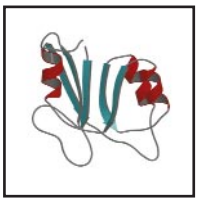
An overview of structural genomics

Stephen K. Burley

With access to sequences of entire human genomes plus those of various model organisms and many important microbial pathogens, structural biology is on the verge of a dramatic transformation. Our newfound wealth of sequence information will serve as the foundation for an important initiative in structural genomics. We are poised to embark on a systematic program of high-throughput X-ray crystallography and NMR spectroscopy aimed at developing a comprehensive view of the protein structure universe. Structural genomics will yield a large number of experimental protein structures (tens of thousands) and an even larger number of calculated comparative protein structure models (millions). This enormous body of structural data will be freely available, and promises to accelerate scientific discovery in all areas of biological science, including biodiversity and evolution in natural ecosystems, agricultural plant genetics, breeding of farm and domestic animals, and human health and disease.



The benefits of combining three-dimensional structural information with whole genome sequences are well-documented by the enormous success of investigator-initiated, hypothesis-driven biomedical research using X-ray crystallography and NMR spectroscopy. To paraphrase influential architects of the early 20th century, "function follows form". Virtually every one of the 12,000 plus protein structures in the Protein Data Bank (PDB, *http://www.rcsb.org/pdb/*; see the article by Berman and colleagues in this issue) has proved useful. On the computational front, protein fold assignment and comparative protein structure modeling represent important bioinformatic tools, which bring mechanistic insights to biology, chemistry, and medicine (see the articles by Gerstein, and Sali and colleagues in this issue).

The impact of comparative protein structure modeling stems in no small part from the rather low complexity of protein fold space, which is quite unlike the Byzantine character of the 100,000 or so human genes. Although there is still some uncertainty regarding the precise numerology, we now appreciate that the universe of compact globular protein folds is relatively small. Current estimates suggest that there are 1,000–5,000 distinct, stable polypeptide chain folds in nature[1]. The PDB contains experimental structures of <700 of these distinct protein folds[2], with some 'popular' folds such as the eight-fold αβ barrel of the triose phosphate isomerase (TIM) type represented by >20 protein sequence families. In eukaryotes, most genes encode proteins with multiple globular domains (the average domain size being 153 ± 87 residues[2]), giving larger proteins the appearance of beads on a string. Typically, a single 'bead' is responsible for carrying out a specialized biochemical task. A significant evolutionary change in gene sequence manifests itself at the level of an individual protein functional unit or domain, which may be regarded as the focus of natural selection. Changes that destabilize the structure of a critical domain within an essential gene product do not endure, whereas active site changes that create beneficial new biochemical activities can persist.

Computational biologists exploit evolution to leverage the output of experimentalists. Each newly determined structure typically provides useful structural information for 15–40 protein sequences for which no data were previously available (see the article by Sali and colleagues in this issue). Comparative protein structure modeling need not be one at a time. Large-scale analyses have been applied to whole genomes and large databases. In practice, homology modeling is currently restricted to protein sequences for which a nearby (defined as amino acid sequence identity >30–35%) experimental template is available. The paucity of the protein structure database currently limits the scope of modeling to ~50% of the open reading frames of the *Saccharomyces cerevisiae* genome. If one considers that only a portion of a given protein can usually be modeled, the situation looks decidedly worse (~18% of all residues in yeast proteins).

Structural biologists can contribute enormously to biology and biomedical research by providing full coverage of protein sequence space (that is, by determining at least one experimental structure for every protein sequence family, defined at the level of 30–35% identity). This achievement will bring all globular proteins within the radius of convergence of current homology modeling tools. All that remains is to define the best way of arriving at this important end point. Even with a business-as-usual paradigm, existing experimental programs will ultimately succeed but it could take decades because protein crystallographers and NMR spectroscopists typically choose their structure determination targets with an eye to addressing particular biological or biochemical questions. (Of the 13,000 plus PDB entries, only ~5,000 represent distinct experimental templates for comparative protein structure modeling.) The spectacular success of the Human Genome Project suggests that a systematic structural genomics effort should be able to do it better, faster, and cheaper. Current estimates[3] suggest that a directed program of structural study focussing on 10,000–20,000 selected targets will yield a comprehensive ensemble of protein structures capable of supporting homology modeling of every globular segment of every protein found in nature. Only structural biologists can do this important job, and I would argue that we have a duty to larger biomedical research community to get it done as soon as possible.

Howard Hughes Medical Institute, Laboratories of Molecular Biophysics, The Rockefeller University, 1230 York Avenue, New York, New York 10021 USA.
email: burley@rockefeller.edu

Tactical considerations (How?)

The feasibility of a large-scale, high-throughput structure determination program is being explored by pilot studies underway in North America, Europe, and Asia (reviewed in the articles by Terwilliger, Heineman, and Yokoyama, respectively, in this issue). Efforts involve both X-ray crystallography and solution NMR spectroscopy, with target lists derived from the genome sequences of archaea, eubacteria, and eukaryotes. Most of these pilot studies have already deposited structures into the PDB, and there is general agreement that all of the necessary technologies are in place (see the articles by Edwards and colleagues, Stevens and colleagues, and Montelione and colleagues in this issue for discussions of protein expression/purification, and X-ray and NMR approaches).

Recently, the National Institute of General Medical Sciences (NIGMS) of the National Institutes of Health (NIH) issued a request for applications for P50 grants to fund expansion of structural genomics pilot studies (*http://www.nih.gov/nigms/funding/psi.html*). Awards have been made to seven consortia (see the article by Terwilliger). In addition, private sector efforts are focused on medically-relevant proteins that represent drugs or drug discovery targets or drugs in their own right (see the article by Harris and colleagues in this issue). Although there are some suggestions in the US press that the private and public structural genomics efforts are on a destructive collision course, there seems no cause for alarm. The US funded effort is aimed at broad coverage of protein structure space, whereas the private efforts will be focused on medically important protein families. There is no *a priori* conflict between breadth and depth.

The NIH-funded pilot studies have been charged with the development of all of the technology required for obtaining protein structures, going from their gene sequences to disseminated structures. The NIGMS is seeking an appreciation of the types of problems, the likely success rate, and the feasibility of large-scale, highly parallel structure determination. In general terms the process involves: (i) PCR amplification of the coding sequence from genomic or cDNA; (ii) cloning the coding sequence into an appropriate expression vector; (iii) expressing the protein at a sufficiently high level; (iv) sequencing the cloned gene to verify that the coding sequence was correctly amplified; (v) confirming the identity of the expressed protein and characterizing it as a prelude to NMR or crystallographic studies; (vi) obtaining the protein in sufficient amounts and purity for either approach; (vii) defining suitable crystallization or NMR solution conditions; (viii) NMR or X-ray measurement (ix) determining and refining the experimental structure; (x) calculating comparative protein structure models using this new template; and (xi) making functional inferences from the structure plus derived models and disseminating the findings. Failures are anticipated at every step, making the process somewhat akin to a funnel.

Strategic considerations (What? Who? How much? To what end?)

Target selection is the most important strategic issue confronting the structural genomics pilot studies. Their respective performances will be measured in terms of the number of structures determined, what fraction contain novel folds, their impact on biology, and the cost per structure. The question of which structures to target is also relevant to the US funding agencies, because Congress will need to be convinced of the social and medical benefits of a public structural genomics initiative. Numerous discussions of target selection have been held[4,5]; (*http://lion.cabm.rutgers.edu/bioinformatics_meeting/*, *http://www.nih.gov/nigms/funding/psi.html*), and there is general agreement that it represents a research problem in its own right. There is no such agreement on the extent to which biomedical criteria should play a role in target selection. Pilot projects run the gamut from exhaustive studies of all proteins found in a model organism (*Methanococcus jannaschii*, *Mycobacterium tuberculosis*) to selectively chosen targets from a large number of different organisms (see the article by Brenner in this issue). Communication of target lists among the various pilot projects has been streamlined with the creation of two web sites (*http://structuralgenomics.org/*, *http://presage.berkeley.edu/*), and the NIGMS has already signaled that it will help coordinate the effort.

Preliminary pilot study results suggest that a structural genomics initiative will need to be a centralized endeavor. While structural genomics centers develop automated tools for cloning, protein expression and purification, sample preparation, X-ray and NMR measurement, structure solution/refinement, homology modeling, annotation, and dissemination, it makes sense to allow other research teams to explore alternate approaches to optimizing the process. Eventually, however, the need to achieve economies of scale and the repetitive nature of an industrial level production phase will require an integrated approach.

Another important strategic issue concerns money. To date, no clear picture has emerged as to how much it will cost to determine 10,000–20,000 experimental structures. Given the experience of the Human Genome Project, we can be certain that economies of scale will bring costs down. Publicly funded pilot studies should provide accurate initial cost estimates, which will be essential for any plausible justification of an expanded effort. NIGMS officials took great pains to explain that they would not be inclined to fund a program that threatened funding of traditional research grants. They were rightly influenced by debates over the Human Genome Project. Early detractors were quick to argue that this 'big' science would pauperize biomedical research. Their alarmist predictions were not borne out, and there is no reason to believe otherwise for the US funded structural genomics initiative.

Unlike genome sequencing, it is no simple matter to decide when a structural genomics initiative will be complete. Using the uncertain 'new fold' criterion, any definitively stated answer would only serve as a lightening rod for criticism. Indeed, many of the early criticisms of the Human Genome Project have resurfaced as objections to structural genomics (that is, unimaginative big science with little or no intellectual or training value, sapping public biomedical research funding). An alternative, but similarly unhelpful, definition of the end point would be experimental structures of all proteins encoded by a particular genome (*S. cerevisiae*, for example). The most pragmatic approach may be an operational definition based on the current performance of homology modeling procedures. As a general rule, 30–35% sequence identity represents a critical threshold for successful modeling[3,6]. Restricting our analysis to the extreme case of all sequences that we will ever know, an estimate of the total number of 30–35% sequence identity protein families is required. Statistical work places that number somewhere between 10,000 and 30,000 (C. Sander, pers. comm.). The accuracy of the foregoing analysis should improve substantially with the release of the entire human genome sequence, making it possible to define a useful end point for the NIGMS structural genomics initiative in short order.

Expected benefits

There are four possible outcomes when a target protein structure is determined, corresponding to all combinations of 'new/old'

fold and 'known/unknown' function (either its own or that of one or more close homologs). New structures will be useful in all four instances, admittedly in different ways. At a minimum, every new structure will permit modeling of a protein family for which no structural information was previously available. Each one of these homology models can serve as a starting point for a rational program of experimentation, such as site-directed mutagenesis, ligand binding studies, enzyme assays, protein–protein interaction studies, and so forth. Should the structure represent a new fold with a known function, it may well be possible to identify regions of the protein responsible for function *in silico* by comparing the newly determined structure with those of structurally distinct yet functionally similar proteins. When the structure proves to be a known fold with a known function, we can expect to learn something about divergent evolution. This experience has played out repeatedly with the TIM barrel enzymes, which catalyze a wide variety of chemical reactions using the same protein fold decorated with different patterns of surface-accessible residues creating functionally distinct active sites[7].

Where we do not know anything about biochemical function, both new and previously known structures should still prove useful. The newly determined structures that are not in fact novel can be compared with their structural homologs, and it may be possible to infer function. The new fold structures for which there is no functional information represent a red flag to some critics of structural genomics. In these cases, however, two courses of action are open. First, novel structures may be functionally characterized by scanning them against a library of all known binding sites and enzyme active sites. Second, the new fold structures represent an excellent vantage point from which to develop testable hypotheses regarding function.

I believe that choosing medically relevant targets will have all of the benefits outlined above, plus a number of important consequences for disease- and patient-oriented research. First, each newly determined structure will be of immediate relevance to academic and/or industrial research teams studying that particular system. Second, by publicizing target lists on the Internet, the pilot studies could attract biological and biochemical expertise from the larger scientific community and even entertain suggestions for additions to their respective target lists. Third, the pilot studies will be able to serve as important sources of new technologies and useful reagents. Fourth, some of these newly characterized proteins may represent protein pharmaceuticals in their own right. Fifth, a database of comparative protein structure models could be used for *in silico* 'docking' with libraries (possibly combinatorial) of small organic molecules. The resulting database of docked structures could permit identification of putative therapeutic leads in advance of high-throughput screening. Finally, the structural consequences of medically important SNPs that occur in coding regions could be examined using comparative protein structure modeling.

Limitations and caveats

Although the pilot study results are encouraging to say the least, there are some regions of protein structure space that will not succumb immediately to either NMR or X-ray methods. Membrane protein crystallization continues to represent a considerable technical challenge, but advances in robotic protein solubilization/purification and crystallization may ease these difficulties. (Five years ago, it would have been difficult to anticipate the speed with which we can now study the structures of single domain globular proteins.) Alternatively, the recent development of the TROSY technology by Kurt Wüthrich and coworkers may offer an approach to solution NMR studies of protein–detergent micelles. There have also been a considerable number of predictions that proteins normally resident in macromolecular complexes cannot be studied in isolation. This contention may be true, but misses the point of the genome-wide philosophy of structural genomics. Somewhere in biology, the same fold will almost certainly be used in a context where it is not inextricably bound up in a large complex. The remainder of protein sequence space is occupied by so-called low complexity regions, which may never adopt stable conformations or remain unstructured until they interact with their respective targets. Clearly these cases are beyond the initial scope of a structural genomics initiative.

Release of data

In closing, I would argue that the most significant caveat for structural genomics concerns getting the issue of atomic coordinate release right. At the 1st International Structural Genomics Meeting, held at the Wellcome Trust Genome Campus in the UK on April 4–6, 2000, there was unanimous agreement that the results of the public structural genomics initiatives should be freely available *via* the PDB. The only contentious discussions of the meeting focused on publication and timely release of the results of public efforts. Two issues emerged. First, there was a clear difference of opinion between the structure solvers and the high-throughput sequencers. The latter group argued strongly in favor of immediate, automated release of all interim results, including unvalidated structural information. Both X-ray crystallographers and NMR spectroscopists were unwilling to support such a plan on the grounds that preliminary structural information did not have the same value as imperfect DNA sequences, and could prove very misleading. Second, there was no clear agreement on when any structure determination could be said to be complete. Some participants called for automatic release of coordinates once certain statistical criteria were met (for example, when the R_{free} statistical measure of quality drops below 30%), but that view did not gain acceptance. The participants ultimately agreed that atomic coordinates should be released immediately on publication (possibly electronic), a measure that conforms to current best practice in structural biology. A description of the outcome of this important first 'Bermuda' meeting (a reference to the early discussions of high-throughput genome sequencing held on the island of Bermuda) for structural genomics is available from *http://www.nigms.nih.gov/news/meetings/hinxton.html*.

Associations with structural genomics

S.K.B. is the the Principal Investigator of the New York Structural Genomics Research Consortium P50 NIH Center Grant, and serves as Chairman of the Scientific Advisory Board of the PDB. In addition, he is a cofounder of Prospect Genomics Inc. and a member of the Scientific Advisory Board of Structural GenomiX Inc.

1. Brenner, S.E., Chothia, C. & Hubbard, T. *Curr. Opin. Struct. Biol.* **7**, 369–376 (1997).
2. Orengo, C.A. *et al. Nucleic Acids REs.* **27**, 275–279 (1999).
3. Sanchez, R. & Sali, A. *Proc. Natl. Acad. Sci. USA* **95**, 13597–13602 (1998).
4. Gaasterland, T. *Trends Genet.* **14**, 135 (1998).
5. Sali, A. *Nature Struct. Biol.* **5**, 1029–1032 (1998).
6. Sanchez, R. & Sali, A. *J. Comp. Phys.* **151**, 388–401 (1999).
7. Burley, S.K. *et al. Nature Genet.* **23**, 151–157 (1999).