Pacific Symposium on Biocomputing 5:3-5 (2000)

Protein Evolution and Structural Genomics

DMITRIJ FRISHMAN
Munich Information Center for Protein Sequences
frishman@mips.biochem.mpg.de

RICHARD A. GOLDSTEIN
Chemistry Department
Biophysics Research Division
University of Michigan
richardg@umich.edu

DAVID D. POLLOCK
Theoretical Biology and Biophysics
Los Alamos National Laboratory
dpollock@lanl.gov

The genomic data available to computational biologists represents the product of the complex processes of evolution. In particular, the forces of mutation, duplication, and selection have acted to sculpt modern protein sequence and structure in the context of changing functional requirements. Just as crystallographers are able to determine protein structures through an analysis of X-ray diffraction patterns, scientists are learning to read the evolutionary history of proteins in order to infer and explain both structures and functions. This pursuit depends on the development of new computational approaches in order to make optimal use of genomic data, and requires interaction with experiment for comparison and verification of computational results.

With the realization that genomes provide a new vantage on protein structure studies, there has also been intense interest in understanding structural biology in a genomic context. Each complete genome codes for a full set of functions necessary for a whole organism. This set of proteins can also be considered as a collection of protein folds sufficient for the required cellular activities such as metabolism, replication, and communication. Structural genomics aims to provide structures and theoretical model for all proteins encoded in completed genomes. These large undertakings will vastly increase our knowledge of structural biology and are poised to give us insight into the functions of many broadly conserved yet presently uncharacterized genes. Computational work is guiding the selection of targets for experimental characterization, and the methods of selection are under active development.

The papers published in this track range from an explicit focus on modeling the evolutionary process in proteins to broad scale categorization of protein structures in a complete genome. The

papers by Yang and by Dimmic and Goldstein reflect the growing interest in modeling the process of evolutionary change in protein sequences. Modeling protein sequence evolution is complicated as the mutations occur at the DNA level yet much of the selective pressure occurs at the amino acid level. Selection at the protein level is itself complicated by the heterogeneous nature of the protein and its environment - the type and degree of selective pressure will vary significantly between different locations in different proteins, in ways that cannot be identified a priori. On the other hand a successful model of protein sequence evolution has the potential of helping us decode the patterns of selective pressure, providing much insight into particular sets of homologous proteins as well as to proteins in general. These two papers deal with attempts to model the process of sequence change, both yielding better results in likelihood tests than traditional substitution models that ignore site heterogeneity.

Yang's model involves two approaches. Firstly, it deals explicitly with the dual nature of the sequence evolution, modeling mutations at the codon level but selection at the amino acid level. In addition, he considers the possibility that different locations in the protein are under different degrees of selective pressure using a "mixture model" of different site classes. A physical-chemical based distance criterion is used to identify conservative and non-conservative substitutions, with more conservative mutations having a larger probability of fixation. The relationship between dissimilarity and substitution rate varies according to the site class. While in Yang's model the fixation probability is dependent upon the similarity of the original and new amino acids, Dimmic and Goldstein describe a different type of model. According to this approach, different amino acids have different propensities for various locations in the protein. The probability of fixation is therefore dependent upon the relative propensities of the two different types of amino acids for that location. In this model, conservative substitutions are seen as resulting from the tendency for amino acids with high propensities being substituted by other amino acids with high propensities. As with Yang, Dimmic and Goldstein propose a set of different site classes with different sets of amino acid propensities, and use a mixture model to model the heterogeneous nature of amino acid substitutions. In contrast to the Yang model, the measure of similarity can itself vary from one site class to another, and is not dependent upon the underlying measurable properties of the amino acids.

We expect that structural genomics data will help to obtain more precise estimates of how protein topology evolves over time, and how this evolution interacts with sequence evolution. It is known not only that protein structures tend to change much more slowly than protein sequences over evolutionary time, but also that different proteins evolve much more slowly in topology and/or sequence than others, and that these overall rates of evolution can change with time. Different sites have different structural and functional contexts, there is almost certainly some degree of interaction between sites, and the adaptive landscape itself probably changes over time as major features of structure and function evolve. The patterns of variation and conservation throughout a homologous

sequence set can provide signals indicating the underlying shared structure, and two papers in this track begin to address these structure/sequence evolution issues directly. Dean and Golding use phylogenetic models and likelihood analysis to address (more precisely than has been done in the past) the extent to which the structural environment can explain variation in the rate of evolution at different sites. Finding the primary effects to be solvent exposure and proximity to the active site, they then define other minor factors and are able to explain a large amount of the site-to-site rate variation. Taverna and Goldstein consider whether, following gene duplication, there might be selection to maintain structure in the absence of selection for any other function. Using a lattice model, they consider the effects of the designability of a structure on the overall process, and determine how such a situation would affect the rate of sequence evolution following the gene duplication. This novel attempt to introduce structural requirements into a classic question of population genetics holds great promise for our understanding of this and related questions in the future, and may provide a needed link to the evolution and proliferation of protein folds throughout genomes.

One of the most productive approaches in computational structural genomics has been the investigation of the genomes' structural census. Studies on phylogenetic distribution of protein folds revealed a number of promiscuous structural patterns that frequently occur in all organisms as well as many others that are specific for a particular organism group. These results are especially valuable because they serve as a road map for high-throughput structure determination projects aimed at solving all existing protein structures. In this track Gerstein et al. present an analysis of the protein structural tendencies in C. elegans, the first multicellular organism with a completely sequenced genome. They identified 36 folds that are specific for C. elegans and are presumably involved in intra-cellular communication. Accurate similarity based structure prediction is often a decisive step in elucidating protein function, even if the sequence identity between the proteins involved is extremely low. Pawlowski et al. describe an objective way to estimate the functional similarity between two proteins based on the overlap of their EC numbers. They also discuss the cases of orthologous gene displacement as well as the cases when similar proteins have apparently different functions. Another challenging problem in genome analysis is coping with the growing flood of published information associated with genetic data. Renner and Aszodi developed a system for automated assignment of gene function based on a combination of standard sequence analysis techniques and a novel linguistic approach to document clustering.

Protein evolution and structural genomics are complicated and diverse topics, from which the papers published here represent only a small sampling. It is evident, however, from the quality and scope of these papers, that those who study these subjects have much to offer each other and the community.