

an integrated aspect of multidisciplinary team research. Moreover, to address fundamental problems in medicine and biology, there is no one tool that can provide all the answers. Thus, in addition to developing specific technologies, a second set of programs is envisioned, focusing on bringing together topic- or disease-based expertise with multiple structural methodologies to work on a key problem in medicine and biology. For these teams, an emphasis would also be placed on developing approaches to foster integration of the structural expertise and technologists with functional analyses of the system under investigation. The objective here is to demonstrate the power of integration by achieving successes in specific areas of high biomedical relevance. This aspect of the PSI would lead naturally to partner-

ships with disease-based institutes at NIH.

How would this modified PSI operate? The key feature is that rather than developing a small number of very large centers, the new PSI strategy would involve a series of tightly coordinated smaller groups, each with a specific focus that feeds into the overall technology development plan. These delocalized and smaller-scale teams would be built around institutional and regional groups of investigators who can interact readily on a daily basis. The smaller size and larger number of research teams will provide for a much greater degree of flexibility and adaptability. In this scenario, there has to be a very strong central organizational structure responsible for maintaining a high degree of coordination of the

research efforts and facile exchange between the various teams. This group could be housed at one of the existing PSI sites or on the NIH campus.

The NIGMS PSI is evolving and the time is ripe to make key modifications in response to new knowledge and the changing environment. Generating a consensus across the structural biology community and focusing on grand structural challenges will enhance the PSI and result in high impact on the biomedical research enterprise.

#### ACKNOWLEDGMENTS

I thank many structural biology colleagues for valuable discussions, in particular, Jim Berger, Juliette Lecomte, Chuck Sanders, Bill Weis, and Erik Zuiderweg.

---

## Comparative Modeling in Structural Genomics

John Moult<sup>1,\*</sup>

<sup>1</sup>Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute, Rockville, MD 20850, USA

\*Correspondence: [moult@umbi.umd.edu](mailto:moult@umbi.umd.edu)

DOI 10.1016/j.str.2007.12.001

The genome sequencing revolution has resulted in a dramatic increase in the demand for structural information, a demand that traditional structural biology is totally unable to meet. The chance that a sequenced protein domain will have an experimental structure is already near 0.1% and that ratio is falling fast. All over the world, molecular and cell biologists are staring in frustration at the sequences of their proteins of interest, wishing they had structure. In recognition of this, a number of structural genomics projects, the NIH Protein Structure Initiative (PSI) in particular, have set out to maximize the number of protein domains for which structural information is available. Inherent in the strategy is leverage of experimental information through comparative modeling of related structures.

There is no doubt that for some purposes an experimental structure, or indeed a whole series of structures, is highly

desirable in order to obtain the necessary understanding of the system of interest. For example, if the goal is structure-based drug design, an experimental high-resolution crystal structure of the protein and relevant complexes is almost essential (although even for this most demanding application, there are a number of reports of drug design on models, for example, [Becker et al., 2006](#)). But such applications are not in fact typical, and the output of the PSI and other structural genomics projects is aimed at a much broader set of possible uses. To assess the utility of its strategy, we must ask how useful the models are to the full range of relevant biologists.

The lowest resolution models, typically produced by remote fold relationship recognition and not refined, have many errors, but are nevertheless usually just fine for such applications as recognizing approximate domain boundaries (often

critical for successful expression; [Tress et al., 2007](#)), assigning approximate function (by identifying members of a known superfamily), or selection of epitopes for vaccine development ([Nassal et al., 2007](#)). Medium resolution models, typically at the limit of detection of a structural relationship using PSI-BLAST, and unrefined, may be used for a range of additional purposes, such as detecting likely sites of protein-protein interactions ([Krasley et al., 2006](#)), identifying the approximate role of disease-associated substitutions ([Ye et al., 2006](#)), or assessing the likely role of alternative splicing in protein function ([Wang et al., 2005](#)). Higher resolution models, derived from relationships with better than about 30% sequence identity or refined from lower resolution starting models, add such uses as molecule replacement in solving a crystal structure ([Qian et al., 2007](#)), providing a detailed interpretation of the impact of

disease mutations and nonsynonymous population SNPs on protein function (Yue and Mout, 2006), identifying orthologous functional relationships, and in favorable cases, assigning detailed aspects of molecular function (Murray and Honig, 2002). (Note that the latter application is often not possible from an experimental structure.) All three classes of model are used to provide a more detailed structure for molecular complexes obtained by cryo-electron microscopy and other lower resolution techniques (Chiu et al., 2002). While there is no question that an experimental structure is always the ideal, these and other applications addressable with models span many of the needs of most biologists.

Because of the strong correspondence between the level of sequence relationship of two proteins and the similarity of their structures, it is often possible to produce a useful model with very simple procedures—mapping the sequence of the protein of interest onto the template provided by a sequence relative. As outlined above, depending on the level of sequence identity to a known structure, these models are often already very useful. They do have limitations both in accuracy and in the fraction of the structure included, and these limitations increase markedly with decreasing sequence relatedness. For this reason, much effort has been devoted to developing more sophisticated methods of comparative modeling, with the goal of approaching as close to the correct structure as possible. The Critical Assessment of Structure Prediction (CASP) experiments have been measuring the effectiveness of modeling methods every two years since 1994. While some of the results from the 1994 experiment were not impressive, the accuracy and power of comparative modeling techniques has improved beyond recognition in the intervening years (Kryshtafovych et al., 2007; Mout, 2005).

The fraction of protein domains for which at least a low resolution model can be obtained has increased greatly, partly because of the large amount sequence information now available for building profiles and the greatly enhanced availability of representative experimental structures (lately mostly generated by structural genomics activities), but also because new computational methods have greatly increased sensitivity. In the most

recent CASP, topologies were recognized for all but two of the target domains with known folds (Kopp et al., 2007). As well, the quality of models at all resolutions has increased dramatically. Three factors contribute to this increase. First, fidelity of aligning the sequence of interest onto a template, a major factor in limiting overall accuracy, has improved steadily, and in CASP7, 60% of template-based best models had more residues correctly aligned than could be deduced from a single best template (Kryshtafovych et al., 2007), implying near perfect alignment. Second, all target residues are not usually represented in a single available template. Modeling of these missing regions, using both alternative templates and template-free methods, has improved steadily over the CASPs, contributing most of the additional aligned residues in the best models. Third, in regions that are covered by a template, the backbone conformation of target and template will not be identical. Recently, methods of refining starting models so they converge toward the experimental structure have begun to be impressively effective (see Figure 5 in Kryshtafovych et al., 2007 for some examples). Neither problem is completely solved, but as a result of these advances, the fraction of CASP targets containing nontrivial added information beyond that provided by a single template has increased from an already impressive 65% in 2002's CASP5 to almost 80% in 2006's CASP7 (Kryshtafovych et al., 2007). Further, a recent report (Qian et al., 2007) demonstrates that refinement methods are now able to improve the accuracy of structures derived from NMR. One should not exaggerate the power of the current methods—the results quoted above are for the best models submitted to CASP. Nevertheless, the impact on model quality has already been very large, and there is every sign the recent improvements will continue. Until recently a legitimate criticism of model usefulness was that no estimate of accuracy is available, and thus a user does not know what to trust. In the most recent CASP experiment this critical aspect of modeling was an area of focus. It transpired that the best methods are already usefully effective at assigning detailed accuracy (Cozzetto et al., 2007), although there is certainly a great deal of room for improvement. Still, here again there is every rea-

son to expect substantial further short-term progress.

In summary, the PSI strategy of obtaining experimental structures that maximize the coverage of protein space in terms of the quality and quantity of models is critical to addressing the structural needs of the vast majority of molecular and cell biologists. The models often do provide the level of accuracy and detail required to address the scientific questions of interest, and modeling methods have improved beyond recognition in the last 15 years and are likely to continue to do so.

#### REFERENCES

- Becker, O.M., Dhanoa, D.S., Marantz, Y., Chen, D., Shacham, S., Cheruku, S., Heifetz, A., Mohanty, P., Fichman, M., Sharadendu, A., et al. (2006). An integrated in silico 3D model-driven discovery of a novel, potent, and selective amidosulfonamide 5-HT<sub>1A</sub> agonist (PRX-00023) for the treatment of anxiety and depression. *J. Med. Chem.* 49, 3116–3135.
- Chiu, W., Baker, M.L., Jiang, W., and Zhou, Z.H. (2002). Deriving folds of macromolecular complexes through electron cryomicroscopy and bioinformatics approaches. *Curr. Opin. Struct. Biol.* 12, 263–269.
- Cozzetto, D., Kryshtafovych, A., Ceriani, M., and Tramontano, A. (2007). Assessment of predictions in the model quality assessment category. *Proteins* 69, 175–183.
- Kopp, J., Bordoli, L., Battey, J.N., Kiefer, F., and Schwede, T. (2007). Assessment of CASP7 predictions for template-based modeling targets. *Proteins* 69, 38–56.
- Krasley, E., Cooper, K.F., Mallory, M.J., Dunbrack, R., and Strich, R. (2006). Regulation of the oxidative stress response through Sit2p-dependent destruction of cyclin C in *Saccharomyces cerevisiae*. *Genetics* 172, 1477–1486.
- Kryshtafovych, A., Fidelis, K., and Mout, J. (2007). Progress from CASP6 to CASP7. *Proteins* 69, 194–207.
- Mout, J. (2005). A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr. Opin. Struct. Biol.* 15, 285–289.
- Murray, D., and Honig, B. (2002). Electrostatic control of the membrane targeting of C2 domains. *Mol. Cell* 9, 145–154.
- Nassal, M., Leifer, I., Wingert, I., Dallmeier, K., Prinz, S., and Vorreiter, J. (2007). A structural model for duck hepatitis B virus core protein derived by extensive mutagenesis. *J. Virol.* 81, 13218–13229.
- Qian, B., Raman, S., Das, R., Bradley, P., McCoy, A.J., Read, R.J., and Baker, D. (2007). High-resolution structure prediction and the crystallographic phase problem. *Nature* 450, 259–264.
- Tress, M., Cheng, J., Baldi, P., Joo, K., Lee, J., Seo, J.H., Lee, J., Baker, D., Chivian, D., Kim, D., et al. (2007). Assessment of predictions submitted for

the CASP7 domain prediction category. *Proteins* 69, 137–151.

Wang, P., Yan, B., Guo, J.T., Hicks, C., and Xu, Y. (2005). Structural genomics analysis of alternative splicing and application to isoform structure

modeling. *Proc. Natl. Acad. Sci. USA* 102, 18920–18925.

Ye, Y., Li, Z., and Godzik, A. (2006). Modeling and analyzing three-dimensional structures of human

disease proteins. *Pac. Symp. Biocomput.* 2006, 439–450.

Yue, P., and Moutl, J. (2006). Identification and analysis of deleterious human SNPs. *J. Mol. Biol.* 356, 1263–1274.

## Harnessing Knowledge from Structural Genomics

Helen M. Berman<sup>1,\*</sup>

<sup>1</sup>Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA

\*Correspondence: [berman@rcsb.rutgers.edu](mailto:berman@rcsb.rutgers.edu)

DOI 10.1016/j.str.2007.12.003

As the central repository for all macromolecular structures, the Protein Data Bank (PDB) started collaborating with the worldwide structural genomics projects from their inception (Berman et al., 2000, 2003). From the beginning, it was clear that structural genomics, including the U.S.-funded Protein Structure Initiative (PSI), would change the ways in which we think about publishing and data sharing. As time has gone on it is becoming clear that these efforts will make a significant impact on how we do structural biology. By creating an appropriate infrastructure in the form of a Knowledgebase, the fruits of the PSI effort can enable a new kind of biology.

Since 1989, it has become the norm to submit coordinates as a condition for publishing articles describing structure determinations (International Union of Crystallography, 1989). For PSI projects, it has been mandatory to deposit and release the coordinate and structure factor data within one month of completing a structure, prior to any journal publication. The impact of this policy raises some interesting questions. Would this mean that PSI research could be no longer published in standard journals? How would journal publication practices change? Two things have emerged so far. First, a PDB entry can itself be thought of as a publication. The PDB now assigns a digital object identifier (DOI) to every structure, and these are beginning to appear as references in published articles. Second, more than 600 papers describing the results of structure determinations have been authored

at PSI centers—subsequent to data release—and many more are in the pipeline. Whether or not this will become a trend for non-PSI structures, which are typically released after journal publication, remains to be seen.

An important aspect of the charter of the PSI is the suggestion of a new paradigm for the information sharing in support of the advancement of science. In addition to sharing the results of structure determinations, the PSI projects provide the sequence as well as information about the status of each target under investigation. It is very unusual in conventional structural biology for these types of data to be made public in advance of publication of the structure. TargetDB (<http://targetdb.pdb.org>) tracks status indicators for each step in the structure determination pipeline (Chen et al., 2004). Along with information such as protocols for protein production, PepcDB (<http://pepcdb.pdb.org>) provides the reasons why work on a particular target has stopped (Kouranov et al., 2006). The information in these resources provides methods to facilitate experimental design, not only for the PSI projects but also for the biological community at large. Data sharing that includes the disclosure of sequences, tracking, and protocol details in advance of publication or deposition into the PDB and the early release of coordinate and experimental data is far ahead of current practices in structural biology. This represents a significant leap from where we were 25 years ago, when some investigators worried about making their coordinates

available to the rest of the research community!

A review of the progress of the PSI since it began in 2001 demonstrates that it has been tremendously successful in achieving the initial goals of selecting, producing, and determining the structures of many novel proteins in a high throughput manner. More than 2700 structures have been determined; most remarkably, about half of these have been determined in the two years since the second phase, PSI-2, began. Of the structures determined, more than 68% are novel, meaning they have less than a 30% sequence identity with those in the PDB. Our understanding of structure space has been transformed in that the conservation of overall polypeptide chain folds is greater than had been anticipated. With the clever targeting of structures for analysis, the coverage of sequence space that can now be modeled is ever-increasing.

In June 2007, I was selected to lead the development of the PSI Structural Genomics Knowledgebase (PSI\_SGKB). The idea was to make the products of the PSI widely available to the broader community of biologists. Although I was very aware of the success of the initiative with respect to the production of many structures, I needed to investigate the full scope of activities of the PSI centers before accepting this new challenge. In reviewing all of the PSI center progress reports and websites, I discovered a treasure trove. Indeed, the PSI projects have done more than simply determine