Informatics Approaches in Structural Genomics: Session Introduction

S.D. Mooney and P.C. Babbitt

Pacific Symposium on Biocomputing 8:176-179(2003)

INTRODUCTION TO INFORMATICS APPLICATIONS IN STRUCTURAL GENOMICS

SEAN D. MOONEY
Stanford Medical Informatics
Department of Genetics, Stanford University
Stanford, CA 94305

PATRICIA C. BABBIT
Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry,
University of California
San Francisco, CA 94143

The goal of structural genomics is to determine representative three dimensional structures of all proteins and macromolecules found in nature. Current efforts in protein structure determination have a major focus in the development of high-throughput methods. This session presents progress and achievements in solving the computational challenges in this field. Topic areas included are (1) Determination of all common scaffolds found in naturally evolved proteins, (2) Structure-based prediction and classification of function and (3) Elucidating structurally defined function and phenotype.

1. Structural Genomics

1.1 Introduction

Structural genomics initiatives aim to determine all of the naturally evolved macromolecular scaffolds. The large numbers of structures resulting from these projects are stored in publicly available databases such as the PDB or the NDB. While these projects are clearly far from finished, researchers have made great strides toward achieving their primary goals. This session focuses on the applications that use the large datasets created by these discovery projects. These datasets can be used to infer function, identify targets and to understand the underlying physical properties that dictate how proteins fold. This introduction is not meant to be a review of the field; see Chance, *et al.*[1] or Norin and Sundstrom[2] for a more thorough treatment.

The three dimensional structure of a protein is typically determined by one of three methods: X-ray crystallography, NMR spectroscopy or modeling based on inferred similarities to homologous proteins or other macromolecules.

Crystallographic methods continue to be the gold standard for protein structure determination and the vast majority of experimentally determined structures are solved using these methods. Structures determined by NMR spectroscopy are equally useful in characterizing in-solution protein behavior, yet their representation in the PDB is still relatively low. Homology modeling methods are becoming more popular, but limitations in the technology continue to hinder widespread adaptation.

Currently there are less than 18,000 structures in the protein databank (PDB)[3]. This comprises a small subset of the more than 500,000 characterized protein sequences and the millions of structures derived from variants in populations. Methods for characterizing the protein structure universe are being developed. Equally important is the structural classification and functional annotation of these protein structures.

The number of protein sequences far outnumbers the number of protein scaffolds. Many sequences share the same fold, with current estimates putting the number of folds between 500 and 2000. A complete set of protein folds will greatly advance our ability to build and store large numbers of protein structures based on comparative models. Progress in this field is excellent, in 2001, nearly 4,000 new structures were deposited in the protein databank and ModBase[4], a database of theoretical models, now contains over 500,000 reliable protein structures. SCOP, the resource for the structural classification of proteins[5], has identified and annotated over 600 folds and nearly one thousand superfamilies comprising over 31,000 domains.

A nearly complete set of scaffolds is a powerful research tool that raises many future challenges and opportunities. These include, 1) macromolecular structure prediction, 2) identification of functionally and structurally important motifs, 3) understanding the relationship between structure and molecular phenotype, and 4) understanding the physical principles that specifiy the structure and dynamics of macromolecules. Here we present a brief summary of the problems and approaches addressed in this session.

1.2 Macromolecular structure prediction

Comparative modeling of protein structures becomes an increasingly important problem as more naturally evolved scaffolds are determined and characterized. Comparative modeling often consists of a four step process: fold identification, alignment between target and scaffold, model building and model evaluation and refinement. Several contributions to this process are presented here. First, Xu and Li present a linear programming method for threading, the process of building an alignment between the target and scaffold. This rapid method can be used for both fold identification and building a high-resolution alignment suitable for model building. Second, Edgar and Sjolander describe a method for building accurate alignments using hidden Markov models. Third, Kersting, *et al*. describe a different application of hidden Markov models to find specific structural motifs in protein sequences. Finally, Ohsen *et al*. demonstrate their use of profile-profile alignments to achieve accurate alignments with highly similar sequences as well as more distant relatives.

1.3 Identifying important motifs from a database of macromolecular structure

An important problem is to understand both the structure and function of many solved or modeled protein (or nucleic acid) structures. Singh and Saha describe a method for identifying known motifs from a set of protein structures. Liang *et al*. introduce a new method for describing and categorizing structural motifs based on characteristic sequence motifs within them. They show that their automatically generated motifs perform similarly to manually determined motifs in sequence alignments and yield better alignments than those based on simple sequence motifs.

1.4 Understanding the underlying physical properties of proteins

Understanding the underlying physical principles that dictate the folding and dynamics of proteins is required for understanding macromolecular function. Song *et al*. present a method that gives sophisticated insight into the folding energy landscape of a protein. Using homologous model proteins, protein G and L, their method captures folding differences between the structurally similar proteins.

Radivojac *et al*. describe a method for predicting the boundaries between intrinsically structured and "disordered" regions in protein sequences. This work continues their previous research suggesting that lack of structure may play a functional role and quantitatively determines where these regions are likely to occur.

2. Conclusions

The underlying theme of our discussion is to translate structural information now becoming available into a functional understanding of a macromolecule's purpose. This process of linking structure to function involves organizing and recognizing structural information at many levels. New methods of recognizing sequence and structural motifs within protein scaffolds, aligning these motifs and finding similarities between homologous folds is required for understanding protein structure. Inferring function from structure requires such sophisticated methods as those that are being developed now both in industry and academia.

References

1. Chance, M., et al. "Structural Genomics: A Pipeline For Providing Structures For The Biologist." Protein Science, 11(4), 723-738 (2002).
2. Sundstrom, M. and M. Norin "Structural Proteomics: Developments in Structure-to-Function Predictions." Trends Biotechnology, 20(2), 79-84 (2002).
3. Berman, H., et al. "The Protein Data Bank." Nucleic Acids Research, 28(235-242 (2000).
4. Sanchez, R. and A. Sali "ModBase: A Database Of Comparative Protein Structure Models." Bioinformatics, in press, (2001).
5. Murzin, A., et al. "SCOP: A Structural Classification Of Proteins Database For The Investigation Of Sequences And Structures." Journal of Molecular Biology, 247(536-540 (1995).