Modeling Transcription Factor Binding Sites with a Supervised Learning Approach

Ericka Mochan (Western New England College) and Kevin Su (University of Pennsylvania)

2010 Bio-Grid REU Program, University of Connecticut, Department of Computer Science and Engineering

Transcription factors (TFs) are proteins that regulate the production of other proteins when bound to DNA at specific sequences. These sequences are called binding sites, and each TF can bind to multiple sites that are of various lengths and nucleotide content. The variable length causes issues in binding site detection, as the ideal manner in which to align the binding sites has yet to be ascertained. A truly optimal alignment is an exponential-time problem, so alignment strategies must be less than ideal to function in a practical setting. The success of these strategies can be quantified with a scoring mechanism. These include Consensus, Centroid, PSSM, and Berg and von Hippel (BvH), among others. Osada et al [1] reviewed these four scoring methods for binding sites of the same length. However, since true binding sites are typically not all of the same length, these mechanisms need to be modified to be applicable to current TF data. This project aims to determine an alignment strategy for binding sites of varied length for a set of training data via a supervised learning approach.

This research is important biologically because understanding the production of proteins is crucial to the understanding of the cell’s functioning as a whole. Proteins are involved in virtually all cell processes, and so comprehension of their production will further understanding of the cell. Therefore, it is important to be able to predict where on a DNA strand the binding sites may occur, and it is equally important to limit the number of false positive binding site sequences. There are several theories on the most successful way to accomplish this. Various alignment strategies and scoring mechanisms will be explored in this research to maximize the success of the alignment while also minimizing the false positive rate. To accomplish this, the leave-one-out cross-validation method will be utilized, which is based on the hypothesis that binding sites of the same TF are more alike than those of different TFs. The leave-one-out method allows testing of the true positive and false positive binding site detection.

The data used in this research will be training data as used by Osada. These data include all known binding sites for four distinct species: Saccharomyces cerevisiae, Drosophila melanogaster, mouse, and human. The program utilized will be a modification of that used by Osada to allow for the various length binding sites to be aligned and scored. The scoring of four species’ TF alignments will allow for a good amount of experimental data and then some conclusions as to the biological significance of the findings. It is assumed that the binding sites contain a “core” region where all of the sites will overlap once aligned, and that this region is important biologically. This research will help support or refute that assumption, which can then aid the binding site detection for other species in future work.