Computational Approaches for Structural Genomics

Steven E. Brenner and Michael Levitt
Department of Structural Biology
Stanford University
Fairchild D-109
Stanford, CA 94305-5126
{brenner, levitt}@hyper.stanford.edu

Structural genomics aims to provide a good experimental structure or computational model of every tractable protein in a complete genome. Underlying this goal is the immense value of protein structure, especially in permitting recognition of distant evolutionary relationships for proteins whose sequence analysis has failed to find any significant homolog. A considerable fraction of the genes in all sequenced genomes have no known function, and structure determination provides a direct means of revealing homology which may be used to infer their putative molecular function. The solved structures will be similarly useful for elucidating the biochemical or biophysical role of proteins that have been previously ascribed only phenotypic functions. More generally, knowledge of an increasingly complete repertoire of protein structures will aid structure prediction methods, improve understanding of protein structure, and ultimately lend insight into molecular interactions and pathways.

We use computational methods to select families whose structures cannot be predicted and which are likely to be amenable to experimental characterization. Methods to be employed included modern sequence analysis and clustering algorithms. A critical component is consultation of the PRESAGE database for structural genomics, which records the community's experimental work underway and computational predictions. The protein families are ranked according to several criteria including taxonomic diversity and known functional information. Individual proteins, often homologs from hyperthermophiles, are selected from these families as targets for structure determination. The solved structures are examined for structural similarity to other proteins of known structure. Homologous proteins in sequence databases are computationally modeled, to provide a resource of protein structure models complementing the experimentally solved protein structures.