Structure, Vol. 11, 1043–1047, September, 2003, ©2003 Elsevier Science Ltd. All rights reserved. DOI 10.1016/S0969-2126(03)00163-1

NIH Workshop on Structural Proteomics of Biological Complexes

Meeting Review

Andrej Sali*
Departments of Biopharmaceutical Sciences
and Pharmaceutical Chemistry and
California Institute for Quantitative
Biomedical Research
University of California, San Francisco
San Francisco, California 94143

Summary

Recently, some 50 biologists and officials from government funding agencies met at the NIH campus in Bethesda, MD to explore the interdisciplinary science and organization of the emerging field of structural proteomics. Structural proteomics aims to discover most macromolecular complexes and characterize their three-dimensional structures and functional mechanisms in space and time. The goal seems daunting, but the consensus was that the prize would be commensurate with the effort invested, given the importance of molecular machines and functional networks in biology and medicine. Identification of assemblies and transient complexes combined with their structural and functional characterization will allow us to understand, control, design, and change the functioning of larger biological systems as well as to contribute to drug target discovery, lead discovery, and lead optimization for treatment of human disease.

Introduction

This meeting review is organized according to the presumed steps in the structural proteomics pipeline: (i) comprehensive identification of macromolecular complexes; (ii) selection of complexes for structure characterization; (iii) preparation of samples; (iv) validation of biological role, with biochemical and structural characterization; and (v) archival and analysis of the structures.

Structural proteomics shares similarities with structural genomics, which aims to determine structures of carefully chosen protein domain targets, such that most of the remaining domains can be modeled by similarity. However, there are also important differences. For example, (i) while the subject of structural genomics are the finite set of proteins encoded by the genomes, it is not yet clear what the complete set of complexes is; (ii) it is also not clear how to select target complexes for structure determination since there is no obvious equivalent of protein domain families in the case of complexes; (iii) structure determination of complexes is likely to involve multiple experimental and computational methods at different levels of resolution, as opposed to primarily X-ray crystallography and nuclear magnetic resonance spectroscopy; and (iv) biochemical and functional characterization of complexes is likely to play a more important role in structural proteomics than structural genomics.

Identification and Characterization of Macromolecular Complexes

David Drubin (Berkeley) reviewed the motivation of cell biologists for describing the structures and mechanisms of macromolecular complexes. He suggested that proteins and their associated complexes be categorized on the basis of their involvement in core biological processes, such as the maintenance of chromosome structure (nucleosomes), replication (DNA polymerase), transcription (RNA polymerase), nuclear transport (nuclear pore complex), protein synthesis (ribosome), protein degradation (proteosome), metabolism (aspartate transcarbamylase), signal transduction, chromosome movement, and segregation (kinetochore). Merits of the genome-wide versus a more targeted approach were discussed, balancing efficiency, bias, and quality. He suggested that interactions observed by proteomics should be validated by independent means, such as microscopy with green fluorescence protein. He also highlighted the power of emerging chemical genomics approaches, which utilize tailor made pairs of small molecule inhibitors and signaling molecules, to parse the contributions of individual signaling pathways to complex biological processes.

Jack Greenblatt (University of Toronto) described a focused tandem affinity purification survey of *E. coli* protein complexes containing ∼190 proteins that are conserved in bacteria and essential in at least one bacterial species. Such proteins are potential targets for new kinds of broad spectrum antibiotics. In addition, yeast protein complexes involved in transcription by RNA polymerase II were identified by tagging and purification followed by clustering of the bait/prey data. Independently, yeast proteins were also clustered by a synthetic genetic array analysis based on double deletion mutant data. Genetic clustering sometimes correlated with the clustering of proteins in space.

Peter Sorger (MIT) presented his studies of yeast kinetochores, which are multiprotein complexes that assemble on centromeres, and their role in the movement of chromosomes along the microtubules. Kinetochores consist of 60–80 subunits and have a molecular mass of ∼8 MDa. Determination of the protein composition of kinetochores was greatly facilitated by liquid chromatography and tandem mass spectrometry as well as by localizing proteins in the cell using light microscopy. Hydrodynamic analyses by size-exclusion chromatography and reconstitution experiments with subcomplexes were also helpful. He pointed out that the studies of function and structure must be linked, and while methods for identifying subunits are excellent, methods and reagents for elucidating biochemical and biological functions remain unsatisfactory. An increase in the resolution of light microscopy from 220 nm by a factor of 6

*Correspondence: sali@salilab.org

was identified as an important and major technological challenge. This aim will likely be achieved not by changes in hardware but by the development of appropriate image processing software that uses knowledge of both the biology and the physics of the instrument to improve data extraction and interpretation.

Rebecca Heald (Berkeley) described a systematic identification and functional characterization of cell division proteins involved in cytokinesis. One hundred fifty-nine proteins were identified in isolated mammalian midbodies, which are the remnant of the cytokinetic ring responsible for dividing the cell in two, by liquid chromatography and tandem mass spectrometry. One hundred seven of these proteins were previously uncharacterized, and their coarse functions were determined quickly by RNA-mediated interference in *C. elegans* embryos. Most of these proteins are conserved in mammals and nematodes, pointing to a common and ancient mechanism of cell division.

John Yates (The Scripps Research Institute) demonstrated the key role of mass spectrometry in the characterization of complexes, including the identification of the protein components, the biological processes in which they are involved, cellular localization, subunit stoichiometry, and posttranslational modifications that are frequently involved in regulation. In particular, emerging sample preparation techniques and new mass spectrometry strategies are directly allowing for the identification of components and their stoichiometries, without the use of the electrophoretic techniques. Although mass spectrometry is a serial technique, simply using more instruments will enable a very high-throughput analysis that is needed for proteomics studies.

Charlie Boone (University of Toronto) is mapping genetic interactions between pairs of yeast genes using an automated system for genetic analysis termed synthetic genetic array (SGA) analysis. Two genes interact genetically if the corresponding double mutant is lethal or slow growing and the single mutants are not. SGA analysis enables a query mutation to be crossed to all 5,000 viable yeast deletion mutants. On the average, a yeast gene has 35 genetic partners, many of which encode proteins within the same complex or pathway. The approach was illustrated by application to the *BNI1* gene, which controls cytoskeletal dynamics and cell polarity. For structural proteomics in particular, it is important to be able to decipher physical relations among gene products. This task may be facilitated by identifying genes with similar patterns of genetic interactions.

Bill Jacobs (Albert Einstein College of Medicine) described an efficient gene knock out technology for *Mycobacterium tuberculosis* called specialized transduction. The system uses a hybrid molecule that replicates as a plasmid in *Escherichia coli* and as a phage in mycobacteria. In addition, the vector has a number of temperature sensitive mutations in the phage replicating genes that allow the mycobacteriophage to replicate at 30°C but not at 37°C. This conditional replication provides a simple means to use the phage to deliver allelic exchange substrates efficiently at nonpermissive temperatures and easily achieve allelic exchanges with typical yields of 95%–100% for nonessential genes. The technology has been used to generate hundreds of allelic exchanges in many different *M. tuberculosis* strains.

Mark Gerstein (Yale University) discussed annotation of protein function. Function of a protein can be predicted by sequence and structure comparisons with already characterized proteins, as well as by clustering mRNA expression profiles. Enumeration of all the interactions of a protein, including small ligands, proteins, and other molecules, is a good way to define its function. Unfortunately, there are relatively small overlaps and large differences between interaction sets derived by a variety of experimental and computational methods. Therefore, it is imperative to develop integrated statistical approaches to identification of protein interactions in order to minimize false positives and negatives. In addition, he suggested that an in-depth characterization of some 100 large complexes would provide a valuable resource for generalization and extension to many other complexes by computational means.

Selection of Complexes for Structural Characterization

No speaker focused on target selection for structural proteomics, presumably because it still requires considerable research before reasonable suggestions can be made. However, a number of schemes were proposed for further consideration. For example, it was suggested that a target list could correspond to the following: (i) the proteins organized into 200–300 core biological processes (D. Drubin); (ii) 100 complex structures of some importance (M. Gerstein); (iii) select subcomplexes, such as those determined by large-scale affinity purification experiments (S. Almo); (iv) the same complex from a number of species to facilitate discovery of evolutionary relationships and functional mechanisms (S. Almo); or (v) a comprehensive set of representative binary domain interfaces that may allow us to model higher order complexes with useful accuracy and efficiency (A. Sali).

Preparation of Samples, Biochemical and Structural Characterization

Obtaining protein samples of sufficient quantity and purity for structural characterization will likely be a bottleneck in the structural proteomics process. However, the existing structural genomics infrastructure, which includes automation and parallelization for cloning, protein expression, purification, and crystallization, will also facilitate structural proteomics.

Paul Matsudaira (MIT) described a comprehensive experimental and computational study of podosomes, which are actin-based adhesion structures. Bioimaging, including confocal microscopy, was an essential tool to validate the biological authenticity of complexes. Bioimaging was presented from the perspective of an informatics science, posing new challenges in data storage, retrieval, and visualization. A point in case is the terabytes of computer memory that are required for storage and handling of dynamic 3D images even at the relatively modest resolution of 512 × 512 pixels per slice. Such images have to be processed to localize the components of complexes in space and time. He suggested that at least some of the needed improvements in bioimaging may be achieved through closer involvement of biologists, engineers, and computer scientists.

Andrej Sali (UCSF) is developing a framework for computing 3D models of a given protein assembly that are consistent with all available information about its composition and structure. In contrast to structure determination of individual proteins, structural characterization of macromolecular assemblies usually requires diverse sources of information, which may vary greatly in terms of their accuracy and resolution, and include data from both experimental and computational methods. The proposal was illustrated by two examples: (i) the use of both high-resolution single particle electron cryomicroscopy (cryo-EM) and comparative protein structure modeling in the structure determination of eukaryotic ribosomes; (ii) the use of low-resolution single particle cryo-EM, immuno-EM, affinity chromatography, and theoretical considerations to model the configuration of proteins in the yeast nuclear pore complex.

Wah Chiu (Baylor College of Medicine) described the latest developments in the application of cryo-EM to the medium resolution structure determination of large complexes. For example, it is now possible to visualize secondary structure segments of large viruses (e.g., rice dwarf virus) and molecular machines (e.g., GroEL) at subnanometer resolutions of 6–9 Å. The areas of cryo-EM that require further improvement include cryo-specimen preparation, instrument automation, and image processing. The computer software that controls or implements many of the operations in structure determination by cryo-EM has to be user friendly. As for light microscopy, there is a need for effective collaborations between biochemists, physicists, engineers, and computational scientists to enable the imaging technology for studying complexes at near atomic resolution.

Ken Downing (Lawrence Berkeley National Laboratory) focused on electron crystallography of monolayer crystals of the membrane and soluble proteins. Structures of bacteriorhodopsin and tubulin were determined at ~3 and ~3.7 Å resolution, respectively. Moreover, it was possible to study tubulin in different states, such as the complex with the anticancer drug taxol. Hierarchical models were shown, including a bird's eye view of the microtubule with an ellipsoid representation of tubulin and a closeup of the interfaces between different molecules of tubulin that relied on the atomic crystallographic structure of tubulin. Eventually, advances in instrumentation and methodology are expected to narrow the resolution gap between electron crystallography and X-ray crystallography.

Ron Milligan (The Scripps Research Institute) also described a hybrid approach that relies on cryo-EM of the entire complex and X-ray crystallography of its components. The approach was illustrated by its application to kinesin motors. It was possible to simulate the movement of the kinesin motor by supplementing the crystallographic structures with cryo-EM maps of the motor-microtubule complexes trapped at various stages in its enzymatic cycle. Therefore, the utility of the hybrid method for revealing functional mechanisms of complexes was demonstrated. He also highlighted a need for automation of cryo-EM to increase its throughput, and a recent success at The Scripps Institute: the test case structure of the tobacco mosaic virus was determined to ~10 Å resolution in a single day with little manual intervention.

Steven Almo (Albert Einstein College of Medicine) described the large-scale determination of individual protein structures as envisioned by structural genomics, including the Protein Structure Initiative of NIH (http://www.rcsb.org/pdb/strucgen.html). The New York Structural Genomix Research Consortium determined some 70 structures so far, resulting in models for ~3,000 proteins of previously unknown structure. Examples of functional annotations based on structures were given. For example, the active site of mevalonate pyrophosphate decarboxylase was localized by mapping the conservation of residues among the family members on its fold. He suggested that the existing structural genomics infrastructure can be extended to provide a pipeline for structural proteomics, which will allow for the systematic structural analysis of subcomplexes and complexes by cryo-EM and in certain cases by X-ray diffraction. As with structural genomics, the preparation of protein samples was identified as the major bottleneck. This bottleneck is likely to be even more acute for structural proteomics, since in the most favorable cases only 10–100 μg of a complex per 10 liters of yeast culture are anticipated.

David Stuart (Oxford University) described a European structural proteomics effort (SPINE) funded by the European Commission Framework V to focus on biomedical targets with a major emphasis on proteins from human pathogens. SPINE includes partners from some 20 European laboratories. Major new resources dedicated to structural genomics and proteomics in the UK include a synchrotron (Diamond) with P3-compatible beamlines to be ready at the end of 2006 and a protein production facility (the OPPF) for human proteins and proteins of human pathogens, oriented toward disease and eukaryotic expression, funded in Oxford by the Medical Research Council. The ultimate goal will be to utilize these resources to address protein complexes. Established techniques have already revealed whole virus structures. As an example, the crystallographic structure of an icosahedral virus with a lipid bilayer, the bacteriophage PRD1, was presented. The crystals have 66 MDa in the asymmetric unit and the structure was phased with a low-resolution cryo-EM map. Layers of ordered genomic DNA packed against the lipid bilayer in the bacteriophage were seen for the first time. This result emphasizes the potential power of the interplay between cryo-EM and protein crystallography. To facilitate the extension of such studies to human pathogens, a facility (the FEIP) housing state-of-the-art cryo-EM in P3 is under construction in Oxford, funded by the U.K. government and the Wellcome Trust.

Archival and Analysis of the Structures

Development of the databases and software will be necessary to store and make available information collected during the process of structural proteomics. In addition to facilitating structural proteomics itself, significant informatics resources will also be needed to cross link the new data with other biological and medical information. These tools will allow researchers to address efficiently a host of questions.

Helen Berman (Rutgers University) highlighted the computational biology challenges in structural proteomics, emphasizing the data management and informatics aspects. Such challenges include complex selection, data management, mining, and integration, structure determination, structure analysis and comparison, structure archiving, visualization, database interoperability, data exchange, and functional analysis. In particular, the merits and disadvantages of the top down and bottom up approaches to data management were discussed. There is a need to define a dictionary of terms and relationships between them to be used in constructing the database schema as soon as possible.

Centers

Lee Makowski (Argonne National Laboratory) introduced "Genomes to Life" (GTL) (http://doegenomestolife.org), an 18-month-old initiative by the Department of Energy (DOE) that focuses on systems biology. The initiative recognizes the need for large facilities to identify protein machines, characterize gene regulatory networks, explore microbial function in ecosystems, and develop computational tools to model and predict responses of cells to environment. The initiative will provide support for four user facilities with hundreds of scientists and budgets of over $150 million per facility. They will specialize in protein production, whole proteome analysis, characterization and imaging of molecular machines, and analysis and modeling of cellular systems.

Michelle Buchanan (Oak Ridge National Laboratory) described Center for Molecular and Cellular Systems, a pilot project in the DOE GTL initiative for identification and characterization of protein complexes in microbial cells. Trifunctional crosslinking reagents with biotin are used to isolate complexes, followed by liquid chromatography and tandem mass spectrometry detection of the components. The Center supports construction and optimization of the structural proteomics pipeline through work on isolation of the complexes, imaging and other analytical tools, and computational tools to assist in data storage, data interpretation, and protein structure prediction.

Alex Kurosky (University of Texas Medical Branch at Galveston) described one of ten Proteomics Centers in the U.S. established by the National Heart, Lung, and Blood Institute (NHLBI) to enhance and develop innovative proteomics technologies. The total budget of this initiative is $157 million over seven years. Identification of complexes involving the RSV virus proteins is in progress, using liquid chromatography and tandem mass spectrometry of immunoprecipitates, confocal microscopy for colocalization studies, and thioaptamer chip methods.

Trisha Davis (University of Washington) outlined the Yeast Resource Center, which is a Biomedical Technology Resource Center supported by NIH's National Center for Research Resources (NCRR). The Yeast Resource Center focuses on the technologies that extract information about macromolecular interactions from the yeast genome sequence, including mass spectrometry, two-hybrid assays, fluorescence microscopy, and protein structure prediction. These technologies are made available to others through $\sim$80 collaborative projects. All four technologies were applied to $\sim$100 uncharacterized essential yeast genes. The importance of advanced protein homology searches and cellular localization in the context of the other data was emphasized.

Scientific and Technical Scope

The final session of the workshop aimed to collect as many useful attributes of structural proteomics as possible. The object of study should include regulation of complexes as well as their dynamics and structures at the highest resolution practicable. It should be genomic in scale to afford massive data collection as well as hypothesis driven to result in most biomedically relevant results. Complexes should be understood in terms of both physics and evolution, which involves 3D modeling and docking based on physical energy functions and comparative analysis of complexes from an evolutionary point of view.

Structural proteomics needs to be comprehensive and complete, which will only be possible by integrating a variety of experimental and computational technologies. Interaction maps, yeast two-hybrid, genetics, production, and biochemical characterization of the complexes are a prerequisite for structural studies. Structure determination will generally be achieved by hybrid multiresolution methods, including X-ray crystallography, nuclear magnetic resonance spectroscopy, cryo-EM, computation, and a number of lower-resolution methods, such as chemical crosslinking and foot printing. All steps of the structural proteomics process must be validated. The large technological effort should encompass technology development, automation, robotics, laboratory information management systems, and informatics.

Resources and Funding

The resources for structural proteomics include databases, instrumentation for multiple technologies, computers, personnel, training activities, collaborative projects, and research resources. Roles for both investigator initiated research and large centers that are capable of developing, testing, and integrating various technologies were recognized. The mechanisms of funding could include pilot grants (e.g., for optimization of protocols), planning grants, glue grants, program project grants, center grants, and training grants. Both DOE's GTL and NIH's Protein Structure Initiative were recognized as existing models.

Rationale and Impact

It was generally agreed that structural proteomics is timely, because of the confluence of the available data as well as experimental and computational methods that can be brought to bear on the problem. There is also a special need for this project because it is highly risky, interdisciplinary, and complex both scientifically and technologically.

Structural proteomics will contribute to structure-based functional annotation of molecular machines and functional networks and thus have an enormous impact on biology and medicine. Structural biology is a great unifying discipline of biology. Thus, structural proteomics may be the way to bridge the gaps between genome sequencing, functional genomics, proteomics, and systems biology. In addition, our understanding of the machinery of life and its evolution will surely result in profusion of new drug targets and better drugs. For example, it is likely that using complexes as drug targets will provide a significant increase in drug specificity relative to protein targets and thus reduce drug side effects.

We are now poised to integrate structural information gathered at multiple levels of the biological hierarchy—from atoms to cells—into a common framework. The goal is a comprehensive description of the multitude of interactions between molecular entities, which in turn is a prerequisite for the discovery of general structural principles that underlie all cellular processes.

Acknowledgments

We thank the speakers, and in particular Wah Chiu and Steven Almo, for commenting on the review. The PowerPoint presentations of the participants are accessible at http://ncmi.bcm.tmc.edu/ncmi/ structural proteomics workshop. The meeting was organized through the National Center for Macromolecular Imaging by Wah Chiu of Baylor College of Medicine, Houston, TX, and Steven Almo of Albert Einstein College of Medicine, Bronx, New York. The workshop was generously supported by the National Center for Research Resources.