This project studied molecular recognition and enzyme function on a genomic scale. A comprehensive database survey showed that most protein functions are carried out by a single fold and most folds carry out only a single function. There are, however, a small number of multi-purpose folds which carry out many functions. Further information is available at http://bioinfo.mbb.yale.edu/genome.
FINAL REPORT

Contract/Grant Number: # N000149710725

Principal Investigator(s): Mark Gerstein

PI Institution: Yale U., Molecular Biophysics & Biochemistry Dept.

Contract/Grant Title: Investigating Molecular Recognition Through Large-scale Analysis of Protein Sequences and Structures

Award Period: 8/1/97-7/31/00

OBJECTIVE: The objective of this project is to study protein sequence-structure relationships through large-scale computational analysis of gene sequences and crystal structure in the databanks. The results of this analysis will be used to help better understand molecular recognition.

APPROACH: A "data-mining" approach was taken where the rapidly increasing amount of data in the publicly accessible databanks was sifted by computational techniques of increasing complexity. The techniques employed will include sequence comparison, structure comparison, packing calculations, molecular simulation, and composition analysis.

ACCOMPLISHMENTS (during entire period of grant):
During the period of the grant I principally worked on the setup of my laboratory. In terms of science, I began to do large-scale database comparison of the protein structures encoded by a number of the recently sequenced genomes, e.g. yeast and E. coli. This work involved extensive recognition of distant homologies to known folds and secondary structure prediction. In particular, I accomplished the following objectives:

* SHARED FOLDS. I have compared the proteins in various major phylogenetic divisions (e.g. plants vs. animals) and a number of the first genomes sequenced in terms of super secondary-structures.

* PREDICTION. Using structure-prediction on the genomes, I found that bacterial genomes have more all-helix super-secondary structures (e.g. more four-helix bundles), eukaryote, more all-strand ones, and archaeon, more mixed ones (e.g. more strand-helix-strand units).

* DATABASE SYSTEM. I have tried to integrate everything I did into a relational database system. I have received equipment grants from Informix and Intel allowing my group to implement a robust and high-throughput system, and we have recently begun designing object-relational schemas to accommodate protein data.
* OPTIMIZE. We have helped optimize high-throughput sample preparation for structural genomics and done retrospective datamining on the results (NAR and NSB papers).

* TREES. We have constructed whole genome trees based on a variety of characteristics (Genome Res. paper)

* EXPRESSION. We have developed a system to analyze whole-genome expression data and relate this to subcellular localization in a Bayesian framework (TIG and JMB paper).

* ANNOTATION-TRANSFER. We have measured the degree to which functional annotation can be transferred as a function of sequence similarity (Wilson et al., JMB).

* LITERATURE. We have put forth a variety of proposals on integrating on-line literature with genome annotation.

CONCLUSIONS: Our initial analyses of genomes have shown that a relatively small number of basic structural parts (i.e. folds and structural superfamilies) are common among all organisms. These parts tend to be metabolic scaffolds, of which the TIM-barrel is an exemplar, that can support multiple functions. They also tend to be highly expressed (in gene-expression studies). Conversely, we have also found unique structural parts in some genomes. With regard to pathogens, these could potentially be useful drug targets.

SIGNIFICANCE: Our studies should help in comparing and understanding microbial genomes, in relating protein function and structure, and in helping with the general progress of structural genomics.

PUBLICATIONS, ABSTRACTS, TECHNICAL REPORTS, PATENTS, AND AWARDS (last 12 months):


