TARGET SELECTION FOR STRUCTURAL GENOMICS

Terry Gaasterland and Gulriz Kurban
Laboratory of Computational Genomics
The Rockefeller University, New York, New York, USA
{gaasterl,gulriz}@genomes.rockefeller.edu

The goal of structural genomics is to carry out high-throughput determination of protein structures sufficient to build reliable 3D molecular models for all proteins in fully sequenced genomes [Gaa98, Sal98]. Structural genomics intersects the fields of genome sequencing, protein cloning and expression, crystallography, nuclear magnetic resonance, and use of high flux synchrotron source X-ray beam lines as well as computational genome analysis, computational protein structure analysis, computational modeling of protein structures, methods for structure prediction, and protein structure and fold databases.

Several structural genomics pilot projects have been initiated with differing specific goals. Some seek to discover novel folds; others seek to solve structures for entire microbial genomes; others seek to solve structures for all proteins associated with specific diseases. The following pilot projects for high-throughput structure determination are underway: structural genomics of cancer-proteins (The Rockefeller University, Einstein Medical College, and Brookhaven National Laboratory); *Pyrobaculum aerophilum* (UCLA, Los Alamos National Laboratory, Caltech); *Methanococcus jannaschii* (UC-Berkeley, Lawrence Berkeley National Laboratory); *Haemophilus influenzae* (CARB, TIGR); metazoan proteins (Rutgers); *Thermus thermophilus* (Japan); *Methanobacterium thermoautotrophicum* (Canada); targeted identification of novel folds (Argonne National Laboratory, The Rockefeller University).

Strategies for selection of protein targets depend on the specific goals of the project. For the purpose of finding novel folds, optimal targets must be selected from protein family clusters so that a target has no overlap in sequence space with proteins of known structure, and so that the target is the furthest outlier within a family cluster yet resides in a large family. For the purpose of generating sufficient crystallographic structures to model all proteins encoded in a genome, targets must be selected from families of proteins within the genome using criteria about their multi-genome families. For proteins associated with disease, likelihood of a new fold is less important than the size and complexity of protein families. For example, for disease-related proteins with large families, the determination of a structure for an outlier protein in a family that already has a known structure can help to elucidate the protein structure when no functional information is otherwise available.

To support target selection for cancer-related proteins, for the discovery of new folds, and for the modeling of proteins across whole genomes, we have designed and built a general system to construct and query protein family clusters from pairwise sequence similarity data. The system is general and can use any source of pairwise data. We are currently using pairwise similarities constructed with PSI-BLAST for all sequences from all fully annotated genomes plus PDB. This dataset was built in collaboration with Andrej Šali and Roberto Sánchez who are using the dataset to support the generation of 3D homology models of proteins across whole genomes, which are managed in the MODBASE database [SS98]. Future work includes the integration of the cluster analysis system with the MODBASE system. We have used the system in the selection of 306 potential targets for the Rockefeller/Brookhaven/Einstein project, the Argonne project, and the Canadian project. To track selected proteins for a particular project, we have constructed a new configuration of MAGPIE [GS96] that displays the protein family and properties related to ease of expression and crystallization for each selected protein. `http://genomes.rockefeller.edu/magpie/aps_proteins/aps4/` shows an example of the proteins selected from *M. thermoautotrophicum* in a collaboration between the Argonne and Canadian projects.

References

[GS96] Gaasterland, T. & Sensen, C. W. (1996). Fully automated genome analysis that reflects user needs and preferences — a detailed introduction to the MAGPIE system architecture. *Biochimie,* **78**, 302–310.

[Gaa98] Gaasterland, T. (1998). Structural genomics: bioinformatics in the driver's seat. *Nature Biotechnology* **16**, 625–627.

[Sal98] Šali, A. (1998). 100,000 protein structures for the biologist. *Nature Structural Biology,* **5**, 1029–1032.

[SS98] Sánchez, R. & Šali, A. (1998). Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. *Proc. Natl. Acad. Sci. USA,* **95**, 13597–13602.