

Predicting Transcription Factor Binding Sites Using a Bayesian Allocation Method

Dustin T. Holloway¹
dth128@bu.edu

Dr. Charles DeLisi¹
delisi@bu.edu

¹ Department of Bioinformatics, Boston University, Boston, MA
U.S.A.

Keywords: gene regulation, transcription factor binding site, Bayesian analysis, Position Weight Matrix, PWM, motif.

1 Introduction

Transcription factor binding sites (TFBS) in gene promoter regions are often predicted by comparing upstream sequence regions to binding site Position Weight Matrices (PWMs). Although PWMs are more reliable than simple consensus string matching in predicting a true binding site, they generally result in high numbers of false positive hits. This study attempts to reduce the number of false positive matches by assigning a probability to a PWM hit based on a Bayesian statistical measure which incorporates several types of biological data.

The recent deluge of genomic information makes such an analysis practical, and this technique is demonstrated here in the *Saccharomyces cerevisiae* genome. So far, two methods have been employed to strengthen the prediction of a true TFBS; binding site redundancy, and phylogenetic conservation. Binding site redundancy refers to the number of times a particular transcription factor's binding motif is discovered in the upstream region of a gene. More instances of a motif in a promoter region is expected to translate into a higher probability of a TF binding an upstream region. Phylogenetic conservation takes into account the number of orthologous upstream regions in closely related species that contain a particular binding site, noting that the higher the degree of conservation, the more a motif sequence is likely to be a true binding site. The publication of large datasets of protein-DNA binding data (ChIP) in the literature allows for the estimation of the number of true positives in the predicted binding data; furthermore, the near complete sequencing of several yeast species of the sensu stricto and sensu lato groups makes accurate conservation analysis of predicted TFBSs possible.

2 Method and Results

Upstream regions of *S. cerevisiae*, *S. bayanus*, *S. mikatae*, and *S. paradoxus* genes were obtained as annotated by Kellis et al. [1] from the supplementary information published on their website [2]. Since the start and stop sites of transcription for these genes were generated from multiple sequence alignments of yeast sensu stricto species, these annotations are expected to be more accurate than previous yeast gene annotations. Upstream regions are defined as proceeding from the transcriptional start codon to the next upstream genome feature (i.e., end of transcription for nearest upstream gene).

All upstream gene regions were then masked using the dust [3] algorithm to exclude low complexity sequences from further analysis. The MotifScanner [4] algorithm was used to scan all upstream regions for transcription factor binding sites using PWM models. This algorithm requires a background sequence model, which in this case is a transition matrix of a 3rd order Markov model generated from the masked upstream regions of each genome individually. The PWMs used were adapted from all Transfac *S. cerevisiae* count matrices [5].

Perl scripts were used for all parsing of MotifScanner output files. Transcription factor-gene binding data was combined from the Transfac database, genome-wide ChIP data [6], and data curated by Lee et al. from the Saccharomyces Genome Database (file available as supplementary information to [6]). For each TF and its associated binding site, the number of redundant binding sites upstream of each gene were counted. Similarly, for each set of orthologous yeast genes, the presence of each TF's binding site in the orthologous sequence was counted. These counts of motifs present in genes and orthologs are each represented in matrices where rows indicate TF motifs and columns represent genes. Numbers in this matrix are counts of how many times a motif is found in the upstream region of a gene for the redundancy data or in how many orthologous sequences a motif is found for the conservation data.

A computational comparison of each of these matrices to the assembled binding data results in probability associations showing the relationship of redundancy or conservation to true binding. For example, the probability that motifs with redundancy of 4 in the upstream region of a gene are actually bound by a TF is 0.3.

Since the goal is to be able to predict from the motif detection whether a promoter with a given motif will be bound by given TF, we can derive this relationship from Bayes rule:

$$F(T|k) = \frac{F(k|T)F(T)}{F(k|\bar{T})F(\bar{T}) + F(k|T)F(T)}$$

where k represents number of motifs in an upstream region (for redundancy data), T indicates true binding, and \bar{T} means no binding. This representation is valid for both the redundancy data and the conservation data, and, in fact, both of these can be combined into one statistical measure:

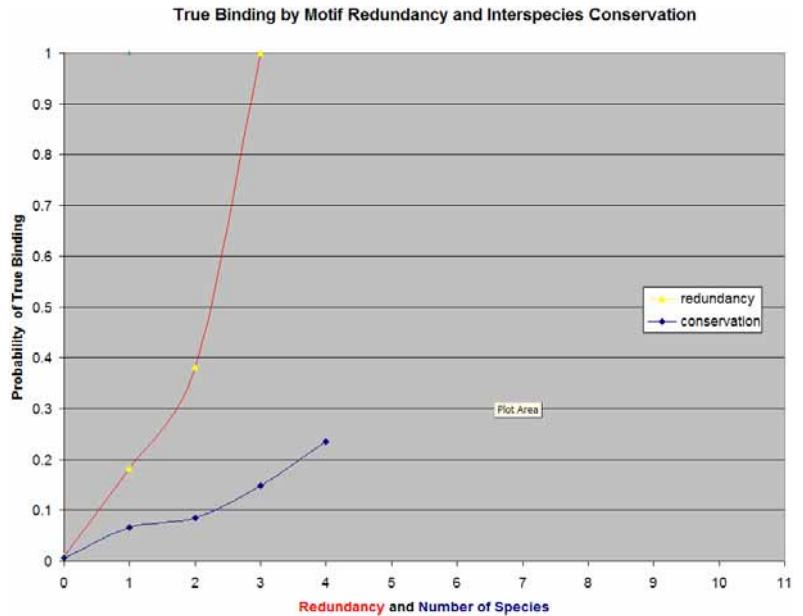
$$F(T|k_1, k_2) = \frac{F(k_1|T)F(k_2|T)F(T)}{[F(k_1|T)F(T) + F(k_1|\bar{T})F(\bar{T})][F(k_2|T)F(T) + F(k_2|\bar{T})F(\bar{T})]}$$

where k_1 equals the number of times a particular motif is present in the upstream region of a gene and k_2 equals the number of orthologous sequences in which the gene is present.

Probability of true transcription factor binding is predicted significantly by interspecies conservation of motif element and motif repetition upstream of genes as shown in Fig 1.

2.1 Figures

Figure 1. Shows probabilities of true binding given a number times a motif is repeated upstream of a particular gene or the number of species in which a given binding site is conserved.



3 Discussions

Here we conclude that the redundancy of a motif in the upstream region of a gene does in fact improve the prediction that the motif is bound by its associated TF. Furthermore, the probability that a discovered binding site is a true binding site increases as the number of genomes in which it is found increases. It is our belief that these methods and others, when combined, will dramatically improve the prediction of true binding over any one method alone. Along these lines, the use of a Bayesian allocation procedure to combine biological evidence can be extended to include any type of data that can conceivably improve the prediction of true binding. In the future it will be possible to incorporate detection of dense motif clusters (e.g., as detected by ClusterBuster[7]) and microarray expression data to further refine and filter the outputs of PWM scanning algorithms.

References

- [1] M. Kellis, N. Patterson, M. Endrizzi, B. Birren and E. S. Lander, Sequencing and Comparison of Yeast Species to Identify Genes and Regulatory Elements, *Nature* 423 241-254,2003.
- [2] M. Kellis and et al., http://www.broad.mit.edu/annotation/fungi/comp_yeasts/, 2003.
- [3] R. L. Tatusov and D. J. Lipman,dust,unpublished work.
- [4] S. Aerts, G. Thijs, B. Coessens, M. Staes, Y. Moreau and B. De Moor, Toucan:Deciphering the Cis-Regulatory Logic of Coregulated Genes, *Nucleic Acids Research* 31 1753-1764,2003.
- [5] V. Matys and et al., TRANSFAC: Transcriptional Regulation, from Patterns to Profiles, *Nucleic Acids Research* 31 374-378,2003.
- [6] I. T. Lee and et al., Transcriptional Regulatory Networks in *Saccharomyces cerevisiae*, *Science* 298 799-804,2002.
- [7] M. C. Frith, M. C. Li and Z. Weng, Cluster-Buster: Finding Dense Clusters of Motifs in DNA Sequences, *Nucleic Acids Research* 31 3666-3668,2003.