Long live structural biology

Raymond C Stevens

Two camps continue to evolve in the field of structural biology—a 'systems-oriented' camp, which studies proteins or complexes carefully one system at a time, and a 'discovery-oriented' one, which studies proteins of entire families, pathways or genomes. The end goals of both camps are the same: to decipher the atomic-resolution structures and mechanisms of biological macromolecules and understand them in the context of the living cell.

In the January issue of *Nature Structural & Molecular Biology*, Stephen Harrison accurately communicates in a commentary[1] the need to expand the length and timescale at which we study biological systems at atomic resolution. Expanding the number of systems being studied will certainly benefit our knowledge of basic science and medicine. Unfortunately, the average time it takes to solve a challenging eukaryotic protein target from clone to structure has been one to three years, and the time investment is even longer, with a higher risk of failure, for viruses, molecular machines or membrane proteins. At this pace, it will take a very long time before enough structures are accumulated so that one can begin to make sense of the different systems in the context of the living cells. Using conventional methods, the throughput of three-dimensional biological structures can only be improved by increasing the number of person-hours of work.

Structural biology also has an impact on the drug discovery process. For example, an estimated 50% of the cost of drug discovery would be saved if a target protein structure were used at an early stage to generate leads of high quality. However, until recently, the structure determination process has been too slow and not sufficiently robust to make a significant impact[2], except in a few isolated albeit highly celebrated cases, such as the inhibitors for angiotensin-converting enzyme[3], HIV-1 protease[4] and influenza virus neuramindase[5]. The high-throughput structural biology approach is being developed in in parallel to several other high-throughput and cell-based biochemical techniques that have recently emerged, including cell-based mass spectrometry[6], activity-based protein profiling[7] and microarrays[8]. All of these methods, including structural genomics, allow us to gain biochemical understanding of the complexity of the entire living cell in a shorter amount of time.

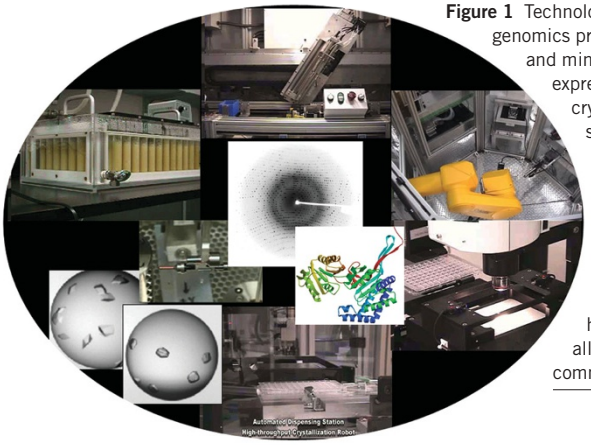

Figure 1 Technologies developed by structural genomics programs include automation and miniaturization of protein expression, purification and crystallization, automation of synchrotron data collection, and software development for rapid structure solution, refinement and improved quality control (figure compiled from the Joint Center for Structural Genomics; http://www.jcsg.org). Almost all of these technologies are now commercially available.

Lessons learned from genomics

Genome sequencing of *Hemophilus influenza* in 1995 (ref. 9) and *Drosophila melanogaster* in 2000 (ref. 10) prepared the stage for the completion of the human genome project in 2001 (refs. 11,12). The sequencing of the genetic materials of several SARS variants increased our understanding of the infection pathway at a rapid rate[13,14]. These and other genome sequencing projects have been valuable for our understanding of biology. In the early 1990s, few researchers would have expected the human genome to have the relatively small number of genes that was eventually reported, the power of comparative genomics to be as powerful as it has become, the similarity among genomes to be as great as they are, or the role of single nucleotide polymorphisms (SNPs) to be as significant as has been discovered. Although genomics was originally met with opposition as being non-hypothesis-driven, it is now being described as discovery-oriented relative to the traditional systems-oriented approach[15]. The discovery-oriented approach is likely to be less biased because it examines, for example, large networks of macromolecules up to the scale of a complete genome and focuses on exploring in the context of the cellular organism. Furthermore, the information gleaned is notably cheaper, more complete and of higher quality, and the rate of information accumulation is faster. The information is also useful

The author is in the Department of Molecular Biology and Chemistry at the Scripps Research Institute, La Jolla, California 92037, USA, and is a member of the Joint Center for Structural Genomics.
e-mail: stevens@scripps.edu

for all researchers in the life sciences. The development of technologies such as new sequencing approaches and instrumentation, as well as novel algorithms to analyze incredible amounts of data, were crucial for the achievements described above.

New technologies for structural biology

Before 1999, the funding of technology development in structural biology by government and industrial entities was minimal, with a focus on the systems-oriented approach. Now that we know the size of many genomes, the goal of understanding the entire cell becomes even more exciting as well as challenging[15,16]. Significant technological developments have recently been made in many aspects of structural biology (**Fig. 1**). For example, cell-free protein expression that lowers the cost, particularly for labeling experiments, is becoming more robust[17]. Nanovolume crystallization, developed to decrease the amount of protein required for experiments, is now repeatedly showing success for proteins that have not crystallized with the traditional microliter volumes[18]. MAD phasing using synchrotron radiation and automation at synchrotron beamlines have significantly increased the efficiency of perhaps one of the most precious resources for crystallographers—synchrotron beamtime[19]. Although only a few beamlines currently have automation, every synchrotron facility worldwide has at least one robotics station planned, and this is almost certain to increase as robotics become more affordable. Automated crystallographic software has not only accelerated data processing, structure solution and refinement, but also increased the assessment of data quality at all stages of structure determination[20]. All of these technical developments are now readily available to the entire scientific community via their rapid commercialization based on demand. Hardware including crystallization, imaging and crystal mounting systems are all available at under $300,000, making them accessible to individual investigators. Most software is available via the web at low or no cost to noncommercial groups.

For bacterial proteins under investigation, the time and cost have been reduced two- to three-fold. However, for eukaryotic proteins, the timeframe is still too long, the cost very expensive and the probability of success relatively low (~10–20%). The field of structural biology still needs to overcome the challenges of eukaryotic protein expression, large protein assemblies and membrane proteins. To make the systems biology approach involving protein structures useful, structure determination of these very challenging proteins will be necessary. Phase 1 of the Protein Structure Initiative (PSI) has done much to advance the ability to make significant progress on prokaryotic proteins. Future support, such as phase 2 of PSI, will be necessary to advance the remaining technology for studying many other exciting systems (for example, human proteins) that are more complex and require different handling considerations.

How complex is the living cell?

One of the most interesting questions in structural genomics programs is evaluating how many proteins in an organism have defined structures by themselves or require cofactors or binding partners to fold. In a high-throughput experiment that took only six months by the Joint Center for Structural Genomics (La Jolla, California, USA), it was determined that 24% of the 1,877 proteins encoded in the *Thermotoga maritima* genome could be expressed and crystallized by themselves[21], whereas another ~35% would probably require a folding partner or small molecule, or are membrane proteins. The remaining 40% of the proteins in the genome are poorly understood or validated at the biochemical or structural level. It has been estimated that a significant number of proteins in a bacterial cell are disordered, and this number could be even larger for eukaryotic proteins[22]. The ordering of these proteins may affect their function or regulation. As an example of eukaryotic cell activity exhibiting disorder-order transition, vesicle fusion is mediated by SNARE proteins. SNAREs are largely unstructured in the cell but become helical upon signaling[23].

Another question being studied by the PSI-funded structural genomics efforts is the number of folds and types of folds in an organism. A thorough understanding of the fold space is necessary to build a model of a complete organism at an atomic level of detail. Notably, the software currently used to predict novel protein folds has not performed sufficiently well in structural genomics efforts, suggesting that this issue is more complex than anticipated. An impressive 70% of structures deposited by PSI centers in the Protein Data Bank have unique sequences (defined as having <30% sequence identity with other protein sequences in the Protein Data Bank), compared with 10% of all structures deposited. However only 12% of those unique sequences have novel folds. Although populating the fold space is one goal of structural genomics, another goal of the more commercial-oriented structural genomics efforts is to characterize entire families of proteins. For example, ~700 human kinases are being studied so that one can begin to understand the specificity of these proteins (and possibly their signaling pathways) at the active site atomic resolution level. Lastly, both academic and industrial structural genomics efforts plan to solve the majority of structures of proteins from pathogenic microbes such as *Plasmodium falciparum* (the malaria parasite), *Mycobacterium tuberculosis* or the SARS virus.

The challenges described above clearly indicate that, even after knowing the genome sequence of an organism, we still have a long way to go to understand the living cell from a biochemical or structural perspective. Delineating molecular-level details for the function of all cellular components could facilitate an in-depth understanding of the workings of integrated cellular machinery. A great deal of data remains to be collected using a variety of high-throughput proteomic techniques before we can describe in molecular detail what is happening inside even the simplest of organisms. Mass spectrometry, activity-based profiling and microarrays have made substantial strides in technology development in this respect. However, detailed structural information is also a critical component of an understanding of the living cell, and the technologies in structural genomics must keep pace with these other approaches.

A union of experimental and computational biology

Currently, the experimental side of structural biology and biochemistry is not integrated well with the computational side of these fields, although great progress has been made over the past few years. Eventually, perhaps within the next decade, this merger will occur, and it is a clear goal of both traditional structural biology and structural genomics efforts. As an example of technology development, structural genomics collects cloning, expression, purification, crystallization and NMR data in a consistent and controlled manner. Such data can be used to increase our understanding of basic protein biophysical behavior. This not only improves the processing of proteins, thereby increasing efficiency and lowering cost, but also generates valuable negative and positive data so that we can understand and possibly predict protein behavior[24]. On the discovery side, an additional benefit of structural genomics is coordinated target selection to avoid the overlap in structural space, use government funds efficiently, and accelerate scientific progress. A similar approach with

coordinated sequencing efforts was critical for the incredibly rich genomic data that we now have access to and benefit from.

Closing remarks

When I was a postdoctoral fellow in William "The Colonel" Lipscomb's laboratory, The Colonel would often say, "to solve a three-dimensional structure is one trophy, but to understand its structure and mechanism is an even greater trophy." This is true for those that study the structural biology of individual proteins and/or complexes, and for those that study protein pathways, families or entire genomes. The goal of the two different camps is the same. The only difference is the path and the amount of time that the field of structural biology will take to get there. Based on the similarities among the genomes of human, chimpanzee, rat and fly, it is certain that the subtle differences at very detailed levels will be needed to understand many of the differences between such species.

DISCLOSURE

The author is a cofounder and member of the scientific advisory board of Syrrx, a rational drug discovery company and a cofounder and Chairman of the Scientific Advisory Board of Sagres Discovery (via MemRx acquisition), a therapeutic antibody company focused on membrane protein oncology targets. Both companies use high throughput structural biology platforms.

ACKNOWLEDGMENTS

I thank I.A. Wilson, P. Kuhn, A. Godzik, R. Page, S. Lesley and J. Canaves for valuable input, A. Walker for manuscript and figure preparation and the Joint Center for Structural Genomics (P50 GM62411) for funding.

1. Harrison, S.C. *Nat. Struct. Mol. Biol.* **11**, 12–15 (2004).
2. Stevens, R.C. *Drug Disc. World* **4**, 35–48 (2003)
3. Ondetti, M.A., Rubin, B., & Cushman, D.W. *Science* **196**, 441–444 (1977).
4. Kaldor, S.W. *et al. J. Med. Chem.* **40**, 3979–3985 (1997).
5. Kim, C.U. *et al. J. Am. Chem. Soc.* **119**, 681–690 (1997).
6. Schirmer, E.C. *et al. Science* **301**, 1380–1382 (2003).
7. Adam, G.C. *et al. J. Am. Chem. Soc.* **126**, 1363–1368 (2004).
8. MacBeath, G. *Nat. Genet.* **32** (suppl.), 526–532 (2002).
9. Smith, H.O. *et al. Science* **269**, 538–540 (1995).
10. Adams, M.D. *et al. Science* **287**, 2185–2195 (2000).
11. Lander, E.S. *et al. Nature* **409**, 860–921 (2001).
12. Venter, J.C. *et al. Science* **291**, 1304–1351 (2001).
13. Marra, M.A. *et al. Science* **300**, 1399–1404 (2003).
14. Rota, P.A. *et al. Science* **300**, 1394–1399 (2003).
15. Ho, Y. *et al. Nature* **415**, 180–183 (2002).
16. Gavin, A-C. *et al. Nature* **415**, 141–147 (2002)
17. Stevens, R.C., Yokoyama, S. & Wilson, I.A. *Science* **294**, 89–92 (2001).
18. Santarsiero, B.D. *et al. J. Appl. Crystallogr.* **35**, 278–281 (2002).
19. Abola E. *et al. Nat. Struct. Biol.* **7**, 973–977 (2000).
20. Adams, P.D. *et al. J. Synchrotron Radiat.* **11**, 53–55 (2004).
21. Lesley, S.A. *et al. Proc. Natl. Acad. Sci. USA* **99**, 11664–11669 (2002).
22. Wright, P.E. & Dyson, H.J. *J. Mol. Biol.* **293**, 321–331 (1999).
23. Fiebig, K.M. *et al. Nat. Struct. Biol.* **6**, 117–123 (1999).
24. Page, R. *et al. Acta Crystallogr. D* **59**, 1028–1037 (2003).