Pacific Symposium on Biocomputing 7:548-551 (2002)

# Proteins: Structure, Function and Evolution

Peter Clote
*Department of Computer Science, Boston College*
*Chestnut Hill, MA 02467* clote@bc.edu

Gavin J.P. Naylor
*Dept. of Zoology and Genetics*
*Iowa State University*
*Ames, Iowa 50011* gnaylor@iastate.edu

Ziheng Yang
*Department of Biology, Galton Lab*
*University College London*
*London NW1 2HE, United Kingdom*
z.yang@ucl.ac.uk

Protein structure prediction is one of the most exciting and difficult problems in computational molecular biology. New computational advances, such as that currently underway with IBM's Blue Gene supercomputer (projected completion in 2005), as well as advances in the understanding of energy potentials and development of statistical methods for recognition of specific folds or protein classes, together promise better methods for protein structure prediction. Indeed, we have recently witnessed enormous progress in the field of protein folding, including successful approaches to computational structure prediction as documented in recent CASP competitions.

On a different front, advances in evolutionary genomics have led to better methods for extracting information from the evolutionary history of proteins, to further our understanding of their structure and function. Just as molecular biologists can study protein function by examining the effects of mutations introduced through site-directed mutagenesis, we can learn to interpret the results of the comprehensive experiment performed by Nature over millions of years of biological evolution. For this purpose novel statistical and computational methods are being developed. Models for detecting amino acid residues under diversifying selective pressure across closely related species are generating interesting hypotheses about the structure and function of the protein, which can be tested in the laboratory. We now have the ability to recreate ancestral proteins, and thus test general trends in protein function and hypothetical correlated functional relationships. In vitro evolution, both random and directed, can attempt to replicate the historical patterns, and further elucidate the details of the adaptive landscape.

There is a growing realization that proper evolutionary analysis is an essential component for optimal extraction of structural and functional prediction from multiple sequences. The patterns of variation and conservation throughout a homologous sequence set provide signals indicating the underlying shared structure. Even neural network methods perform better when multiple and diverse sequences are included in the analysis. Analyzing the presence of co-evolving sites in proteins is beginning to make important contributions to the solution as analytical methods improve, as are more refined estimates of the tendencies of different secondary structures and hydrophobic environments to have different substitution rates. Furthermore, it is equally important to consider structural information when devising evolutionary models of sequence change. In particular, accounting for the heterogeneity of the evolutionary process among sites in the sequence is known to lead to much better fit to the model. We hope that by bringing together scientists working on structural and functional prediction as well as evolutionary analysis, this session will help mutually-beneficial interactions between these fields.

In the current session, *Proteins: Structure, Function and Evolution*, new and cutting-edge approaches are introduced for the determination of protein structure and function; additionally, evolutionary models are introduced, which provide new vantage points for these problems. In "Screened charge electrostatic model in protein-protein docking simulations", J. Fernandez-Recio, M. Totrov and R. Abagyan introduce a new method for treating solvation effects in calculating electrostatics for protein docking, and report the success of their method in screening near-native states from false positives. In "The spectrum kernel: An SVM-string kernel for protein classification", C. Leslie, E. Eskin and W. Stafford Noble introduce a new, easily computable string kernel for support vector machine classification. Their "spectrum kernel" essentially adds up uniform contributions of how many size $k$ subwords are shared (independent of position of subword) between given sequences $X, Y$, and provides an $O(n \log n)$ algorithm for classifying whether a given protein $X$ belongs to a protein family. In "Detecting positively selected amino acid sites using posterior predictive $P$-values", R. Nielsen and J.P. Huelsenbeck use a Bayesian approach with the development of "posterior predictive $p$-values", to identify amino acid residues that are under positive Darwinian selection. Those sites exhibit an excess of replacement (amino acid-altering) substitutions relative to silent (synonymous) substitutions and might be important for functional divergence. In "Improving sequence alignments for intrinsically disordered proteins", P. Radivojac, Z. Obradovic, C.J. Brown and A.K. Dunker measure the performance of different scoring matrices for 55 disordered protein families, and develop an iterative algorithm for realigning sequences and recalculating matrices. Investigating a wide range of gap penalties, the authors obtain an improvement n the ability to detect and discriminate related disordered proteins, when average sequence identity with other members of the same family is below 50%. In “Ab initio folding of multiple-chain proteins”, J.A. Saunders, K.D. Gibson, and H.A. Scheraga extend their UNRES force field and Conformational Space Annealing algorithm to handle multiple-chain proteins, illustrating the success of this approach on two homo-oligomeric systems, both of which were targets in the CASP3 experiment (3rd Critical Assessment of Techniques for Protein Structure Prediction). In “Investigating evolutionary lines of least resistance using the inverse protein-folding problem”, J. Schonfeld, O. Eulenstein, K Vander Felden and G. Naylor present a polynomial time algorithm for approximating the solution of the inverse protein folding problem using the Sun-Brem-Chan-Dill grand canonical model. The authors give an improvement of J. Kleinberg’s application of maximum weighted bipartite matching in this context, and apply their algorithm to the PDB, in order to explore the genotype-phenotype mapping. In “Using evolutionary methods to study *G*-protein coupled receptors”, O. Soyer, M.W. Dimmic, R. Neubig, and R. Goldstein develop a model of heterogeneous substitution patterns among partitions of sites in a protein, where the fitness values of amino acids are different in different partition classes. One possible interpretation of the model is that different structural categories (alpha helix or beta sheet, exposed or buried) are under different selective pressures and thus have different substitution rate matrices. The authors find that, in agreement with data, transmembrane regions of G-coupled protein receptors are strongly correlated with hydrophobicity, while non-transmembrane regions are positively correlated with flexibility and negatively correlated with hydrophobicity. In “Progress in predicting protein function from structure: Unique features of O-glycosidases”, E.W. Stawiski, Y. Mandel-Gutfreund, A.C. Lowenthal, and L.M.Gregoret describe unique structural features of O-glycosidases, enzymes which hydrolyze O-glycosidasic bonds between carbohydrates. Using these structural characteristics, the authors show that accurate prediction of O-glycosidase function is possible. In “Support vector machine prediction of signal peptide cleavage using a new class of kernels for strings”, J.-P. Vert develops a class of SVM kernels which interpolate between the diagonal and product kernel and applies this approach in retrieving up to 47% more true positives in signal peptide cleavage site recognition than that obtained using classical weight matrices.

In “Constraint-based hydrophobic core construction for protein structure prediction in the face-centered-cubic lattice”, S. Will develops a new algorithm using constraint programming in order to compute the optimal hydrophobic core for the HP-model on the FCC lattice. The FCC lattice, where each lattice point has 12 nearest neighbors, is much more natural than the 3-dimensional cubic lattice, and though the problem is $NP$-complete, the author reports in benchmark tests that his algorithm can correctly thread large HP-sequences to a core of size up to 100 within 20 seconds. In "Detecting native protein folds among large decoy sets with hydrophobic moment profiling", R. Zhou and B.D. Silverman use the "hydrophobic ratio" (ratio of radii from the protein centroid where the second order hydrophobic moment and the zero order moment vanishes), as a measure of the extent of a protein's hydrophobic core in successfully distinguishing native protein folds from decoy sets.

We would like to thank the authors of this session for reporting their exciting work on protein evolution and protein structure and function determination, as well as the many other authors, whose submissions could not be reported in the current proceedings. We would like to collectively thank many persons involved in the anonymous reviewing process, and finally to thank Dr. Helen Frame Peters, Dean of the Carroll School of Management at Boston College, for additional financial support for the current session.