nature structural & molecular biology

PSI—phase 1 and beyond

Ten years ago, the growing impact of protein structure on biomedical research and significant advances in genome sequencing ushered in the new field of structural genomics. In 2000, following years of investments in structural biology through individual R01 grants, the National Institute of General Medical Sciences (NIGMS) expanded its commitment to the field by establishing the Protein Structure Initiative (PSI) (http://www.nigms.nih.gov/psi). As the first phase of the PSI nears an end, we examine the progress and challenges faced by the initiative.

The ultimate goal of the PSI is to make feasible the prediction of accurate three-dimensional structures of proteins based on their sequences alone. This will require a database of structures with at least one high-resolution experimental structure for each protein family—defined as a group of proteins sharing >30% sequence identity. To populate the so-called 'fold space,' it is necessary to determine these structures as quickly as possible. Thus, the goal for the first five years of the PSI, the pilot phase (2000–2004), is to establish the high-throughput technologies and the pipeline for structural determination in the second 'production phase.'

Whether the pilot phase has achieved its goal depends on how one measures success. Qualitatively, one can examine the technological developments and the ability to overcome bottlenecks in protein production and crystallization. Quantitatively, one can ask how many unique structures have resulted from this significant investment of tax dollars. The pilot phase focused on development of tools and streamlining of processes rather than structure determination. Nevertheless, from 2001 to 2003, ~420 structures were solved. During the same time period, the average cost of each structure decreased from $650,000 to ~$240,000. The numbers are promising.

The impact of these 420 structures on biomedical research is difficult to assess. But even in its pilot phase, the PSI is, by percentage, contributing more new folds to the Protein Data Bank (PDB) than are individual investigators. Specifically, in 2002–2003, 70% of the structures determined by the PSI were for proteins with unique sequences. This is in comparison to only 10% of all structures deposited in the PDB during the same period of time. Overall, 12% of the PSI structures identified a new fold, as compared with only 3% from other sources. It's evident that the PSI structural data is less redundant than that of the PDB. However, the PSI structures are dominated by structures of single domains, primarily from prokaryotic proteins. Production and crystallization of eukaryotic proteins have proven difficult for various reasons, including their requirement for post-translational modification or protein partners. Since one goal of the PSI is high throughput, proteins that 'misbehave' during any phase of expression and structure determination are set aside for future trials with improved technologies. There is also the expectation that other researchers, independent of the PSI, will fill in the gaps.

Such researchers would benefit from databases that are managed by the PDB for the PSI. For example, the complete listing of PSI targets, along with weekly progress updates, is collected in TargetDB (http://targetdb.pdb.org). This database was established to facilitate coordination of efforts between the nine PSI centers. TargetDB has grown into a valuable resource for scientists pursuing nonstructural research programs, as the site is also accessed by many outside the PSI centers. Furthermore, a new database that includes all expression, purification and crystallization trials (including negative results) for all PSI targets is under construction. This database will be a welcome addition for scientists, especially biochemists and molecular biologists, wishing to expand on the structural work of the PSI.

In phase 2, the nine centers currently funded by PSI and any new ones will compete for a second round of PSI grants. Target selection will continue to focus on providing structural coverage for all protein families but will be under tighter control to prevent overlap and duplication. Overall, the production phase is expected to increase the number of unique structures in the PDB by ~6,000–8,000, with each center producing over 200 structures per year at a cost of ~$50,000 per structure. This effort is projected to culminate in a 40% increase in structural information of sequenced genes, an important achievement if it can be attained.

To ensure that phase 2 and the PSI as a whole stay on course, it is essential that NIGMS carefully evaluate the progress of the initiative versus its impact, and this will occur in the last year of each phase. From preliminary evaluations of the pilot phase, it is evident that the ambitious goal of phase 2 will require significant improvements in the current strategy of high-throughput structure determination. For example, it is still unclear how bottlenecks for eukaryotic and membrane protein structure determination will be overcome. The PSI structural data would clearly be most useful if one could integrate the structures of individual domains into full-length proteins and ultimately into relevant complexes. Perhaps other NIGMS and/or NIH initiatives will tackle these unresolved issues (for example, see http://nihroadmap.nih.gov) but this is uncertain, at least in the current funding climate.

Given that we agree that population of fold space and the ability to predict three-dimensional protein structures are likely to generate testable hypotheses about functions, it is somewhat surprising that the initiative does not provide the PSI centers with funds to study function. However, scientists with existing R01 grants can apply for an administrative supplement to determine the function of a structure determined by a PSI center (http://www.nigms.nih.gov/funding/psi_supplements.html). Therefore, the understanding of molecular mechanisms remains in the hands of individual investigators, and as always we hope that you'll send the best of these mechanistic studies to *Nature Structural & Molecular Biology*. ■