Pacific Symposium on Biocomputing 4:480-481 (1999)

Protein Structure Prediction in the Post Genomic Era

Jeffrey Skolnick
Department of Molecular Biology, TPC5
The Scripps Research Institute
10550 North Torrey Pines Road
La Jolla, California 92037
skolnick@scripps.edu

Richard H. Lathrop
Department of Information and Computer Science
University of California
Irvine, CA 92697-3425
rickl@uci.edu

Protein structure prediction from sequence remains a major unsolved problem despite decades of work by the best minds in science. However, the availability of genome-scale data is changing the basis for the computational analysis of biological systems, and organizational patterns previously obscured by limited data sets now are becoming apparent. The opportunity to relate sequence, structure, and function at the full organism level is at hand.

The papers in this session further our understanding of the relationship between conservation and structure; combine approaches for ab initio prediction; and present a general method for improving predictive accuracy using parameterized objective functions.

Protein Evolution and Protein Folding: Non-Functional Conserved Residues and Their Probably Role, by Ptitsyn, explores the structural role of conserved residues. The paper looks at two very ancient, large, and diverse protein families, the cytochromes *c* and the globins. Despite an evolutionary history that has erased most primary sequence conservation, these families have retained their overall 3-D structure. Of the few residues that are well conserved within each family across a wide range of modern organisms, some clearly are conserved because of their crucial role in protein function (e.g., binding or coordinating the heme). Setting these residues aside, a small handful of conserved residues remain that appear to have no obvious functional role (four in the cytochromes *c* and six in the globins).

The hypothesis that these remaining residues are conserved for their structural role is intriguing. Based on a wide variety of computational, evolutionary, and experimental support, the author characterizes these conserved residues as a *folding cluster*, and hypothesizes that they form folding nuclei within the framework of the nucleation-growth mechanism of protein folding.

A Combined Approach for Ab Initio Construction of Low Resolution Protein Tertiary Structures From Sequence, by Samudrala, Xia, Levitt, and Huang, approaches protein structure prediction from the opposite extreme. Rather than start with a few critical residues as a folding nucleus, they begin with an exhaustive enumeration of all possible compact conformations on a tetrahedral lattice (about 10 million). These are then successively refined using increasingly fine-grained structural representations, augmented with increasingly informative sources of knowledge about protein structure, and ranked and filtered according to increasingly accurate objective functions. At each stage, fewer possible conformations are retained, and these few are more detailed and supported by more persuasive evidence.

Ultimately this results in a single predicted structure. The method was applied to twelve small proteins. For five of the twelve, the correct protein topology was identified and the conformations produced were within ≈ 6 Å of the experimental structure.

Application of Parameter Optimization to Molecular Comparison Problems, by Lemmen, Zien, Zimmer, and Lengauer, considers another opposite end of structure prediction: improving the objective functions. They approach the problem as an exercise in optimization. Here there are two dual problems. One is, given an unknown structure and an objective function with a large number of fixed parameters, optimize the predicted structure relative to the fixed parameters. The dual problem is to consider the structure as fixed and the parameters as unknown or variable, and optimize the parameters relative to the structure, i.e., improve the objective function so that the fixed or correct structure is recovered by the prediction.

These methods were applied to the protein threading task of ranking a set of possible folds, given a sequence. Six parameters in the objective function were optimized across the three major types of secondary structure, resulting in eighteen variable values to assign. The result was improved predictive performance on a test set disjoint from the training set used to fix parameters.

Acknowledgments

The authors gratefully thank the session referees, whose careful reviews of the submitted papers and insightful, judicious suggestions for improvement are materially reflected in the high quality of the presented papers. Special thanks to all crystallographers who deposited their coordinates in the international scientific databases.

RHL was supported in part by CAREER grant IRI-9624739 from the U.S. National Science Foundation.