PROTEIN STRUCTURE PREDICTION

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 premiere computational problem for modern molecular biology. The papers in this session address *ab initio* folding simulations and theoretical studies, the use of known structures as a basis for prediction, and analysis of how and what we know (or don't) about those known structures.

Ab initio and theoretical folding studies

The "thermodynamic hypothesis" asserts that proteins fold to their global free energy minimum. To compute this accurately is difficult because a folded protein results from the delicate energetic balance of complicated, poorly understood atomic forces; the vast number of possible conformations poses a formidable computational barrier; and the existence of numerous local minima renders the search intractable. To make these challenges approachable, researchers have sought simpler folding models (e.g., lattices) and simpler force fields (e.g., statistical potentials).

Combined Multiple Sequence Reduced Protein Model Approach to Predict the Tertiary Structure of Small Proteins, by Ortiz, Kolinski, & Skolnick, seeks to incorporate additional knowledge about protein folding into the folding simulation. This knowledge takes the form of restraints based on predicted secondary structure and tertiary contacts. The restraints operate on a lattice based reduced protein model of folding, under the influence of a statistical potential extracted from known protein structures, driven by a simulated annealing folding simulation. Beginning with a fully extended initial conformation, they achieve impressive results on small proteins.

Using Constraint Programming for Lattice Protein Folding, by Backofen, simplifies the representation even further in order to study general properties of protein folding from a theoretical perspective. Dill's HP-lattice model simplifies residues to *hydrophobic* (H) and *polar* (P) types embedded in a regular lattice. This clearly exposes the conformational search problem. The paper presents a clever constraint programming approach which is able to discover the minimal energy conformation for a given sequence more rapidly than before by efficiently pruning the search tree.

Using known structures for prediction

One important alternative approach is to use the wealth of information contained in already-known protein structures. The structures can serve as spatial folding templates, impose constraints on possible folds, and provide geometrical and chemical information. This is an attractive strategy because proteins exhibit recurring patterns of organization; there are estimated to be only around 1,000 to 10,000 different protein structural families.

Are Binding Residues Conserved, by Ouzounis, Pérez-Irratxeta, Sander, & Valencia, addresses the important issue of what is conserved, and what is not, across homologous structures. Binding residues represent a particular challenge, because active sites typically are composed of conserved functionally important residues in structural environments they otherwise might not occupy. The authors are able to demonstrate a remarkable dichotomy between a residue type's propensity to be conserved and its propensity to participate in binding sites.

Linear programming based approach to the derivation of a contact potential for protein threading, by Akutsu & Tashimo, tackles the fundamental problem of finding an objective function to guide protein structure prediction when the known proteins are used as structural templates. They use a linear programming approach to derive a score function from constraints requiring the native threading to be preferred, while accommodating alignment gaps in the derivation. An important advantage is that the method can be used when only a small number of training proteins is available.

A Protein Conformational Search Space Defined by Secondary Structure Contacts, by Parisien, Major, & Peitsch, explores a coarse-grained decomposition of the protein tertiary structure conformational search space. Recognizing that the core of a protein is composed largely of tightly packed secondary structure, they investigate the tertiary structure search space that results from considering possible secondary structure contacts and show that it contains the X-ray crystal structures.

Analysis of known structures

Ultimately, all of our knowledge of protein structure is derived from the relationship between native sequences and structures. The final papers in this session consider the extent to which protein sequences necessarily do form a unique structure, and the process of determining that structure from crystallographic data.

Thousands of Proteins Likely to Have Long Disordered Regions, by Romero, Obradovic, Kissinger, Villafranca, Garner, Guilliot, & Dunker, presents strong evidence that some amino acid sequences cause regions of proteins to be disordered rather than structured, and that such disordered regions are commonly involved in function. Using a neural network predictor trained to recognize disordered regions in known structures, they identify thousands of putative disordered regions in proteins of unknown structure. The disordered-to-ordered transition is hypothesized to have important implications for protein function.

Protein Model Determination from Crystallographic Data, by Edgecombe, Ableson, Baxter, Chiverton, Glasgow, & Fortier, describes an approach to elucidating X-ray crystal structures by incorporating both crystallographic and sequence data. They use a machine vision approach to understanding the electron density image, and derive potential scene models for a protein structure. Experimental results demonstrate the viability of the approach.

Acknowledgments

The author gratefully thanks 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.

The author was supported in part by a CAREER grant from the U.S. National Science Foundation.