Inferring Function from Structural Genomics Targets: Session Introduction

P.C. Babbitt, P. Bourne, and S.D. Mooney

Pacific Symposium on Biocomputing 10:319-321(2005)

INTRODUCTION TO INFORMATICS APPROACHES IN STRUCTURAL GENOMICS: MODELING AND REPRESENTATION OF FUNCTION FROM MACROMOLECULAR STRUCTURE

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

PHILIP E BOURNE
Department of Pharmacology
The University of California San Diego
San Diego, CA 92093-0505 USA

SEAN D. MOONEY
Department of Medical and Molecular Genetics
Center for Computational Biology and Bioinformatics
Indiana University School of Medicine
Indianapolis, IN 46202 USA

Despite the advantages provided by the enormous recent increases in the availability of structural information, functional assignment for the large number of proteins represented in the sequence and structural genomics projects remains a pressing problem for genomic era biology. This section describes work relevant to this problem from several perspectives, including new approaches that take advantage of combined structure and sequence-based classification. Leveraging of genomic context and evolutionary information to improve classification and predictive power is a second prominent theme in the papers represented here. Finally, issues in building a database for linking sequence, structural, and functional information are explored.

1.1. *Issues in Functional Inference for Structural Genomics*

As the structural genomics projects move beyond the initial implementation phase, functional assignment for *Initiative* targets and their homologs is an increasingly more important and frequent problem. For one Center for Structural Genomics, it has been estimated that approximately one-third of the solved structures are hypothetical proteins for which no functional information is available in sequence databases [1]. Currently, there are 1,227 structures in the Protein Databank [2] that are submitted from structural genomics projects, of which approximately 50% have unassigned functional classification

(http://targetdb.rcsb.org and http://pdbbeta.rcsb.org). Moreover, because the core *Initiative* currently does not include functional characterization, there is no large-scale infrastructure to address the critical need for functional annotation of functionally uncharacterized protein structures. Thus, there is both a compelling need and an important opportunity for the biocomputing community to address the issue of modeling and representation of function from macromolecular structure.

Until recently, predicting the functions of gene products and identification of functionally important regions has largely been performed using sequence information alone. More sophisticated approaches that integrate sequence and structural data for functional characterization are being increasingly employed, however. Given the rise of structural genomics projects and related data, approaches that utilize three-dimensional structural information for inference of functional characteristics are an especially important area for research and represent a major theme in the work described here. This session is represented by five papers that address various issues in functional prediction and that directly or indirectly use three-dimensional structural information. In addition, the extended capabilities provided by incorporating evolutionary or other types of genomic context are explored in this session, along with issues in representation of structure-function linkage for database development.

1.2. *A short description of papers in this section*

Structural and functional similarity search methods enable functional annotation from a structural perspective. Chen*, et al*. describe a method for the comparison of structural motifs, or functional sites, in protein structures called the "Match Augmentation Algorithm." The authors find that statistically significant matches between two structures have functional significance, and this method significantly improves performance over geometric hashing methods.

An important problem in functional annotation is subfamily annotation. Brown, et al. show that subfamily hidden Markov models can classify the members of a family better than a single family HMM, and report a low expected error rate. They then apply the method to the serotonin receptors as a proof of concept.

Structural analysis of short protein segments can improve structure annotation and prediction methods. Tang, et al. provide an analysis of clusters of short protein segments. They find that their model improves both tertiary and secondary structure prediction methods.

Support vector machines continue to show improvement over other classification methods in bioinformatics problems. Nguyen and Rajapakse

apply a multi class support vector machine to the problem of secondary structure prediction, showing improved results on several datasets.

In the last paper in this section, Pegg *et al*., describe a new "Structure-Function Linkage Database," designed to aid in functional prediction from sequence and structure. Here, the authors compile sequence, structure, and functional information available for related proteins in large superfamilies of enzymes in order to associate and the sequence/structure variation in families within each superfamily with the similarities and differences among them. The result is a hierarchical system that distinguishes aspects of function common across all members of a superfamily from those common only to subgroups or families within the set. Computational issues in representing such structure-function linkage are also explored.

References

1. Laskowski, R. A., J. D. Watson, et al. *J. Struct. Funct. Genomics* **4**, 167- (2003).
2. H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne. Nucleic Acids Research, 28, 235-242 (2000).