Pacific Symposium on Biocomputing 13:438-440(2008)

PROTEIN-NUCLEIC ACID INTERACTIONS: INTEGRATING STRUCTURE, SEQUENCE, AND FUNCTION

MARTHA L. BULYK

Brigham & Women's Hospital and Harvard Medical School
Boston, MA 02115

ALEXANDER J. HARTEMINK

Duke University
Durham, NC 27708

ERNEST FRAENKEL

Massachusetts Institute of Technology
Cambridge, MA 02139

YAEL MANDEL-GUTFREUND

Technion-Israel Institute of Technology
Haifa, Israel 32000

Over the last several years, various groups have been developing methods to incorporate information about the three-dimensional structure of proteins, DNA, and RNA into algorithms for analyzing high-throughput genomic and proteomic data. In particular, these methods have been shown to significantly improve predictions of a wide range of functional properties, including the regulatory targets, of nucleic acid binding proteins. These approaches are likely to become increasingly important in analyzing the many different types of data that can now be collected on a genome-wide and proteome-wide scale, including DNA sequence from various genomes, gene expression data, protein-protein and protein-ligand interactions, and protein-DNA and protein-RNA binding data.

This emerging paradigm builds on recent technological advances in data collection and computational developments in diverse areas including DNA binding site motif discovery, modeling of transcriptional regulatory networks, multiple sequence alignments, structural genomics, and structural and evolutionary studies of proteins and nucleic acids. While each of these specific aspects of protein-nucleic acid interactions has been studied previously, these different aspects have just recently begun to be considered together. This PSB session focuses on methods that bridge structure, sequence, and function to infer previously undiscovered associations between these different aspects of protein-nucleic acid interactions.

Methods that employ structure and sequence as they relate to function have several key advantages. First, structural data alone often do not permit the inference of biological function. Second, experimental genomic datasets often contain errors or noise due to imperfections in the applied technology. Third, functional studies typically do not connect function to structure. Indeed, only a small body of work addresses how to take advantage of these currently separate areas of research on protein-nucleic acid interactions. We anticipate that combining these different types of data will allow us to identify essential biological associations, and ultimately to model and predict these interactions.

This year there are six papers in this session. In the first paper, McCord and Bulyk reveal that numerous families of transcription factors from bacteria, yeast, fly, and mouse that contain the same type of DNA binding domain have similar functions and/or regulate genes of similar function. The observed correlations between transcription factor structural classes and the regulatory roles of the transcription factors themselves suggest that structural information could be useful for predicting the functions of transcription factors and their regulatory targets.

In their paper, Gordân and Hartemink report that experimentally determined transcription factors' DNA binding sites in yeast are significantly biased toward regions of higher predicted DNA duplex stability. By incorporating information about helix destabilization energy—which can be calculated directly from DNA sequences—as a Bayesian prior, they are able to markedly improve the accuracy of transcription factor DNA binding site motif discovery.

Lusk and Eisen introduce an evolutionarily based approach for choosing an appropriate position weight matrix cutoff when identifying transcription factor binding site motif matches in genomic DNA. They find that yeast transcription factors appear to fall into different categories of cutoff stringency, suggesting that different transcription factors may have been under pressure to maintain binding sites of varying stringency.

Pan and coauthors introduce a parametric mixture model for estimating the targets of a transcription factor genome-wide by combining evidence from assays of transcription factors' DNA binding (such as from ChIP-chip experiments), assays of target co-expression, and presence of transcription factors' DNA binding site motifs in target promoters. By combining this evidence in a joint mixture model, they present a method that is at once both simple and effective.

Two of the papers in this session examine methods for predicting the protein residues that make contact with DNA and RNA. Kauffman and Karypis use mutual information to systematically analyze the relationship between various

sequence and structural properties of amino acids and their role in binding DNA in a set of almost 250 protein-DNA complexes. Lee and colleagues combine threading and machine learning methods to identify residues that contact RNA and DNA in the catalytic subunit of human and yeast telomerase.

Progress in these areas may further improve the ability to predict the functions, targets, and regulatory mechanisms of DNA- and RNA-binding proteins. In addition, numerous other challenges remain in this nascent research area aside from those addressed in the accepted papers for this session. Future work will need to address questions such as:

- Do certain types of domains of DNA/RNA binding proteins confer particular biophysical properties, either in terms of kinetics or ligand specificity?
- How is RNA structure involved in interaction with proteins, and what are the regulatory or other functional consequences of those interactions?
- How are affinities of protein-DNA interactions tied to function?
- What are the relative contributions of biophysical constraints and evolutionary history in shaping the functional roles of proteins sharing a common domain structure?
- Can a fully predictive (energetic) model of protein-nucleic acid interactions be developed?

As more types of data become widely available, integrative approaches will become increasingly important in computational approaches for understanding regulation.

Acknowledgments

We are grateful to those who submitted manuscripts for consideration for inclusion in this session, and we thank the numerous reviewers for their valuable expertise and time throughout the peer review process.