Impact of Structures from the Protein Structure Initiative

Wayne A. Hendrickson1,2,*
1Howard Hughes Medical Institute
2Department of Biochemistry and Molecular Biophysics
Columbia University, New York, NY 10032, USA
*Correspondence: wayne@convex.hhmi.columbia.edu
DOI 10.1016/j.str.2007.11.006

The collective efforts of the Protein Structure Initiative (PSI) have so far produced over 2700 protein structures (http://kb.psi-structuralgenomics.org/KB/, Metrics Module). This is an impressive statistic, but what of interest have we learned from these structures and what might be the future impact of this knowledge? In the case of conventional structural biology, most structural investigations are undertaken in the context of substantial prior knowledge about the subject molecules; typically, the structural studies are accompanied by functional analyses inspired by structural results, and the findings are published in papers (although months or years after structures are solved). Here in the case of PSI, and most notably so for the production efforts of large-scale centers of PSI-2, many of the subject molecules are previously uncharacterized and oftentimes they are not even annotated by homology. Moreover, apart from collaborative efforts supported with other funds, functional analyses are outside the scope of PSI. Publication as a PDB deposition is immediate, but papers appear for only a fraction of the PSI structures. Thereby, outcomes from this structural genomics effort differ qualitatively from those from conventional structural biology.

Although the majority of PSI structures are deposited without an accompanying publication, the PSI has nevertheless produced approximately 500 structural publications and a comparable number of methodological publications (http://olenka.med.virginia.edu/psi/). The structure papers are dominated by straightforward descriptions of the structures in relation to relevant functional information from the literature, but a few also incorporate data from functional experiments done in a solicited collaboration (e.g., Mougous et al., 2006; Forouhar et al., 2007) or with other support to a PSI laboratory (e.g., Yu et al., 2006). While there are some abbreviated structure notes, the bulk of PSI structure publications are full-fledged papers. Structures of unknown function typically have only PDB deposits and no accompanying PSI publication, but such structures are now being taken up for study in the scientific community (see below).

There are numerous examples of PSI results that compare with the output from conventional structural biology, and this is most evident from work on biological themes and community-nominated targets. But the main premise of the PSI venture, the very mission of PSI-2, is the broader goal of providing a pan-genomic picture of protein structures, using homology modeling to flesh out sequence families. The explicit focus here is on novelty and the work of large-scale centers is on those families for which a representative structure is not already available in the PDB, thus many PSI-2 structures are from families that have not heretofore been subjects of intense biological investigation. Such PSI structures therefore necessarily have greater value prospectively than in explanation of known biology, as in conventional structural biology. The aim here is to facilitate biological discovery rather than only to follow it.

Despite the focus on prospective uses of structure, analogous to prospective uses of gene sequences, the value of PSI structures is oftentimes already evident from structural similarities that are cryptic in sequences, from structural features evocative of active sites, or from the presence of adventitious ligands (37% of PSI structures contain a ligand). Such analyses provide for structure-based assignments of likely function (e.g., Jin et al., 2006) or corrections of mistaken annotations (e.g., Keller et al., 2002). Indeed, the very existence of a protein structure minimally moves the commonplace annotation of “hypothetical protein” to “protein of unknown function”. Moreover, irrespective of function, the very substantial contribution of PSI structures to the universe of novel structures (>40% of novel structures deposited into the PDB in Year 2 of PSI-2) is having an important impact on understanding of the protein universe. Besides its intrinsic value, this information also has utility in bioinformatic identification of remote relationships.

The ultimate aim for PSI structural information is to provide a comprehensive and unencumbered basis for future biological investigations. The fruits of PSI are just beginning to be exploited in this manner, and a recently reported study on a protein of unknown function from Thermotoga maritima (Tm0936) is perhaps a poster child for this approach. Hermann et al. (2007) studied this putative enzyme by first docking hypothetical transition states of predicted substrates and then testing the most favorable substrate candidates in enzymatic assays and crystal structures. Although a PSI participant performed the confirmatory structure determination and a PSI lab provided a clone to facilitate the work, this investigation was initiated as a completely
independent exploitation of PSI results. It is notable, however, that the relevant PSI structures were deposited in July 2002 and April 2003, but it took until August 2007 for the resulting publication to appear. Clearly, it will take some time for the vast reservoir of opportunity in the PSI Knowledge-base to be fully realized.

As for sequence genomics, there is a special value in structural completeness. This cannot be managed in the same way for structures as for genome sequences, of course, but we do aspire in the pan-genomic setting of PSI to obtain representative structures for a substantial portion of all sequences from all of life. All of life, as we know it from sequences, is a rapidly expanding entity with the recent flood of metagenomic data; this can complicate the PSI effort, but it also presents exciting opportunities and new structural results relevant to human health are now emerging from nascent PSI-2 efforts on the human gut microbiome. Despite the challenge, substantial inroads are also being made for specific organisms. A recent paper describes the progress in determinations of structures from Mycobacterium tuberculosis (Baker, 2007), including a dominant fraction from the TB Structural Genomics Consortium started with PSI-1. Similarly, led by the concerted effort of another PSI-1 center, there has been substantial movement toward completeness of structures from Thermotoga maritima, and the results provide novel insight into metabolic pathways and networks whereby, for example, a new structural solution emerged for an old function and this provided a drug discovery opportunity (Agrawal et al., 2004).

Beyond the overarching mission to make atomic-level structural information easily obtainable for most proteins from their corresponding DNA sequences, the PSI also has a significant component from biological theme projects within the various centers and from community-nominated targets. These projects take advantage of structural genomics approaches and high-throughput infrastructure in the centers, but many of these projects have outcomes like those from conventional structural biology. In addition, in the course of developing approaches for addressing challenging targets such as membrane proteins and eukaryotic proteins, PSI-2 specialized centers are also now contributing structures of substantial significance.

Finally, the structures of PSI have brought with them advances in the methods for protein production and for structural analysis. These technological developments are an indirect, but nevertheless substantial, benefit from PSI structures. Enhanced technology improves efficiency for PSI, reducing the cost per structure, but much of this technology is also broadly applicable to studies in individual laboratories. Many of the advances are improvements on ideas that pre-date PSI (e.g., reductive methylation for increased crystallization probability), but the scale of PSI tests places conclusions about efficacy on a statistically sound basis freed from the tyranny of anecdotal accounts. Other enabling technologies (e.g., autoinduction for protein expression) are directly attributable to PSI even though they are already so engrained in general practice that PSI origins may be forgotten.

In sum, PSI structures have a multifaceted impact that is manifest at present; but, and perhaps more importantly, the comprehensive resource that is being compiled also has a latent impact for future realization.

ACKNOWLEDGMENTS

I thank other participants in the Protein Structure Initiative for their contributions to the effort and to the material for this commentary. My participation is supported in part by NIH Cooperative Agreement U54-GM075026 for the New York Consortium on Membrane Protein Structure (NYCOMPS).

REFERENCES


