

AN ANALYSIS OF INFORMATION CONTENT PRESENT IN PROTEIN-DNA INTERACTIONS

CHRIS KAUFFMAN AND GEORGE KARYPIS*

*Department of Computer Science, University of Minnesota
117 Pleasant St. SE
Minneapolis, MN 55455, USA
E-mail: {kauffman,karypis}@cs.umn.edu*

Understanding the role proteins play in regulating DNA replication is essential to forming a complete picture of how the genome manifests itself. In this work, we examine the feasibility of predicting the residues of a protein essential to binding by analyzing protein-DNA interactions from an information theoretic perspective. Through the lens of mutual information, we explore which properties of protein sequence and structure are most useful in determining binding residues with a particular focus on sequence features. We find that the quantity of information carried in most features is small with respect to DNA-contacting residues, the bulk being provided by sequence features along with a select few structural features. Supplemental information for this article is available at <http://www.cs.umn.edu/~kauffman/supplements/psb2008>

1. Introduction

Complex behaviors of the genome are now beginning to be understood in terms of feedback network models in which regulatory elements promote or inhibit transcription of genes and are themselves affected by the transcription of other elements. Key to this system are interactions between DNA, the main storage unit for genetic information, and proteins, which are both products and managers of transcription. To that end, a plethora of computational methods have been presented to predict which proteins will bind to DNA^{1,15}, what parts of a protein will bind to DNA^{2,11,17,18}, and which segments of DNA a protein will favor for binding. These methods have yet to reach a performance plateau and researchers continue to apply machine learning and statistical techniques in an attempt reach the

*Work supported by the NIH Training for Future Biotechnology Development grant, NIH T32GM008347

highest accuracy and sensitivity supported by available information.

We endeavor in this study to provide some insight into the inherent difficulty of predicting protein-DNA interactions. From a thermodynamic perspective, the interactions have been found to be quite sensitive. Binding is marginally favored when considering the whole complex⁷. This leaves very little in the way of individual contributions for each residue requiring methods that predict binding residues to make shrewd use of any available features to achieve accuracy. Predicting binding residues would benefit genome studies as mutating them to less favorable analogues gives a mechanism to affect a protein's role in the system. In particular, prediction of binding residues from sequence alone is desirable as it would open the door to a wide variety of experiments involving transcription regulatory elements which have not been co-crystallized with DNA and for which CHIP-Chip experiments¹⁰ are not feasible.

In this paper we focus on sequence and structure features of single protein residues and how they may describe a residue's contributions to the DNA-binding event. We lay out an information theoretic framework in which to conduct the study, illustrate the features of interest, and report the most likely candidates for use in prediction methods.

2. Methods and Materials

2.1. *Mutual Information (MI)*

The main tool we employ for analysis is mutual information (MI)^{5,14}. The MI between two random variables is a measure of how easily the value of one may be predicted given the other's value. That is, mutual information measures how much information two variables carry about one another. In the discrete case, it is defined for random variables X and Y as

$$I(X; Y) = \sum_{x,y} p(x, y) \log_2 \frac{p(x, y)}{p(x)p(y)} \quad (1)$$

where x and y are the discrete values or classes which random variables X and Y can take on and $p(x, y)$ is the probability of x and y occurring together. Due to the base-two logarithm, mutual information in this paper is reported in bits.

2.2. *Features*

In our setting, each residue of a protein has associated with it features that are represented by random variables. The first feature considered is always whether the residue is DNA-contacting or not, a binary feature, while the

