Structure *14*, 1211–1217, August 2006 ©2006 Elsevier Ltd All rights reserved DOI 10.1016/j.str.2006.06.005

Meeting Report

Outcome of a Workshop on Archiving Structural Models of Biological Macromolecules

Helen M. Berman,[1,*] Stephen K. Burley,[2] Wah Chiu,[3] Andrej Sali,[4] Alexei Adzhubei,[5] Philip E. Bourne,[6] Stephen H. Bryant,[7] Roland L. Dunbrack, Jr.,[8] Krzysztof Fidelis,[9] Joachim Frank,[10] Adam Godzik,[11] Kim Henrick,[12] Andrzej Joachimiak,[13] Bernard Heymann,[14] David Jones,[15] John L. Markley,[16] John Moult,[17] Gaetano T. Montelione,[18] Christine Orengo,[19] Michael G. Rossmann,[20] Burkhard Rost,[21] Helen Saibil,[22] Torsten Schwede,[23] Daron M. Standley,[24] and John D. Westbrook[1]

[1] The Research Collaboratory for Structural Bioinformatics Protein Data Bank
Rutgers, The State University of New Jersey
Piscataway, New Jersey 08854
[2] SGX Pharmaceuticals, Inc.
San Diego, California 92121
[3] Department of Biochemistry & Molecular Biology
National Center for Macromolecular Imaging
Baylor College of Medicine
Houston, Texas 77030
[4] Department of Biopharmaceutical Sciences
University of California, San Francisco
San Francisco, California 94143
[5] The Biotechnology Centre of Oslo
University of Oslo
Blindern
N-0317 Oslo
Norway
[6] The Research Collaboratory for Structural Bioinformatics Protein Data Bank
University of California, San Diego
San Diego Supercomputer Center
La Jolla, California 92093
[7] National Center For Biotechnology Information
National Library of Medicine
National Institutes of Health
Bethesda, Maryland 20894
[8] Institute for Cancer Research
Fox Chase Cancer Center
Philadelphia, Pennsylvania 19111
[9] University of California, Davis
Genome and Biomedical Sciences Facility
Davis, California 95616
[10] Howard Hughes Medical Institute
Department of Biomedical Sciences
Wadsworth Center
New York State Department of Health
Albany, New York 12201
[11] Bioinformatics and Systems Biology Program
The Burnham Institute for Medical Research
La Jolla, California 92037
[12] Macromolecular Structure Database
EMBL Outstation Hinxton
European Bioinformatics Institute, Hinxton
Cambridge CB10 1SD
United Kingdom
[13] Department of Structural Biology Center
Argonne National Lab
Argonne, Illinois 60439
[14] Laboratory of Structural Biology Research
National Institute of Arthritis and Musculoskeletal and Skin Diseases
National Institutes of Health
Bethesda, Maryland 20892
[15] Department of Computer Science
Bioinformatics Unit
University College London
WC1E 6BT London
United Kingdom
[16] BioMagResBank
Department of Biochemistry
University of Wisconsin-Madison
Madison, Wisconsin 53706
[17] Center for Advanced Research in Biotechnology
University of Maryland Biotechnology Institute
University of Maryland
Rockville, Maryland 20850
[18] Rutgers, The State University of New Jersey
Department of Molecular Biology and Biochemistry
Center for Advanced Research in Biotechnology
Piscataway, New Jersey 08854
[19] Biomolecular Structure & Modeling Unit
Department of Biochemistry & Molecular Biology
University of College London
London WC1E 6BT
United Kingdom
[20] Department of Biological Science
Purdue University
West Lafayette, Indiana 47907
[21] Department of Biochemistry & Molecular Biophysics
Columbia University
New York, New York 10032
[22] Department of Crystallography
Birkbeck College London
Bloomsbury Centre for Structural Biology
London WC1E 7HX
United Kingdom
[23] Division of Bioinformatics
Biozentrum
University of Basel
CH-4056 Basel
Switzerland
[24] Protein Data Bank Japan
Institute for Protein Research
Osaka University
Osaka 565-0871
Japan

*Correspondence: berman@rcsb.rutgers.edu

This paper describes the outcome of a "Workshop on Biological Macromolecular Structure Models" held in November 2005 in which experimentalists and modelers discussed the best way to archive models of biological macromolecules.

Background and Goals of the Workshop

We have entered a new era in structural biology in which many methods will be used in concert to derive structural information for proteins. The National Institute of General Medical Sciences (NIGMS)-funded Protein Structure Initiative (PSI) promises to place thousands of new structures into the public domain, each of which will be representative of many homologous protein sequences. In addition, the emergence of cryo-electron microscopy (cryo-EM) as a powerful method for structure determination of macromolecular complexes has highlighted the central role of homology models for interpretation of cryo-EM density maps. Moreover, it is not straightforward to draw a bright line between experimentally determined structures and computed structural models. Indeed at some level, every experimental structure in the Protein Data Bank (PDB) archive is a model, albeit a model based on structural measurement. Thus, there needs to be a clear distinction between classes of structural models: ab initio, homology, and experimentally derived structures/models.

The PDB policy regarding archiving of theoretical models has been ambiguous. Although models have been accepted during the long history of the PDB (Berman et al., 2000, 2003; Bernstein et al., 1977), it is only recently that these models have been put in separate data storage areas. In addition, there has never been a clear policy for how best to validate these models.

In order to help resolve the many issues surrounding the archiving of and access to models, a workshop was held in November 2005 at Rutgers, The State University of New Jersey, under the sponsorship of the Research Collaboratory for Structural Bioinformatics Protein Data Bank. Participants included experimental biologists, scientists with expertise in structure determination by X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and cryo-EM, and computational biologists with expertise in modeling. The goal of this workshop was to open a dialog between experimentalists and modelers so that the PDB can most usefully meet the needs of the scientific community regarding models of biological macromolecules.

At the workshop, the following issues were explored: needs of the modeling community with respect to the archiving of models, needs of the experimentalists with respect to the availability of models, current limitations of theoretically derived models, how the scientific community at large can be best served with respect to the availability and annotation of models, and quantitative measures that should be used to assess the accuracy of models.

In the following sections, we first describe the current state of modeling, how models are used in interpreting electron density maps derived from cryo-EM, and the impact of structural genomics on models and vice versa. We then present key recommendations that emerged from this workshop as well as some ideas as to how best to implement them.

Introduction

Types of Theoretical Models and Current Resources

Protein structure prediction methods include many types (Baker and Sali, 2001; Marti-Renom et al., 2000; Petrey and Honig, 2005) that differ in terms of the input information used and the aspects of protein structure predicted. Some examples follow: (1) the secondary structure can be predicted from a protein sequence (Rost, 2003), (2) an atomic and reduced representation model of a domain can be obtained from the sequence alone by ab initio or de novo prediction methods (Schueler-Furman et al., 2005), (3) fold assignment and sequence-structure alignment can be achieved by threading against a library of known folds (Godzik, 2003), (4) an atomic model of a protein can be calculated on the basis of known template structures by using comparative protein structure or homology modeling (Madhusudhan et al., 2005), and (5) atomic and reduced representation models of protein complexes with small ligands and other macromolecules, such as nucleic acids, can be derived with various physics-based docking methods (Shoichet, 2004). Increasingly, hybrid methods rely on more than one type of information, especially for the structural characterization of protein assemblies (Alber et al., 2004). For example, some methods for flexible docking into density maps from cryo-EM depend on physics-based scoring and comparative modeling (Topf and Sali, 2005). Certain hybrid methods begin to blur the distinction between models based primarily on theoretical considerations and those based primarily on experimental findings from the characterized system.

Various prediction methods are often available as standalone computer programs and increasingly as web servers (see the *Nucleic Acid Research* Database Issue [2006] for examples). Correspondingly, models can be calculated automatically, although manual intervention can still provide an advantage, especially in more difficult cases (Kryshtafovych et al., 2005). Time invested in the preparation of a model varies from seconds or minutes of CPU time to many weeks of CPU and human time. So-called web metaservers can take an input sequence, submit it to a variety of protein structure prediction servers, collate the results, and return an ensemble of the modeling results to the user (Ginalski et al., 2003; Rost et al., 2004). Other helpful resources are databases of precalculated comparative protein structure models for protein sequences (Kopp and Schwede, 2006; Pieper et al., 2006). Modeling of protein structures is also greatly facilitated by various databases available on the Internet, including primary databases of protein sequences (Bairoch et al., 2005), experimentally determined structures (Berman et al., 2003), structural classifications, and sequence alignment (Andreeva et al., 2004; Marti-Renom et al., 2001; Pearl et al., 2005). Finally, modeling is also supported by programs and web servers intended for assessing different prediction methods using the experimentally determined structures as the reference (Koh et al., 2003; Rychlewski and Fischer, 2005) and for predicting errors in a model when the actual structure is unknown (Hooft et al., 1996; Luthy et al., 1992; Melo et al., 1997; Sippl, 1995; Zhou and Zhou, 2005).

Cryo-electron Microscopy and Models

Cryo-EM is an emerging structural technique for studying three-dimensional structures of multicomponent macromolecular complexes with masses >0.5 million Daltons (Chiu et al., 2005; Frank, 2002). Electron cryotomography is a promising tool for visualizing molecular

landscapes inside a living cell in its native state (Baumeister, 2004). The structural resolution possible with cryo-EM, which ranges from 2 to 100 Å, can reveal corresponding details ranging from the polypeptide backbone and secondary structural elements to gross molecular size and shape. In the highest resolution studies with two-dimensional crystalline membrane protein arrays, water and lipid molecules can be also visualized (Gonen et al., 2005). Examples of cryo-EM studies of macromolecular complexes include membrane proteins, cytoskeletal complexes, ribosomes, quasi-spherical viruses, molecular chaperones, flagella, ion channels, and oligomeric enzymes.

This imaging technique is complementary to X-ray crystallography and NMR spectroscopy in that it is better suited to the study of large complexes in different physiological or chemical states. Numerous investigations have shown how cryo-EM can capture different conformations of a complex undergoing a physiological process (Gao et al., 2003; Heymann et al., 2003; Jiang et al., 2003; Subramaniam and Henderson, 2000). Modeling of individual components of the macromolecular complex is an important method for extracting maximum structural knowledge from a cryo-EM structural study. Various modeling approaches that utilize cryo-EM density maps as a constraint in deriving a pseudo atomic model of the molecular components within a large complex include:

- Rigid body fitting of crystal structures of components to an cryo-EM density map (Kuhn et al., 2002)
- Flexible fitting of crystal structures of components to a cryo-EM density map of a complex (Mitra et al., 2005; Tilley et al., 2005)
- Sequence-based modeling of components such as homology or ab initio modeling combined with cryo-EM density map restraints (Topf et al., 2006)
- Integration of bioinformatics, biochemical and biophysical properties, and cryo-EM density map for model building (Zhou et al., 2001)

Generation of a model and fitting it within a cryo-EM density map can be carried out either manually through direct visualization, or computationally with quantitative evaluation. Various software packages with figures of merit such as R factor, scoring function, correlation function, and goodness of fit have been developed for measuring the best fit of the model with the experimental cryo-EM density map. The accuracy of any cryo-EM-based model must be validated, not only by quantitative estimates but also by direct visualization and by estimates of consistency with various biophysical and biochemical findings. Because of the significant likelihood of conformational differences between the crystal and biological states, additional research leading to the development of reliable methods for validating cryo-EM-based models is essential. The important structural information from models of authentic biological structures derived from cryo-EM studies is expected to increase sharply in coming years as fueled by progress in structural biology and structural genomics.

Structural Genomics and Models

Worldwide structural genomics efforts aim to expand our structural knowledge of proteins (Stevens et al., 2001). Structural genomics focuses on high-throughput structure determination of novel proteins and takes advantage of genome sequence data to select proteins for structural studies. Genome sequencing efforts continue to add rapidly novel protein sequences, and new protein families continue to grow along with the added new genomes. Structural characterization of novel proteins accelerated considerably in recent years, mainly through structural genomics contributions, although this effort continues to lag behind sequence data. Hence, structure determination of all proteins encoded by sequenced genomes appears to be an unrealistic goal at present, and other approaches to the development of useful structural models need to be considered.

The NIGMS-supported PSI was recently funded for the second stage (http://sg.pdb.org/funding.html). The PSI applies structure determination pipelines to a large number of protein sequence families for which no structural information is available and proposes to determine structures of several thousand novel proteins over the next 5 years. The methods used for structure determination are X-ray crystallography and NMR spectroscopy in which the atomic coordinates are determined directly from experimental data. Current PSI efforts support coarse granularity coverage of large protein families with the goal of covering as many protein sequences as possible. The current view is that the novel protein structures determined by structural genomics efforts will serve as valuable templates to generate a large number of three-dimensional homology models by applying advanced computational approaches to all sequence-related proteins found in nature. It is recognized that the current limitations of homology modeling are overcome by the availability of accurate three-dimensional structures of homologous proteins, especially for "high-value" proteins of biomedical importance. Current PSI strategies for target selection are aimed at maximizing homology modeling output. At present, this approach appears to be the most cost-effective strategy for providing fairly accurate structural models for a significant proportion of proteome sequences.

Therefore, it is anticipated that in the near future a large number of three-dimensional protein models, based upon limited experimental data, will be generated by homology modeling. The number of three-dimensional homology models already far exceeds the number of experimental structures deposited in the PDB archive. Thus, the structural biology community must address this issue and develop proper policies and appropriate strategies for handling and distributing such data. Three major issues raised by structural genomics and other homology modeling efforts pertain to the structural biology community and the PDB: (1) how to accommodate the three-dimensional homology models within the current system of databases and provide uniform public access, (2) how to consistently assess the quality of these three-dimensional models, and (3) where to store these models for public access and what role the PDB should play.

These questions need to be addressed as quickly as possible. Public discussion of the role of the PDB was initiated during the workshop with the goal of proposing initial guidelines for consideration by the structural biology community.

Recommendations

Recommendation 1

PDB depositions should be restricted to atomic coordinates that are substantially determined by experimental measurements on specimens containing biological macromolecules.

The PDB archive is a global resource that serves as a freely available, unified archive of experimentally determined three-dimensional structures of biological macromolecules and associated primary data. Although three-dimensional structural models are derived from a combination of experimental and theoretical techniques, workshop participants unanimously agreed that the PDB should contain only models substantially based on experimental measurements such as X-ray crystallography, NMR spectroscopy and cryo-EM; these structures include those obtained by docking or modeling atomic structures into cryo-EM maps. In addition, there was unanimous agreement that models derived principally by theoretical techniques should be made publicly available via mechanisms other than the PDB.

Recommendation 2

A central, publicly available archive (or technical equivalent thereof) or portal should be established for models that are the explicit subject of peer review. A central mechanism of access (a portal) should be established to permit systematic interrogation of 3D macromolecular structural information (both models and experimental structures and their provenance).

Overview of the Portal. A portal is synonymous with a gateway and is a World Wide Web site that is or proposes to be a major starting site for users by offering an array of resources and services, such as most of the traditional search engines. What is envisioned here is a model-orientated niche or vertical portal (or multiple portals adhering to comparable data standards) permitting access to information on three-dimensional structures of biological macromolecules and biological systems derived from experimental data and theoretical modeling. The portal would primarily be a collection of descriptions of resources, and pointers to those resources on model data and an essential starting position would be to define a data standard. The portal should be flexible and allow various modes of accepting/presenting data. The portal needs to be sensitive to new ideas and methods emerging in the community. Information served will come from various sites, including the PDB and current holdings of the various research centers that produce structural models. The sources of information should be acknowledged. Each model should be accompanied by an estimate of its accuracy. It is recommended that authors who use models in their publications (either created by themselves or obtained from a modeling site) deposit these models in a publicly available archive (to be established) to ensure access for peer review. This archive will also be accessible from the portal. It is envisioned that validation sites could also be associated with the portal as users of models; successful assessment methods developed by these validation sites could become accepted methods for evaluating models.

Data Standards. The portal and associated sites will utilize standards endorsed by the community. Models will consist of three-dimensional coordinates and information on how the model was derived, date of creation, authors, and estimated accuracy (overall, per segment, per residue, per atom). Each model should have a unique static identifier. Standardization of resource descriptors can be similar to the Dublin Core Metadata Initiative (http://dublincore.org/) as used in part by The UK National Crystallography Service Grid Facility (Coles et al., 2006) (see http://www.ncs.chem.soton.ac.uk/) and the UK eBank Project (Hey and Trefethen, 2005) (see http://www.ukoln.ac.uk/projects/ebank-uk/schemas/).

Specific Features of the Model Archive. Each model submitted to the model archive will be curated. Models and metadata will be checked for proper nomenclature and quality assessment requirements. Each model will be issued a stable, unique identifier that can be included in the publication.

Recommendation 3

It was unanimously agreed that methods for assessing model quality are essential for the integrity and long-term success of any publicly available model portal, either from a central repository or a set of linked resources. It was, however, acknowledged that currently there was no consensus as to which single method or group of methods should be applied.

Identifying the most appropriate existing evaluation methods or developing novel methods was recognized as a challenging research problem that must be addressed. Workshops are already being planned in this area by the Protein Structure Prediction Center (University of California, Davis), and further workshops will be needed in the near future.

Accuracy assessment metrics can be derived for both entire models and for segments or individual residues. Both clearly have value. At present, there are three main approaches to deriving both local and global quality metrics for a nonexperimental protein model.

The first ensemble of approaches are based on statistical treatments wherein models are evaluated on the basis of expected structural parameters, such as main chain stereochemistry, long-range protein contacts and solvation properties (e.g., PROSA II [Sippl, 1993], ProQ [Wallner and Elofsson, 2003], MODCHECK [Pettitt et al., 2005]). These methods have the advantage that they can be applied to a three-dimensional structure without any additional supporting information. Benchmarking studies have shown that these methods give limited estimates of model quality and cannot be relied upon to accurately rank models under all circumstances. This area needs substantial additional research.

The second set of approaches used to derive model quality estimates are based on supporting information used in the modeling process, such as alignment quality, number of homologs in PDB or the sequence databases, secondary structural features, and agreement with existing experimental data. This approach is frequently used by individual authors or individual modeling methods to derive internal quality estimates. Deriving a single approach that could evaluate supporting information for diverse modeling methods and different combinations of supporting information would be a significant if not impossible challenge.

The third approach is to derive "community-based" statistics, wherein models from different sources are compared and regions of agreement are identified and

evaluated. From previous CASP results, this approach has proven very effective in metaprediction approaches such as 3D-Jury (Ginalski et al., 2003) or Pmodeller (Wallner et al., 2003) and may well form part of the assessment protocol. Nevertheless, this approach suffers from a number of problems, including how to handle outlier methods that produce models substantially different from those falling within the group consensus. It was acknowledged that in some circumstances, these outlying predictions might be accurate.

It is clear that a core set of, as yet to be agreed upon, validation methods must be applied to all models presented by the portal. Ideally, concurrence with this philosophy should be a prerequisite for participation in the portal, but it was also felt that groups with novel approaches should not be discouraged. In short, we advocate quality assessment, but not a quality veto.

Community consensus on this core set of evaluation methods will be needed, and regular review thereof will be required. The basis of this review could be the continuous assessment statistics provided by the portal itself. Such continuous assessment would not only evaluate individual modeling resources but also various assessment methods. Effective presentation of evaluation statistics would be critical to the success of the portal.

Calculation of quality metrics for potentially large number of models would present a significant computational challenge. A possible solution would be to delegate part of the responsibility for metric calculation to the model contributors themselves, but ideally the majority of the assessment should be carried out by the portal itself to facilitate evolution of new assessment methods for application to archived models.

Implementation of Recommendations

Relationship between a "Central" Portal of Models and the PDB

The proposed portal aims to provide easy, transparent access to macromolecular structural data derived from various sources of experimental, observational, and simulation data and kept on a multitude of systems and sites. We present here one vision for such a portal.

We believe that this would be best realized as a grid, with PDB being one of many nodes on the grid, playing additional role as one of foundation management groups that would set the standards and issues covered by the data portal. Modbase (Sanchez et al., 2000), SwissModel (Kopp and Schwede, 2004), and other to-be-developed resources contributing and exchanging model metadata via a common standard would form other nodes in the grid. The model depository or archive could be also one of such nodes. The portal will work as a broker between the scientists, the facilities, the data, and other services. Furthermore, it will provide links to other web/grid services, which will allow the scientists to further use the selected data as shown (Figure 1, based on the CCLRC Data Portal [Matthews and Sufi, 2002], http://tiber.dl.ac.uk:8080/).

The portal would have similarities to caBIG (cancer Biomedical Informatics Grid, http://cabig.nci.nih.gov/), a voluntary network or grid connecting individuals and institutions to enable the sharing of data and tools to create a World Wide Web of cancer research and BIRN (Biomedical Informatics Research Network, http://www.nbirn.net/), a National Institutes of Health initiative that fosters distributed collaborations in biomedical science by utilizing information technology innovations. Currently, the BIRN includes 21 universities and 30 research groups that participate in one or more of three test bed projects centered on brain imaging of human neurological disorders and associated animal models. The mission of the BIRN is to accelerate discovery science by creating and fostering a new biomedical collaborative culture and infrastructure.

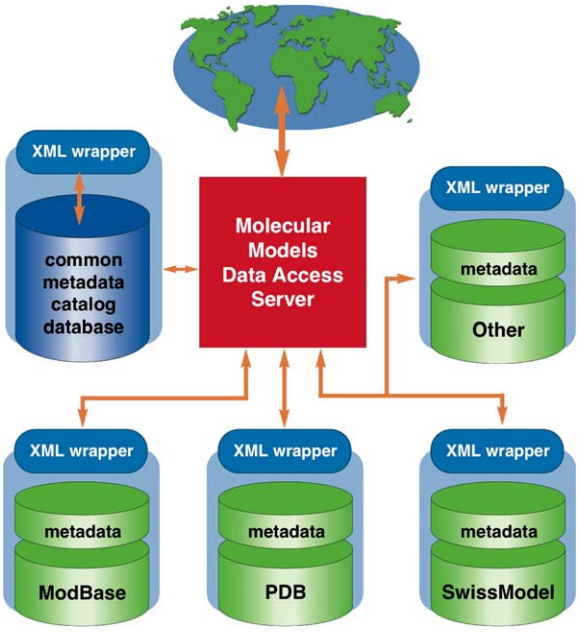

Figure 1. A Schematic for the Proposed Portal

The portal is envisioned to be a modular web services architecture achieved by using an implementation of the Simple Object Access Protocol (SOAP, http://www.w3.org/TR/soap/) that allows for seamless data exchange between the portal and all registered contributors. SOAP is a lightweight protocol for exchange of information in a decentralized, distributed environment. It is an XML-based protocol, which defines a framework for representing remote procedure calls and responses. The XML wrappers basically map the individual contributing metadata format into an XML format that the data portal understands (Drinkwater et al., 2004). These services will be platform and language independent allowing other services (other portals or clients) to communicate with the data portal (see http://www.e-science.clrc.ac.uk/web/projects/dataportal/).

This type of network of connected components, wherein a component can be an application (e.g., assessor software developers) described in an XML schema, is similar to portal developments at the University of Indiana (http://www.extreme.indiana.edu/).

Data Standards for Models Coming from a "Central" Portal

Theoretical models have traditionally required significant time to curate. The expected increase in the number of new models will require automated curation/deposition. Such automation, and the need for the portal to communicate with modeling servers, will require well-defined data standards. The portal will need a new conceptual data model that will cover existing and new types of structural models. This model will incorporate common concepts from the PDB Exchange Dictionary (Westbrook et al., 2005) and evolve to include

new concepts required to describe new kinds of models. This approach has already been used in the modeling extension (MDB) and cryo-EM exchange (Bsoft) dictionaries (see http://mmcif.pdb.org/).

There should be a policy of "minimal data standards" plus the flexibility to support new types of models. The portal should require compliance with its data standard for both depositors and communicating databases. By requiring compliance, the portal will facilitate automatic annotation and validation and the ability to query all data. This measure will ensure transparent interoperability between the portal and the PDB. Where possible, the portal will adopt existing conventions from other appropriate data resources, including the PDB and the sequence databases.

Access Models for a Central Portal of Models

The implementation of the recommended portal would require extensive planning and resources obtained through the grant mechanisms. Given below are some of the specifications that were discussed during the workshop.

The definition of the portal is a single point of entry to a set of local and distributed information resources for structural models not based on experimental measurements (see Recommendation 1). The minimum contents for this portal require a unique identifier for each model registered with the system, each model's polypeptide chain sequence, and quality assessment information (Recommendation 3).

Additional information should be available, including: keywords, structural motifs, standard test sets of data, bound ligands, domains, flexibility, surface electrostatic properties, coding and noncoding SNPs, alternative splicing, oligomeric state, macromolecular interactions, literature references, subcellular localization, pathways, transcript profiling, and drugability.

Access to these data should be free and constantly available to a diverse worldwide user community of both model producers and users. Several levels of access are required for the different levels of users of the portal.

Data producers should have the ability to automatically upload multiple models efficiently, if desired. By using the data portal, data users should be able to:

- Immediately access models from local or remote archives
- Comment on the contents of the resource
- Query one or more remote modeling resources based on the queries defined by the content described above
- Automatically access metaservers for calculating models on the fly
- Perform simple and advanced review of models, which includes scoring analysis and visualization of one or multiple models—implies simplified views of large volumes of data
- Download (both manual and automated) contents of the local resource as well as remote content (as defined by that resource)

Applications should have access to the complete contents of the local archive as well as a defined mechanism to download content from remote resources. The portal should also track usage statistics to help, in part, to enable an understanding of how modeling resources are being used by the biological community.

Acknowledgments

This workshop was funded through a supplement to the Research Collaboratory for Structural Bioinformatics Protein Data Bank. The Research Collaboratory for Structural Bioinformatics Protein Data Bank is supported by funds from the National Science Foundation, National Institute of General Medical Sciences, the Office of Science, Department of Energy, the National Library of Medicine, the National Cancer Institute, the National Center for Research Resources, the National Institute of Biomedical Imaging and Bioengineering, and the National Institute of Neurological Disorders and Stroke.

References

Alber, F., Eswar, N., and Sali, A. (2004). Structure determination of macromolecular complexes by experiment and computation. In Practical Bioinformatics, Volume 15, J. Bujnicki, ed. (New York: Springer), pp. 73–96.

Andreeva, A., Howorth, D., Brenner, S.E., Hubbard, T.J., Chothia, C., and Murzin, A.G. (2004). SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res. *32*, D226–D229.

Bairoch, A., Apweiler, R., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., et al. (2005). The Universal Protein Resource (UniProt). Nucleic Acids Res. *33*, D154–D159.

Baker, D., and Sali, A. (2001). Protein structure prediction and structural genomics. Science *294*, 93–96.

Baumeister, W. (2004). Mapping molecular landscapes inside cells. Biol. Chem. *385*, 865–872.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res. *28*, 235–242.

Berman, H.M., Henrick, K., and Nakamura, H. (2003). Announcing the worldwide Protein Data Bank. Nat. Struct. Biol. *10*, 980.

Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977). Protein Data Bank: a computer-based archival file for macromolecular structures. J. Mol. Biol. *112*, 535–542.

Chiu, W., Baker, M.L., Jiang, W., Dougherty, M., and Schmid, M.F. (2005). Electron cryomicroscopy of biological machines at subnanometer resolution. Structure *13*, 363–372.

Coles, S.J., Frey, J.G., Hursthouse, M.B., Light, M.E., Milsted, A.J., Carr, L.A., DeRoure, D., Gutteridge, C.J., Mills, H.R., Meacham, K.E., et al. (2006). An e-science environment for service crystallography—from submission to dissemination. J. Chem. Inf. Model *46*, 1006–1016.

Drinkwater, G., Sufi, S., Manandhar, A., Blanshard, L., Kleese, K. (2004). Data management with the CCLRC data portal. In Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications, PDPTA '04, June 21–24, 2004, Las Vegas, Nevada, USA, Volume 2, H.R. Arabnia and J. Ni, eds. (CSREA Press), 815–821.

Frank, J. (2002). Single-particle imaging of macromolecules by cryo-electron microscopy. Annu. Rev. Biophys. Biomol. Struct. *31*, 303–319.

Gao, H., Sengupta, J., Valle, M., Korostelev, A., Eswar, N., Stagg, S.M., Van Roey, P., Agrawal, R.K., Harvey, S.C., Sali, A., et al. (2003). Study of the structural dynamics of the *E. coli* 70S ribosome using real-space refinement. Cell *113*, 789–801.

Ginalski, K., Elofsson, A., Fischer, D., and Rychlewski, L. (2003). 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics *19*, 1015–1018.

Godzik, A. (2003). Fold recognition methods. Methods Biochem. Anal. *44*, 525–546.

Gonen, T., Cheng, Y., Sliz, P., Hiroaki, Y., Fujiyoshi, Y., Harrison, S.C., and Walz, T. (2005). Lipid-protein interactions in double-layered two-dimensional AQP0 crystals. Nature *438*, 633–638.

Hey, T., and Trefethen, A.E. (2005). Cyberinfrastructure for e-Science. Science *308*, 817–821.

Heymann, J.B., Cheng, N., Newcomb, W.W., Trus, B.L., Brown, J.C., and Steven, A.C. (2003). Dynamics of herpes simplex virus capsid maturation visualized by time-lapse cryo-electron microscopy. Nat. Struct. Biol. *10*, 334–341.

Hooft, R.W., Vriend, G., Sander, C., and Abola, E.E. (1996). Errors in protein structures. Nature *381*, 272.

Jiang, W., Li, Z., Zhang, Z., Baker, M.L., Prevelige, P.E., and Chiu, W. (2003). Coat protein fold and maturation transition of bacteriophage P22 seen at sub-nanometer resolutions. Nat. Struct. Biol. *10*, 131–135.

Koh, I.Y., Eyrich, V.A., Marti-Renom, M.A., Przybylski, D., Madhusudhan, M.S., Eswar, N., Grana, O., Pazos, F., Valencia, A., Sali, A., and Rost, B. (2003). EVA: evaluation of protein structure prediction servers. Nucleic Acids Res. *31*, 3311–3315.

Kopp, J., and Schwede, T. (2004). The SWISS-MODEL Repository of annotated three-dimensional protein structure homology models. Nucleic Acids Res. *32*, D230–D234.

Kopp, J., and Schwede, T. (2006). The SWISS-MODEL Repository: new features and functionalities. Nucleic Acids Res. *34*, D315–D318.

Kryshtafovych, A., Venclovas, C., Fidelis, K., and Moult, J. (2005). Progress over the first decade of CASP experiments. Proteins *61* (Suppl.), 225–236.

Kuhn, R.J., Zhang, W., Rossmann, M.G., Pletnev, S.V., Corver, J., Lenches, E., Jones, C.T., Mukhopadhyay, S., Chipman, P.R., Strauss, E.G., et al. (2002). Structure of dengue virus: implications for flavivirus organization, maturation, and fusion. Cell *108*, 717–725.

Luthy, R., Bowie, J.U., and Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. Nature *356*, 83–85.

Madhusudhan, M.S., Marti-Renom, M.A., Eswar, N., John, B., Pieper, U., Karchin, R., Shen, M.-Y., and Sali, A. (2005). Comparative protein structure modeling. In The Proteomics Protocols Handbook, J. Walker, ed. (Totowa, NJ: Humana Press), pp. 831–860.

Marti-Renom, M.A., Stuart, A.C., Fiser, A., Sanchez, R., Melo, F., and Sali, A. (2000). Comparative protein structure modeling of genes and genomes. Annu. Rev. Biophys. Biomol. Struct. *29*, 291–325.

Marti-Renom, M.A., Ilyin, V.A., and Sali, A. (2001). DBAli: a database of protein structure alignments. Bioinformatics *17*, 746–747.

Matthews, B., and Sufi, S., eds. (2002). The CLRC scientific metadata model, version 1(http://epubs.cclrc.ac.uk/work-details?w=29024).

Melo, F., Devos, D., Depiereux, E., and Feytmans, E. (1997). ANOLEA: a www server to assess protein structures. ISMB *5*, 187–190.

Mitra, K., Schaffitzel, C., Shaikh, T., Tama, F., Jenni, S., Brooks, C.L., Ban, N., and Frank, J. (2005). Structure of the *E. coli* protein-conducting channel bound to a translating ribosome. Nature *438*, 318–324.

Pearl, F., Todd, A., Sillitoe, I., Dibley, M., Redfern, O., Lewis, T., Bennett, C., Marsden, R., Grant, A., Lee, D., et al. (2005). The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. Nucleic Acids Res. *33*, D247–D251.

Petrey, D., and Honig, B. (2005). Protein structure prediction: inroads to biology. Mol. Cell *20*, 811–819.

Pettitt, C.S., McGuffin, L.J., and Jones, D.T. (2005). Improving sequence-based fold recognition by using 3D model quality assessment. Bioinformatics *21*, 3509–3515.

Pieper, U., Eswar, N., Davis, F.P., Braberg, H., Madhusudhan, M.S., Rossi, A., Marti-Renom, M., Karchin, R., Webb, B.M., Eramian, D., et al. (2006). MODBASE: a database of annotated comparative protein structure models and associated resources. Nucleic Acids Res. *34*, D291–D295.

Rost, B. (2003). Prediction in 1D: secondary structure, membrane helices, and accessibility. Methods Biochem. Anal. *44*, 559–587.

Rost, B., Yachdav, G., and Liu, J. (2004). The PredictProtein server. Nucleic Acids Res. *32*, W321–W326.

Rychlewski, L., and Fischer, D. (2005). LiveBench-8: the large-scale, continuous assessment of automated protein structure prediction. Protein Sci. *14*, 240–245.

Sanchez, R., Pieper, U., Mirkovic, N., de Bakker, P.I., Wittenstein, E., and Sali, A. (2000). MODBASE, a database of annotated comparative protein structure models. Nucleic Acids Res. *28*, 250–253.

Schueler-Furman, O., Wang, C., Bradley, P., Misura, K., and Baker, D. (2005). Progress in modeling of protein structures and interactions. Science *310*, 638–642.

Shoichet, B.K. (2004). Virtual screening of chemical libraries. Nature *432*, 862–865.

Sippl, M.J. (1993). Recognition of errors in three-dimensional structures of proteins. Proteins *17*, 355–362.

Sippl, M.J. (1995). Knowledge-based potentials for proteins. Curr. Opin. Struct. Biol. *5*, 229–235.

Stevens, R.C., Yokoyama, S., and Wilson, I.A. (2001). Global efforts in structural genomics. Science *294*, 89–92.

Subramaniam, S., and Henderson, R. (2000). Molecular mechanism of vectorial proton translocation by bacteriorhodopsin. Nature *406*, 653–657.

Tilley, S.J., Orlova, E.V., Gilbert, R.J., Andrew, P.W., and Saibil, H.R. (2005). Structural basis of pore formation by the bacterial toxin pneumolysin. Cell *121*, 247–256.

Topf, M., and Sali, A. (2005). Combining electron microscopy and comparative protein structure modeling. Curr. Opin. Struct. Biol. *15*, 578–585.

Topf, M., Baker, M.L., Marti-Renom, M.A., Chiu, W., and Sali, A. (2006). Refinement of protein structures by iterative comparative modeling and CryoEM density fitting. J. Mol. Biol. *357*, 1655–1668.

Wallner, B., and Elofsson, A. (2003). Can correct protein models be identified? Protein Sci. *12*, 1073–1086.

Wallner, B., Fang, H., and Elofsson, A. (2003). Automatic consensus-based fold recognition using Pcons, ProQ, and Pmodeller. Proteins *53* (Suppl 6), 534–541.

Westbrook, J., Henrick, K., Ulrich, E.L., and Berman, H.M. (2005). The Protein Data Bank exchange data dictionary. In International Tables for Crystallography, S.R. Hall and B. McMahon, eds. (Dordrecht, The Netherlands: Springer), pp. 195–198.

Zhou, H., and Zhou, Y. (2005). SPARKS 2 and SP(3) servers in CASP 6. Proteins *61* (Suppl.), 152–156.

Zhou, Z.H., Baker, M.L., Jiang, W., Dougherty, M., Jakana, J., Dong, G., Lu, G., and Chiu, W. (2001). Electron cryomicroscopy and bioinformatics suggest protein fold models for rice dwarf virus. Nat. Struct. Biol. *8*, 868–873.