From Motif-Finding to Promoter Structure

Chun (Jimmie) Ye [1] and Eleazar Eskin [2]

Keywords: motif-finding, transcription factor binding sites, promoter regions

1 Motif-Finding

The discovery of transcription factor binding sites (TFBSs) by the analysis of promoter sequences of motif-finding is one of the most well studied problems in computational biology. Motif-finding algorithms discover statistically significant signals in the promoter regions. These signals are typically short patterns[3] or profiles[1] representing a motif or candidate TFBS. These signals are often statistically significant based on several different criteria. The signals may be over represented based on some background model. The signals may have stronger conservation in aligned genomes then what is expected or may be enriched in a set of genes with respect to a functional group or gene expression data. Each of these criteria, implicitly assume a different underlying model for the motif-finding problem. Each of these models finds different types of signals and their underlying statistical tests of often incompatible.

Once a motif-finder discovers a set of over represented signals, these are used as candidate TFBSs. However, while the algorithms for discovering over represented signals are often very sophisticated, the generation of a set of candidate TFBSs from a set of over represented signals is often done in a very ad-hoc way. The over represented signals are clustered together and heuristics are used to determine the TFBSs length.

In this project, we propose a set of post-processing techniques to make predictions for a set of TFBSs from a set of over represented signals. These techniques include statistical significance tests which incorporates many different models for motif significance. These statistical tests assign p-values for the motif under different models using different types of information such as a background set of sequences, gene expression data, positional tendencies of the signals, spacial relations of the signal with other known or predicted motifs and finally over representation in aligned genomes. This combination of statistical tests gives a complete picture of the evidence for whether a motif is an actual biological signal or just an artifact of the motif-finding algorithm.

We apply a variant of the MITRA[2] algorithm to discovering transcription factor binding sites to efficiently evaluate every possible pattern using each of the statistical tests.

In addition, in many cases, especially with signals that consist of patterns, often there are many closely related over represented signals. These signals may be either shifted variants of each other, signals of different length, or slightly different signals. We propose techniques for deciding how to merge multiple signals into a single prediction for a TFBS.

2 Promoter Modeling

Often when a researcher is interested in discovering TFBSs in a set of co-expressed genes, the researcher is interested in both known and unknown TFBSs. Typically, the a motif-finder is used to discover these TFBSs which look for any motifs.

[1] University of California, San Diego. E-mail: `yimmieg@ucsd.edu`

[2] University of California, San Diego. E-mail: `eeskin@cs.ucsd.edu`

The statistical models for known and unknown motifs are fundamentally different. This is because the number of known motifs is much smaller than the number of possible motifs. The statistical models for motif-finding algorithms make the assumption that the the motifs that they are looking for are unknown. Intuitively, if we are looking for a known transcription factor binding site, a weaker signal may still be strong evidence for the presence of the motif.

In this project, we introduce a similar set of statistical tests for evaluating the significance of observing a known TFBS in a set of sequences using various types of additional information.

By combining the algorithm and statistics for finding novel motifs with the algorithm and algorithm for finding the over represented known motifs we obtain a much richer picture of the promoter region and can quantify with p-values every part of our prediction.

3 Software Availability

The algorithm and statistical tests are available via webserver at `http://www.calit2.net/compbio/mitra`.

References

[1] T. L. Bailey and C. Elkan. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning*, 21:51, 1995.

[2] E. Eskin and P. A. Pevzner. Finding composite regulatory patterns in dna sequences. In *Special Issue Proceedings of the Tenth International Conference on Intelligent Systems for Molecular Biology (ISMB-2002) Bioinformatics.*, pages 1:S354–63, 2002.

[3] M. Sagot. Spelling approximate or repeated motifs using a suffix tree. *Lecture Notes in Computer Science*, 1380:111–127, 1998.