Informatics Approaches in Structural Genomics: Session Introduction

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

Pacific Symposium on Biocomputing 9:372-374(2004)

INTRODUCTION TO INFORMATICS APPLICATIONS IN STRUCTURAL GENOMICS

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

P.E. BOURNE
The San Diego Supercomputer Center
The University of California San Diego
San Diego, CA 92093

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

1. Structural Genomics

Structural genomics initiatives aim to determine all of the naturally evolved macromolecular scaffolds of proteins, RNA and DNA. In this introduction, we introduce several recent advances in the computational methods that support structural genomics. These include improvements at all levels of structure analysis, from fold identification of a target sequence and structure prediction, to structure evaluation and classification. The reader is referred to Goldsmith-Fischman and Honig[1] for a thorough treatment on computational methods in structural genomics and to Bourne, *et al.* in this volume for the status of target structure determination.

Improvements in computational methods for structural genomics are facilitating the identification of new, previously uncharacterized targets with novel fold classifications and predicted functions. These computational methods support the structural genomics pipeline by identifying targets, storing assay data, and by analyzing results in a statistically sound manner. The six papers presented here address many aspects of this diverse topic.

One of the primary ways of identifying the function of an unknown structure is to identify its most similar structural neighbors. These "nearest neighbor" structural classification methods have proven to be powerful tools for identifying unknown function. For example, the Structural Classification of Proteins project, SCOP, is an effort to classify all protein domains. SCOP classification is performed using both human intervention and through automated methods. Therefore, the challenge for fully automated computational methods is to correctly classify protein domains and to produce results similar to that of methods or databases that rely on human annotation. In this volume, Huan, *et al.* apply an information theoretic approach to identify coherent subgraphs in graphs that represent protein structures. They test their method on several families and find that their classifications correlate well with SCOP.

Another challenge for computational structural bioinformatics methods is macromolecular structure prediction. A common approach to predicting the structure of an amino acid sequence is to apply comparative modeling methods, by modeling an unknown sequence upon a structure having a similar sequence. Comparative modeling is often performed in a four-step process: fold identification, threading, model building and evaluation with refinement of the structure.

Fold identification and threading remain significant challenges. A target sequence may have little sequence similarity to any known scaffold. This volume presents two papers aimed at improving the identification of the appropriate fold for a target protein sequence through experimental intervention. First, Potluri *et al.* present a method for discriminating well predicted structures from poorly predicted ones using chemical cross linking data. Second, Qu *et al.*, present a method for identifying the fold of a sequence using the NMR technique of residual dipolar coupling. Their program, RDCthread, identifies structural homologs of a target protein using RDC data and secondary structure prediction.

Although most structural genomics techniques aim at studying protein structures, similar techniques have been applied to RNA structure prediction. For a review of structure prediction techniques as applied to RNA structure, see Schuster, *et al*[2]. In this volume, Nebel presents a method for identifying good predictions of RNA secondary structure, thereby improving secondary structure prediction overall.

Finally, one of the most exciting activities in structural genomics is studying the many structures that are now stored in public databases such as the Protein Databank (PDB)[3]. Peng, *et al.* apply contrast classifiers to explore bias in the PDB. When they compared the distributions of proteins in SWISS-PROT and the PDB, they found that transmembrane, signal, disordered and low complexity regions are poorly represented in the PDB. They reason that contrast classifiers can be used to select important targets for structural genomics initiatives.

Successes in structural genomics initiatives continue to be accompanied by the development of computational methods that apply sophisticated analyses from such diverse fields as information theory, clustering methods, and novel experimental techniques. As a result, novel structures continue to be added to our structural repertoire, giving new biological insight in this post-genomic era.

Acknowledgements

We would like to Giselle Knudsen for her advice in preparing this document.

References

1. Goldsmith-Fischman S and Honig B (2003) "Structural genomics: computational methods for structure analysis" Protein Science 12(9):1813-21
2. Schuster, P., Stadler, P.F., and Renner, A. (1997) "RNA structures and folding: from conventional to new issues in structure predictions" Current Opinions in Structural Biology 7(2):229-35.
3. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. (2000) "The Protein Data Bank" Nucleic Acids Research 28(1):235-42.