Industrializing Structural Biology

Raymond C. Stevens and Ian A. Wilson

Since Röentgen's discovery of x-rays, crystallography has been the method of choice to determine molecular structure (*1*, *2*). The utility of this technique toward proteins was first demonstrated in 1960 with the determination of the myoglobin structure (*3*) and has since been applied to much more complex macromolecular structures and assemblies, such as viruses (*4*, *5*), membrane proteins (*6*), and the ribosome (*7*, *8*). The gradual accumulation of structures has allowed a more complete representation of three-dimensional (3D) protein folds, enabling more successful homology modeling of unknown structures (*9*). Breakthroughs in genome sequencing projects have underscored the need for concomitant advances in structural biology if we hope to determine the extent of protein folding-space and elucidate how an assembly of proteins constitutes a cellular organism. As a result, structural genomics programs have sprung up both nationally and internationally, with significant funding from the National Institutes of Health in the United States (www.nigms.nih.gov).

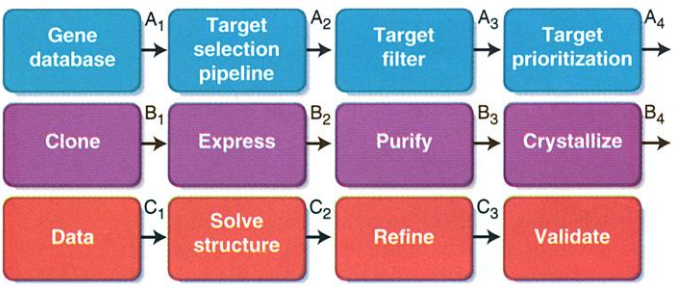
Determining protein structures on a genome-wide scale is a formidable task. Traditionally, each target gene is cloned into an expression vector, expressed with the use of a single set of conditions, and the resulting protein is then purified. With this protein in hand, a number of basic screens for crystallization are used, followed by further screening, if necessary, to optimize crystal quality. Lastly, the best crystal(s) are used for diffraction analysis. Such isolated investigations on individual proteins are being replaced with parallel sample processing approaches for structure determination in both the public and private sectors. Public efforts, at least in the United States, are being directed toward elucidating the universe of protein structural families. In the private sector, structural data are also being harnessed for biochemical function prediction and as initial 3D templates in drug discovery programs.

Using conventional methods, the throughput of macromolecular 3D structure determination can be improved only by increasing the person-hours of work. As a consequence, academic and industry-based researchers have initiated research and development programs to develop high-throughput (HT) structure determination process pipelines, as diagrammed in the figure. The stage is now being set for the industrialization of structural biology in much the same way that Henry Ford revolutionized the auto industry. Structures will no longer be determined one at a time. "Assembly line" structure determination approaches are being developed that can cope with HT. Large multi-institutional conglomerates of expertise have coalesced into factory-like consortiums to provide the wide range of methodologies needed to convert gene sequences into validated protein structures.

HT structural biology requires development of methods and reagents to streamline and automate the process of protein structure determination. Given the large number of gene sequences, their protein products, and corresponding 3D structures, all procedures must be automated if the goal of providing a structure corresponding to every known gene in an organism is to come close to reality. In addition, eliminating bottlenecks in the pipeline will substantially increase the rate of sample processing and will improve the efficiency of solving protein structures.

The early steps in the pipeline can capitalize on the HT technologies developed during genomic sequencing efforts. Robotic liquid handling and colony picking procedures, as well as automated sequencing, can easily be adapted for the cloning and expression steps. But as these impediments are removed, other processes become rate-limiting. For example, once sufficient quantities of purified protein are obtained, crystals suitable for x-ray diffraction must be grown. The substantial improvements in HT protein crystallization with the use of robotic workstations could eliminate this bottleneck. Lastly, the diffraction data collection and analysis steps must be automated, streamlined, and made user-independent. The numerous groups currently working on HT structural biology will undoubtedly come up with ways to realize these goals (*10*).

Over the past decade, bigger, more powerful computers have also enabled industrialization of structural biology, not only by collecting data faster, but also with increasingly sophisticated data analyses. In a substantial change in philosophy, both positive and negative data are now being archived in order to correlate successes and failures with specific steps in the structure determination process.



Pipeline flow from cDNA to structure. Each step along the path has been a bottleneck in the past. Currently, the biggest bottleneck is finding the protein variant from which suitable diffraction quality crystals can be obtained. The parameters A_1 to C_3 for each path are positive and negative data collection points now being analyzed to improve the efficiency of the cDNA to structure pipeline.

Estimates of the cost of a determining protein structure by traditional methods range from $100,000 to $250,000 for a soluble protein with a previously unknown fold; membrane protein structures are much more expensive to determine due to the greater complexity of their environment. Reducing this cost by an order of magnitude, therefore, must also be a goal of HT methodologies. A HT combinatorial approach simultaneously applied to multiple protein targets requires the parallel utilization of multiple expression constructs and purification schemes. Standard expression systems, such as *Escherichia coli* or eukaryotic baculovirus-infected insect cell systems, are currently the major workhorses (*11*, *12*). To simplify purification, an affinity tag, such as polyhistidine to facilitate isolation of the desired protein, is usually incorporated into the expression clone. Expression trials are run in parallel on multiple expression clones, using a variety of different expression constructs and experimental conditions, to improve overall success rates. The combinatorial application of different methods of expression, including re-engineering by mutation and truncation and improving the protein chemistry processing steps (e.g., main-

The authors are in The Joint Center for Structural Genomics, Department of Molecular Biology and the Skaggs Institute for Chemical Biology, The Scripps Research Institute, La Jolla, CA 92037, USA.

taining disulfide integrity or avoiding aggregation and side-chain decomposition), enhances the number of protein targets that are likely to reach the end of the protein production pipeline.

Crystallization trials typically require between 2 and 20 mg of a purified protein for 100 to 1000 trials, using 2 μl of a 10 mg/ml protein solution. To improve success rates, the protein sample should be microscopically homogeneous, as determined by mass spectrometry or isoelectric focusing electrophoresis (*13*). A helpful indicator of successful crystallization is that the sample contains a single oligomeric state. This property can be analyzed by size exclusion chromatography or light scattering (*14*). HT application of these analytical techniques, even before crystallization, can help guide the protein production process.

Traditionally, protein crystallization has been a labor-intensive manual procedure using 24-well tissue culture plates in a hanging drop vapor diffusion configuration (*15*). Automation of the protein crystallization step coupled to miniaturization (nanoliter drop volume dispensing) could rapidly generate a range of crystal choices for diffraction analysis. With current robotic systems, throughputs of 2,500 to 100,000 protein crystallization experiments per day can be achieved. High-density 96-well and 1536-well plate configurations have allowed for the use of smaller volumes in each crystallization experiment. Miniaturization has resulted in easier automated storage and has allowed for manipulation of much larger numbers of samples, with a substantial reduction in material costs (*16*, *17*).

HT crystallization is likely to make the largest impact of all the new technologies available. Nanodrop crystallization trials reduce the required protein sample size, allowing for more complete exploration of crystallization parameter space, as well as for improved rates of drop equilibration (*18*) and successful crystal production. Faster crystal formation permits faster feedback, and, perhaps more surprisingly, can lead to improved crystal quality due to decreased decomposition and degradation. And the most obvious advantage to automation is that it removes operator error.

Due to the large number of crystallization experiments generated in a HT environment, automated imaging analyses are necessary to monitor and record the results of the crystallization trials. The nanodrop crystallization trials have made possible robotic HT imaging for detection of crystal formation, and numerous commercial imaging devices are becoming available. The automated detection of crystal formation is a much more difficult challenge than one might anticipate, given the ease with which the human brain can process such visual information.

HT collection of x-ray diffraction data is also based on recent technological advances, such as the construction of reliable and dedicated second- and third-generation synchrotron sources and the use of cryo-cooling to reduce sample degradation during data collection (*19*). Parallelization of the phase determination step using the multi-wavelength anomalous dispersion (MAD) technique (*20*) can also help automate the structure-determination process (e.g., by selenomethionine incorporation in the expressed protein samples). Numerous groups are developing robotic and automation systems for data acquisition and structure solution. Manual harvesting, freezing, and crystal mounting steps need to be converted to a HT format. Automated crystal centering, diffraction image collecting, data processing, phase determination and interpretation of electron density maps, model building, refinement, validation, and deposition of final coordinates (in the Protein Data Base) are pipeline process steps already being addressed (*21–24*). Importantly, major advances in robotic manipulation of crystals at synchrotron beamlines have already been achieved (*10*).

Automation, miniaturization, and parallelization of process steps are reinventing structural biology, determining structures faster and at lower cost. Similar improvements in nuclear magnetic resonance (NMR) methodologies will also enhance the pace of structure determination. The application of these HT processes will certainly enable the determination of many novel and interesting structures. In the beginning, the majority of new structures may represent the so-called "low-lying fruit" that are represented in the genome sequence databank, such as small stable proteins from thermophilic bacteria. But more intractable protein targets, such as membrane proteins, and, indeed, the more complex proteins of human and other higher organisms, will eventually also have their structures determined rapidly from the learning factory approach to structural genomics.

Does this, then, signal the end of conventional macromolecular crystallography? Most likely, yes. A fundamental shift to "industrial" structural biology will accelerate protein structure determination and increase the use of structural information in biological research. Protein structure determination per se will no longer be the main obstacle in elucidating the molecular basis of biology; this challenge will be passed back to the biologist. HT structural biology will facilitate comprehensive studies of, for example, complete metabolic pathways or how the complex assembly of protein components provides a blueprint for the operation and function of a cell. Structural genomics will help enable such identification of function from structure. The similarities of fold, the identification of bound ligands or protein partners, and the ability to screen libraries of small molecules will all facilitate elucidation of the open reading frame (ORF) function and will further the biomedical uses of the human genome efforts, as well as those of other organisms. Currently, there is renewed interest in using microbial genome information to target new antibacterial drugs. Future young scientists entering structural biology will address more complex biological structures, a trend already obvious with the recent nucleosome (*25*), proteasome (*26*), DNA polymerase (*27*), and ribosome structures (*7*, *8*, *28–30*). The industrial age of structural biology has clearly arrived.

References and Notes

1. W. C. Röentgen, *Wurzburg Physical Medical Society* (1895).
2. J. D. Bernal, D. Crowfoot, *Nature* **133**, 794 (1934).
3. J. C. Kendrew *et al.*, *Nature* **185**, 442 (1960).
4. S. C. Harrison *et al.*, *Nature* **276**, 368 (1978).
5. K. M. Reinisch *et al.*, *Nature* **404**, 960 (2000).
6. J. Deisenhofer *et al.*, *Nature* **318**, 618 (1985).
7. N. Ban *et al.*, *Science* **289**, 905 (2000).
8. B. T. Wimberly *et al.*, *Nature* **407**, 327 (2000).
9. J. M. Thornton, *Science* **292**, 2095 (2001).
10. R. F. Service, *Science* **292**, 187 (2001).
11. A. M. Edwards *et al.*, *Nature Struct. Biol. Suppl.* **7**, 970 (2000).
12. R. C. Stevens, *Structure Fold Des.* **8**, R177 (2000).
13. B. Lorber, R. Giege, in *Crystallization of Nucleic Acids and Proteins: A Practical Approach* (Oxford Univ. Press, New York, 1999), pp. 17–43.
14. Y. Georgalis, W. Saenger, *Sci. Prog.* **82**, 271 (1999).
15. A. McPherson, in *Crystallization of Biological Macromolecules* (Cold Spring Harbor Press, New York, 1999).
16. R. C. Stevens, *Curr. Opin. Struct. Biol.* **10**, 558 (2000).
17. U. Mueller *et al.*, *J. Biotechnol.* **85**, 7 (2001).
18. D. Diller, W. G. J. Hol, *Acta Crystallogr. D* **55**, 656 (1999).
19. R. M. Sweet, *Nature Struct. Biol. Suppl.* **5**, 654 (1998).
20. S. E. Ealick, *Curr. Opin. Chem. Biol.* **4**, 495 (2000).
21. E. Abola *et al.*, *Nature Struct. Biol. Suppl.* **7**, 973 (2000).
22. S. W. Muchmore *et al.*, *Structure Fold Des.* **8**, R243 (2000).
23. V. S. Lamzin, A. Perrakis, *Nature Struct. Biol. Suppl.* **7**, 978 (2000).
24. A. Perrakis, R. Morris, V. S. Lamzin, *Nature Struct. Biol.* **6**, 458 (1999).
25. K. Luger *et al.*, *Nature* **389**, 251 (1997).
26. J. Löwe *et al.*, *Science* **268**, 533 (1995).
27. P. Cramer *et al.*, *Science* **288**, 640 (2000).
28. V. Ramakrishnan, P. B. Moore, *Curr. Opin. Struct. Biol.* **11**, 144 (2001).
29. M. M. Yusupov *et al.*, *Science* **292**, 883 (2001).
30. J. M. Ogle *et al.*, *Science* **292**, 897 (2001).

31. We thank NIH for funding through grant 1P50 GM062411, and we gratefully acknowledge the valuable contributions of all of the members of the Joint Center for Structural Genomics, the Genomics Institute of the Novartis Research Foundation (especially P. G. Schultz), and Syrrx.