100,000 protein structures for the biologist

Andrej Šali

Structural genomics promises to deliver experimentally determined three-dimensional structures for many thousands of protein domains. These domains will be carefully selected, so that the methods of fold assignment and comparative protein structure modeling will result in useful models for most other protein sequences. The impact on biology will be dramatic.

Recently, some 200 structural biologists from both academia and industry gathered in Avalon, New Jersey to explore the science and organization of the new field of structural genomics[1]. The aim of the structural genomics project is to deliver structural information about most proteins. While it is not feasible to determine the structure of every protein by experiment, useful models can be obtained by fold assignment and comparative modeling for those protein sequences that are related to at least one known protein structure. Structure determination of 10,000 properly chosen proteins should result in useful three-dimensional models for hundreds of thousands of other protein sequences. In other words, structural genomics will put each protein within a comparative modeling distance of a known protein structure (Eaton Lattman, Johns Hopkins University).

Knowing the structures of biological macromolecules is almost always useful for predicting, interpreting, modifying and designing their functions. The high-throughput, coordinated structure determination efforts should be more efficient for obtaining 10,000 structures than the current distributed and largely uncoordinated efforts in individual structural biology laboratories. Structural genomics will build on the human genome project and will be integrated with the functional genomics project. Here, I review the recent meeting in Avalon. Due to space restrictions, the review omits many of the presentations. However, a complete list of speakers may be found on the meeting web site[1].

The structural genomics project will deliver improved methods and a process for high-throughput protein structure determination. The process involves: (i) selection of the target proteins or domains, (ii) cloning, expression, and purification of the targets, (iii) crystallization and structure determination by X-ray crystallography or by nuclear magnetic resonance (NMR) spectroscopy, and (iv) archiving and annotation of the new structures. Several pilot projects in structural genomics that implement all of these steps are already underway (Table 1). It is feasible to determine at least 10,000 protein structures within the next five years. No new discoveries or methods are needed for this task. The average cost per structure is expected to decrease significantly from ~$200,000 per structure to perhaps less than $20,000 per structure. These savings will be achieved by the economies of scale, new or improved methods, and their parallelization in most of the steps in the process.

Target selection

The targets for structure determination generally will be individual domains rather than multi-domain proteins (Chris Sander, Millenium Information). Such targets will be easier to determine by X-ray crystallography or NMR spectroscopy than the more flexible multi-domain proteins, although it would be beneficial to determine whole proteins whenever possible. Target selection will begin by pairwise comparison of all protein sequences, followed by clustering into groups of domains. The targets for the structural genomics project will be the representatives of the groups without structurally defined members. The groups of domains should contain members that share at least ~30% sequence identity (30-seq families), rather than those that correspond to superfamiles or fold families.

This relatively high granularity is needed because of the limitations in the current methods for comparative modeling and for inferring function from structure. Errors in the comparative models become relatively small as the sequence identity increases above 30% (Andrej Šali, Rockefeller University) and function is generally conserved within families (John Moult, Center for Advanced Research in Biotechnology, Rockville, Maryland); Christine Orengo, University College, London). Given the clustering into the 30-seq families, the total number of targets for experimental structure determination has been estimated to be on the order of 10,000. As an alternative, it has also been suggested that 100,000 protein structures need to be determined by experiment (David Eisenberg, UCLA); this would allow calculation of models with higher accuracy than is possible with only 10,000 known structures. After a list of the 30-seq families is obtained, the representatives will be picked and prioritized. A variety of different criteria likely will be used for this task; for example, size of the family, biological knowledge about the family, distribution of the members among various organisms, the pharmaceutical relevance, the likelihood of a successful structure determination, and so forth .

A preliminary analysis suggests that among recent protein structures determined in the hope of discovering a new fold, more than two-thirds actually have exhibited an already known fold (Steven Brenner, Stanford). While this unveils weaknesses in the fold assignment methods, it is not a waste of resources. In practical terms, the structure of a protein is worth determining if we do not know it, for whatever reason. In fact, it may even be an advantage when the newly determined protein is related to other known structures because such structure is likely to be more informative about its function than in the case of structural 'orphans'.

Given a fixed number of experimentally determined protein structures, the quality and the number of models for the rest of the proteins will be influenced strongly by the methods for fold assignment and comparative protein structure modeling. The impact of expected improvements in these methods may be equivalent to several years of the experimental protein structure output (Andrej Šali). A number of different approaches to fold assignment have been described, applied, and evaluated (Ron Levy, Rutgers University; George Rose, John Hopkins University; Barry Honig, Columbia University; Manfred Sippl, CAME, Salzburg; David Eisenberg; Eugene Koonin, NCBI; Steven Brenner). The methods that rely on multiple

sequence information, such as PSI-BLAST, appear surprisingly sensitive in detecting remote relationships (Liisa Holm, EMBL-EBI, Cambridge; Eugene Koonin; Steven Brenner).

Cloning, expression and purification

Obtaining protein samples for crystallization trials and NMR measurements likely will be the bottleneck in the structural genomics process. Although the current methods are sufficiently powerful to begin the project, advances are expected in parallelization and automation of the existing methods for cloning, expression, and purification of proteins. In addition, entirely new technologies may arise, such as an *in vitro* system that already has allowed researchers to produce proteins up to concentrations of 6 mg ml^{-1} (Shigeyuki Yokoyama, Tokyo University).

Crystallization and X-ray crystallography

Stephen Burley (Rockefeller University) described how to test protein samples for their propensity to form useful crystals. Crystals have a higher chance of being formed when the protein species in solution is conformationally homogeneous. This can be evaluated by dynamic light scattering, as well as by limited proteolysis combined with mass spectroscopy. For example, candidates that are monodisperse under standard aqueous conditions have a probability of at least 75% to yield crystals suitable for X-ray studies, in comparison to 10% for polydisperse samples.

Wayne Hendrickson (Columbia University) described several pivotal advances that have matured only recently and form the basis of crystallographic analysis at a markedly accelerated level. These include: undulator sources at synchrotrons, CCD detectors, cryo-crystallography, MAD phasing analysis, and selenomethionyl proteins[2]. It is now possible to sustain a data collection throughput of four structures per day on a single undulator beam line. Assuming 180 operational days per year and that only one out of three data sets results in a useful final structure, this gives a conservative estimate of approximately 360 structures per year per single undulator beam line. This means that only several dedicated undulated beam lines are needed over the course of five years to achieve the goal of determining 10,000 new protein structures.

Tom Terwilliger (LANL) has developed the SOLVE computer program for obtaining electron density maps of proteins directly from multiwavelength or multiple isomorphous replacement X-ray diffraction data. An optimization of a scoring function replaces a complex rule-based process followed manually by crystallographers. This system has already allowed crystallographers to obtain images of proteins within hours of the end of X-ray data collection. The procedure has been successfully applied to MAD data with up to 26 selenium sites per derivative.

Sung-Hou Kim (University of California, Berkeley) described results from the first structural genomics pilot project, focusing on the *Methanococcus janaschii* genome (Table 1). He presented the structures of two proteins, an ATP dependent molecular switch with a new ATP binding motif, and a nucleotide pyrophosphatase, with a new fold. For each protein, the type of a ligand bound to it, and thus its biochemical function, were predicted from the protein structure alone. The predictions were subsequently confirmed by direct experiment. These results support the idea that protein function can be anticipated from structural information alone and thus that high-throughput structure determination will make a valuable contribution to biology.

Table 1 List of pilot projects in structural genomics[1]

Pilot project organizers	Organism
S. Burley, A. Šali, J. Sussman	*S. cerevisiae*
A. Edwards	*M. thermoautotrophicum*
D. Eisenberg, T. Terwilliger	*P. aerophilum*
S.-H. Kim	*M. janaschii*
G. Montelione, S. Anderson	Metazoa
J. Moult	*H. influenzae*
S. Yokoyama	*T. Thermophilus, HB8*

[1]These projects were described at the meeting. Each project generally involves several independent laboratories. Only speakers at the meeting are listed in the first column. Their addresses can be found at http://lion.cabm.rutgers.edu/bioinformatics_meeting/.

Nuclear magnetic resonance spectroscopy

Kurt Wüthrich (ETH, Zürich), Gaetano Montelione and Gerhardt Wagner (Harvard Medical School) described recent and prospective advances in NMR spectroscopy that will allow structure determination of larger proteins as well as increase the accuracy and speed of the analysis. These advances include partial deuteration of protein samples, new magnets (up to 1 GHz), superconducting probes, the TROSY detection method, triple resonance methods for resonance assignment, spatial information from residual dipolar coupling, and software for resonance assignments and structure determination. For example, Montelione and colleagues have developed the program AUTOASSIGN to perform automated backbone assignments in a few minutes of CPU time for proteins smaller than 200 residues. This tool, integrated with the program AUTO_STRUCTURE for automatic analysis of NOESY data and structure generation, decreased the time needed for structure refinement of a Z-domain from several months to several hours. If the NMR data collection for a 120 residue protein takes 4–5 weeks on a 600 MHz spectrometer and if automated data processing takes 1 day, 120 structures can be determined in 1 year by 12 machines (Montelione), approximately at a cost of using one synchrotron beam line. By using higher field 800 MHz to 1 GHz magnets and other sensitivity-enhancement methods, this throughput could be significantly increased.

Anticipating the usefulness of NMR for structural genomics, Shigeyuki Yokoyama of Tokyo University announced the Japanese Genomic Sciences Center that will include a large NMR facility. The center will also support computational analysis of the human genome, sequencing and mapping of the mouse genome, and collaborations with high-throughput protein crystallography efforts elsewhere in Japan. The center will have 20 NMR instruments by 2001 and will be open to international collaboration. A 1 GHz magnet is being constructed in collaboration with the Tsukuba laboratory, the engineers of the magnets used for the bullet train.

NMR spectroscopy has become increasingly more useful and efficient. However, given the synchrotrons, X-ray crystallography is likely to retain an edge in throughput, accuracy, maximal protein size, and possibly cost per protein structure determination. Nevertheless, NMR spectroscopy will be indispensable for maximizing the benefits of structural genomics because of its unique ability to obtain information about the internal dynamics of a protein on multiple time scales as well as about its ligand binding properties. For example, Kurt Wüthrich described the importance of flexible tails for the function of several proteins. Another example is the routine use of chemical shift perturbation to enumerate

binding sites on a protein for a thousand ligands per day from a library of 15,000 compounds, with up to 1 mM sensitivity in the dissociation constant (Steven Fesik, Abbott Laboratories). Also potentially feasible is NMR structure determination of membrane protein structures, which are exceptionally difficult to crystallize.

Databases

Several different databases will be useful in facilitating the structural genomics project (Table 2). These databases will be used for target selection, coordination of the experimental efforts, archiving of the new structures, dissemination of the results, and for performing new research with the old and new data. Helen Berman of Rutgers University described a new macromolecular structure database that will replace the Brookhaven Protein Data Bank[3]. Liisa Holm, Christine Orengo, and Steven Brenner described CATH, FSSP, and SCOP databases respectively. These databases classify protein structures in a hierarchical manner, generally consisting of families (proteins that share at least ~30% sequence identity), superfamilies (proteins that are probably related by divergent evolution but can have very little sequence similarity), and fold families (proteins that share similar folds). Orengo pointed out a great need for an exhaustive, systematic and computer friendly database of protein functional annotations. Such a database is necessary for the development and use of the methods that correlate structure to function. Andrej Šali described ModBase, a large database of comparative protein structure models. This database is calculated with program Modeller and currently contains protein models for five genomes. It will continue to grow with the availability of new protein structures and sequences, as well as the improvements in the modeling software. Steven Brenner introduced PRESAGE, a database that stands to become the central nervous system for coordinating the efforts in structural genomics. PRESAGE includes entries for proteins; each entry contains information about the protein's homologs, current status of structure determination, links to three-dimensional models, and so forth. The entries in PRESAGE will be contributed by the community at large. Biologists will use it as a portal for accessing the most up-to-date information about proteins of interest. For example, PRESAGE would have eliminated the duplicated structure determination of elongation factor 5A by two structural genomics pilot projects (Sung-Hou Kim; Tom Terwilliger).

Using the structures

Development of the databases and software will be necessary to integrate raw structural information and structure-derived results with other information, such as genetic and biochemical analysis, gene expression patterns, mutational analysis, binding data from protein chips, interaction data from two-hybrid systems, metabolic pathway information, and so forth (Chris Sander). These tools will then allow researchers to address efficiently a host of new questions. Some such studies, illustrating but the tip of the iceberg, have already been performed and were described at the meeting. For example, questions about the distribution of different structures among the genomes, and the implications for models of evolution, have been asked by Mark Gerstein (Yale University) and Eugene Koonin. In addition, David Eisenberg focused on understanding the extreme thermostability of proteins in thermophilic organisms in terms of the amino acid composition and structural features of the proteins. The study was based on a comparison of proteins that have homologs in both thermophilic and mesophilic organisms and which have homologs of known structure. It appears that thermostability of proteins originates in part from salt bridges and from deletions of loop regions.

Jeff Skolnick (Scripps Research Institute) described a method for predicting protein function based on structure prediction and three-dimensional motif identification. This method could push the detection of protein function much further into the twilight zone of sequence identity. First, the tertiary structures of the sequences of interest are predicted either *ab initio* or by threading. Then, the resulting models are screened using a library of three-dimensional descriptors of protein active sites, termed 'fuzzy functional forms'. If the geometry and residue types in the predicted structures match a three-dimensional motif, then the protein is

Table 2 Structural genomics web sites

Pilot Projects	
DOE	http://proi3.lanl.gov/structural_genomics/
New York	http://genome5.bio.bnl.gov/Proteome/
CARB/TIGR	http://s2f.carb.nist.gov
Databases	
NCBI	http://www.ncbi.nlm.nih.gov/
PDB	http://www.pdb.bnl.gov/
MSD	http://www.rcsb.org/
DALI	http://www2.ebi.ac.uk/dali/
CATH	http://www.biochem.ucl.ac.uk/bsm/cath/
SCOP	http://scop.mrc-lmb.cam.ac.uk/scop/
PRESAGE	http://presage.stanford.edu
ModBase	http://guitar.rockefeller.edu/modbase/
GeneCensus	http://bioinfo.mbb.yale.edu/genome
Fold assignment	
PhD	http://www.embl-heidelberg.de/predictprotein/predictprotein.html
THREADER	http://globin.bio.warwick.ac.uk/~jones/threader.html
123D	http://www-lmmb.ncifcrf.gov/~nicka/123D.html
UCLA-DOE	http://www.doe-mbi.ucla.edu/people/frsvr/frsvr.html
PROFIT	http://lore.came.sbg.ac.at/
Comparative modeling	
COMPOSER	http://www-cryst.bioc.cam.ac.uk/
CONGEN	http://www.cabm.rutgers.edu/~bruc
DRAGON	http://www.nimr.mrc.ac.uk/~mathbio/a-aszodi/dragon.html
Modeller	http://guitar.rockefeller.edu/modeller/modeller.html
PrISM	http://honiglab.cpmc.columbia.edu/
SWISS-MODEL	http://www.expasy.ch/swissmod/SWISS-MODEL.html
WHAT IF	http://www.sander.embl-heidelberg.de/vriend/
Miscellaneous	
Target selection	http://structuralgenomics.org/
Funding NIH	http://www.nih.gov/nigms/
Funding NSF	http://www.nsf.gov/home/grants.htm
Pedro's list	http://www.public.iastate.edu/~pedro/research_tools.html
Roberto's list	http://guitar.rockefeller.edu/~roberto/tools/tools.html

predicted to have the specified molecular function. The idea and its power relative to local sequence pattern searches were illustrated by searching for proteins with a disulfide oxidoreductase active site in three bacterial genomes.

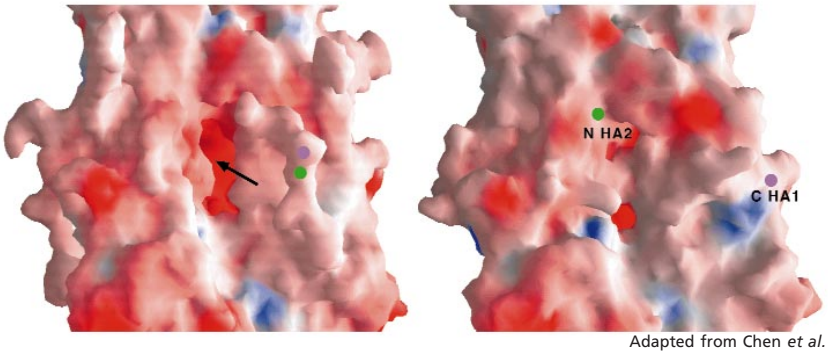
Impact of structural genomics

Suggestions about how best to proceed with the structural genomics project were discussed. Ideas ranging from continuing 'research as usual' to establishing large centers for structural genomics were explored. The pilot projects will help shape the full scale effort. It is likely that relatively large centers will provide the infrastructure for X-ray crystallography and NMR spectroscopy, allow the economy of scale, foster collaboration, and also ensure that the technical and repetitive jobs are not performed by students and postdocs in the academic laboratories. Centers will make it possible for academic and pharmaceutical researchers to obtain protein samples and research information, such as purification protocols and crystallization conditions. Such an infrastructure will enable the individual laboratories to focus on more challenging research problems, including functional characterization of their favorite proteins, structure determination of disease targets, large proteins and protein complexes, study of post-translational modifications, and systemic questions involving networks of proteins. Structural biologists in the pharmaceutical industry will increase their impact on the drug discovery process because they will be in a position to determine the structures of many more protein–ligand complexes, at a lower cost and faster, using the structures, samples and protocols from the high-throughput structure determination. The impact of structural genomics on the structural and functional characterization of proteins, as well as on our understanding of the machinery of life and its evolution, will surely result in profusion of new drug targets and better drugs. Structural genomics will refocus biology, just as the advent of of high throughput DNA sequencing machines freed researchers to focus on more elaborate experiments rather than on accumulating sequence data.

Andrej Sali is with the Laboratories of Molecular Biophysics, The Rockefeller University, New York, New York 10021, USA. email: sali@rockefeller.edu

1. Structure-based functional genomics, October 4–7, Avalon, New Jersey (1998). http://lion.cabm.rutgers.edu/bioinformatics_meeting/. The meeting was organized by Gaetano Montelione, Stephen Anderson, Edward Arnold, and Ann Stock of the Center for Advanced Biotechnology and Medicine at Rutgers University (CABM), with the backing of the New Jersey Commission on Science and Technology, CABM, the Merck Genome Research Institute, and the Burroughs Wellcome Fund.
2. Synchrotron supplement. *Nature Struct. Biol.* **5**, 614–656 (1998).
3. Editorial, *Nature Struct. Biol.* **5**, 925–926 (1998).

Fight the flu



Adapted from Chen *et al.*

Every year, we are reminded of the power of the influenza virus, especially when severe and deadly outbreaks occur — such as last season's Hong Kong 'bird flu' in which 6 of the 18 confirmed cases in humans were fatal. What makes one influenza virus more virulent than another? In the case of the Hong Kong virus, one factor was probably a variation in its hemagglutinin (HA) precursor protein, which had a five residue insertion. Hemagglutinin, a glycoprotein that is held in the viral membrane by a C-terminal transmembrane sequence, is required for viral membrane fusion during infection. Two events must occur for hemagglutinin to become active: (i) the precursor protein, HA0, must be cleaved to create two disulfide linked subunits HA1 and HA2 and (ii) cleaved HA must undergo a low pH-induced conformational change in endosomes. The new N-terminus of HA2 is the fusion peptide that becomes embedded in the target membrane, leading to infection. The pH-induced conformational change projects the fusion peptide to the end of a long coiled coil, placing it close to the HA C-terminal transmembrane domain and thus bringing the viral and the target membranes into close proximity. The five-residue insertion in the Hong Kong HA0 hemagglutinin variant occurred at the HA0 cleavage site. Why would such a mutant lead to higher virulence? Researchers have found that HA0 variants with insertions in this region are more susceptible to proteases, leading to a greater proportion of cleaved HA molecules and hence a higher infectivity.

Hemagglutinin is a logical drug target because of its key role in the success of an infection. Researchers could aim to develop drugs that would block the cleavage step or prevent the conformational change. Useful information for such endeavors is provided by a recent structure of the HA0 precursor (Chen, J. *et al. Cell* **95**, 409-417; 1998), which shows the conformation of the intact cleavage site. A comparison with the previously determined cleaved HA structure suggests that proteolysis may be playing a direct role in setting the low pH trigger for the conformational change — in addition to playing more obvious roles in exposing the fusion peptide and giving conformational flexibility to the protein. In the HA0 precursor, the cleavage site (left image, between the dots) is exposed on a surface loop. After proteolysis, the new N-terminus of HA2 fills a hole in the protein (arrow), burying several ionizable Asp and His residues that were previously exposed in the cavity (right image). In theory, this could set a sensitive trigger: protonation of the His residue at low pH could severely destabilize the structure and lead to the dramatic conformational change that is necessary for infection. TS