Commentary

Structural genomics: An approach to the protein folding problem

Gaetano T. Montelione*

Center for Advanced Biotechnology and Medicine, Department of Molecular Biology and Biochemistry, Rutgers University, and Department of Biochemistry, Robert Wood Johnson Medical School, University of Medicine and Dentistry of New Jersey, Piscataway, NJ 08854-5638

The large-scale genome sequencing projects present tremendous new opportunities for structural biology and molecular biophysics. This explosion of biological information provides novel insights into molecular evolution and molecular genetics, new reagents for molecular biology, and exciting new avenues for molecular medicine. However, to fully realize the value of these genetic blueprints, further investment is required to characterize the biological functions and three-dimensional structures of the corresponding gene products. These efforts, broadly characterized as functional and structural genomics, have the potential to provide a unified understanding of molecular biology from atomic to cellular levels.

During the last few years, several international efforts have been initiated with the common goal of genomic-scale three-dimensional (3D) protein structure determination (for a summary of international structural genomics centers and consortia, see http://www.rcsb.org/pdb/strucgen.html#Worldwide). Driven by the availability of many complete genome sequences, recent technological advances in rapid 3D structure analysis (1–5), and the integrative thinking of bioinformatics (6–10), these efforts aim to provide a coarse sampling of the space of 3D protein structures. Clustering proteins into homologous sequence families, it has been estimated that high-resolution structure determinations of some 15,000–20,000 carefully selected proteins will enable accurate modeling of hundreds of thousands of protein structures (10). As well as being useful in their own right, such models can provide the basis for rapid analysis of x-ray crystallographic or NMR data, facilitating experimental high-resolution structure determinations. A recent issue of PNAS includes a report (11) from the New York Structural Genomics Research Consortium (NYSGRC) describing the x-ray crystal structures of two proteins involved in sterol/isoprenoid biosynthesis and the amplification of these structural data by homology modeling. This study is particularly noteworthy as a model of the kinds of information and analyses that will be available as recently funded structural genomics centers and consortia around the world (12–15) come up to speed.

Although the vision of structural genomics is laudable, the feasibility of such an undertaking is, at the very least, controversial. It remains to be demonstrated that "high throughput" protein production and 3D structure analysis is feasible, that the resulting structures and biological insights are unique relative to ongoing traditional structural biology efforts, and that approaches for amplifying the resulting structural information are valuable. Fundamental to the scientific validity and impact of structural genomics is the target selection process (6–10). Ideally, structural genomics efforts focus on targets with high leverage value, either as members of large protein families across which the structural information can be amplified, or selected on the basis of functional genomics criteria for which broad biological information is available. Obviously, targets should be selected for which structural information is limited or which complement in valuable ways the structural information already available for that family. This target selection process generally involves significant input from bioinformatics and/or functional genomics analyses (9, 10).

Although the vision of structural genomics is laudable, the feasibility of such an undertaking is, at the very least, controversial.

In their current work, Bonnano *et al.* (11) describe high-quality x-ray crystal structures of the 396-residue yeast mevalonate-5-diphosphate decarboxylase (MDD) and 182-residue *Escherichia coli* isopentenyl diphosphate isomerase (IDI) enzymes. Both structures were determined by using multiwavelength anomalous diffraction (MAD; ref. 1), a rapid method of x-ray structure determination that exploits multiwavelength synchrotron x-ray radiation together with unique diffraction properties of certain atoms to efficiently determine the phases of the diffraction data required to determine the protein structure. In this study, MAD was enabled by biosynthetic incorporation of selenomethionine (SeMet) residues into the proteins, and data were collected at the National Synchrotron Light Source at Brookhaven National Laboratories in Upton, NY, or the Cornell High Energy Synchrotron Source in Ithaca, NY. MAD techniques using synchrotron radiation (1–3) represent a critical enabling technology for high throughput structure analysis by x-ray crystallography, underpinning the feasibility of the proposed genomic-scale structure projects (12–15).

MDD and IDI function at sequential steps in the biosynthetic pathway of sterols and other natural products. MDD catalyzes the last of three sequential ATP-dependent reactions that convert mevalonate to isopentenyl diphosphatase, whereas IDI catalyzes interconversion of isopentenyl diphosphate and dimethylallyl diphosphate, which condense in the next step of this biosynthetic pathway. Other enzymes in this pathway exhibit sequence similarities with MDD, suggesting potential structural and evolutionary relationships.

It is especially noteworthy that the NYSGRC has focused on multiple proteins from a common biosynthetic pathway. This is an important theme for structural genomics activities. Having structures and protein reagents in hand, the group is now in a position to further leverage their structural studies in experimental and computational analyses of protein–protein interactions, studies of the functional complementarity of these

See companion article on page 12896 in issue 23 of volume 98.

*To whom reprint requests should be addressed at: Center for Advanced Biotechnology and Medicine, Rutgers University, 679 Hoes Lane, Piscataway, NJ 08854-5638. E-mail: guy@cabm.rutgers.edu.

enzymes, and in structural/functional studies of other members of the pathway. Another important feature of the current analysis involves the use of homologous structures to circumvent practical challenges presented by sample preparation. Although the yeast IDI protein could be produced and purified, it exhibited aggregation properties that confounded crystallization efforts. The group was able to circumvent these problems by producing and crystallizing the *E. coli* IDI homologue. This structure provided a useful model of the yeast IDI protein. Moreover, although the yeast IDI protein was not crystallized, the group now has access to both MDD and IDI yeast protein samples, facilitating downstream biochemistry efforts that might require multiple members of the pathway from the same organism.

These results clearly demonstrate that in some cases it will be necessary to intelligently sample multiple experimental structures from each protein superfamily.

The structural information for MDD and IDI were amplified by homology modeling, using methods pioneered by Sali and colleagues (16, 17). The paradigm of structural genomics depends on this crucial bioinformatic approach. A key criticism of this step of the process is that it is essential to have reliable measures of structure quality to evaluate the accuracy of the predicted structures. Useful structure quality measures have been developed by Sali (16, 17) and others. However, there are still no universally accepted standards. In the cases of MDD and IDI, modeling was supported by experimental x-ray crystal and NMR structures of other members of the corresponding protein superfamilies. Structural models with good "structure quality scores" were generated for 379 proteins, spanning a substantial fraction of both superfamilies.

Particularly instructive was the modeling of the GHMP kinase superfamily based on the MDD and homologous *Methanococcus jannaschii* homoserine kinase (HSK) x-ray crystal structures. These two x-ray crystal structures allow generation of good quality models for 181 proteins, belonging to 2 of 19 discrete GHMP subfamilies. To provide models for most of the remaining members of this superfamily, an additional 17 or more experimental x-ray or NMR structures will be required. These results clearly demonstrate that in some cases it will be necessary to intelligently sample multiple experimental structures from each protein superfamily. On the other hand, reliable models could be generated for two other enzymes of the same yeast sterol/isoprenoid biosynthesis pathway, mevalonate kinase and phosphomevalonate kinase, which work together with MDD in converting mevalonate to isopentyl diphosphate. Remarkably, these three enzymes share common folds, exhibit similar surface properties, and catalyze phosphorylation of similar substrates, suggesting that they have evolved from a common ancestor to function cooperatively in this biosynthetic pathway.

Although still in its early days, structural genomics is a promising approach to one of the great challenges of modern biology—the protein folding problem. Rather than predicting small protein structures *de novo*, it is already possible to model tens of thousands of protein structures from high-quality experimental structures. The expanding database of high-resolution experimental protein structures, combined with improved methods for accurate homology modeling and/or rapid experimental data analysis using these structures as starting points, provides a *bona fide* avenue for generating reliable 3D structural information on a genomic scale.

Although this vision is very exciting, it by no means addresses some of the more challenging systems for which structural information is required. Large swaths of the structural landscape are inaccessible to the rapid data collection and analysis strategies currently being developed for structural genomics. For example, these opportunistic methods cannot yet be applied to the very important classes of integral membrane proteins or in structure analysis of large macromolecular complexes that require dedicated research efforts to reconstitute. Even some "simple" proteins will not be tractable to "high throughput" crystallization or NMR screening methods, and can only be addressed by specific research programs focused on those particular systems. However, for the broad class of relatively small soluble proteins that can be produced efficiently and that are tractable to rapid x-ray crystallography or NMR, the kinds of approaches outlined and demonstrated by Burley and coworkers (11, 15) are sure to have a significant impact by rapidly expanding the value of genomic sequence data and connecting it with biochemical and biophysical aspects of protein function.

G.T.M. is Director of the National Institutes of Health-funded Northeast Structural Genomics Corsortium (P50-GM62413).

1. Hendrickson, W. (1991) *Science* **254,** 51–58.
2. Terwilliger, T. C. & Berendzen, J. (1999) *Acta Crystallogr. D* **55,** 849–861.
3. Abola, E., Kuhn, P., Earnest, T. & Stephens, R. C. (2000) *Nat. Struct. Biol.* **7,** 973–977.
4. Moseley, H. N. B. & Montelione, G. T. (1999) *Curr. Opin. Struct. Biol.* **9,** 635–642.
5. Prestegard, J. H., Valafar, J., Glushka, J. & Tian, F. (2001) *Biochemistry* **40,** 8677–8685.
6. Murzin, A. G., Brenner, S. G., Hubbard, T. & Chothia, C. (1995) *J. Mol. Biol.* **247,** 536–540.
7. Holm, L. & Sander, C. (1996) *Science* **273,** 595–602.
8. Gerstein, M. (1997) *J. Mol. Biol.* **274,** 562–576.
9. Brenner, S. E. (2000) *Nat. Struct. Biol.* **7,** 967–969.
10. Vitkup, D., Melamud, E., Moult, J. & Sander, C. (2001) *Nat. Struct. Biol.* **8,** 559–566.
11. Bonanno, J. B., Edo, C., Eswar, N., Pieper, U., Romanowski, M. J., Ilyin, V., Gerchman, S. E., Kycia, H., Studier, F. W., Sali, A. & Burley, S. K. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 12896–12901.
12. Terwilliger, T. C. (2000) *Nat. Struct. Biol.* **7,** 935–939.
13. Heinemann, U. (2000) *Nat. Struct. Biol.* **7,** 940–942.
14. Yokoyama, S., Hirota, H., Kigawa, T., Yabuki, T., Shirouzu, M., Teerada, T., Ito, Y., Matsuo, Y., Kuroda, Y., Nishimura, Y., *et al.* (2000) *Nat. Struct. Biol.* **7,** 943–945.
15. Burley, S. K., Almo, S. C., Bonnano, J. B., Capel, M., Chance, M. R., Gaasterland, T., Lin, D., Sali, A., Studier, F. W. & Swaminathan, S. (1999) *Nat. Genet.* **23,** 151–157.
16. Sali, A. & Blundell, T. L. (1993) *J. Mol. Biol.* **234,** 779–815.
17. Sanchez, R. & Sali, A. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 13597–13602.