Pacific Symposium on Biocomputing 6:101-102 (2001)

DNA Structure, Protein-DNA Interactions, and DNA-Protein Expression

P.F. Baldi, R.H. Lathrop
Information and Computer Science
University of California, Irvine
Irvine, CA 92612-2748, USA

The session explores computational approaches to understanding the fundamental cycle of DNA ↔ Protein ↔ DNA by which DNA and proteins co-exist, co-regulate, and co-create each other. There is mounting evidence that DNA structural properties beyond the double helix significantly affect its interactions with proteins and play an important role in a number of biological phenomena ranging from gene regulation to triplet repeat expansion diseases. The selected papers address the underlying structural and functional basis by which these interactions can be understood, modeled, and predicted.

The study of DNA structure, protein-DNA interactions, and DNA-protein expression addresses the structural and functional basis underlying many areas of current significance, such as gene expression arrays, gene networks, genetic regulation, motif discovery, molecular recognition, and ultimately the regulated control of all life processes. As examples, the protein-DNA interactions of regulatory proteins provide the physical basis for the expression patterns studied by gene expression arrays; triplet repeat expansion diseases such as Huntington's disease are due to the abnormal expansion of DNA triplets which have very extreme structural properties; and the viral packaging of nucleotides into the protein shell provides the physical basis for some of the most virulent of human pathogens. The interplay between DNA and proteins is the most fundamental of biological interactions, and has pervasive implications in biology, medicine, pharmacology, and biotechnology.

A Topic Balanced Between String and Structure

The DNA↔Protein↔DNA cycle, by which DNA and proteins co-exist, co-regulate, and co-create each other, is a delicate balance between string and structure. The heteropolymer nature of both DNA and proteins allows for their compact representation and manipulation as sequences of characters drawn from alphabets of four and twenty characters respectively. Sequence analysis has been one of the most fruitful areas of computational biology, and much of our knowledge in areas as diverse as phylogenetic ancestry and DNA binding has been obtained by clever string manipulations. On the other hand, ultimately the characters of the string represent full-atom placements of structures in space. It is these molecular interactions that give the string characters their properties, and in addition lead to other properties not well represented as strings. The papers in the session reflect this balance between string and structure, making contributions that span both.

The paper by Benham analyzes the structure of duplex DNA for its stress transmission and stabilization properties, and relates these to transcriptional regulation. Computational methods that predict destabilization regions as a function of DNA sequence and superhelicity are applied to three transcriptional regulatory events, and the computational predictions made have been experimentally verified in each case.

Benos et al. and Mandel-Gutfreund et al. both take a structure-based approach to building a string-based predictor of DNA regulatory motifs. Crystallographic information from one or more protein-DNA complexes is used to define a set of putative amino acid ↔ DNA base interactions. An objective function (score function) is supplied by knowledge-based or probabilistic potentials, which have had much success in protein structure prediction. Here the objective functions are specialized to model DNA-protein interactions. The protein model is scanned down the DNA sequence computationally, putative DNA-protein contacts are assignd from the crystal data, and the site is scored using the knowledge-based or probabilistic potentials. Predictions are demonstrated for both helix-turn-helix and zinc-finger DNA binding motifs.

Pavlidis et al. and Liu et al. both take a string-based approach to recognizing DNA regulatory motifs, based in both cases on advances in hidden Markov models. Pavlidis et al. use motif-based hidden Markov models to provide a Fisher kernel for a support vector machine. Liu et al. extend the Gibbs sampling strategy using 0th to 3rd order Markov models, including extensions to gapped motifs and palindromic patterns. These are used to predict the classification of unannotated promoter regions in yeast and bacteria.

Structures at the molecular level hold the keys to molecular function, while strings hold the essence of encoded biological information. The papers in this session show the structural and sequence viewpoints working together to help understand and predict DNA structure, protein-DNA interactions, and DNA-protein expression.

Acknowledgments

We gratefully thank the session blind reviewers, who helped in the difficult task of choosing the best contributions from many excellent submissions.