

Discovering Motif Interactions in Gene Promoter Regions

Ben Szekely¹ and Nir Friedman²

Keywords: *cis*-regulatory motifs, hidden markov model

1 Introduction

We seek to understand the rules governing the organization of *cis*-regulatory elements in promoter regions of *saccharomyces cerevisiae* (yeast). Starting with a library of known motifs, we construct a Hidden Markov Model where states correspond either to positions within different motifs or to background states. By learning the transition probabilities of this HMM we uncover preferences in ordering and orientations between *cis*-regulatory elements.

2 Model and Approach

We model each type of motif by a fragment of HMM that consists of two parallel chains, one for each strand the motif may appear on (forward and reverse compliment). Each chain has a state per motif position and a specific background state whose purpose is to remember that we have just emitted that motif (Figure 1). To model a network of transcription factors within a promoter region, we combine these fragments to form a larger HMM, where we allow transitions from motif-specific background states and to the start of other motifs. The probabilities of these transitions represent the likelihood that some motif follows immediately after another (where we ignore the intermediate background positions). This model is reminiscent to several recent proposals in the literature [4, 2]. However, most HMM based approaches use a single background state, and thus lose information about the order of elements in promoters.

In our application, we use the library of motifs published by Harbison et al [3] (see <http://jura.wi.mit.edu/fraenkel/download/>). Constructing an HMM from this library results in nearly 2000 states. And thus full Forward-Backward or Viterbi dynamic programming is prohibitively slow. However, we take advantage of the sparsity of our transition matrix as well as the limited number of learned transition parameters to prune dynamic programming iterations. In addition, we exploit the fact that the motif library is fixed. Thus we can prune from consideration during dynamic programming many (position,state) pairs where the likelihood of a motif is extremely small (compared to background states). This pruning can be done in a preprocessing step and then used in all subsequent parameter training.

3 Statistical Validation

To test the statistical significance of our findings, we compare our model to several variations. The null comparison model captures a situation where motifs are not inter-related. This model has states for the motifs but no motif specific background. Comparison with this model will indicate whether motif networks are statistically more significant than the presence of motifs alone. A second, more refined test, examines where the transitions following a

¹Department of Computer Science, Harvard University. E-mail: bszekely@fas.harvard.edu

²Bauer Center for Genomics Research, Harvard University. E-mail: nir@cgr.harvard.edu

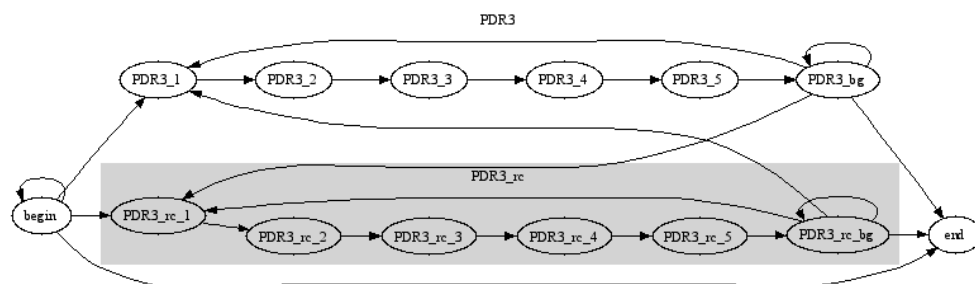


Figure 1: A Sample Model Instance

particular motif A are significant. To do so, we compare to a null model where upon exiting the states of motifs the HMM transitions to a global background data. Hence in the null model there is no memory that A was omitted.

In both tests we estimate significant by using the likelihood ratio between the learned model and the null model. We then use synthetic datasets generated from the null model to get an empirical estimate of this statistic under the null hypothesis.

4 Visualization and Results

To visualize our discovered motif networks, we generate a graph with the motifs as nodes and transition probabilities as directed edges for significant transitions. To visualize significant individual motif to motif transitions we generate a colored array.

Thus far, we have experimented with motifs and yeast 5', 500bp UTR as a whole compared to a model of the promoters of genes within specific GO [1] annotations. Our visualizations have confirmed some known motif occurrences in the cell-cycle sequences and have uncovered some interesting motif transitions. Our immediate future work will involve extensive testing to discover some real biological phenomena.

References

- [1] Ashburner, M. et al. "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium" *Nat. Genet.* **25**:25-29 2000.
- [2] Bailey et al. "Searching for statistically significant regulatory modules" *Bioinformatics*, **29** Suppl 2 2003
- [3] Harbison et al. Transcriptional regulatory code of a eukaryotic genome. *Nature*. **431**:99-104. 2004
- [4] Xing et al. *LOGOS, a modular Bayesian model for de novo motif detection*. IEEE Computer Society Bioinformatics Conference, CSB2003.