGT-UDP-Mg^{2+} structures from crystals grown in the same conditions except the concentration of magnesium ions. Crystal of BGT-UDP-Mg^{2+} at 20mM diffracts at 2.5Å resolution while crystal of BGT-UDP-Mg^{2+} at 40 mM diffracts at 2Å resolution. Both crystals belong to 2121 space group but cell parameters are different. Both structures contain one magnesium ion in the UDPG binding site. The presence of a second Mg^{2+} ion far from the active site in the structure with 40mM Mg^{2+} could explain the difference of crystal packing between these two structures. Here, we present the metal site of BGT and from these two models, we propose a role of Glu163 in the catalytic mechanism of BGT.

Reference

1. Moréra, S., Imberty, A., Aschke-Sonnenborn, U., Rüger, W., and Freemont, P. 1999. T4 phage ß-glucosyltransferase: substrate binding and proposed catalytic mechanism. J. Mol. Biol. 292:717-730.

Prediction of Functional Sites in Proteins

Patrick Aloy,[1,2] Enrique Querol,[1] F. Xavier Avilés,[1] and Michael J.E. Sternberg[2] [1]Institut de Biologia Fonamental, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain; [2]Biomolecular Modelling Laboratory, Imperial Cancer Research Fund, Lincoln's Inn Fields, London WC2A 3PX, UK. patrick@luz.uab.es

An ultimate goal of genome analysis is to determine the biological function of all the gene products in a genome. Typically, prediction of function is based on sequence similarity with proteins of known function, but even in a simple organism with a small genome, like Mycoplasma genitalium, more than 40% of the ORFs have no homologous sequences in the databases. Moreover, the new initiatives in Structural Genomics will lead to the determination of the three-dimensional structure of proteins prior to knowledge of their function. All this has created the need for new methodologies for the analysis and prediction of protein function from its sequence and structure analysis. Here, we present a method to predict the location of functionally important sites in proteins.

A non-redundant database (<25% homology) of 107 protein chains was built using all the entries with detailed information of the functional residues in the Brookhaven protein databank (~1,800). An exhaustive analysis of residue type, secondary structure, and solvent accessibility preferences of the active site residue was performed. The next step was to develop a fully automated method following the ideas given by the Evolutionary Trace method (Lichtarge et al., 1996). The standard database search method WU-BLAST (Altschul et al., 1990) was used to pick up all the proteins that matched our probe sequence and a multiple alignment was then carried out with the program CLUSTALW (Thompson et al., 1994). A hierarchical clustering algorithm was implemented to build up a phylogenetic tree and the consensus sequence was extracted for each branch (subfamily). If our probe protein, or at least one of the homologous, has known structure, the consensus sequence was mapped onto the protein structure and techniques of double spatial clustering were applied in order to identify those residues that are not only conserved but close in the space as well. A sphere containing all the residues that have clustered together was then built and defined as the Functional Site.

The results show that it is possible to predict Functional Sites automatically in a given protein, with 60% accuracy and 99% significance, when we can find enough homologous sequences in databases and, at least, one of them has known structure. The method also gives us a clue to identify cases of divergent evolution with still high homology levels. The information obtained from the functional sites was used to filter putative protein-protein and protein-DNA docking complexes generated by FTDock (Gabb et al., 1997) giving similar results than when experimental information from mutagenesis experiments is used.

Crystallographic Studies of a Flavoprotein from *E coli*

A.L. Lovering, Protein Stucture Function Group, School of Biosciences, University of Birmingham, UK

A flavoprotein from E. coli, with potential use for gene-directed enzyme prodrug therapy (GDEPT), has been purified and crystallized, with the eventual aim of protein engineering. X-ray diffraction data for three different crystal forms were collected and the structures solved by MAD, SAD, and MR to a maximum resolution of 1.6Å. The overall fold of the protein and environment of the FMN cofactor with substrate analog are presented.

Homology Model of Glutamate Receptor

Indira H. Shrivastava and Mark S.P. Sansom, Laboratory of Molecular Biophysics, Rex Richards Building, University of Oxford, South parks Road, Oxford OX1 3QU, UK

Glutamate receptors that are permeable to Na^+, K^+, and $Ca2^+$ and potassium channels, selectively permeable to K^+, are two important families of ion channels. The crystal structure of KcsA, a potassium channel, revealed the structure of the protein in the transmembrane region.[1] The glutamate receptor, Glur0, from Synechocystis PCC 6803 binds glutamate and forms a potassium-selective channel.[2] The transmembrane region of Glur0 was found to have a high sequence similarity to the KcsA protein from Streptomyces lividans, particularly in the selectivity filter region (TVGYG) and in the putative gating region. Hence, the KcsA structure was used as a template to develop a homology model for Glur0. The alignment was done such that the overlap is maximal between the selectivity filter region and the pore-lining residues of KcsA and Glur0. This model was then inserted in a fully solvated palmitoyl oleyl phosphatidylcholine (POPC) lipid bilayer. Three K^+ ions were placed in the model protein, two in the selectivity filter and one in the cavity, and a molecular dynamics simulation trajectory was generated for 1ns. The model is seen to be stable over the period of 1ns, in a membrane environment with no obvious collapse in any region. The ions were seen to enter into the intracellular side from the channel pore. The interaction of the ions in the selectivity filter and the cavity with the protein and water molecules are compared to those of similar interactions in KcsA.[3]

References

1. Doyle et al. 1998. Science 280:69-77.
2. Chen et al. 1999. Nature 402:817-821.
3. Shrivastava and Sansom. 2000. Biophys. J. 78:557-570.

Calcium Binding by Proteins

Koen Bossers, Centre for Molecular and Biomolecular Informatics, University of Nijmegen, The Netherlands

Many proteins have calcium ions as part of their structure. The number of oxygen atoms involved in binding this ion (coordination number) can vary. The most frequently found coordination numbers are 6 and 7. With different coordination numbers, the calcium binding site adapts different geometrical conformations. In case of a 7-coordination site, the geometry is a pentagonal bipyramid, while a 6-coordination site adapts a octahedral geometry. The bidentate ligand (capable of providing two oxygen atoms for calcium binding, only glutamic acid, an aspartic acid, is capable of doing this) is essential for the pentagonal bipyramidal structure. Detailed statistical analysis and superposition of sites that adapt the same geometry provide valuable information for homology modelling and structure validation.

Experimental Confirmation of the Growth Hormone / Proteinase Function as Discovered by a Threading Approach of a Novel Gene

Cristina Mitsumori, Henrik T. Yudate, Keiichi Nagai, Yasuhiko Masuho, and Hisashi Koga, Helix Research Institute, Yana 1532-3, Kisarazu, Chiba 292-0812, Japan, Tel: +81-438-52-3951; Fax: +81-438-52-3952. cris@hri.co.jp

Protein structure determination has revealed an unexpected conservation and divergence of function both within and between families (Thornton et al., 1999), but shows the feasibility of function prediction of novel genes based on the tertiary structure of the coded proteins. We used THREADER (Jones et al., 1992) to analyze novel genes that had no sequence similarity to proteins with known functions. We focused on cytokine/growth factors because clones screened from a full-length cDNA library possessing signal-peptides were available and studies of cytokines have revealed they can be grouped into different structural families despite lack of sequence similarity (Rozwarski et al., 1994).

The THREADER program lists many relevant Z scores, but in the final selection, we used the score of the weighted sum of pairwise and solvation energies from the structure search (Z-13), and the score for shuffling of the sequence on the candidate using pairwise and solvation energies (Z-7).

Among the full-length cDNAs analyzed by THREADER, the human lung type-I cell membrane-associated protein hT1a-2 (160 aa) (Ma et al., 1998) had a very significant Z-7 score to the human growth hormone (PDB-ID: 1huw, 191 aa) and very significant Z-13 and Z-7 scores for proteinase A (PDB-ID: 2sga, 181 aa).

The first step of the experimental strategy consisted in sub-cloning of the hT1a-2 cDNA gene in a pCDNA3.1(-) Myc/His expression vector. The 30 kDa hT1a-2 protein expressed in mammalian cells was subsequently purified using a Ni column. Bioassay of the hT1a-2 activity as a growth hormone was tested using Nb2 rat lymphoma cell line. Nb2 can not grow in FBS-free medium, but exogenous application of hT1a-2 (700 ng/ml) induced a 60% increase in the growth rate, in comparison to FBS-treated Nb2 cell. Proteinase activity was tested using 5 peptide-4 methylcoumarin amide (MCA) substrates and release of 7-amino-MCA was determined fluorometrically. The purified hT1a-2 was observed to cleave the factor Xa- and trypsin-specific substrate Boc-Ile-Glu-Gly-Arg-MCA.

The above results show that hT1a-2 possesses activity both as a growth hormone and as a proteinase, as has also been described for the growth factor from Spirometra mansonoides (Phares and Kubik, 1996). In order to determine its main function in vivo, we used the phage display method and isolated a single phage expressing a 24 residue oligopeptide that specifically binds to the hT1a-2 protein. This oligopeptide exhibited a 43% sequence identity to the signal peptide cleavage region of the human insulin-like growth factor precursor, indicating that the basic function of hT1a-2 in vivo is probably as a proteinase that cleaves specific sequences.

We have used a bioinformatics approach to make a priority of which genes to analyze in the laboratory, and this is the first work confirming the function of hT1a-2 as estimated from annotation to the structures obtained from the threading approach.

3D Structure of Mammalian Thioredoxin Reductase: Comparison of TRR with Other Nucleotide-disulfide Oxidoreductases

Tatyana Sandalova, Medical Biochemistryu and Biophysics, Karolinska Institute, 17177 Stockholm, Sweden

Mammalian thioredoxin reductase (mTRR) is a member of pyridine nucleotide-disulfide oxidoreductase (PNDO) family that contains glutathione reductase (GR), lipoamide dehydrogenase (LAD), trypanothione reductase, mercuric ion reductase, and some other enzymes. All of them are flavoproteins and contain active disulfide that is reduced by NAD(P)H and then transfers the reducing equivalents to a substrate.

Thioredoxin reductase catalyses the reduction of the thioredoxin, a widely expressed 12-kDa protein that participates in many processes in the cells. Reduced thioredoxin is a hydrogen donor for ribonucleotide reductase and some other enzymes, it also controls the thiol-disulfide redox balance as well as being involved in the regulation of various transcription factors. It was shown that TRR is overexpressed in certain tumor cells and down-regulated in apoptotic cells. In addition, TRR participates in ascorbate recycling; it is inhibited by auranofin — the therapeutics of rheumatoid arthritis; and it is involved in the protection of the skin from UV radiation.

Surprisingly, mTRR is more similar to human GR than bacterial or plants TRR if size, sequence, and position of active cysteines are compared. TRR has broad substrate specificity, in addition to the reduction of thioredoxin, it catalizes the reduction of lipoic acid, protein disulfide isomerase, and many other substrates but not a glutathione. Unlike all PNDO, mammalian TRR contains an essential selenocysteine residue. SeCys is penultimate residue of a 16 amino acid long extension of mTRR that is absent in GR or LAD. Mutation of SeCys498→Cys greatly decreases the activity of TRR, however, it allows to overexpress the protein. The recombinant SeCys498→Cys mutant of rat TRR was crystallized as a complex with NADP$^+$. 3D structure of mTRR was solved and refined to 3Å (R/R$_{free}$= 23.0/ 29.5%). It is the first 3D structure of mTRR. Here, the comparison of 3D structure of mTRR with other members of nucleotide-disulfide oxidoreductase family is presented.

The 3D structure of human GR is the most similar to that of TRR: the alignment of 3D structure shows that these two proteins have 153 identical residues at the corresponding positions, rmsd is 1.4Å for 407 residues of one subunit, or 1.7Å for 810 Ca atoms of the dimer. TOP server found other similar proteins, all of them belong to PNDO family: trypanothione reductase (1aog.pdb) with rmsd 1.4Å for 410 residues (145 of them are identical), and dihydrolipoamide dehydrogenase (1ebd.pdb) with rmsd 1.5Å for 387 residues (108 residues are identical). The main difference is a 16-residues extension at C-terminus with essential Cys497-SeCys498 pair. The extension is located at the place, occupied by GSSG in GR; the distance between essential residue His472 and CysB498 is about 6.5Å. The detailed comparison of the active site structure of TRR with all other members of NDO-family is presented.

One-step Derivatization in Proteome Analysis

Francesco Brancia,[2] Simon J. Hubbard,[1] Simon J. Gaskell,[3] and Stephen G. Oliver,[1] [1]University of Manchester, 2.205 Stopford Building, Oxford Road, M13 9XX,UK; [2]Michael Barbert Centre for Mass Spectrometry, Department of Chemistry, UMIST, Sackvill Street, Manchester, UK; [3]Department of Biomolecular Sciences, UMIST

The identification of individual protein species within Saccharomyces cervisiae proteome has been optimized by increasing the information produced from mass spectra analysis through the chemical derivatization of tryptic peptides. Matrix assisted laser desoprtion ionization time-of-flight mass spectrometry (MALDI-TOF-MS) is commonly employed to analyze samples obtained from tryptic digestion; such proteolytic enzyme allows formation of digest fragments, well-suited to be protonated in the mass spectrometer. Recent studies have shown the strong dominance of signals belonging to arginine-containing peptides in the peptide mass fingerprinting spectrum. This behavior is readily explicable considering the different proton affinities of the two C-terminal residues. Conversion of lysine residues into homoarginine containing peptides increases the number of peptides in the spectra, rendering more detectable the lysine terminal peptides; such an improvement of signal response provides additional data for searching the database. Novel bioinformatic tools have been developed so as to exploit the further information obtained with guanidination in conjunction with other chemical derivatization.

A Bayesian Network Model for Protein Fold and Remote Homolog Recognition

D.L. Wild,[1] A. Raval,[1] and Z. Ghahramani,[2] [1]Keck Graduate Institute of Applied Life Sciences, 535 Watson Drive, Claremont, CA 91711, USA. david_wild@kgi.edu, alpan_raval@kgi.edu; [2]Gatsby Computational Neuroscience Unit, University College London, Queen's Square, London, UK. zoubin@gatsby.ucl.ac.uk

We describe Bayesian network models for protein folds and superfamilies that incorporate both primary sequence and structural information, with applications in the identification of remote homologs during the selection of potential targets for structure determination and in the classification of newly determined structures from structural genomics projects.

The Bayesian network approach is a framework that combines graphical representation and probability theory, that includes hidden Markov models (HMMs). HMMs trained on amino acid sequence or secondary structure data alone have been shown to have potential for addressing the problem of protein fold and superfamily classification. This poster describes a novel implementation of a Bayesian network that simultaneously learns amino acid sequence, secondary structure, and residue accessibility for proteins of known three-dimensional structure. An awareness of the errors inherent in predicted secondary structure may be incorporated into the model by means of a confusion matrix. Training and validation data have been derived for a number of protein superfamilies from the Structural Classification of Proteins (SCOP) database. Results using posterior probability classification indicate that the Bayesian network performs better in classifying proteins of known structural superfamily than a hidden Markov model trained on amino acid sequences alone. These results will be compared to classifications obtained using predicted secondary structure and residue accessibility information, and to a Fisher kernel (Support Vector Machine) method of scoring.

Crystal Structure of Tetradecameric *Mycobacterium tuberculosis* Chaperonin-10

Michael M. Roberts,[1] Alun R. Coker,[2] Anthony R.M. Coates,[1] and Steve P. Wood,[2] [1]Department of Medical Microbiology, St. George's Hospital Medical School, Cranmer Terrace, London SW17 0RE, UK; [2]Division of Biochemistry and Molecular Biology, School of Biological Sciences, University of Southampton, Bassett Crescent East, Southampton SO16 7PX, UK

Heptameric chaperonin 10 (cpn10) and tetradecameric chaperonin 60 (cpn60) interact to catalyse intracellular protein folding.[1] The crystal structure of Mycobacterium tuberculosis chaperonin 10 (Mtcpn10) has been solved to 2.8Å resolution. The heptameric Mtcpn10 substructure is similar to the cpn10 structures of E. coli (GroES)[2] and Mycobacterium leprae.[3] Each Mtcpn10 subunit has a wedge-shaped ß-barrel structure with a mobile loop from residues 17-35. A smaller loop from residues 51-56 at the other end of each subunit forms an acidic cluster of sidechains defining an 8Å hole at the roof of the dome-shaped heptamer. The mobile loops extend from the base of the heptamer like a jellyfish. Two Mtcpn10 heptamers complex through these mobile loops to form a tetradecamer with 722 symmetry and a spherical cage-like structure. The hollow interior enclosed by the tetradecamer is lined with hydrophilic residues and is 30Å perpendicular to and 60Å along the seven-fold axis and could therefore encapsulate a small folded protein. Within this chamber, difference maps show electron density that matches to mobile loop peptides of Mtcpn10 enclosed under the dome of each heptamer. This is confirmed by mass spectrometry, that reveals the peptides to be cleaved from Mtcpn10 on prolonged incubation in the crystallization buffer. Furthermore, as determined by the enzyme active site searching program TESS,[4] the Glu52 and Asp53 sidechains at the roof of the dome match the stereochemistry of active site sidechains in N-acetylglucosaminidases that can cleave bacterial cell wall peptidoglycan. This implies a mechanism for Mtcpn10 secretion and could explain the significance of the Mtcpn10 mobile loop in bone resorption through the stimulation of osteoclast proliferation[5] and the stimulation of the T-cell response,[6] since the mobile loop peptides would be transported by Mtcpn10 outside the cell for presentation to other cell receptors. The existence of the tetradecamer has been confirmed in solution for both GroES and Mtcpn10 by dynamic light scattering. Therefore, other tetradecameric cpn10 structures may be biologically significant in vivo as a mechanism for transporting a folded protein out of the cpn60 cavity for association with another folded protein. For example, two Mtcpn10 heptamers encapsulating two folded subunits would complex to form a dimer from those subunits. The crystallization conditions, data collection, and molecular replacement solution with GroES have been described.[7] The Mtcpn10 model was refined with NCS restraints on the 14 subunits to an R-factor of 21.3% (R_{free} = 25.3%) by simulated annealing, torsion angle, and positional refinement in X-PLOR[8] and CNS[9] in between rounds of model-building with QUANTA97 (MSI) and SwissPdb-Viewer.[10] PROCHECK[11] shows 91% of residues in the allowed regions and an overall G-factor of 0.1.

References

1. Horwich, A.L., Weber-Ban, E.U., and Finley, D. 1999. Proc. Natl. Acad. Sci. USA 96:11033-11040.
2. Hunt, J.F., Weaver, A.J., Landry, S.J., Gierasch, L., and Deisenhofer, J. 1996. Nature 379:37-45.
3. Mande, S.C., Mehra, V., Bloom, B.R., and Hol, W.G. 1996. Science 271:203-207.
4. Wallace, A.C., Borkatoti, N., and Thorton, J.M. 1997. Protein Science 6:2308-2323.
5. Meghji, S. et al. 1997. J. Exp. Med. 186:1241-1246.
6. Rosenkrands, I. et al. 1999. Infect. Immun. 67:5552-5558.
7. Roberts, M.M., Coker, A.R., Fossati, G., Mascagni, P., Coates, A.R.M., and Wood, S.P. 1999. Acta. Crystallogr. D. Biol. Crystallogr. 55:910-914.
8. Brunger, A.T. 1998. J. Mol. Biol. 203:803-816.
9. Brunger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.S., Kuszewski, J., Nilges, M., Pannu, N.S. et al. 1998. Acta Crystallogr. D. Biol. Crystallogr. 54:905-921.
10. Guex, N. and Peitsch, M.C. 1997. Electrophoresis 18:2714-2723.
11. Laskowski, R.A., McArthur, M.W., Moss, D.S., and Thornton, J. 1993. J. Appl. Crystallogr. 26:283-291.

Exploiting Protein Metal Interactions to Develop NMR Approaches to Structural Genomics

Joanne C. Ladds, Lesley K. Machlachlan, and Julia A. Hubbard, Computational and Structural Sciences, SmithKline Beecham Pharmaceuticals R&D, New Frontiers Science Park (North), 3rd Avenue, Harlow, Essex, CM19 5AW, UK

The goal of structural genomics is to understand the structure and function of proteins on a genomic scale. Many, possibly the majority, of proteins exist or are active in the cell in multi-protein complexes. Thus it is vital to understand protein-protein interactions on a structural level in order to fully understand protein function.

Two major approaches to structure determination are NMR and X-ray crystallography. NMR has developed rapidly over the last few years and is no longer an approach that is limited to proteins of 20 kD. The use of isotopic labelling, improved instrumentation, and novel pulse sequences initially increased this limit to approximately 30 kD. Around this size, structure determination dependent on large amounts of nOe data runs into difficulty. Protein-protein interactions are frequently also difficult to study due to the dearth of nOes in the interaction sites. Now a range of distance and orientation dependent NMR parameters promise to extend both the size and resolution of structure that can be determined and decrease dramatically the time that initial folds for proteins can be produced. These approaches will increase the utility of NMR for understanding protein recognition at a molecular level.

One important class of these new approaches uses NMR parameters that become available when a paramagnetic ion (either metal or nitroxide spin label) is attached to a protein. In cases where a native metal binding site is present, metal ions with appropriate properties may be substituted for the native metal. So far this has been limited to either Fe (often via heme) or Ca binding sites. It may be possible to produce generic metal-binding sites using recombinant techniques. It may also be possible that these approaches are not limited to the very strong metal-binding sites present in current published studies.

We discuss paramagnetic lanthanides in a range of proteins with different affinities for metal ions to obtain distance and orientation data and probe how this can be extended to understand protein-protein interactions by selection of the appropriate metal ion.

Assigning Sequences to Pfam Domains by Comparision of Medlars Documents

Benjamin J. Stapley and Michael J.E. Sternberg, Biomolecular Modelling Lab, Imperial Cancer Research Fund, 44 Lincoln’s Inn Field, London WC2A 3PX, UK

Functional annotation of proteins is an ever more pressing requirement for successful exploitation of genomic information. Here, we use textual information from Medlars to aid in the assignment of sequences to Pfam alignments. From a Pfam seed alignment, we extract Medlars documents that are cited in the SwissProt entries of sequences in the alignment. “Log Entropy” term weighting is applied and Pfam documents are generated by concatenation of the relevant Medlars documents. We then attempt to assign remote homologs — not included in the orginal alignments — to their respective Pfam alignments but measuring cosine similarities of their cited Medlars documents to the Pfam documents. Successful assignment to one of a subset of 100 Pfam alignments is achieved with up to 40% accuracy (25% recall). In addition to aiding in the correct functional assignment of sequences, generated Pfam documents allow textual information retrieval of Pfam domains with much higher recall. We have also applied the method to annotating Pfam alignments of indeterminant function.

Modelling of the Structure and S1 Specificity Pocket of a Potato Leaf Roll Virus Protease

Tomasz Cierpicki,[1] Jolanta Grembecka,[2] Filip Jeleń,[1] Marek Juszczuk,[3] and Jacek Otlewski,[1] [1]Institute of Biochemistry and Molecular Biology, University of Warclaw; [2]Institute of Organic Chemistry, Biochemistry and Biotechnology, Warclaw University of Technology; [3]Department of Biochemistry and Molecular Biology, Institute of Biochemistry and Biophysics, PAS, Warsaw, Poland

The amino acid sequence of 27 kDa domain of potato leaf roll virus protease (PLRVP) does not exhibit any detectable homology to known proteins deposited in Protein Data Bank (PDB). BLAST search within non-redundant protein data base allowed us to find a similarity to few proteins described as serine proteases. Further analysis by use of the fold recognition server 3D-PSSM showed the highest similarity of PLRVP to chymotrypsin-like serine proteases. Sequence comparison of PLRVP to serine proteases revealed that their most similar regions lie close to the catalytic triad residues.

The modelling of PLRVP scaffold was attempted based on the presence of structural similarity of chymotrypsin-like serine proteases, exhibiting inherent little primary structure similarity. Because they show very low sequence similarity to PLRVP, the conventional homology methods were useless. Therefore, we used the simulated annealing calculations based on the structural restraints derived from five selected proteases of similar fold (ETA, neuropsin, 2A, SVCP, and NS3). A set of the restraints, generated for structurally conserved core residues, included upper and lower distance ranges between α-carbons, backbone dihedral angles, and conserved hydrogen bonds.

The simulated annealing calculations were started from random conformations. Finally, 10 out of 30 structures with the lowest energies were selected. The calculated model included ß-strands close to the catalytic His and Asp residues (strands 2', 3', and 6') and second ß-barrel (strands 1', 2', 3', 4', 5', and 6') involving catalytic Ser residue. The modelling of the S1 specificity pocket was based on additional restraints for the loop connecting strands 3' and 4' extracted from ETA, ETB, and Glu-SGP structures. The interactions of PLRVP with peptide ligands were modelled based on Glu-SGP-inhibitor and SGPB-OMTKY3 complexes.

Preliminary kinetic studies, using Suc-Ala-Ala-Pro-Xaa-pNA, showed low proteolytic activity with some specificity for P1 Leu. Our modelling studies indicate that the S1 specificity pocket of PLRVP is primarily built of hydrophobic residues: Phe, Leu, and Thr. The hydrophobic S1 pocket prefer nonpolar residues (Leu), in agreement with kinetic studies.

Crystal Structure of a Bacteriophage T7 Endonuclease I: A Holliday Junction Resolving Enzyme

J.M. Hadden,[1] M.A. Convery,[2] A. Declais,[2] D.M.J. Lilley,[2] and S.E.V. Phillips,[1] [1]Astbury Centre for Structural Molecular Biology, School of Biochemistry and Molecular Biology, University of Leeds, Leeds, LS2 9JT, UK; [2]CRC Nucleic Acid Structure Group, Department of Biochemistry, University of Dundee, DD1 4HN, UK

Genetic recombination is a fundamental process in the evolution of all living organisms. This process results in the exchange of sequences between DNA segments and plays a fundamental role in the production of new genetic variants. The four-way DNA (Holliday) junction is an important intermediate in the recombination process.

Bacteriophage T7 encodes a 149 amino acid residue protein, endonuclease I, that has been shown to bind and cleave four-way DNA junctions in vitro. A number of mutants of endonuclease I have been isolated that bind, but do not cleave DNA junctions, and these are particularly useful for studying the binding process.

We have solved the crystal structure of one such mutant (residues 12-149) to 2.1Å resolution using selenomethionine substituted protein and the MAD technique. Unfortunately, the form of endo I used to grow crystals does not contain any methionine residues. For this reason, we introduced a single methionine residue into the protein by site-directed mutagenesis (I92M, one methionine per 138 residues) and following substitution, we were easily able to solve the protein structure.

The structure of the isolated protein shows endo I is an unusual homodimer arranged in two domains. Each domain is composed of approximately 1/6 of the residues from one monomer and approx. 5/6 of the residues from the other monomer. The domains are connected by a small inter-domain bridge. Details concerning the techniques used to solve the structure of the protein together with a full description of the protein topology will be presented.

Unconventional Crystallization Techniques That Have Produced High Quality Protein Crystals

J.M. Hadden and S.E.V. Phillips, Astbury Centre for Structural Molecular Biology, School of Biochemistry and Molecular Biology, University of Leeds, Leeds, LS2 9JT, UK

Unconventional crystallization techniques have been used successfully to grow crystals of two proteins. It has not previously been possible to produce crystals of these proteins suitable for X-ray diffraction studies.

The first technique highlights how a slow drop in temperature has been used to induce nucleation in a microbatch experiment. Once the correct level of nucleation had been achieved further nucleation was prevented, and a suitable crystal growth rate was achieved, by a small elevation in temperature. Crystals that diffracted to beyond 1.8Å have been produced using this technique. The structure of the protein has now been solved.

The second technique outlines the effect of varying the composition of grease/oil used to seal a vapor diffusion experiment. The effect of drop surface area to volume ratio was also investigated. By choosing the correct combination of crystallization well sealing material and drop surface area to volume ratio, large single crystals of protein were produced. These crystals diffracted X-rays to 2.1Å and the structure of the protein has now been successfully solved.

The Structure of the Transmembrane Segment of Vpu from HIV-1: Modelling and Simulations Studies

W.B. Fischer, F. Cordes, and M.S.P. Sansom, Laboratory of Molecular Biophysics, Oxford University, South Parks Road, Oxford OX1 3QU, UK. wolfgang@bioch.ox.ac.uk

The genome of the enveloped virus HIV-1 encodes an 81-residue auxiliary phospho-protein, composed of a N-terminal hydrophobic transmembrane (TM) domain (amino acids 1-27) and a hydrophilic 54-residue cytoplasmic domain. Vpu is not found in the envelope of the virus particle but is expressed in the membranes of sub-cellular compartments of the infected cell. Vpu has two major roles in the life cycle of the virus: (i) it controls the release/secretion of virus particles from the cell surface and (ii) mediates the degradation of the CD4 protein in the ER.

There is reasonable evidence that Vpu can form ion channels. Studies on Vpu expressed in Xenopus oocytes using whole cell voltage clamp technique revealed cation selective conductance. A synthetic peptide corresponding to the putative TM segment of Vpu also showed channel activity. NMR- and FTIR-spectroscopy show that the TM segment reconstituted in a lipid bilayer is predominantly α-helical. Also X-ray reflectivity data on Vpu containing monolayers indicate α-helical structure.

Self-assembly is a characteristic feature of Vpu in vivo as well as in vitro. Until now, the exact number of the homo-oligomers is not known. We have generated five bundles each consisting of five TM segments of Vpu (AIV A^{10} LVVAIIIAI V^{20} VWSIVIIE). Simulations for 2 ns were run for bundles obtained from a global molecular dynamics search protocol with restrains to experimental values.[1] In one of the bundles, all tryptophans were pointing into the pore; in the other model, they were pointing outwards. In comparison, bundles based on the same criteria for tryptophan orientation were created by using a simulated annealing protocol combined with a short molecular dynamics simulation (SA/MD).[2] In addition, a structure was generated driven by the idea that hydrophilic residues are facing the pore. This last model preserves its bundle-like structure throughout the simulation and seems to be the model of choice for the proposal of the bundle structure.

References

1. Kukol, A. and Arkin, I.A. 1999. Biophys. J. 77:1594-1601.
2. Kerr, I.D., Sankararamakrishnan, R., Smart, O.S., and Sansom, M.S.P. 1994. Biophys. J. 67:1501-1515.

Structure-based Design of New Strong Inhibitors of Leucine Aminopeptidase

J. Grembecka,[1] W.A. Sokalski,[2] P. Kafarski,[1] [1]Institute of Organic Chemistry, Biochemistry and Biotechnology; [2]Molecular Modelling Laboratory, Institute of Physical and Theoretical Chemistry, Wroclaw University of Technology, Wybrzeże Wyspiańskiego 27, 50-370 Wroclaw, Poland

The Ligand Design (LUDI/Insight_97.0) program[1] was applied in order to design leucine aminopeptidase inhibitors, predict their activity, and analyze the interactions with the enzyme. The investigation was based on the crystal structure of bovine lens leucine aminopeptidase (LAP, 1lcp) complexed with its inhibitor — the phosphonic acid analog of leucine (LeuP), a lead compound in our studies. The LUDI Link mode was used to obtain new inhibitors of the enzyme that were designed by the modification of LeuP structure. More than 50 potential leucine aminopeptidase inhibitors were obtained, including the most potent aminophosphonic LAP inhibitors with experimentally known activity.[2] Several of new designed inhibitors were synthesized and their activity towards the enzyme was measured. All of the tested compounds appeared to be strong LAP inhibitors. Most of them are significantly more active than already known inhibitors of the enzyme, containing phosphorus atom in the structure. The most active among the tested amino acid analogs is the phosphonic analog of homophenylalanine (K_i=0.14 μM for the DL mixture), while the phosphinic analog of leucylleucine (K_i=0.11 mM for the mixture of 4 diastereomers) is the most active among the peptide analogs. A reasonable agreement between theoretical and experimental activities has been observed for most of the studied inhibitors. Our results confirm that LUDI is a powerful tool for the design of enzyme inhibitors as well as in the prediction their activity.

In addition, for inhibitor-active site interactions dominated by the electrostatic effects, it is possible to improve binding energy estimates using more accurate description of inhibitor charge distribution.[3] For this purpose, we applied the another method, developed in our laboratory, that is based on ab initio calculations of the interaction energy in ligand–receptor system.[4] This permitted us to obtain more precise inhibitory activity estimates than using LUDI scoring function for several known LAP inhibitors differing with the electronic structure of functional groups.[5]

References

1. Gubernator, K. and H.J. Böhm (eds.). Structure-based ligand design. In Methods and principles in medicinal chemistry, R. Mannhold, Kubinyi, H., Timmerman, H. (ed.). Vol. 6. 1998, Wiley-VCH, Weinheim, p. 153.
2. Grembecka, J., W.A. Sokalski, and P. Kafarski. 2000. Computer-aided design and activity prediction of leucine aminopeptidase inhibitors. J. Comp. Aided Mol. Design 14:531-544.
3. Sokalski, W.A., Kedzierski, P., Grembecka, J., Dziekoński, P. Strasburger, K., In Computational Molecular Biology, J. Leszczyński (ed.), Elsevier Science, Amsterdam, p. 369-396, 1999.
4. Grembecka, J., P. K?dzierski, and W.A. Sokalski. 1999. Nonempirical analysis of the nature of the inhibitor-active site interactions in leucine aminopeptidase. Chem. Phys. Lett. 313:385-392.
5. Grembecka, J., W.A. Sokalski, and P. Kafarski. Submitted.

Gearing Individual Optimization Methods Towards High Throughput

Naomi E. Chayen and Emmanuel Saridakis, Biological Structure and Function Section, Division of Biomedical Sciences, Imperial College School of Medicine, London SW7 2AZ, UK

High throughput screening crystallization trials are already under way in several laboratories worldwide, but optimization, which is the more difficult part, has yet to be adapted to cope with the volume of experiments required by the Genome Projects.

The first multiple experiments for both screening and optimization were done as microbatch trials under oil.[1] This procedure lends itself for adaptation to high-throughput crystallization.

The use of oil has established a unique way of producing crystals, making the experiments more efficient and saving time and materials. Oil affects the accuracy, cleanliness, and reproducibility of crystallization experiments as well as providing a reliable environment for controlling nucleation and growth.[2]

We have designed the following optimization methods which have resulted in production of better-ordered crystals compared to those grown by conventional methods:

- Container-less crystallization, in which a crystallization drop is suspended between two oils of different densities resulting in reduction of heterogeneous nucleation.[3] High throughput is achieved by replacing one of the oils with a gelled hydrophobic surface onto which the drops are automatically dispensed.
- A means to slow down the equilibration rate and approach supersaturation more slowly, in order to avoid crystal "showers," is accomplished by placing a layer of oil as a barrier over the reservoir of a hanging or sitting drop trial.[3] The advantage of this technique is that no change is required to the crystallization conditions nor to the method used — it can be applied in Linbro, VDX, Cryschem, or any other vessel, and it can also be automated.
- Inducing nucleation in microbatch by evaporation through a thin oil layer, and then arresting it by variation of the oil layer thickness over time.
- Automatic dispensing of gelled microbatch drops.
- Special filtration of samples.

Producing better ordered crystals by de-coupling nucleation and growth can be achieved in either microbatch[4] or hanging drops.[5] In the case of hanging drops, the coverslips holding the drops are transferred after incubation for some time at conditions normally giving many small crystals, over reservoirs at concentrations that normally yield clear drops. In microbatch, the drops are diluted by automated means after incubation.

This poster will present examples of successful crystallization of several proteins as well as ways to automate and adapt all these methods to high-throughput applications.

References

1. Chayen, N.E. et al. 1990. J. Appl. Cryst. 23:297-302.
2. Chayen, N.E. 1998. Acta Cryst. D54:8-15.
3. Chayen, N.E. 1997. Structure 5:1269-1274.
4. Saridakis, E.E.G. et al. 1994. Acta Cryst. D50:293-297.
5. Saridakis, E. and Chayen, N.E. 2000. Prot. Sci. 9:755-757.

Inferring Protein Quaternary Structure from X-ray Crystallographic Data

Hannes Ponstingl,[1] Kim Henrick,[1] and Janet M. Thornton,[1,2] [1]EMBL Outstation, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK; [2]Biomolecular Structure and Modelling Unit, Biochemistry and Molecular Biology Department, University College London and Crystallography Department, Birkbeck College, London, UK

The physiologically relevant macromolecular assembly of multimeric proteins often cannot be derived without ambiguity from crystallographic studies of protein structure. An automatic tool is being developed to differentiate physiologically relevant intermolecular contacts from contacts that are artifacts of the crystalline state. We compare the performance of simple structural features in identifying the macromolecular assembly. The comparison is based on a non-redundant set of protein structures with known multimeric states prevalent in solution. Non-parametric statistical methods are used for error assessment.

Protein-Protein Interaction Networks

F. Pazos, C. Blaschke, J.C. Oliveros, and A. Valencia. Protein Design Group, CNB-CSIC, Cantoblanco, Madrid 28049, Spain. Tel: +34-1-585 45 70, Fax: +34-1-585 45 06. valencia@cnb.uam.es

The increasing knowledge about individual protein components (genome sequences, structural genomic initiatives, high throughput functional genomics) is making clear the need of integrating information in superior orders of complexity. Protein-protein interaction is the obvious next step in this direction. We present here three complementary computational efforts for the study of protein-protein interactions.

The first approach is based on the study of the patterns of variation in multiple sequence alignments. The rationale behind these approaches is that proteins that have evolved to form specific molecular complexes would have accumulated during evolution compensatory substitutions that can be detected in current protein families. We have previously demonstrated that the analysis of the patterns of variation is sufficient to single out the right inter-domain docking solution amongst many wrong alternatives in two-domain proteins[1] and tested the predictions of interacting regions in different experimental systems.[2,3] The extension of this method to the detection of interacting partners in large collections of multiple-sequence alignments shows quite promising results in terms of coverage, related with the number of interactions predicted in complete genomes, and accuracy, defined as the quality of the predicted interactions when compared with known molecular complexes.[4]

The second approach is based on the application of text retrieval techniques[5,6] to the extraction of information about protein interactions directly from the scientific literature (Medline abstracts). Our current system is automatically able to detect networks of functional interactions, by identifying protein names and the actions linking them.[7] We will discuss the results of the application of the system to different complex biological systems.

Finally, the application of clustering techniques[8] and text retrieval methods[9] to the available expression array results leads to a new avenue for the discovery of relations between genes. It can be considered complementary information to the predicted and detected protein interactions, and represents promising new technologies to be combined with other experimental approaches like yeast two hybrid systems.

References

1. Pazos et al. 1997. J. Mol. Biol. 272:1-13.
2. Gässler et al. 1998 Proc. Natl. Acad. Sci. USA 95:15229-15234.
3. Azuma et al. 1999. J. Mol. Biol. 289:1119-1130.
4. Pazos et al., submitted.
5. Andrade, M.A. and Valencia, A. 1997. ISMB 5:25-32.
6. Andrade, M.A. and Valencia, A. 1998. Bioinformatics 14:600-607.
7. Blaschke et al. 1999. ISMB 7:60-67.
8. Herreros et al., submitted.
9. Blaschke et al., submitted.

Drug Discovery: From Genes to Leads

Stanley R. Krystek and Jonathan S. Mason, Bristol-Myers Squibb Pharmaceutical Research Insititute, Princeton, NJ 08543 USA

The integration of genomics information with drug discovery is expected to identify, in the next few years, thousands of novel protein targets. This presentation will describe how combining structural genomics methodologies and structure-based drug design can be used to prioritize drug discovery projects. The application of the following methods to disease targets allows for the rapid generation of potential lead compounds.

- Database mining
- Protein fold recognition
- Protein function identification
- Protein modelling
- Virtual screening (structure- and pharmacophore-based)

Molecular Basis of the Specificity Requirements of Arginase and Agmatinase, Two Enzymes with a Common Evolutionary Origin: Homology Modelling and Site Directed Mutagenesis of *Escherichia coli* Agmatinase

Mónica Salas,[1] Rolando Rodríguez,[2] Elena Uribe,[1] P. Herrera, Vasthi López,[1] and Nelson Carvajal,[1] [1]Departamento de Biología Molecular, Facultad de Ciencias Biológicas, Universidad de Concepción, Chile; [2]CIGB, La Habana, Cuban and EMBL-Heidelberg, Germany

Arginase (EC 3.5.3.1) and agmatinase (EC 3.5.3.11) catalyse the production of urea from closely related substrates. In fact, agmatine results from decarboxylation of arginine by arginine decarboxylase. On the other hand, several highly conserved residues are detected in the amino acid sequences of these enzymes. For these reasons, they are considered as members of the arginase family of proteins. The idea is that they diverged from a common evolutionary origin to reach their particular substrate specifities. The crystal structures of rat liver and Bacillus caldovelox arginases are available, and specific roles have been assigned to several active site residues, including His101, His126, His141, and Asp128 (according to their positions in the rat liver sequence). These roles also have been validated by chemical modification and site-directed mutagenesis of the rat liver and human liver arginases. A critical role for His163 (His141 for rat liver arginase) also has been deduced from chemical modification and site-directed mutagenesis of Escherichia coli agmatinase.

At present, a crystal structure for agmatinase is not available. We have, therefore, used molecular modelling, by analogy with B. caldovelox arginase as a reference, to obtain a model for the structure of E. coli agmatinase. The model thus obtained gives an accurate description of the interaction of agmatinase with Mn2^{+} and the existence of a binuclear metal center in fully-activated enzyme. It also suggests a significant role for a loop, that include C159, Y155, and F16 in agmatine binding to agmatinase. Since this loop differs from that for arginase, which is bigger, a knowledge of these regions would explain the differences in specificity between arginase and agmatinase. To test the validity of these conclusions, site directed mutagenesis was used to introduce changes in the loop for agmatinase. Replacing these residues by the corresponding residues in the sequence of arginase, the single-mutants C159S, Y155N, and F161N, the double-mutants C159S/Y155N and Y155N/F161N, and the triple-mutant Y155N/C159S/F161N were constructed. Interestingly, alteration in the entire conformation of this region, produced in the triple-mutant, was required for total loss of agmatinase activity. The single-mutant C159S and the double-mutant C159S/Y155N were even more active than wild-type agmatinase (~2-4 fold) and the other species were almost equally active than wild-type enzyme. In conclusion, our results emphasize the importance of one specific loop region in substrate recognition by agmatinase. For a better understanding of the significance of these loops regions for the specificity requirements of arginase and agmatinase, insertions are now being introduced in the agmatinase sequence.

In Silico Structural Analysis of Bacterial Virulence Factors

Kelly Paine and Darren Flower, Bioinformatics Group, Edward Jenner Institute for Vaccine Research, Compton, Newbury, Berks, RG20 7NN, UK, Tel: 01635 577954. kelly.paine@jenner.ac.uk

Pathogens can be distinguished from their avirulent counterparts by the presence of specific gene clusters or pathogenicity islands that convey the virulence necessary for infection.[1] These can be acquired through lateral transfer between distinctly related species in evolution, and are dubbed virulence factors. Exotoxins secreted by such pathogens are an excellent example; any function inferred from the tertiary protein structure can be used in the development of new drugs.

For example, the recently published structure of an invasive Streptococcus pyogenes SpeB cysteine protease[2] revealed a hitherto unknown homology to the papain protease family. This gave a new insight into the mechanisms of virulence carried by the protease. An important human integrin-binding motif was also discovered, hinting that the exotoxin may have multiple functions. Most importantly, an invariant finger loop at residues 19-42 on the mature protease was identified as a potential therapeutic antibody-binding site.

Our focus is on virulence factors as potential candidate vaccines. The structure determination of prototypic virulence factors will facilitate the prediction of their potential antibody binding sites and the delineation of escape mutations, as well as allowing the design of new antibiotics and anti-microbial drugs. We have developed new approaches to the problem of selecting sequences for structural analysis and are currently applying them to our database of virulence factors. Dissimilarity searching algorithms, coupled to in-depth analyses of protein families using the PRINTS methodology,[3] have been used to select candidates for structural analysis.

References

1. Mecsas, J.J. and Strauss, E.J. 1996. Molecular mechanisms of bacterial virulence: type III secretion and pathogenicity islands. Emerg. Infect. Dis. 2:270-288.
2. Kagawa, T.F, Cooney, J.C., Baker, H.M., McSweeney, S., Liu, M., Gubba, S., Musser, J.M., and Baker, E.N. 2000. Crystal structure of the zymogen form of the group A Streptococcus virulence factor SpeB: An integrin-binding cysteine protease. Proc. Nat. Acad. Sci. USA 97:2235-2240.
3. Attwood, T.K., Croning, M.D.R., Flower, D.R., Lewis, A.P., Mabey, J.E., Scordis, P., Selley, J., and Wright, W. 2000. PRINTS-S: the database formerly known as PRINTS. Nuc. Acids Res. 28:225-227.

The Identification of Novel, Putative Metallo-lactamase-like Metal Binding Domains and Folds

Brian P. Clarke and Mike G. Tennant, SmithKline Beecham Pharmaceuticals, Computational and Structural Sciences, New Frontiers Science Park (North), Third Avenue, Harlow, Essex, CM19 5AW, UK

Using state-of-the-art sequence searching programs, we have identified a broad class of proteins which mediate a zinc-dependent hydrolytic reaction. Included in this superfamily are bacterial metallo-lactamase (MBL), glyoxylase-II, arylsulphatase, sdsA, CPSF, phnp, and cpdP proteins. By examining the sequences and the known structures of MBL proteins in this family, we conclude that these proteins have evolved through divergent evolution and maintain hydrolytic function and that an archetypal protein fold exists for this family.

Designing Sequence Profiles from an All-atom Force Field

Alfonso Jaramillo, Stephany Hery, Lorenz Wernisch, and Shoshana J. Wodak, Service de Conformation de Macromolecules Biologiques et de Bioinformatique, Universite Libre de Bruxelles, av F.D. Roosevelt 50 - CP 160/16, B-1050 Bruxelles, Belgium, Tel: 32-2-6505200, 6502013, Fax: 32-2-6488954. alfonso@ucmb.ulb.ac.be

Understanding the mechanism of protein folding and the factors that govern the stability of the protein native state remains a major goal in molecular biology. “Which sequences are compatible with a given fold?” is another formulation of the same problem, also termed the inverse folding problem, which may have useful practical application in de novo protein design. With the aim of answering this question, we implemented DESIGNER, a versatile procedure for selecting sequences that are compatible with a given backbone structure. The sequence selection is done by computing the folding free energy difference, between the corresponding models for the folded and unfolded states. We use our interface to the CHARMM program and the force field comprises all the classical non-bonded energy terms of CHARMM, and a implicit solvation free energy term. We illustrate the application of DESIGNER to the design of core and surface residues in three proteins (the SH3 domain, protein G, and Ubiquitin) and full designs of SH3 domain-related proteins. In the core and surfaces designs, DESIGNER is shown to select sequences that are much more similar to the native sequence than any other available method. The full designs are used to evaluate the influence of backbone flexibility in protein design. It is also shown to generate native-like sequence profiles, which offer the opportunity of investigating many interesting questions, including how the structure constraints the amino acid sequence.

Analysis of Multiple Gene Expression Responses to Salt Stress in the Halophyte *Mesembryanthemum Crystallinum* Using Microarray Technology

João P. Maroco,[1,*] Christine B. Michalowski,[2] M.A. Cushman,[1] Hans J. Bohnert,[2] David Galbraith,[3] and John C. Cushman,[1] [1]Dept. of Biochemistry and Molecular Biology, Oklahoma State University, Stillwater, OK USA; [2]Dept. of Biochemistry, The University of Arizona, Tucson, AZ USA; [3]Dept. of Plant Sciences, The University of Arizona, Tucson, AZ USA. *Present address: Lab. Ecofisiologia Molecular. IBET-ITQB. Av. Republica, EAN. 2784-505 Oeiras, Portugal

Plant responses to environmental stresses are mediated through the coordinate action of many hundreds of genes each with distinct expression profiles. Thus far, gene expression studies have been limited to one or a few genes at a time due to methodological limitations. Recently, microarray technology has made it possible to study the expression profiles of thousands of genes simultaneously. In this communication, we report a microarray-based evaluation of multiple gene induction by salt stress using more than 1,000 expressed sequence tags (ESTs) from the halophytic plant, Mesembryanthemum crystallinum. We found that approximately 20% of genes associated with CAM metabolism and osmotic stress resistance exhibit induced (>2-fold) expression, whereas a somewhat lower number of genes associated with CO_2 fixation showed down-regulated patterns of expression. In addition, 63 new genes with no previously described function are up- or down-regulated by more than two-fold during salt stress. Determination of these expression patterns represents the first functional information about this set of anonymous ESTs. This analysis of expression profiles of known and unknown genes provides the first integrated assessment of coordinated gene expression patterns in a higher plant undergoing salt stress.

The Uses of Gels for Protein Crystallization

J. Lopez-Jaramillo, J.M. Garcia-Ruiz, M.A. Hernandez-Hernandez, J.A. Gavira, Gonzalez-Ramirez, and F. Otalora, Laboratorio de Estudios cristalograficos (CSIC-UGRA), Facultad de Ciencias, Campus Fuentenueva, Granada, Spain

It is well known that removing convection from the crystallization reactor yields crystals of higher quality. One way to achieve it and assure a mass transport scenario governed by diffusion is the use of gels or high viscosity non-Newtonian fluids. We present here a new crystallization technique, termed Gel Acupuncture MEthod (GAME), based on the counter diffusion of protein and precipitating agent solutions, and exploits the properties of gels.

The counter-diffusion arrangement allows us to screen a continuous range of crystallization conditions in one single experiment consuming as few as 2 µl of gelled protein solution. To fully exploit the advantage of counter-diffusion, it is mandatory to use a long protein chamber. Then, it is possible to obtain a sequence of precipitation pattern starting from amorphous precipitation and finishing with faceted large crystals of the highest quality. Thus, our technique automatically finds the best crystallization conditions and yields isolated crystals immobilized by the gel matrix.

We demonstrate that the experiments can be performed inside the same X-ray capillaries that will be used later for data collection at both room and low temperature without any post-crystallization manipulation.[1] In addition to the improvement in crystal quality, this method has among others the following practical advantages:

- Minimizes the volume of protein solution (less than a drop volume)
- No need of crystal mounting (i.e., no damage of crystals)
- Easy transport to synchrotron facilities (crystals are inside the capillary and immobilized by the gel)
- Crystals can be tested in the home X-ray source, and those of interest can be diffracted at synchrotron in cryo with no post-crystallization manipulation

We will also present another application of gels to protein crystallography: the direct use of electrophoretic gels for screening crystallization conditions.[2] Crystals grown from gels after native electrophoresis and isoelectric focusing will be presented.

References

1. F.J. Lopez-Jaramillo, J.M. Garcia-Ruiz, J.A. Gavira, F. Otalora. J. Appl. Cryst. Submitted.
2. J.M. Garcia-Ruiz, M.A. Hernandez-Hernandez, F.J. Lopez-Jaramillo, and B.Thomas. J. Crystal Growth. In press.

Monte Carlo Envelopes for Removing the Uncertainty in the Evolutionary Trace Method

Mark K. Dean, Richard E. Smith, Graham J.G. Upton,[1] Paul D. Scott.[2] and Christopher A. Reynolds, Department of Biological Sciences, University of Essex, Colchester, Essex, CO4 3SQ UK [1]Department of Mathematics, [2]Department of Computer Sciences

The evolutionary trace method[1] is potentially a very powerful method for studying protein-protein interactions. It involves determining the conserved and conserved in class residues in a multiple sequence alignment for a protein family with a common fold. The conserved in class residues are conserved within all the subgroups, defined by a dendritic tree, for a given partition identity cutoff (PIC) and residue position. The ET method involves plotting these ET residues onto a space-filling structure for increasing PIC values as long as the ET residues cluster. The arbitrary step in the process involves deciding when the ET residues cease to cluster, but rather become distributed randomly over the surface of the protein. To remove this arbitrary step, a cluster score is calculated for the ET distribution at each PIC value, which is compared to the cluster score for 99 equivalent random distributions. The ET analysis is therefore continued until the cluster score for the ET distribution is comparable to that for the random distributions. The performance of this new method is assessed through applications on a number of systems including G-protein coupled receptor dimers,[2] heterotrimeric G-proteins (RGS4-AlF_4^- -activated G_{ia1} complex), the Cylin A -CDK2 complex, and the Beta-trypsin-pancreatic trypsin inhibitor complex.

References

1. Lichtarge, O., Bourne, H.R., and Cohen, F.E. 1996. Proc. Natl. Acad. Sci. USA 93:7507-7511.
2. Gouldson, P.R., Higgs, C., Smith, R.E., Dean, M.K., Gkoutos, G., and Reynolds, C.A. 2000. Neuropsychopharmacology. 24:Oct. issue, ~1 Sept.

Mapping Disease-causing SNPs to Protein Structure

Carles Ferrer-Costa, Modesto Orozco, Xavier de la Cruz, Unitat de Modelatge Molecular i Bioinformatica, Departament de Bioquimica i Biologia Molecular, Facultat de Quimica, Universitat de Barcelona, c/ Marti i Franques, 1, 08028 Barcelona, Cataluynia, Spain

It is well known that variations in the consensus sequence of a protein can cause dramatic alterations in its function, leading to disease. Our work is focused in describing, in structural terms, those single nucleotide polymorphisms (SNPs) that cause pathological effects in humans. To this end, we analyzed a set of human proteins for which disease-associated SNPs are known. Every pathological variant was described in terms of secondary structure, solvent accessibility, and situation in surface cavities. This was done mainly at two levels of structure, tertiary monomeric structure based on the PDB monomer coordinates and quaternary oligomeric structure, using the PQS structure prediction from PQS database. In addition, we analyze changes in physicochemical properties due to mutation according to their location in structure. We studied free energy variations derived from the partition coefficients of the amino acids, and variations in secondary structure propensities. In this poster, we show previous analysis and suggest some general characteristics of pathological mutations.

Sequence and Structural Analysis of the Human MHC Class III Region

Ranjeeva D. Ranasinghe,[1] Geoff J. Barton,[2] Alan J. Bleasby,[1] Jon C. Ison,[1] John B.C. Findlay,[3] Begoña Aguado[1] and R.D. Campbell,[1]
[1]MRC Human Genome Mapping Project Resource Centre, Hinxton, Cambridge, CB10 1SB, UK; [2]European Bioinformatics Institute, Hinxton, Cambridge, CB10 1SD, UK; [3]School of Biochemistry and Molecular Biology, University of Leeds, Leeds, LS2 9JT, UK

The complete sequence of the Major Histocompatibility Complex (MHC) in human is known. The MHC is divided into three regions, class I, II, and III. We focus on the characterization of fold by sequence analysis and homology modelling of the predicted proteins encoded in the genes located in the class III region. The class III region spans approximately 1.1Mb on human chromosome 6p21.3 and has been predicted to contain 59 genes. Three sequence searching methods were applied to MHC III gene products: standard sequence search using BLAST against a non-redundant sequence database that contained sequences of known 3D structures (NRL3D), searching a domain profile databases (Pfam, and SMART), and a fold recognition method (GenTHREADER). The conservation of predicted secondary structure to that of known structural families was also assessed.

A key requirement for current homology modelling techniques is that the protein being modeled must share high sequence similarity to a protein of known structure. One of the findings from this study is that MHC class III gene products show low (20%) sequence similarity to annotated sequences of known structure. Furthermore, the fold recognition analysis showed that the number of proteins whose fold can be correctly identified was low (36%). Moreover, in many cases (13%) where the regions were shown to be similar, the similarity only spanned a short region of the query protein, from which no prediction could be made of its potential 3D fold or function. These results have highlighted the need for the development of new software tools for the alignment and detection of fold and function of remote protein homologs.

We will extend the method of protein signatures, which has been applied to characterize several families. We will apply our methods to the Ig superfamily and its various subgroups, as these are particularly important for the MHC class III proteins. We will construct a library of sparse signatures for each type of known Ig domain. The signatures will be derived from key residue positions taken from the literature, alignment, and analysis of correlated mutations. The existing algorithm will be adapted for correlated mutational data and optimal parameters for handling of gaps and residue variability will be established. The usefulness of the library for detecting known immunoglobulin domains and for detecting new family members will be established. Further, we will test whether it is possible to generate signatures that are characteristic of distinct functional and molecular interaction properties. The result will be a library of signatures that are diagnostic of structural and functional properties of the immunoglobulin domains.

BASIC: Bilaterally Amplified Sequence Information Comparison

Leszek Rychlewski and Janusz M. Bujnicki, Bioinformatics Laboratory, International Institute of Molecular and Cell Biology, Warsaw, Poland. http://bioinfo.pl

Several features of a protein can be inferred based on sequence similarity or assumed homology with other proteins. These include 3D structure or general functional description. Sequence alignment is the most common approach used for the assertion of homology. The predictive power and utility of homology-based prediction methods increases with the continuously growing database of proteins with annotated structure or functions. Additional increase in predictive power can be attributed to the improving accuracy and sensitivity of sequence comparison methods. Consideration of sequence information deduced from the family of proteins closely related to the query protein, as performed by PSI-Blast, enabled a dramatic boost in the predictive power of sequence comparison methods. The approach of amplification of sequence information by incorporation of evolutionary related sequence is being pursued further in a bilateral fashion. The BASIC program represents a prediction method utilizing the evolutionary information on both sides of the comparison: the query and the template. The current descendants of this approach, FFAS and ORFeus, are presented in this work.

Miniaturization of Protein Crystallization

Maxim E. Kuil, Flip J. Hoedemaeker,* Jan Pieter Abrahams, Leiden Institute for Chemical Research, Leiden University, PO Box 9502, 2300 RA Leiden, The Netherlands *Present address: Crystallics B. V., Zekeringstraat 29, 1014 BV Amsterdam, The Netherlands

Testing of crystallization conditions in small volumes has significant advantages: the total number of conditions that is tested can increase even when the total amount of material needed is reduced. We aim to test crystallization conditions in one nanoliter. At present, most crystallization trials are done in µL volumes. Robotics can be used to further increase the number of tests to say a few thousand conditions. The presented project aims at both miniaturization and automated screening of protein crystallization conditions. Preliminary crystallization conditions are screened using the variation of the following parameters: protein concentration, buffer solution and pH, solution ionic strength and ionic species, the type of precipitant (e.g., PEG, MPD, ammonium sulphate), temperature, various surfactants, additives (e.g., cofactors, inhibitors), and the gravitational field. We show the production of a suitable array of containers made of PDMS, each with a volume of one nanoliter. We succeeded to crystallize lysozyme in these nanowells by flood filling using conditions that yield protein crystals in bulk. Paraffin oil was used to prevent evaporation of the small droplets. Using inkjet technology, it is possible to dispense droplets with a volume of 10 to 250 pL reliably. At present we are designing a system that is capable of filling nanowells with a volume of 0.2 to 1 nL in which the wells are separated by 5 to 20 µm with 16 individually controlled liquids including the protein solution. If we index the 96-well plate as the first generation well plate, the 1536-well plate is the third generation and our target density corresponds to roughly the sixth generation micro- or rather nano-well plate. In summary, this technology will allow us to efficiently search for protein crystallization conditions using far less than a milligram of purified protein. High-resolution phase diagrams of proteins in solution can be determined, improving our understanding of protein crystallization. The demands on protein over-expression will become less tight and truly high throughput screening of protein crystallization conditions will be possible.

UPCOMING MEETINGS

December 2000

6th Australian Molecular Modelling Workshop (MM2000)

December 5-8, 2000
RMIT University
Melbourne, Australia

Topics
- Colloid and surface chemistry
- Drug design and discovery
- Medicinal and pharmaceutical chemistry
- Surfactants
- Molecular modelling
- Genomics

Session Titles
- First principles molecular dynamics
- Non-equilibrium molecular dynamics
- Quantum chemistry
- Atomistic simulations of solids
- Surfaces and interfaces
- Polymer modelling
- Drug design
- Protein modelling
- Genomic analysis
- Bioinformatics

Contact

Molecular Modelling Workshop, Anjani Singh
Department of Applied Physics
RMIT, GPO Box 2476V
Melbourne
Victoria 3001, Australia
Tel: 61 3 9925 2600
Fax: 61 3 9925 5290
mm2000@rmit.edu.au
http://www.ph.rmit.edu.au/mm2000

Critical Assessment of Techniques for Free Energy Evaluation

December 7-8, 2000
Asilomar Conference Center
Asilomar, CA USA

Topics

Biological Chemistry, techniques for predicting energies of binding and solvation

CATFEE is a challenge and meeting designed to assess and discuss the techniques employed in computer assisted drug design efforts. CATFEE will provide the opportunity to modelers and experimentalists to test new algorithms and evaluate established techniques used for the prediction of binding free energies.

Contact

CATFEE
Tel: 323-995-6599
catfee@uqbar.ncifcrf.gov
http://uqbar.ncifcrf.gov/~catfee/

Structure-based Drug Design

December 13-15, 2000
St. Catherine's College
Oxford, UK

Session Titles

- Principles and Applications of Structure-based Drug Design
- Advances in Computational Methods
- Advances in Experimental Methods
- New Drug Targets
- Combinatorial Chemistry, Genomics, and Proteomics

Contact

Rebecca Wade
The Molecular Graphics and Modelling Society, UK
wade@embl-heidelberg.de
http://www.mgms.org/oxford2000/

Pacifichem 2000

December 14-19, 2000
Honolulu, Hawaii

Over 6,000 registrants from more than 50 countries are expected to attend. The meeting is being cosponsored by the American Chemical Society, Chemical Society of Japan, Canadian Society for Chemistry, the New Zealand Institute of Chemistry, and the Royal Australian Chemical Institute. All areas of chemistry are covered.

Contact

http://www.acs.org/meetings/pacific2000

GRID 2000: International Workshop on Grid Computing in Conjunction with 7th International Conference on High Performance Computing

December 17-20, 2000
Bangalore, India

Topics

- Computational chemistry
- Grid fabrics and architectures

Sponsors

- IEEE Computer Society
- ACM SIGARCH

GRID 2000 is an international meeting that brings together international grid computing researchers, developers, practitioners, and users. The aim of GRID 2000 is to serve as a forum to present current and future work as well as to exchange research ideas in this field.

Contact

Rajkumar Buyya
GRID
Monash University Clayton Campus
School of Computing Science and Software Engineering
Melbourne, Australia
http://www.buyya.org/Grid2000/

January 2001

Integrative Bioinformatics High-Throughput Interpretation of Pathways and Biology

January 24-26, 2001
Swissôtel Zurich
Am Marktplatz Oerlikon
CH-8050 Zurich
Switzerland

Gene Function Prediction

GeneRAGE: Algorithm for Sequence Clustering and Domain Detection
Dr. Cristos Ouzounis, EBI

Automatic Discovery of Regulatory Patterns in Promoter Regions Based on Whole Cell Expression Data and Functional Annotation
Dr. Steen Knudsen, Technical University of Denmark

Target Gene Identification from Expression Array Data by Promoter Analysis
Dr. Thomas Werner, Genomatix Software GmbH

Combining Gene-finding Programs to Increase Prediction Accuracy
Dr. Cecilia Hammar, University of Skövde (tentative)

Functional Gene Networks: A Case Study of a Novel Data Management Approach for Bioinformatics
Dr. Stephan Heymann, Kelman Gesellschaft für Geninformation mbH

Stability and Flexibility of Cellular Pathways
Dr. Rajan Kumar, Sarnoff Corporation

Structural Genomics

A Comprehensive Database of Protein Structure Models for Drug Discovery
Dr. Tod M. Klingler, Prospect Genomics, Inc.

The Evolution and Structural Anatomy of the Small Molecule Metabolic Pathways in Escherichia coli
Dr. Sarah A. Teichmann, University College London

Structural Pattern Localization Analysis by Sequential Histograms
Speaker to Be Determined, IBM T.J. Watson Research

Center (tentative)

Visualizing Gene Expression Data

Microarray Analysis as Module of LION's Integrated Life Science Informatic Platform
Dr. Jan Michel, LION Bioscience AG

A Decision Analytics Framework for Bioinformatics
Dr. Mark Demesmaeker, Spotfire Inc.

High-Throughput Research Using Microarray Gene Expression Information
Dr. Frank A. White, InforMax, Inc.

Gene Profiling for Target Identification

New Bioinformatics Approaches in Functional Genomics
Dr. Jean-Michel Claverie, CNRS

Human Adult Skeletal Muscle Transcriptional Profile Reconstructed by Novel Computational Approach
Prof. Gian Antonio Danieli, University of Padua

Evaluation of Single Nucleotide Polymorphism Typing with Invader on PCR Amplicons and Its Automation
Dr. Charles A. Mein, University of Cambridge (invited)

Protein Expression

Proteins Have a Diverse and Distinct Repertoire of Interactions
Dr. Michael Lappe, European Bioinformatics Institute

Analysis of the Transcriptome by a Combined Approach
Dr. Alon Amit, Compugen

Mining Mass Spectrometric Data for Disease Biomarkers
Dr. Pierre Huyn, SurroMed, Inc.

Computational Genomics

Genome-to-Genome Comparisons Using Terablast
Mr. Marty Gollery, TimeLogic, Inc.

Fast Probabilistic Analysis of Sequence Function Using Scoring Matrices
Dr. Craig Nevill-Manning, Rutgers University (tentative)

System for Integrating Data on GABA Receptors
Dr. Hannah Hong Xue, The Hong Kong University of Science and Technology

Genomic Discoveries Quickly Converted into Small-Molecule Drug-Discovery Programs
Speaker to Be Determined, Iconix Pharmaceuticals, Inc.

Contact

Christina Lingham
1037 Chestnut St.
Newton Upper Falls, MA 02464 USA
Tel: 617-630-1364, Fax: 617-630-1325
clingham@healthtech.com
www.healthtech.com/2001/bne/index.htm

February 2001

Exploiting Molecular Diversity: Refining Small Molecule Libraries

February 12-14, 2001
San Diego, CA USA

Combinatorial chemistry is a powerful tool for rapidly providing highly diverse compounds for hit discovery and for more focused lead optimization. What are the types of projects and specific examples of where this technology has provided the best results? How can chemoinformatic approaches such as library design or diversity assessment be used to better match a library to biological targets? How can diversity be further extended or the time for iterative cycles of design, synthesis, testing, and analysis be shortened?

Topics

- Chemoinformatics
- Approaches for diversity analysis
- Improved library design
- Integrating virtual screening and library design
- Extracting knowledge from screening results
- End user experiences
- Lead-finding for specific biological targets
- Enzyme inhibition
- Receptor antagonists
- Inhibition of protein-protein interaction
- Results with libraries for a family of targets
- Using combichem for SAR studies
- Combining Combichem and Med Chem for lead optimization
- Integrating Combichem with structure-based design Experience with Tech Transfer

Contact:

John Rodolewicz
Cambridge Healthtech Institute
1037 Chestnut St.
Newton Upper Falls, MA 02464
Tel: 617-630-1352
Fax: 207-493-4573
johnr@healthtech.com
http://www.healthtech.com/2001/mld/index.htm

March 2001

Cutting Edge Approaches to Drug Design

March 13, 2001
Scientific Societies Lecture Theatre
Saville Row, London UK

This meeting addresses methods of drug design at the cutting edge: molecular simulation, knowledge management, and informatics. A key objective of the meeting is to explore how synergistic collaboration between informaticians and the experimental chemist can transform the speed and profitability of the pre-clinical pharmaceutical research process. The meeting is organized by the Royal Society of Chemistry through the Biological and Medicinal Chemistry Sector and Molecular Modelling Group.

Provisional Speakers

- Prof. Sir Tom Blundell, FRS
- Andy Lyall, Oxford GlycoSciences

- Andy Davis, AstraZeneca
- Darren Green, GlaxoWellcome
- Iain Mclay, GlaxoWellcome
- Dave Brown, Pfizer

Contact:

Dr. Darren Flower
Edward Jenner Institute for Vaccine Research
Compton, Berkshire, RG20 7NN, UK
Tel: +44 1635 577954
Fax: +44 1635 577954
darren.flower@jenner.ac.uk

Modelling and Simulation of Microsystems

March 19-21, 2001
Hilton Head Island, SC USA

Keynote Lectures

Computational Chemistry
William Goddard, Caltech

Computational Materials
Roberto Car, Princeton University

Computational Biology
Amos Bairoch, Swiss Institute of Bioinformations

Nano-Structure Simulation: From Thin Oxides to Biological Ion Channels
Karl Hess, University of Illinois at Urbana-Champaign

Session Topics

- Structure Based Drug Design: Theory, Computation and Practice
  Fred Cohen, University of California at San Francisco
  Dirksen Bussiere, Chiron Corporation
- Structural Genomics
  Kurt Krause, University of Houston
- Atomic and Molecular Scale Modelling of Materials
  Niels Gronbech-Jensen, University of California Davis and Berkeley Lab
- Nanoscale Modelling of Front-end Processing in Silicon
  Wolfgang Windl, Motorola
- Quantum Mechanics and Computational Modelling of Soft Matter
  Lawrence Pratt, Los Alamos National Laboratory
  Stephen Paddison, Motorola

Workshop

- Simulation Techniques for Micromachined Devices
  Jacob White, Massachusetts Institute of Technology

Tutorial

- Interdisciplinary Design and Simulation Methods for Micro- and Biomedical-fluidic Applications
  Steffen Hardt, Institute of Microtechnology, Mainz, Germany
- MEMS Simulation Tools
  ANSYS, Inc
  Hewlett Packard, Inc.

Contact

Sarah Wenning
MSM & ICCN 2001 Operations Director
4847 Hopyard Road, Suite 4-381
Pleasanton, CA 94588 USA
Fax: 925-847-9153
wenning@dnai.com
www.cr.org

Discovery 2001: Innovative Advances in Drug Discovery

March 26-29, 2001
The Hilton San Diego Resort
San Diego, CA

Topics

- Bioinformatics
- Pharmacogenomics
- DNA chips and biochips, assays, miniaturization and microfabrication
- Natural products drug discovery
- Lead optimization, validation, and characterization
- Virtual combinatorial libraries
- ChemoInformatics
- Proteomics and functional genomics

Contact

Lisa Baumann
Institute for International Research
708 Third Avenue
New York, NY 10017 USA
Tel: 212-661-3500, Fax: 212-599-2192
lbaumann@iirny.comhttp://www.iir-ny.com/conference.cfm?EventID=P0631&

April 2001

221th American Chemical Society (ACS) National Meeting

April 1-6, 2001
San Diego, CA USA

COMP Symposium Titles

- Advances in 3D Searching and Pharmacophores (Division of Chemical Information, Organizer: Osman F. Guner, Molecular Simulations)
- Artificial Intelligence in Computational Chemistry (Organizer: Curt M. Breneman, Rensselaer Polytechnic Institute)
- Award Symposium: ACS Award for Computers in Chemical and Pharmaceutical Research — Invited papers only
- Computational Chemistry and Molecular Modelling Instruction (Division of Chemical Education)
- Computational Studies of Molecular Electronic Devices (Organizer: Olaf G. Wiest, University of Notre Dame)

- Computational Studies of Reaction Mechanisms and Enzyme Modes of Action (Organizer: Tim Clark, Computer-Chemie-Centrum)
- Computers in Chemistry Posters (Organizer: Ralph A. Wheeler, University of Oklahoma)
- Computers in Chemistry: General Biochemical
- Computers in Chemistry: General Theoretical
- Designing Focused Libraries for Drug Discovery: Hit to Lead to Drug (Organizer: Charles H. Reynolds, R.W. Johnson Pharmaceutical Research Institute)
- Poster: Energy Landscapes of Proteins, Glasses and Clusters Poster Session. (Organizer: Jose Nelson Onuchic, University of California at San Diego)
- Oral: Energy Landscapes of Proteins, Glasses, and Clusters: Dynamics, Folding, Function and Prediction (Organizers: Jose Nelson Onuchic, University of California at San Diego, Charles L. Brooks III, The Scripps Research Institute, David J. Wales, Cambridge University, Richard M. Stratt, Brown University
- Methods for Addressing Time and Length Scale Problems in Molecular Simulations (Organizer: Matt Challacombe, Los Alamos National Laboratory)
- New Computer Architectures in Chemistry — Challenges and Benefits (Organizer: Andrew C. Pineda, Albuquerque High Performance Computing Center, University of New Mexico)
- Structure-based Data Mining (Division of Chemical Information, Organizer: Robert W. Snyder, MDL Information Systems)
- Visualizing Chemistry: Using Animations, Graphics, and Modelling to Teach Chemistry (Organizer: Renée S. Cole, University of Wisconsin — Madison)

Contact

http://www.acs.org/meetings/sandiego2001/

Second Joint Sheffield Conference on Chemoinformatics: Computational Tools for Lead Discovery

April 9-11, 2001
Stephenson Hall
University of Sheffield, UK

The conference will cover all aspects of lead discovery including:

- 3D databases, including docking and pharmacophore analysis
- Assay QC and its influence on data mining
- Chemical data mining
- Descriptor validation
- Design of leadlike combinatorial libraries
- Design of screening collections
- e-business to facilitate lead discovery
- Novel software and hardware systems for lead discovery
- Selective compound acquisition from in house and commercial suppliers
- Similarity and clustering methods
- Structure-activity methods for lead identification and early optimization
- Structure-based design for lead identification and early optimization
- Virtual screening
- Case histories

Contact

Val Gillet
v.gillet@sheffield.ac.uk
www.shef.ac.uk/cisrg/shef2001

CHI's Structure-based Drug Design

April 11-12, 2001
University Park Hotel at MIT
Cambridge, MA USA

Efforts in structural genomics, and advances in computation, have allowed structure-based drug design to emerge as a valuable tool in medicinal chemistry. Combinatorial chemistry and high throughput approaches shifted attention away from structure-based methods, but large-scale determination of protein structures is bringing structure-based design to the forefront. Integration of structure-based methods, virtual screening, and combinatorial chemistry will provide the basis for more efficient drug design.

Topics

- Developing receptor models for use in docking studies
- Computational analysis of protein-ligand complexes
- Flexible ligand docking to estimate binding affinities
- Advances in automatic docking software
- 3D pharmacophore fingerprints for virtual screening and library design
- Structure-based virtual screening protocols
- Integration of combinatorial chemistry and structure-based design
- Design of peptidomimetics

Contact

Edel O'Regan
Tel: 617-630-1323, Fax: 617 630-1325
eoregan@healthtech.com
http://www.healthtech.com/2001/sbd/index.htm

Magnetic Resonance in Chemistry and Biology

April 20-27, 2001
Zvenigorod
Russian Federation

The aim of the XIth conference is to bring together scientists interested in the development of Magnetic Resonance techniques and experimental methods and their applications in chemical and biological research.

Session Titles

- Nitric oxide in chemistry and biology
- Advanced NMR imaging in biomedical fields
- The Environment and Magnetic Resonance research

Contact

Dr. Vladimir Sharygin
Institute of Chemical Physics of Russian Academy of Sciences
Kosygin's Street 4
Moscow
Russian Federation
117977
http://www.chem.msu.su/eng/events/zvenig01.html

Research in Computational Biology RECOMB2001

April 21-24, 2001
Montréal, Canada

Invited Speakers

- Hunger for New Technologies, Metrics, and Spatio-Temporal Models in Functional Genomics
  George Church, Department of Genetics, Harvard Medical School, Boston, MA USA
- RNA Biology and the Genome
  Philip Sharp, Massachusetts Institute of Technology, Cambridge, MA USA
- Quantitative Proteome Analysis: New Technology, Applications and Challenges
  Ruedi Aebersold, Department of Molecular Biotechnology, University of Washington, Seattle, WA USA
- Mark Adams, Vice President Genome Programs, Celera Genomics Rockville, MD USA
- Roger Brent, Molecular Sciences Institute Berkeley, CA USA
- Comparative Analysis of Organelle Genomes, a Biologist's View of Computational Challenges
  Franz Lang, Université de Montréal, Montréal, Canada
- Genetics and Genomics: Impact on Drug Discovery and Development
  Klaus Lindpaintner, Roche Genetics F. Hoffmann-La Roche AG, Basel, Switzerland
- The Role of Computational Chemistry in Translating Genomic Information into Bioactive Small Molecules
  Yvonne Martin, Abbott Laboratories, Abbott Park, IL USA
- Mark Ptashne, Memorial Sloan Kettering Cancer Center, New York, NY USA

Topics

- Genomics
- Molecular sequence analysis
- Recognition of genes and regulatory elements
- Molecular evolution
- Protein structure
- Structural genomics
- Gene expression
- Gene networks
- Drug design
- Combinatorial libraries
- Computational proteomics
- Structural and functional genomics

The origins of the conference came from the mathematical and computational side of the field, and there remains to be a certain focus on computational advances. However, the effective use of computational techniques to biological innovation is also an important aspect of the conference.

The conference program includes between 30 and 40 contributed papers that are peer-reviewed. Full versions of a selection of the papers are published annually in the Journal of Computational Biology.

Contact

recomb2001@gmd.de
http://recomb2001.gmd.de

July 2001

Intelligent Systems for Molecular Biology

July 21-25, 2001
Copenhagen, Denmark

Contact

Johanne Keiding
johanne@cbs.dtu.dk

August 2001

222th American Chemical Society (ACS) National Meeting

August 26-31, 2001
Chicago, IL USA

Contact

American Chemical Society Meetings Department
1155 Sixteenth Street, N.W.
Washington, DC 20036 USA
Tel: 202-872-4396
Fax: 202-872-6128
natlmtgs@acs.org
http://www.acs.org/meetings/chicago2001.html