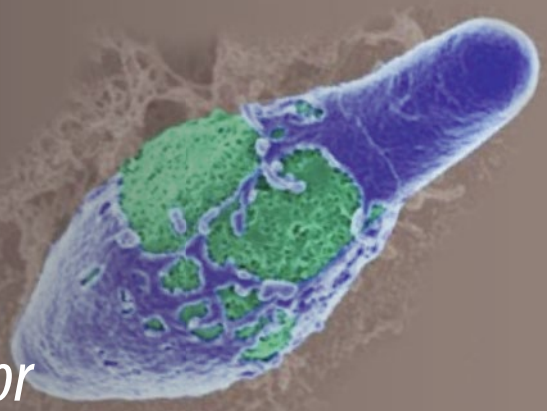
Methods in Molecular Biology 1140

Springer Protocols



Wayne F. Anderson *Editor*

Structural Genomics and Drug Discovery

Methods and Protocols

Humana Press

METHODS IN MOLECULAR BIOLOGY

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

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

Structural Genomics and Drug Discovery

Methods and Protocols

Edited by

Wayne F. Anderson

Center for Structural Genomics of Infectious Diseases,
Midwest Center for Structural Genomics,
Northwestern University Feinberg School of Medicine, Chicago, IL, USA

Humana Press

Editor
Wayne F. Anderson
Center for Structural Genomics
of Infectious Diseases
Midwest Center for
Structural Genomics
Northwestern University Feinberg
School of Medicine
Chicago, IL, USA

ISSN 1064-3745 ISSN 1940-6029 (electronic)
ISBN 978-1-4939-0353-5 ISBN 978-1-4939-0354-2 (eBook)
DOI 10.1007/978-1-4939-0354-2
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2014931091

Printed on acid-free paper

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

Dedication

For Caroline, who has been a wonderful partner for nearly 50 years and made this, as well as everything else, possible.

Preface

"Structural Genomics" as an area of investigation arose from the recognition that genome sequence information could be combined with improved methods for macromolecular structure determination to allow high-throughput structure determination. One of the early justifications for developing the field was the potential to make use of the structural information in drug discovery efforts. All three of these areas, genome sequencing, macromolecular structure determination, and structure-aided drug discovery, have seen dramatic improvements in technology and methodology.

This volume focuses on high-throughput structure determination methods and how they can be applied to lay the groundwork for structure-aided drug discovery. The methods and protocols that are described can be applied in any laboratory interested in using detailed structural information to advance the initial stages of drug discovery. Due to the advances in technology and methodology that have occurred during the past 10–15 years, even the nonspecialist can apply structural biology to most biomedical problems. The methods and approaches that distinguish structural genomics from "classical" structural biology have been decreasing as more and more research groups adopt high-throughput methods and apply them to their specific biological research problems.

In some respects, structure-aided drug discovery is very specific to the one particular protein target being studied and the approaches of structural genomics would not seem to be appropriate. However, if one looks at the problem broadly, there often is more than one protein that could be targeted, and when multiple proteins are being investigated, the advantages of carrying out most of the steps in parallel can increase productivity.

The initial chapters deal with bioinformatics and data management because selecting target proteins and planning how the large amount of diverse data will be handled are the first steps. Following these are the chapters on high-throughput methods for cloning, expression and solubility testing, protein production, purification, crystallization screening, and screening for suitability for NMR structure determination. One of the continuing problems faced by structural genomics efforts is the limited success rate that, not surprisingly, accompanies increased throughput and the associated reduction in individual attention to each protein. Although there is no panacea, a number of chapters describe methods that can rescue, or salvage, target proteins that are failing as they proceed through the pipeline. Finally, the concluding chapters describe methods that use the proteins that have been produced in order to identify initial small molecule hits. These hits can then feed into drug discovery efforts. At this point in the process, the number of technically and biologically suitable targets will have been reduced and each protein, together with the hits that have been generated, will require individual attention.

The structural genomics approach provides an efficient initial step toward drug discovery and the methods described will be useful to anyone interested in moving in this direction.

Chicago, IL, USA *Wayne F. Anderson*

Contents

Contributors

Hao An • *Biosciences Division, Midwest Center for Structural Genomics, Argonne National Laboratory, Argonne, IL, USA*

Wayne F. Anderson • *Center for Structural Genomics of Infectious Diseases, Northwestern University Feinberg School of Medicine, Chicago, IL, USA; Midwest Center for Structural Genomics, Northwestern University Feinberg School of Medicine, Chicago, IL, USA*

Aleksandar Antanasijevic • *Department of Biochemistry and Molecular Genetics, University of Illinois at Chicago, Chicago, IL, USA*

Cheryl H. Arrowsmith • *Division of Cancer Genomics and Proteomics, Northeast Structural Genomics Consortium (NESG), Ontario Cancer Institute, Toronto, ON, Canada*

Gyorgy Babnigg • *Center for Structural Genomics of Infectious Diseases, Computational Institute, University of Chicago, Chicago, IL, USA; Biosciences Division, Midwest Center for Structural Genomics, Argonne National Laboratory, Argonne, IL, USA*

Emily T. Beebe • *Department of Biochemistry, University of Wisconsin-Madison, Madison, WI, USA*

Lance Bigelow • *Biosciences Division, Midwest Center for Structural Genomics, Argonne National Laboratory, Argonne, IL, USA*

T. Andrew Binkowski • *Computation Institute, Center for Structural Genomics of Infectious Diseases, University of Chicago, Chicago, IL, USA*

Michael Caffrey • *Department of Biochemistry and Molecular Genetics, University of Illinois at Chicago, Chicago, IL, USA*

Changsoo Chang • *Biosciences Division, Midwest Center for Structural Genomics, Argonne National Laboratory, Argonne, IL, USA; Biosciences Division, Structural Biology Center, Argonne National Laboratory, Argonne, IL, USA*

Gekleng Chhor • *Biosciences Division, Midwest Center for Structural Genomics, Argonne National Laboratory, Argonne, IL, USA*

Maksymilian Chruszcz • *Department of Chemistry and Biochemistry, University of South Carolina, Columbia, SC, USA*

Marianne Cuff • *Biosciences Division, Midwest Center for Structural Genomics, Argonne National Laboratory, Argonne, IL, USA; Biosciences Division, Structural Biology Center, Argonne National Laboratory, Argonne, IL, USA*

Zbigniew Dauter • *Synchrotron Radiation Research Section, Macromolecular Crystallography Laboratory, National Cancer Institute, Argonne National Laboratory, Argonne, IL, USA*

Douglas R. Davies • *Emerald Bio, Bainbridge Island, WA, USA*

Zygmunt S. Derewenda • *Department of Molecular Physiology and Biological Physics, University of Virginia School of Medicine, Charlottesville, VA, USA*

Benoit H. Dessailly • *Center for Structural Genomics of Infectious Diseases, Department of Structural and Molecular Biology, University College London, London, UK*

Marcin J. Domagalski • *Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, VA, USA; Center for Structural Genomics of Infectious Diseases (CSGID), University of Virginia, Charlottesville, VA, USA; Midwest Center for Structural Genomics (MCSG), University of Virginia, Charlottesville, VA, USA; New York Structural Genomics Research Consortium (NYSGRC), University of Virginia, Charlottesville, VA, USA; Enzyme Function Initiative (EFI), University of Virginia, Charlottesville, VA, USA*

Aiping Dong • *Structural Genomics Consortium, University of Toronto, Toronto, ON, Canada*

Sara F. Dunne • *Center for Structural Genomics of Infectious Diseases, High Throughput Analysis Laboratory and Department of Molecular Biosciences, Northwestern University, Evanston, IL, USA*

Aled Edwards • *Structural Genomics Consortium, University of Toronto, Toronto, ON, Canada; Center for Structural Genomics of Infectious Diseases, Toronto, ON, Canada*

David Eisenberg • *UCLA-DOE Institute for Genomics and Proteomics, Howard Hughes Medical Institute, University of California, Los Angeles, CA, USA*

Jonathan M. Elkins • *Nuffield Department of Clinical Medicine, Structural Genomics Consortium, University of Oxford, Oxford, UK*

Elena Evdokimova • *Department of Chemical Engineering and Applied Chemistry, University of Toronto, Toronto, ON, Canada*

Brian G. Fox • *Department of Biochemistry, University of Wisconsin-Madison, Madison, WI, USA*

Daved H. Fremont • *Center for Structural Genomics of Infectious Diseases, St. Louis, MO, USA; Department of Pathology and Immunology, Washington University, St. Louis, MO, USA; Department of Biochemistry and Molecular Biophysics, Washington University, St. Louis, MO, USA*

Maite Garcia • *Division of Cancer Genomics and Proteomics, Northeast Structural Genomics Consortium (NESG), Ontario Cancer Institute, Toronto, ON, Canada*

Elizabeth M. Glass • *Center for Structural Genomics of Infectious Diseases, Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL, USA*

Lukasz Goldschmidt • *UCLA-DOE Institute for Genomics and Proteomics, Howard Hughes Medical Institute, University of California, Los Angeles, CA, USA*

Marek Grabowski • *Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, VA, USA; Center for Structural Genomics of Infectious Diseases (CSGID), University of Virginia, Charlottesville, VA, USA; Midwest Center for Structural Genomics (MCSG), University of Virginia, Charlottesville, VA, USA; New York Structural Genomics Research Consortium (NYSGRC), University of Virginia, Charlottesville, VA, USA; Enzyme Function Initiative (EFI), University of Virginia, Charlottesville, VA, USA*

Catherine Hatzos-Skintges • *Biosciences Division, Midwest Center for Structural Genomics, Argonne National Laboratory, Argonne, IL, USA*

Wei Jiang • *Argonne Leadership Computing Facility, Argonne National Laboratory, Argonne, IL, USA*

Grazyna Joachimiak • *Biosciences Division, Midwest Center for Structural Genomics, Argonne National Laboratory, Argonne, IL, USA*

Andrzej Joachimiak • *Center for Structural Genomics of Infectious Diseases, Computational Institute, University of Chicago, Chicago, IL, USA; Biosciences Division, Midwest Center for Structural Genomics, Argonne National Laboratory, Argonne, IL, USA; Biochemistry and Molecular Biology, University of Chicago, Chicago, IL, USA*

Youngchang Kim • *Center for Structural Genomics of Infectious Diseases, University of Chicago, Chicago, IL, USA; Midwest Center for Structural Genomics, Argonne National Laboratory, Argonne, IL, USA; Structural Biology Center, Argonne National Laboratory, Argonne, IL, USA*

Keehwan Kwon • *Center for Structural Genomics of Infectious Diseases, J. Craig Venter Institute, Rockville, MD, USA*

Hyun Lee • *Center for Pharmaceutical Biotechnology, University of Illinois at Chicago, Chicago, IL, USA*

Chung A. Lee • *Department of Pathology and Immunology, Washington University, St. Louis, MO, USA*

Hui Li • *Midwest Center for Structural Genomics, Argonne National Laboratory, Argonne, IL, USA*

Samuel H. Light • *Center for Structural Genomics of Infectious Diseases, Northwestern University Feinberg School of Medicine, Chicago, IL, USA*

Chi-Hao Luan • *Center for Structural Genomics of Infectious Diseases, High Throughput Analysis Laboratory, Department of Molecular Biosciences, Northwestern University, Evanston, IL, USA*

Jamey Mack • *Biosciences Division, Midwest Center for Structural Genomics, Argonne National Laboratory, Argonne, IL, USA*

Elizabeth M. MacLean • *Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, VA, USA; Center for Structural Genomics of Infectious Diseases (CSGID), University of Virginia, Charlottesville, VA, USA; Midwest Center for Structural Genomics (MCSG), University of Virginia, Charlottesville, VA, USA; New York Structural Genomics Research Consortium (NYSGRC), University of Virginia, Charlottesville, VA, USA; Enzyme Function Initiative (EFI), University of Virginia, Charlottesville, VA, USA*

Shin-ichi Makino • *Department of Biochemistry, University of Wisconsin-Madison, Madison, WI, USA*

Magdalena Makowska-Grzyska • *Center for Structural Genomics of Infectious Diseases, University of Chicago, Chicago, IL, USA*

Natalia Maltseva • *Center for Structural Genomics of Infectious Diseases, University of Chicago, Chicago, IL, USA*

John L. Markley • *Department of Biochemistry, University of Wisconsin-Madison, Madison, WI, USA*

William H. McCoy • *Department of Pathology and Immunology, Washington University, St. Louis, MO, USA*

Shahila Mehboob • *Center for Pharmaceutical Biotechnology, University of Illinois at Chicago, Chicago, IL, USA*

Karolina Michalska • *Biosciences Division, Midwest Center for Structural Genomics, Argonne National Laboratory, Argonne, IL, USA; Biosciences Division, Structural Biology Center, Argonne National Laboratory, Argonne, IL, USA*

George Minasov • *Center for Structural Genomics of Infectious Diseases, Northwestern University Feinberg School of Medicine, Chicago, IL, USA; Midwest Center for Structural Genomics, Northwestern University Feinberg School of Medicine, Chicago, IL, USA*

Wladek Minor • *Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, VA, USA; Center for Structural Genomics of Infectious Diseases (CSGID), University of Virginia, Charlottesville, VA, USA; Midwest Center for Structural Genomics (MCSG), University of Virginia, Charlottesville, VA, USA; New York Structural Genomics Research Consortium (NYSGRC), University of Virginia, Charlottesville, VA, USA; Enzyme Function Initiative (EFI), University of Virginia, Charlottesville, VA, USA*

Rory Mulligan • *Center for Structural Genomics of Infectious Diseases, Computational Institute, University of Chicago, Chicago, IL, USA*

Peter J. Myler • *Seattle Structural Genomics Center for Infectious Disease, Seattle, WA, USA; Seattle Biomedical Research Institute, Seattle, WA, USA; Department of Global Health, University of Washington, Seattle, WA, USA; Department of Medical Education and Biomedical Informatics, University of Washington, Seattle, WA, USA*

Christopher A. Nelson • *Center for Structural Genomics of Infectious Diseases, Washington University, St. Louis, MO, USA; Department of Pathology and Immunology, Washington University, St. Louis, MO, USA*

Boguslaw Nocek • *Biosciences Division, Midwest Center for Structural Genomics, Argonne National Laboratory, Argonne, IL, USA; Biosciences Division, Structural Biology Center, Argonne National Laboratory, Argonne, IL, USA; Center for Structural Genomics of Infectious Diseases, Computational Institute, University of Chicago, Chicago, IL, USA*

Christine A. Orengo • *Center for Structural Genomics of Infectious Diseases, Department of Structural and Molecular Biology, University College London, London, UK*

Jerzy Osipiuk • *Biosciences Division, Midwest Center for Structural Genomics, Argonne National Laboratory, Argonne, IL, USA; Biosciences Division, Structural Biology Center, Argonne National Laboratory, Argonne, IL, USA; Center for Structural Genomics of Infectious Diseases, Computational Institute, University of Chicago, Chicago, IL, USA*

Khan Tanjid Osman • *Structural Genomics Consortium, University of Toronto, Toronto, ON, Canada*

Scott N. Peterson • *Inflammatory and Infectious Disease Center, Sanford-Burnham Medical Research Institute, La Jolla, CA, USA*

Isabelle Q.H. Phan • *Seattle Structural Genomics Center for Infectious Disease, Seattle, WA, USA; Seattle Biomedical Research Institute, Seattle, WA, USA*

Kanagalaghatta Rajashankar • *NE-CAT and Department of Chemistry and Chemical Biology, Cornell University, Argonne National Laboratory, Argonne, IL, USA*

Benjamin E. Ramirez • *Center for Structural Biology, University of Illinois at Chicago, Chicago, IL, USA*

Kiira Ratia • *Department of Medicinal Chemistry and Pharmacognosy, University of Illinois at Chicago, Chicago, IL, USA; Research Resources Center, University of Illinois at Chicago, Chicago, IL, USA*

Benoit Roux • *Department of Biochemistry and Molecular Biology, University of Chicago, Chicago, IL, USA*

Alexei Savchenko • *Center for Structural Genomics of Infectious Diseases, Department of Chemical Engineering and Applied Chemistry, University of Toronto, Toronto, ON, Canada*

Anthony Semesi • *Division of Cancer Genomics and Proteomics, Northeast Structural Genomics Consortium (NESG), Ontario Cancer Institute, Toronto, ON, Canada*
Ludmilla Shuvalova • *Center for Structural Genomics of Infectious Diseases, Northwestern University Feinberg School of Medicine, Chicago, IL, USA; Midwest Center for Structural Genomics, Northwestern University Feinberg School of Medicine, Chicago, IL, USA*
Tatiana Skarina • *Department of Chemical Engineering and Applied Chemistry, University of Toronto, Toronto, ON, Canada; CSGID, Toronto, ON, Canada*
Robin Stacy • *Seattle Structural Genomics Center for Infectious Disease, Seattle, WA, USA; Seattle Biomedical Research Institute, Seattle, WA, USA*
Kemin Tan • *Biosciences Division, Midwest Center for Structural Genomics, Argonne National Laboratory, Argonne, IL, USA; Biosciences Division, Structural Biology Center, Argonne National Laboratory, Argonne, IL, USA; Center for Structural Genomics of Infectious Diseases, Computational Institute, University of Chicago, Chicago, IL, USA*
Christine Tesar • *Biosciences Division, Midwest Center for Structural Genomics, Argonne National Laboratory, Argonne, IL, USA*
Yufeng Tong • *Structural Genomics Consortium, University of Toronto, Toronto, ON, Canada*
Amy Wernimont • *Structural Genomics Consortium, University of Toronto, Toronto, ON, Canada*
Xiaohui Xu • *Department of Chemical Engineering and Applied Chemistry, University of Toronto, Toronto, ON, Canada; Midwest Center for Structural Genomics, University of Toronto, Toronto, ON, Canada*
Corin Yeats • *Center for Structural Genomics of Infectious Diseases, Department of Structural and Molecular Biology, University College London, London, UK*
Adelinda A. Yee • *Division of Cancer Genomics and Proteomics, Northeast Structural Genomics Consortium (NESG), Ontario Cancer Institute, Toronto, ON, Canada*
Min Zhou • *Center for Structural Genomics of Infectious Diseases, University of Chicago, Chicago, IL, USA; Midwest Center for Structural Genomics, Argonne National Laboratory, Argonne, IL, USA*
Matthew D. Zimmerman • *Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, VA, USA; Center for Structural Genomics of Infectious Diseases (CSGID), University of Virginia, Charlottesville, VA, USA; Midwest Center for Structural Genomics (MCSG), University of Virginia, Charlottesville, VA, USA; New York Structural Genomics Research Consortium (NYSGRC), University of Virginia, Charlottesville, VA, USA; Enzyme Function Initiative (EFI), University of Virginia, Charlottesville, VA, USA*

Chapter 1

Data Management in the Modern Structural Biology and Biomedical Research Environment

Matthew D. Zimmerman, Marek Grabowski, Marcin J. Domagalski, Elizabeth M. MacLean, Maksymilian Chruszcz, and Wladek Minor

Abstract

Modern high-throughput structural biology laboratories produce vast amounts of raw experimental data. The traditional method of data reduction is very simple—results are summarized in peer-reviewed publications, which are hopefully published in high-impact journals. By their nature, publications include only the most important results derived from experiments that may have been performed over the course of many years. The main content of the published paper is a concise compilation of these data, an interpretation of the experimental results, and a comparison of these results with those obtained by other scientists.

Due to an avalanche of structural biology manuscripts submitted to scientific journals, in many recent cases descriptions of experimental methodology (and sometimes even experimental results) are pushed to supplementary materials that are only published online and sometimes may not be reviewed as thoroughly as the main body of a manuscript. Trouble may arise when experimental results are contradicting the results obtained by other scientists, which requires (in the best case) the reexamination of the original raw data or independent repetition of the experiment according to the published description of the experiment. There are reports that a significant fraction of experiments obtained in academic laboratories cannot be repeated in an industrial environment (Begley CG & Ellis LM, Nature 483(7391):531–3, 2012). This is not an indication of scientific fraud but rather reflects the inadequate description of experiments performed on different equipment and on biological samples that were produced with disparate methods. For that reason the goal of a modern data management system is not only the simple replacement of the laboratory notebook by an electronic one but also the creation of a sophisticated, internally consistent, scalable data management system that will combine data obtained by a variety of experiments performed by various individuals on diverse equipment. All data should be stored in a core database that can be used by custom applications to prepare internal reports, statistics, and perform other functions that are specific to the research that is pursued in a particular laboratory.

This chapter presents a general overview of the methods of data management and analysis used by structural genomics (SG) programs. In addition to a review of the existing literature on the subject, also presented is experience in the development of two SG data management systems, UniTrack and LabDB. The description is targeted to a general audience, as some technical details have been (or will be) published elsewhere. The focus is on "data management," meaning the process of gathering, organizing, and storing data, but also briefly discussed is "data mining," the process of analysis ideally leading to an understanding of the data. In other words, data mining is the conversion of data into information. Clearly, effective

---

Matthew D. Zimmerman and Marek Grabowski have contributed equally to this work.

Wayne F. Anderson (ed.), *Structural Genomics and Drug Discovery: Methods and Protocols*, Methods in Molecular Biology, vol. 1140, DOI 10.1007/978-1-4939-0354-2_1, © Springer Science+Business Media New York 2014

data management is a precondition for any useful data mining. If done properly, gathering details on millions of experiments on thousands of proteins and making them publicly available for analysis—even after the projects themselves have ended—may turn out to be one of the most important benefits of SG programs.

Key words Databases, Data management, Structural biology, LIMS, PSI, CSGID

1 Introduction

1.1 Data in Structural Biology

Both structural genomics consortia and individual structural biology laboratories produce tremendous amounts of data, and having accurate, complete and consistent data is critical for reproducibility of biomedical research [1]. A single trip to a synchrotron for data collection by a productive crystallographic lab can generate hundreds of datasets totaling around 2 TB of raw data [2]. Modern data processing software can reduce, on the fly, a raw set of diffraction images into a single file that contains a description of every diffraction peak: Miller indices, intensity, and experimental uncertainty (sigma). These data are further reduced into one relatively small file that contains scaled and merged diffraction intensities. However, each file has to be associated with a particular sample (protein crystal) and the description of the experiment, which is usually written in the header of the diffraction image. These data are further used for structure determination and/or for function–structure relation studies.

To perform these studies the experimenter needs information about the protein (at a minimum, the protein sequence), crystallization conditions, and, for functional studies, protein production details. If this information is available, the process described above is simple to implement. Data harvesting from structure determination is relatively straightforward. The whole process following the placement of a crystal in the X-ray beam can be entirely controlled and captured by computer.

However, while this is very simple in theory, this simplicity has not yet been translated into practice. Analysis of the Protein Data Bank (PDB) [3, 4] shows that the number of data collection parameters marked as "NULL" in the header information (i.e., the detailed description of the experiment) is still significant [5, 6]. Moreover, data in the header are sometimes self-contradictory, contradictory to the experimental description in the paper citing the structure, or both [7, 8]. In that case, contacting the authors of the deposit and paper may be the only way to resolve the arising problems. Taking into account that only a small fraction, about 13 % [9], of structures determined by high-throughput consortia are converted (reduced) to peer-reviewed papers, the correctness of data uploaded to various databases like TargetTrack [10], TargetDB [11], and data banks like PDB is absolutely critical (see below).

1.2 Large-Scale Initiatives Create New Databases: TargetDB/ PepcDB/TargetTrack

Since their inception, many structural genomics efforts have adopted policies that experimental data produced by member consortia should be made available to the community from the moment of target selection. This has been particularly true for the two large initiatives from the National Institutes of Health (NIH): the Protein Structure Initiative (PSI) established in 2000 by the National Institute of General Medical Sciences (NIGMS) and the SG centers focusing on infectious diseases established in 2007 by the National Institute of Allergy and Infectious Diseases (NIAID). Even some partially privately funded SG efforts like the Structural Genomics Consortium (SGC) have established policies to release some experimental data to the general public [12] (typically only after the structure is determined and deposited). In the specific case of the centers funded by NIGMS and NIAID, the NIH established the target registration database, TargetDB [11], and required that all member consortia deposit data on the progress of their targets. Subsequently many other SG centers worldwide have deposited some of their experimental data as well.

Initially, the main purpose of TargetDB was the prevention of duplication of effort between different SG centers and maximization of the structural coverage of the protein fold space. The scope of the data was very modest. It included protein identification information (sequence, organism) and the timeline of changes in experimental status for each target. Status events included target selection, cloning, expression, purification, as well as crystallization, diffraction, determination of crystal structure, and PDB deposition (for targets studied by X-ray crystallography) or obtaining the HSQC spectra, determination of NMR structure, and BMRB/PDB deposition (for targets studied by NMR).

However, even the modest amount of data available in TargetDB permitted interesting analyses of the overall SG structure determination pipeline [13, 14]. In particular, the overall efficiency of the pipeline—the ratio of solved structures to clones—was found to be below 10 % even in the most productive centers. The two steps that contributed most to the failure of a target in the pipeline were production of soluble protein and diffraction-quality crystals. Not surprisingly, the success ratio depended very strongly on the type of protein as well as the methodology used by particular centers. There was not a single overall bottleneck factor. In 2004, TargetDB was extended to the Protein Expression, Purification, and Crystallization Database (PepcDB) [15] which in addition to simple status history included multiple trials, tracking of failed as well as successful experiments, and more detailed descriptions of protocols.

In 2010, PepcDB and TargetDB were merged into a single new database, TargetTrack, part of the new PSI-Structural Biology Knowledge Base (PSI-SBKB) [10, 16]. The new repository

extended the definition of a target to include protein–protein complexes and incorporated tracking of biological assays needed in the PSI:Biology phase. As of January 2013, TargetTrack contained data on over 300,000 targets and over 1,000 protocols.

1.3 Diverse Approaches to Data Management in SG Centers

Development of effective data management systems was a necessity for the large-scale SG centers, not only in order to provide the data to the scientific community but also particularly to effectively handle the huge amounts of experimental data, plan experiments, adjust experimental approaches (e.g., choice of cloning vectors, sequence truncation, crystallization conditions, structure determination procedures), and prioritize targets. These needs required gathering far more data than what was being required by TargetTrack.

In general, two levels of data management are needed in high-throughput, high-output structural biology programs: the *target tracking* level and the *experiment tracking* level. The target tracking level comprises target selection, overall experimental status of each target, center-wide efficiency statistics, and generation of reports to the public and to other databases such as TargetTrack. Almost all SG centers have a separate target-tracking database, though some functionality (e.g., target selection) can be "offloaded" to other specialized databases. The primary audience for the target-tracking level is everyone interested in a "high-level" view of the data produced by the center: the center's scientists and administrators as well as members of the scientific community with interest in the targeted proteins. This level is typically not designed for uploading new data or providing all details of individual experiments; these tasks are better handled at the experimental tracking level.

The experimental tracking level comprises the tools used to collect the results of experiments performed in the laboratory. This type of tool is generally known as a "laboratory information management system" or LIMS. LIMSs are typically used day to day by the researchers conducting the experimental work of a laboratory and may be highly customized to the protocols and work flow of a particular laboratory. LIMSs may also provide tools to help design experiments, operate laboratory equipment, semiautomatically harvest data, track the use of resources, etc. As a result, the primary audience for the LIMS is composed of those interested in a "low-level" view of the data, the center researchers themselves. As compared to the target-tracking level, it is not uncommon to use more than one LIMS in a single SG center, as different systems may be used in different laboratories.

It should be noted that splitting the data management system of a typical SG center into two distinct levels, "high-level" target tracking and "low-level" experiment tracking, is somewhat arbitrary. Some data are natural candidates to be kept at the LIMS

level only, for example, the location in the freezer where a particular clone is stored or the particular lot of a reagent or a crystallization buffer. Conversely, some data may only apply at the target-tracking level, for example, the number of publications referencing a given protein. In principle, it is possible for a single database and/or data management system to fully implement both levels. However, in practice, it seems that solutions where the two levels are implemented as separate systems/databases appear to be more common, especially for the larger scale projects.

There have been several "top-down" attempts to design a general framework for SG data management systems in the form of data dictionaries [17] or a protein production UML data model [18]. The latter has been implemented by several systems, such as HalX [19] or the Protein Information Management System (PiMS) [20] used by a number of European SG labs. However, most of the SG centers set up data management systems in a more ad hoc, "bottom-up" manner. Initially, some centers attempted to use commercial LIMS, but often these solutions were not flexible enough or even robust enough, and most SG centers developed their own solutions "in-house." There are exceptions to this rule. For example, the Structural Genomics Consortium uses two commercially available software systems: the Beehive LIMS (Molsoft LLC; http://www.molsoft.com/beehive.html) and Electronic Laboratory Notebook (now iLabber; Contur Software; http://www.contur.com/home/). It should be noted however that unlike many SG consortia, SGC does not deposit the results of its experiments to PepcDB or TargetTrack. Several of the SG-developed data management systems have been described in the literature [21–23], but to our knowledge, none of these systems have been fully commercialized.

One comprehensive data SG management system that has gained wider use is Sesame, developed by Zsolt Zolnai at Center for Eukaryotic Structural Genomics (CESG) [22]. It has been adopted by a number of labs and specialized centers.

The data management system for the Joint Center for Structural Genomics (JCSG) was developed by the center's programming team in parallel with the construction of the physical pipeline. The LIMS part of the system functions as a hub of information, recording all pipeline steps from target selection to deposition. The tracking database uses Oracle as its engine and tracks 424 experimental parameters, organized into 130 tables [24]. The tools and interfaces to the database contain approximately 360,000 lines of code, which illustrates the level of complexity of this and similar systems.

The Northeast Structural Genomics (NESG) consortium's data management system is organized as a "federated database framework," comprising a set of distributed, interconnecting databases [21]. The main target-tracking database, SPINE, serves

as an analysis system, utilizing data mining and machine learning tools. In particular, decision trees are used for predicting chances for protein solubility, successful purification, and crystallization. These predictions are used in directing targets to X-ray crystallography or NMR studies [14].

The other two large-scale PSI:Biology centers—the Midwest Center for Structural Genomics (MCSG) and the New York Structural Genomics Research Consortium (NYSGRC)—use the data management system developed in the Minor Lab at the University of Virginia. In both cases, the system is based on a collection of customized LIMS in each site laboratory and a central database (UniTrack, described below) that curates and unifies data obtained by various laboratories. In the case of MCSG, several different LIMSs are used in different laboratories, including LabDB, Mnemosyne, and ANL-DB. In NYSGRC, two different instances of LabDB are used. Similar systems are also deployed in the Center for Structural Genomics of Infectious Diseases (CSGID) and the Enzyme Function Initiative (EFI).

2 A Centralized Target Management System: UniTrack

The central, public system comprising the target-tracking level of the SG management system developed by the Minor Lab at the University of Virginia is named UniTrack. As mentioned above, the MCSG, NYSGRC, CSGID, and EFI consortia are all driven by variants of the UniTrack system. The system comprises a core abstraction based on 10 years of experience in SG data management, with a common database architecture and set of tools for managing target and experimental data. Each site is based on the UniTrack core but is then highly customized for the needs of the particular center or consortium of research laboratories. In each case, the UniTrack-derived system comprises the central tracking database and a set of auxiliary databases and applications, which collect and integrate experimental data and are provided by distributed LIMSs deployed in participating laboratories (Fig. 1). Experimental data from different LIMSs are combined and incorporated into UniTrack via a standard protocol. In the most basic case, each LIMS generates XML files in a predefined format, which are parsed by UniTrack tools. An alternative (and more efficient) method, where a LIMS directly communicates with the tracking database, has also been developed. The LIMSs can be very diverse; however, they all must be able to provide the minimum set of required data for cloning, expression, purification, and crystallization experiments.

The experimental pipeline starts with target selection and validation, which is specific for a particular center. The validation process is performed automatically and typically involves checking

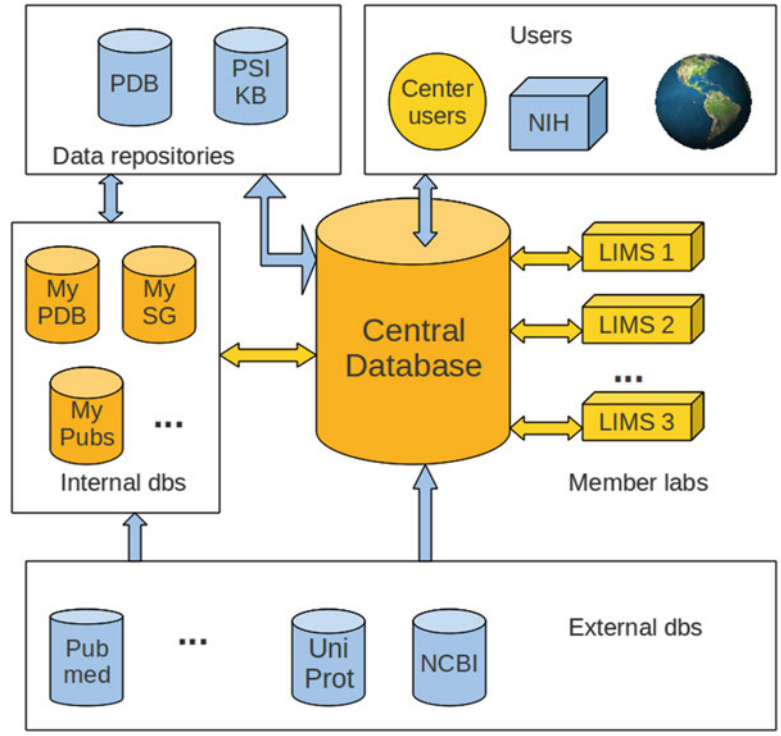


Fig. 1 The architecture of the UniTrack data management system. The central database interacts with LIMSs distributed in member labs. A number of auxiliary databases are used to store data from the PDB, data from other SG centers, and SG publications. The central database is responsible for producing reports for external data repositories such as PSI-SBKB. UniTrack databases are synchronized with external data sources such as NCBI GenBank, UniProt, and PubMed via custom scripts. Users interact with the system via a web interface

the accuracy of the amino acid and the nucleotide sequences as well as checking if the selected protein is homologous to proteins with structures in the PDB or to targets selected by other SG centers. Validated targets are inserted into the tracking database. Protein annotations and related data are automatically imported from external databases such as NCBI GenBank [25], Uniprot [26], PDB, and the PSI-SBKB. Depending on the needs of a particular center, between 30 and 80 attributes of any given protein target are stored in UniTrack.

UniTrack keeps a history and the results of the experiments for each target (Fig. 2). About 400 distinct data attributes are used to describe an experimental trial, from the cloning of a target through the determination of its structure. Almost all protein production and crystallization data can be automatically imported from the local LIMS or equipment database. However, smaller labs that do not have a LIMS deployed can still contribute data to UniTrack by entering it manually using the customized interface. Diffraction

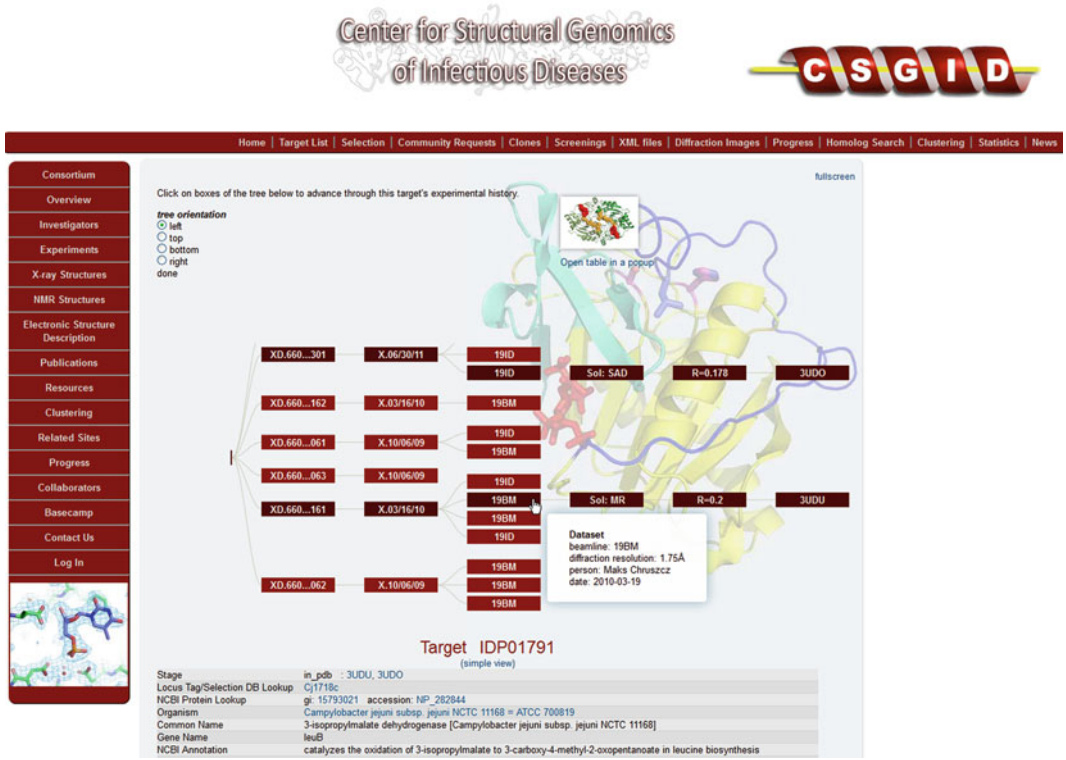


Fig. 2 Fragment of an experiment tree displayed in the UniTrack-based CSGID interface. *Boxes* represent particular experiments: purification (P), crystallization drop (XD), crystal harvest (X), data collection (beamline name), structure solution (Sol), refinement (R factor), and PDB deposit (PDB id). *Paths* in the tree represent trials for a particular sample. The *white box* that appears when the cursor hovers over an item displays additional details about a particular step. In addition, clicking on any of the boxes display all the data known about this step stored in the database

and structure determination data is currently imported automatically only from the LabDB instances that have the *hkldb* module enabled [27]. Researchers in other labs upload scaling logs and refinement files manually via the interface.

The tracking database also generates real-time internal reports and statistics as well as the XML files that are being submitted to the TargetTrack repository. In addition, the periodic reports required by various bodies are generated in real time from the database and accessible to the general public. In some sense, all of the portions of UniTrack that generate publicly accessible web pages serve as reports.

The customized instances of UniTrack for each center drive dynamic parts of the centers' corresponding web portals. The web interfaces are implemented using the Model–View–Controller (MVC) architecture, with separate layers for data retrieval (model), "business logic" (controller), and web page rendering (view).

Even with the use of the CakePHP MVC framework (http://cakephp.org) the customized web interfaces for the centers are quite complicated; as an example, the implementation of the CSGID web interface contains over 50,000 lines of source code.

2.1 The LabDB "Super-LIMS"

LabDB is a modular "super-LIMS," originally developed to track the structure determination pipeline from cloning to structure determination (Fig. 3). The central component of the system is a PostgreSQL database server coupled with a web-based framework, along with two specialized tools: *Xtaldb*, for designing and tracking crystallization experiments, and *hkldb*, a module of the HKL-2000/3000 system [27] for incorporating information from crystallographic data collection and structure determination. *hkldb* and Xtaldb can also be used with stand-alone databases.

One of the fundamental design goals of LabDB is to harvest data automatically or semiautomatically from laboratory equipment whenever possible. To that end, the system has modules to import data from a variety of different types of laboratory equipment, including chromatography systems (GE Healthcare AKTA systems), electrophoresis documentation and separation systems

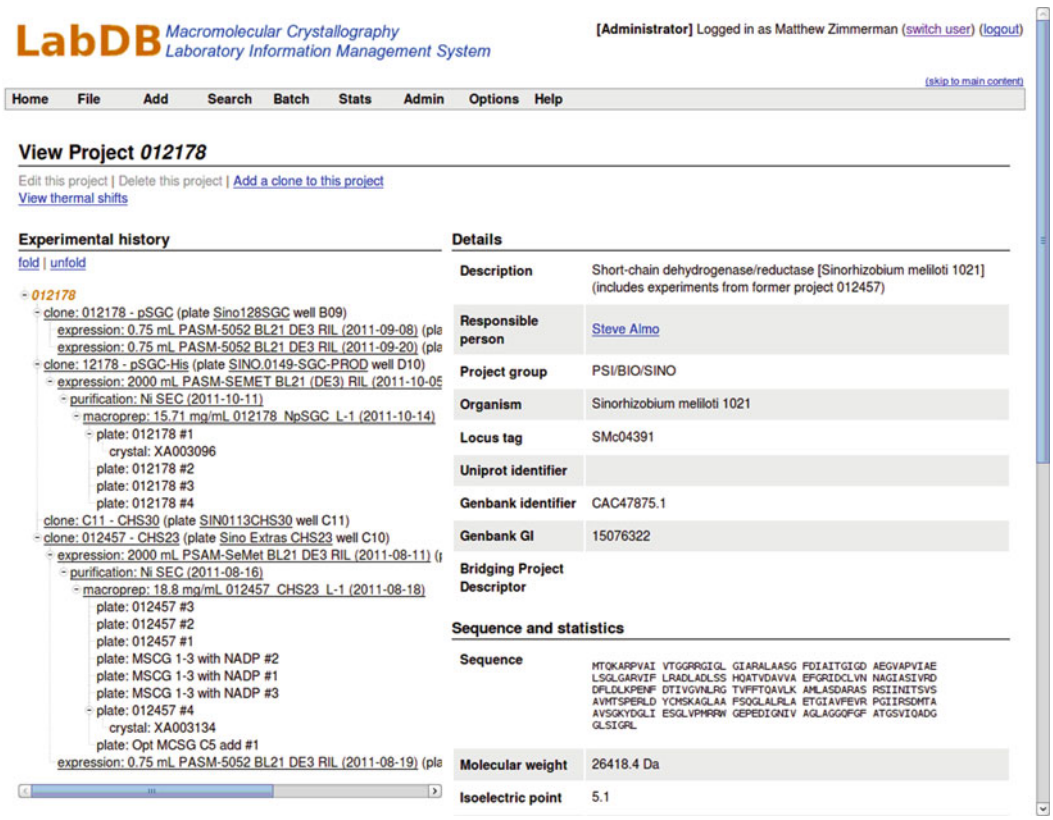


Fig. 3 A typical target overview page in the LabDB LIMS