DISCOVERING, LEARNING, ANALYZING AND PREDICTING PROTEIN STRUCTURE

A. Keith Dunker
Department of Biochemistry/Biophysics
Washington State University
Pullman, WA 99164
Dunker@mail.wsu.edu

Richard H. Lathrop
Department of Information and Computer Science
University of California
Irvine, CA 92717
rickl@ics.uci.edu

This track focuses on the understanding, and ultimately, the prediction, of protein structure. Papers describe discovery, learning, or analysis approaches that lead to testable predictions, and that quantify or compare predictive accuracy. The papers accepted this year span a wide range of methods that illustrate or investigate most of the approaches currently being used in this field.

Innovative Representations

A very productive strategy for attacking a difficult problem is to change the representation to a form that more perspicuously lays bare the problem's essential features. These may suppress extraneous details, highlight the constraints governing acceptable solutions, or transform a problem into a more tractable or more easily understood form.

Protein structure comparison using representation by line segment sequences, by Akutsu & Tashimo, presents an efficient algorithm for representing tertiary protein structures by a series of line segments. This enables fast and effective comparison of two protein structures, and provides a clean representation of the overall fold as abstracted from atom-level detail.

All possible protein folds at low resolution, by Crippen & and Maiorov, proposes another reduced representation of the tertiary structure of a folded protein. They generate a parameterized family of continuous curved three-dimensional figures, and use these to estimate the number of different folds at different levels of resolution.

Multfractals, encoded walks and the ergodicity of protein sequences, by Dewey & Strait, presents an unusual fractal analysis of protein sequences. They use this analysis with an ensemble of representative protein sequences to establish their ergodicity.

Circular clustering of protein dihedral angles by minimum message length, by Dowe et al., investigates clustering dihedral angles using the minimum message length approach, which seeks to inductively construct the description which represents the optimal trade-off between approximation error and over-fitting.

Folding and Dynamics

An important aspect of protein structure prediction is the attempt to predict specific folding patterns of specific amino acid sequences. Quantum mechanics provides a solution in principle, but the computation becomes intractable when confronted with the many thousands of atoms comprising a protein. Consequently there has been an immense amount of effort to find tractable and effective strategies.

An algorithm for prediction of structural elements in small proteins, by Kolinski et al., proposes a method for predicting surface turns and loops, and for assigning the intervening secondary structure. They show good agreement on ten proteins of known structure.

Conformational evolution of a model polymer that folds to a specified target conformation, by Judson, investigates the integration of a state transition matrix approach to dynamics with a computational genetics algorithm which is used to evolve the free parameters needed.

A high performance system for molecular dynamics simulations of biomolecules using a special-purpose computer by Komeiji et al., describes a high-performance special-purpose computer designed especially for molecular dynamics. This is aimed at removing the immense computational burden imposed by classical molecular dynamics when applied to large macromolecules.

Specific Systems

One of the most intriguing aspects of computational molecular biology is the interplay between experiment and theory, in which specific experimental systems shed light on the general underlying principles which are encoded in computational algorithms, and simultaneously the computational leverage so obtained enables more effective progress on specific experimental systems.

Gaussian quadrature calculations of binding free energy difference for N184D mutation in D-xylose isomerase by Hu et al., uses a Gaussian quadrature method to calculate the binding free energy difference of an Asn to Asp mutation. Their careful handling of electostatic interactions leads to a good agreement between calculated and experimental values, which is particularly

impressive because the mutation requires modelling the transition from a polar amino acid to a charged one.

Prediction of the quaternary structure of coiled-coils: GCN4 leucine zipper and its mutants by Vieth et al., addresses the difficult problem of the formation of quaternary structure by studying the equilibrium between different oligomeric species of coiled coils in the leucine zipper. They report good agreement between calculated and experimental values, and suggest specific driving processes of multimer formation.

Homology Modeling and Inverse Folding

One of the most successful approaches to protein structure prediction has been homologous extension modeling. In order to extend the homology modelled core to a full structure accurate enough to support detailed atomic binding studies and drug discovery programs, however, the loops and side-chains also must be placed. A major recent effort has been to extend homology modeling to sequences with lower and lower identity. This has led to sequence-structure alignment techniques, also known as protein "threading" or "inverse" protein structure prediction, that attempt to align a sequence directly to a structural model using any of a variety of empirically derived score functions.

How similar must a template protein be for homology modeling by side-chain packing methods, by Chung & Subbiah, presents a very careful analysis of side-chain packing accuracy at and above the "twilight zone" of marginal primary sequence similarity. They quantify the extent to which measures of sequence similarity, r.m.s. deviation, and side-chain packing prediction accuracy all degrade simultaneously as the "twilight zone" is approached.

Analysis, clustering and prediction of the conformation of short and medium size loops connecting regular secondary structure, by Rufino et al., attacks the other major problem of homology modeling by investigating the loops that connect conserved regions. They cluster a large number of observed loops, and derive templates for both sequence preferences and the relative orientations of the flanking secondary structures.

Assessing the performance of inverted protein folding methods by means of a comprehensive benchmark, by Fischer et al., makes an important step toward rigorous quantitative evaluation of predictive accuracy among the various sequence-structure alignment methods. They describe a large and diverse benchmark set of proteins for testing such methods, as well as the results of a number of methods when tested against the set.

Fast protein fold recognition via sequence to structure alignment and contact capacity potentials, by Alexandrov et al., explores sequence-structure alignment using an empirical potential function designed explicitly to

account for the hydrophobic contribution to the free energy. They show good agreement between their potentials and residue hydrophobicity.

Statistical Geometry Analysis of Proteins: Implications for inverted structure prediction, by Tropsha et al., investigates a Delaunay tesselation of protein structure. The Delaunay simplices induce a spatial "neighbor" structure among the protein amino acids, and these are used to derive a set of empirical amino acid pair potentials for scoring sequence-structure alignments.

Acknowledgments

The session co-chairs 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: K. Albrecht, G. Arnold, P. Argos, A. Aszodi, J. Bajorath, P. Bieganski, B. Bryant, S. Bryant, P. Cheeseman, R. Chen, C. Chothia, K. Chou, D. Covell, K. Dill, I. Dubchak, R. Elber, J. Evans, J. Fetrow, J. Finney, P. Fleming, S. Galactionov, J. Garnier, J. Gibrat, L. Gregoret, A. Grosberg, N. Harris, J. Hart, T. Havel, S. Havlin, J. Hellinga, A. Holtzer, B. Honig, T. Itoh, J. Janin, M. Johnson, D. Jones, C. Kim, K. Kitamura, C. Lawrence, M. Levitt, L. Liebovitch, D. Loshin, T. Kuntz, T. Mattson, R. Miller, H. Nakamura, R. Nambudripad, P. O'Hara, M. Paulsen, N. Richards, M. Rooman, M. Saxton, L. Shoop, T. Smith, S. Thompson, J. Thornton, R. Unger, S. Wherland, Y. Zheng.

We also gratefully thank sponsors whose contributions enabled the participation of authors who otherwise would have been unable to attend: Amgen, Molecular Kinetics, and Zymogenetics.