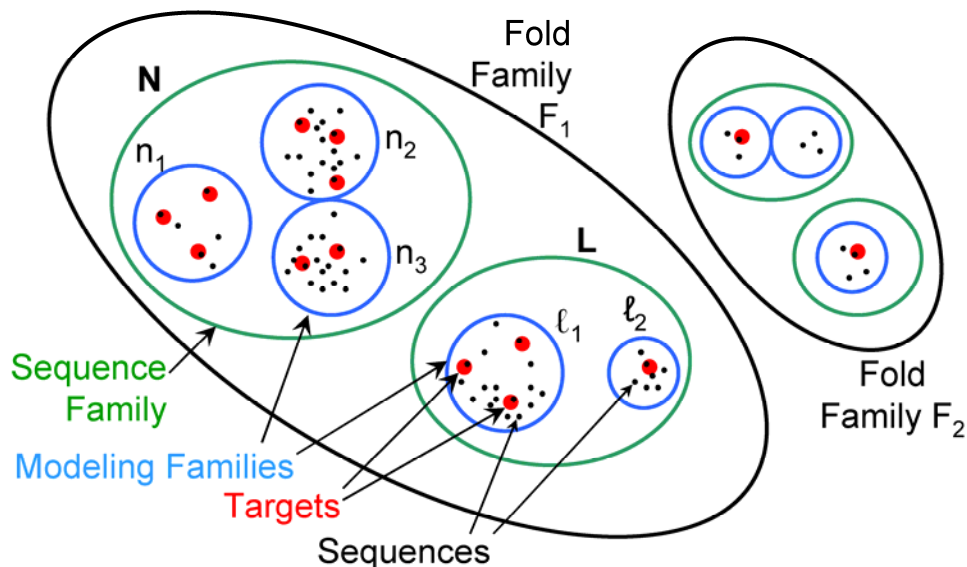


## Target Selection Strategy for the Large-Scale Production Centers of the Protein Structure Initiative

### Background

The overarching goal of the Protein Structure Initiative (PSI) is to make three-dimensional atomic-level structures of most proteins easily obtainable from knowledge of their corresponding DNA sequences. This objective can be realized by selection and successful structural characterization of proteins chosen to be representative of large sequence families, and subsequently leveraging of such structures across sequence families by homology modeling combined with quantitative model assessment. The limits of homology modeling define the boundaries of “modeling families” around groups of targets and guide clustering of large sequence families. The approach to clustering sequences of protein domains into sequence and modeling families and determining one (or sometimes a few) experimental structures from each modeling family is illustrated schematically below.



In some cases, experimental structures emanating from the PSI represent the first examples of particular polypeptide chain topologies (or folds) (e.g., fold families  $F_1$  and  $F_2$ ). More generally, however, a newly determined structure from a given protein domain sequence family (N) is similar to one or more previously determined structures from another protein domain sequence family (L). Such observations represent the norm, because of the way in which evolution re-utilizes protein domain folds to support both similar and diverse biochemical/biological functions. Recognition of unanticipated structure-structure relationships among seemingly diverse protein sequence families is invaluable and can provide significant new insights into biological function. Structural characterization of these sequence families may also serve to improve our understanding of and ability to detect distant evolutionary relationships between proteins that cannot be recognized with currently available bioinformatics tools that rely solely on gene sequence information.

Notwithstanding claims to the contrary from some quarters, the number of distinct sequence families of protein domains is not increasing in proportion to the ever growing number of available gene sequences. Our understanding of protein evolution argues strongly that the vast majority of protein domain sequences found in nature belong to a few thousand distinct protein fold families and that, as our understanding of protein sequence/structure relationships improves, the apparent number of distinct sequence families corresponding to protein domains should decline as we re-sort and remap the protein sequence universe based on our knowledge of three-dimensional structure. Ultimately, we would expect that most protein domain sequences found in nature can be attributed to one or other protein domain fold family, which would encompass a multitude of subfamilies reflecting both sequence and structural diversity therein.

This relative simplicity is not mirrored by the way in which discrete protein domains occur within the polypeptide chain products of genes. Many of the proteins found in nature, particularly those found in eukaryotes, are composed of multiple protein domains with distinct folds or three-dimensional structures, each typically supporting distinct biochemical/biological functions. Detailed examination of the ever growing number of available gene sequences has revealed staggering, seemingly unbounded, complexity in terms of the number and diversity of domain combinations. The central goal of the PSI aims to help make sense of this complexity by providing either an experimental structure or a homology model for the globular regions of most polypeptide chains found in nature. Such information should greatly facilitate our understanding of biological processes by making structural sense of the complexity afforded by protein evolution.

## **PSI-2 Structure Determination Target Selection and Productivity**

During the second phase of the PSI (PSI-2; July 1<sup>st</sup> 2005—June 30<sup>th</sup> 2010), the four Large-Scale Production Centers are required to commit 70% of their effort on centrally chosen PSI 'Network' targets, 15% effort on targets nominated by the broader research community, and 15% effort on the specific biomedical themes chosen independently by each Center.

**Network Targets:** During the first pilot phase of the PSI (PSI-1; September 1<sup>st</sup> 2000—June 30<sup>th</sup> 2005), structurally uncharacterized protein sequence families, typically containing tens to hundreds of members, were targeted for structure determination. All targets were disclosed publicly, and overlaps arising among PSI-1 centers were dealt with informally. With the transition to PSI-2, a formal, centralized process was introduced for selecting the so-called PSI Network Targets and apportioning them to the four Large-Scale Production Centers. This process was used to identify targets to further the principal goal of the PSI. It also served to reduce any overlap in targets among the four centers. Network Targets were selected following similar principles adopted during PSI-1. With the rapid growth in protein sequence information coming from myriad genome sequencing programs, the purview of PSI-2 Network Targets was expanded to include very large sequence families that exhibit enormous phylogenetic/functional variation yet have only limited structural coverage (insufficient to

support homology modeling across most family members). Of late, large-scale sequencing efforts have gone beyond individual organisms to examine large communities of microbes present in different environmental niches, such as the human gut or in soils, oceans, or acid-mine drainages. The resulting metagenomic protein sequences have come to represent another important source of structure targets for the PSI-2.

Most of the 84,479 Network Targets currently pursued by the four centers come from families that encompass tens to thousands to tens-of-thousands of sequences. These domain families often include representatives from both prokaryotic and eukaryotic organisms. Many of these domain families are very challenging as they represent families where conventional structural biology efforts have been tried and failed. As described earlier and depicted schematically above, each PSI-2 Network Target family typically contains many homology “modeling subfamilies” of protein sequences sharing more than ~30% sequence identity. Experimental structure determination of targets selected from distinct homology modeling subfamilies will further the PSI in its principal goal of making structural information of most proteins readily available from knowledge of their corresponding gene sequences.

**Community Nominated Targets:** Each Large-Scale Production Center entertains target nominations from the greater scientific community. Within each center, nominated targets are vetted for feasibility and consistency with the overall PSI goals, and most are accepted for structure determination. Collectively, the four centers are currently pursuing structural studies of 1483 Community Nominated targets.

**Biomedical Theme Targets:** Each of the four Large-Scale Production Centers independently chose targets of known biological and biomedical significance. Biomedical Themes include targets from protein-protein interaction and metabolic networks, networks associated with human cancers and developmental biology, proteins conserved across all kingdoms of life, proteins from pathogenic microorganisms, the repertoire of human and pathogen protein phosphatases, and proteins implicated in tumor metastasis. Collectively, the four centers are currently pursuing structural studies of 13,103 Biomedical Theme targets

**Productivity Summary:** Since the inception of PSI-2, the Large-Scale Production Centers have determined nearly 1500 new protein structures of very high quality at an average cost/structure of approximately US\$ 66,000. During past year, the four centers together contributed 73% of the novel structures (defined as <30% identical in sequence to an extant structure) deposited to the Protein Data Bank (PDB; <http://www.pdb.org>) from all U.S. sources. Many of these structures come from protein families of high biological and biomedical interest. In aggregate, PSI laboratories have cloned nearly 100,000 genes encoding proteins from a significant fraction of all large protein sequence families and purified nearly 25,000 diverse proteins, creating a unique resource of reagents and associated experimental data that is freely available for the scientific community.

## PSI-2 Target Selection Plans Forward

During the second half of PSI-2, the Large-Scale Production Centers plan to retain their collective focus on broad structural coverage of proteins from large sequence families. To further optimize target selection for both homology modeling coverage and biological/biomedical impact, we will systematically prioritize additional Network Targets from two sources. First, we will select targets from the 300 largest protein superfamilies found in nature, which encompass members from all kingdoms of life. Within each of these superfamilies, we will identify the 2000 largest, experimentally tractable modeling subfamilies that have no structural representatives in the PDB. Second, we intend to select targets from the 100 largest protein superfamilies identified by metagenomics sequencing efforts. Again, within each of these superfamilies, we will attempt to identify up to 400 large, experimentally tractable, modeling subfamilies lacking structural representatives in the PDB. For all 2400 planned modeling subfamilies, priority will be given to the top “structure/function clusters” (i.e., the largest subfamilies that have functional annotations in Gene Ontology and other protein function annotation databases). We will continue to apportion these Network Target modeling subfamilies among the four centers so as to avoid target overlap. For each modeling subfamily, the responsible center will be free to use its own bioinformatics and functional annotation tools to prioritize structure determination targets therefrom.

## Community Participation

To encourage involvement of the scientific community in our ongoing target selection activities, the PSI will publicize the 400 selected large superfamilies and 2400 selected modeling subfamilies *via* the PSI Knowledgebase (PSI SGKB; <http://kb.psi-structuralgenomics.org/KB/>). We will continue to seek community input to further prioritize and annotate this evolving target list. In addition, the scientific community will be encouraged to nominate other large families, modeling families, or sets of proteins of high biological/biomedical interest for structure determination. It is the intent of the PSI to ensure that the community actively participates in the target selection process and provides both insights and experimental data to enhance annotation of the resulting structures.

PSI target lists are accessible in TargetDB (<http://targetdb.pdb.org/>) *via* keyword and sequence searches. Expression clones generated by PSI centers will be available for a nominal fee *via* the PSI SGKB website (<http://kb.psi-structuralgenomics.org/KB/>) from the PSI Materials Depository (<http://www.hip.harvard.edu/PSIMR/index.htm>) established by the Harvard Institute of Proteomics.

## Expected Outcomes/Evaluation

As we have done throughout PSI-2, a periodic (approximately twice yearly) evaluation of the impact of our target selection activities on the goals of the PSI will be conducted. In the interim, structure determination and homology modeling coverage statistics are continuously updated and made publicly available by the PSI SGKB.

The Large-Scale Production Centers are committed to achieving the principal goal of the PSI, while providing substantial advances in structural biology and insights into protein structure and function. This discovery-based approach complements traditional hypothesis-driven research and represents a potentially rich source of new ideas and concepts. We believe that the output of the PSI will help generate a plethora of new hypotheses and enable future generations of scientists to make advances in our understanding of protein function and evolution.

It cannot be over emphasized that the goal of the PSI never was to “fill in” unexplored regions of protein sequence/structure space without reference to main stream biological and biomedical research. Achieving a better understanding of the relationships between protein sequence and structure represents a critically important challenge. We are confident that this goal can be accomplished in the context of a systematic, cost-effective structure determination effort that not only emphasizes biologically and biomedically important protein families, but also explores fundamental sequence, function, and structure diversity of protein families, informs our understanding of molecular evolution, and provides exciting new opportunities for future innovations and progress in the scientific endeavor.

### Target Selection Related Information:

- [PSI Recommendations on Target Selection for PSI-2 Large-scale Research Centers](#) from NIGMS.
- Protein Structure Initiative target selection is coordinated by the Intercenter Bioinformatics team (<http://psi-big4.org/>).
- [Current Target Selection Strategy for the Large-scale Production Centers of the Protein Structure Initiative.](#)