

Current Target Selection Strategy for the Large-scale Production Centers of the Protein Structure Initiative

March 31, 2008

Background: The overarching goal of the Protein Structure Initiative is to make 3D atomic-level structures of most proteins readily obtainable from knowledge of their corresponding DNA sequences. This goal can be realized by appropriate selection and successful structural analysis of proteins representative of large sequence families and leveraging of these structures across these protein sequence families by homology modeling and model assessment. Target selection is an ongoing process that requires continual reassessment and reevaluation. In the second half of the PSI-2, we will continue our concerted approach to cover protein sequence space with experimental 3D structures and models by selecting primarily targets from large protein sequence families that are structurally uncharacterized or inadequately covered, in addition to our 30% effort on proteins nominated by the research community and our internally-selected targets of biomedical relevance.

Very large “MEGA-families”, with enormous phylogenetic and functional variation but with limited structural coverage, serve as a major source of targets. Structural representatives from subfamilies within these MEGA superfamilies will provide new and extensive information about the evolution of structural and functional diversity. Similarly new “META-families” from large communities of organisms (such as the human gut, oropharynx, the genitor-urinary tract, inner ear, respiratory system and other microbiomes) with implications for human disease, or from the environment (e.g. soil, ocean, acid-mine drainage) will constitute an important source of novel targets for the remainder of PSI-2. We also intend to continue our efforts to provide structural coverage of large sequence families (Pfam or ‘BIG’ families) for which no structural representatives are known and especially for those families whose “fold” type cannot be predicted. Biomedical Theme targets and Community proposed targets will also continue to provide an important source of targets of known biological and biomedical significance. Biomedical Themes include targets from protein interaction and metabolic networks, including networks associated with human cancers, proteins conserved across all kingdoms of life, proteins from pathogenic microorganisms, the repertoire of human protein phosphatases and proteins implicated in metastasis. The four Large-scale Production Centers expend 70% of their effort on the Communal Network PSI-2 targets, 15% effort on targets nominated by the broader research community, and 15% effort on the specific Biological/Biomedical themes chosen by each Center.

The six Specialized Centers develop innovative methods, approaches, and technologies for producing and determining the structures of proteins that traditionally have been difficult to study (membrane proteins, complexes, human proteins). These centers select protein targets for their studies using approaches that are complementary to the Large-scale Production Centers to support new technology development for the most challenging protein classes.

In PSI-2, target selection activities have had very high priority, and the strategies developed over the past two plus years have been highly successful. This procedure has allowed the Large-scale Production Centers to determine nearly 1500 protein structures and increase the contribution of novel structures (defined as proteins with less than 30% identity to any PDB structure at the time of deposition into the PDB) from less than 60% to over 75% of structures submitted each year. In the past year, the Large-scale Production Centers contributed 73% of the novel structures deposited to the PDB from all laboratories in the USA. Although the main focus of the Large-scale Production Centers is on high-throughput experimental protein structure determination and methods development these Centers have comprehensively sampled the entire prokaryotic protein sequence space more broadly than ever before. The PSI has cloned nearly 100,000 genes of proteins from significant fraction of all large protein sequence families and purified nearly 25,000 highly diverse proteins creating a unique resource available for the biology community. Many of these structures represent protein families of high biological and biomedical interest, and many have provided novel functional hypotheses that impact the research programs of a broad range of investigators.

Approach: In the second half of PSI-2, we will retain this focus on broad coverage of large families of proteins as the main theme of the Large-scale Production Centers Network Activity. In order to optimize target selection and to maximize the homology modeling coverage and biological and biomedical insights that can be derived from the PSI-2 efforts, we will evaluate the largest 300 superfamilies (defined from our BIG and MEGA target groups) that contain representatives from all kingdoms of life and rank (by size) “modeling subfamilies” within these very large families that have no structural representatives in the PDB. We will also analyze the top 100 families overrepresented in microbiomes (“META” families) and rank by size their “modeling subfamilies” with no structural representative in PDB.

The subfamily sequences will be further analyzed to select specific protein targets that are more likely to lead to structure determination, using predictions of cellular localization, conformational disorder, and other bioinformatics analysis methods developed in PSI. Priority will be given to the top “structure/function clusters”; i.e. the largest subfamilies that have functional annotations in Gene Ontology (GO) and similar protein function databases, for which no representative 3D protein structure is yet available. The 2000 largest modeling subfamilies from these 300 BIG and MEGA superfamilies, and 400 largest subfamilies in the 100 META-derived superfamilies will be distributed to the large-scale production centers using a draft pick mechanism.

Each center will prioritize selected sequence subfamilies using available resources (such as protein-protein interaction and other genomic and functional databases) to assess family diversity along with perceived biological and biomedical relevance. These data will be used to refine selection of specific targets for sample production and structure determination.

Community involvement: The PSI will publicize the selected superfamilies and modeling subfamilies through the PSI Knowledgebase, and actively seek biological community input to further prioritize and annotate the list. In addition, the biological community will be encouraged to submit suggestions and supporting rationale for large families to be targeted for structure determination. In this way, the community can actively participate in the target selection process and provide experimental data to enhance annotation of the solved structures.

The final target list will be accessible in TargetDB *via* keyword and sequence search. Clones generated through the PSI will be available from the Harvard Institute of Proteomics for a nominal fee.

Production Centers Network Activity (70% of total Center effort):

75% - 300 largest MEGA/BIG superfamilies

25% - 100 largest META superfamilies

These 400 superfamilies include several thousand modeling subfamilies which do not yet have structural representatives. 2400 of these subfamilies will be targeted during March – July 2008 using the priorities outlined above:

Time table for target selection:

April 30, 2008 - 1000 MEGA/BIG subfamilies distributed (BIG4 – Christine Orengo coordinator)

June 30, 2008 - 400 META subfamilies (BIG4 – Adam Godzik coordinator)

July 31, 2008 - 1000 MEGA/BIG subfamilies distributed (BIG4 – Andras Fiser/Burkhard Rost coordinators)

Evaluation process and expected outcomes :

As in the past, we will periodically (every 6 months) evaluate how our strategy is addressing the PSI-2 main mission and goals. In this way, the Large-scale Production Centers will advance the PSI-2 and provide substantial new value and insights into structural biology and functional annotation compared to traditional R-01-funded research. This discovery-based approach is highly complementary to traditional hypothesis-driven approaches and will provide a wealth of new ideas and concepts that will result in generation of a myriad of new hypotheses on protein structure, function and evolution that can then be tested in individual investigators' laboratories supported by non-PSI funding. It is important to stress that the PSI is not simply filling in unexplored sequence and structure space, which is important in itself, but investigating the co-evolution of protein structure and function. The PSI project is discovering previously unrecognized relationships between disparate sequences and families, providing testable hypotheses regarding biological functions of proteins, and informing phylogenetic and evolutionary analyses. The definitions of Pfam families, as well as BIG, MEGA and META superfamilies, are by nature fluid. As relationships are uncovered among sequences and structures we are redefining these boundaries, reassessing the meaning of family definitions, and using this new knowledge to inform

subsequent target selection. In this effort, it is important to have a wide mix of targets, some of which have known biological activity or significance and others of currently unknown function so that we can make such connections and discover exciting new biological activities and structure/function information that define the evolution and unique characteristics of specific organisms and, hence, define self. These data will also provide advances in understanding mechanisms for acquisition of novel biological function and the process of evolving new or diverse biological systems. In this way, we can truly build on the constantly growing foundation laid by the highly successful genome and metagenome sequencing projects that are enriching and expanding our knowledge of the protein universe and, hence, enable us to explore the diversity of life forms that constitute our fragile planet.

Definitions

Pfam families are protein sequence families defined by unique hidden Markov model (HMM) profiles that are curated, verified and deposited in the Pfam database.

BIG families are new large protein sequence families defined by unique HMM profiles that are structurally uncharacterized (these include annotated families as well as sequence families with poor or no annotation in the Pfam database).

MEGA families are very large protein sequence families with enormous phylogenetic variation as defined by unique HMM based profiles.

META families are Pfam, BIG or MEGA protein sequence families that are overrepresented in large communities of organisms discovered in specific microbiome environments. META families are also defined by a unique HMM profile.

Modeling subfamilies are the collection of sequence subfamilies that comprise the Pfam, BIG, or MEGA families, including META families, whose members differ from other subfamilies by less than 30% sequence identity radius.

Structure/function clusters are modeling subfamilies with proposed molecular functions which do not have a representative 3D structure.