Pacific Symposium on Biocomputing 6:151-164 (2001)

PROMOTER REGION-BASED CLASSIFICATION OF GENES

P. PAVLIDIS
Columbia Genome Center, Columbia University, pp175@columbia.edu

T. S. FUREY
Department of Computer Science, UC Santa Cruz, booch@cse.ucsc.edu

M. LIBERTO
Department of Biological Sciences, Columbia University, ml391@columbia.edu

D. HAUSSLER
Department of Computer Science, UC Santa Cruz, haussler@cse.ucsc.edu

W. N. GRUNDY
Department of Computer Science, Columbia University, bgrundy@cs.columbia.edu

Abstract

In this paper we consider the problem of extracting information from the upstream untranslated regions of genes to make predictions about their transcriptional regulation. We present a method for classifying genes based on motif-based hidden Markov models (HMMs) of their promoter regions. Sequence motifs discovered in yeast promoters are used to construct HMMs that include parameters describing the number and relative locations of motifs within each sequence. Each model provides a Fisher kernel for a support vector machine, which can be used to predict the classifications of unannotated promoters. We demonstrate this method on two classes of genes from the budding yeast, *S. cerevisiae*. Our results suggest that the additional sequence features captured by the HMM assist in correctly classifying promoters.

1 Introduction

The regulation of transcription is largely dependent on the complex interactions of DNA binding proteins with regulatory sequence elements in the promoter regions of genes. While there is a wealth of information regarding promoters and the basis of their influence on gene transcription, it is usually very difficult to identify coregulated genes. The availability of the sequences of entire genomes, coupled with computational methods for sequence analysis and classification, provides an opportunity to perform this task automatically on large numbers of genes. In this paper, we address the problem of identifying coregulated genes based on promoter sequences.

Any approach to this problem should account for the various kinds of information present in promoters. A simple approach to classifying promoters would rely on the presence of short (6-12 base pairs), highly conserved motifs that function as binding sites for proteins called transcription factors. Many computational techniques for identifying such motifs exist, including several recently reported methods.[1,2,3,4,5] However, the small size and degeneracy of these binding sites make classification based on these sequences difficult. It is well known that transcription regulatory elements include features in addition to these binding sites that could be exploited by a more sophisticated classifier. For example, the promoters of many genes contain poly(dA-dT) elements, which, while not comprising a specific motif, have been shown to be critical for the regulation of transcription.[6] The presence of multiple copies or combinations of motifs can also be critical to normal transcriptional regulation, as is the spacing between two motifs that form a binding site for a heterodimeric transcription factor.[7] Finally, the DNA sequences flanking highly conserved binding motifs can have profound effects on the ability of the transcription factor to bind efficiently and regulate gene transcription.[8] Thus, the features of a promoter region that function in transcriptional regulation include important elements in addition to the canonical transcription factor binding motifs. We hypothesize that classification can be improved by capturing these additional features.

Here we describe a method for automatically classifying promoters based upon the presence and relative positions of one or more transcription factor binding sites. The method consists of first building a motif-based hidden Markov model (HMM) from a collection of transcription factor binding site (TFBS) motifs. In addition to the nucleotide distributions at each position within the motifs, the HMM includes parameters that capture the number and relative locations of motif occurrences within each sequence. This model then acts as the kernel function for a support vector machine (SVM).[9,10,11] SVMs have previously proven useful in classifying proteins from primary sequence [12] and in classifying genes from microarray expression data.[13] In the current application, the SVM uses the motif-based HMM to learn to discriminate between a given set of promoters from coregulated genes and a second set of "negative example" promoters. We demonstrate this method on two classes of genes in *S. cerevisiae.*

2 Algorithm

The process of building a promoter-based classifier consists of five steps. First, TFBS motifs are identified from the promoter regions of a given set of genes that are either known or predicted to be coregulated. Second, the motifs and the promoter region sequences are used to build a motif-based HMM. Third, the promoter regions of a larger set of genes, which includes the coregulated genes plus a large number of genes that are known not to be coregulated, are compared to the HMM, and the gradients of the model parameters are computed with respect to each sequence. Fourth, these gradient vectors are used to train an SVM. In the final step, the trained SVM is used to search a database of un-annotated genes. For each gene, the SVM predicts whether it belongs in the original class of coregulated genes. Thus, the algorithm takes as input two sets of promoter sequences (positives and negatives) and an unannotated database. The output is a set of labels (positive or negative) for each promoter sequence in the database. The following section describes in detail each step of this method.

Motif models can be generated using any motif-discovery algorithm or can be constructed by hand from the literature. In this work, we generate motif models using Improbizer, which is part of the cis-Site Seeker software package (www.cse.ucsc.edu/~kent/improbizer). Similar to MEME,[14] Improbizer uses expectation-maximization to discover motifs. We use Improbizer mainly due to its ability to use a first-order Markov background model, thus reducing the likelihood of finding low-complexity motifs such as polyA tracts. For each class of genes, we instruct Improbizer to find five motifs, each of which occurs at most twice within a gene, allowing motif occurrences to be present on either strand. Each motif is represented as a matrix, where each column specifies the probability of a nucleotide (A,C,G,T) occuring at that position in the motif. A motif cannot be less than seven nucleotides in length, but a strong preference is made for shorter, highly conserved sequences, controlled by setting the *Restrain Expansionist Tendencies* parameter to 5.0. We use all positive and negative sequences to create a first-order Markov background model. A score for each promoter sequence relative to a particular motif is calculated by summing the log-probability scores for the two subsequences that best match the motif. Each of the five motifs is then scored by averaging the promoter sequence scores for that motif. Of the five motifs, the two with the highest scores are retained for use in the next step.

In step two, we use Meta-MEME [15] to construct a motif-based hidden Markov model from the TFBS motif matrices generated by Improbizer. Each model consists of a collection of five completely connected motif models: a forward and reverse-complement version of each of the two Improbizer motifs, as well as a "background" motif. The reverse complement motifs are important because transcription factor binding sites can occur on either strand of DNA. The background motif consists of two states, each with emission probabilities the same as the base frequencies of the nucleotides in all of the promoter sequences. This motif permits modeling of sequences that do not contain any of the TFBS motifs. The five fixed-length motif models are connected via variable-length spacer states. The emission probabilities at the spacer states are the same as the background motif. The self-transition probability at each spacer state is 0.999, and the exit transition is 0.001. Thus, a typical path through this model for a sequence involves alternating between emitting bases from a spacer state and emitting bases from the states in a motif. Though it is possible to further train the parameters of the Meta-MEME model via expectation-maximization, doing so in conjunction with an SVM can lead to over-training, because the same sequences would be used to train the HMM and the SVM. Therefore, we use the Meta-MEME model without further training.

The third step consists of comparing the Meta-MEME model to the complete set of promoter sequences and computing the model parameter gradient vector with respect to each sequence. This technique was introduced by Jaakkola and Haussler,[11] who subsequently showed it to be highly effective for detecting remote protein homologies.[12] Here we extend this method to classifying promoter region sequences. The gradient, or Fisher score, vector for observation X given a generative probability model H with parameters Θ is defined as $\vec{\mathbf{U}} = \nabla_\theta \log P(X|H,\Theta)$. For an HMM, the components of this vector associated with the parameters for the emission probabilities in the states can be computed as follows:

$$\vec{\mathbf{U}}_{ij} = \frac{E_j(i)}{e_j(i)} - \sum_k E_j(k),$$

where $E_j(i)$ is the number of times that nucleotide i is observed in state j, and $e_j(i)$ is the emission probability for nucleotide i in state j. Likewise, for transition probabilities from states, the components are calculated as:

$$\vec{\mathbf{V}}_{ij} = \frac{T_j(i)}{t_j(i)} - \sum_k T_j(k),$$

where $T_j(i)$ is the number of times a transition to state i is taken from state j, and $t_j(i)$ is the transition probability for transitions to i from state j.[12] The counts $E_j(i)$ and $T_j(i)$ can be computed easily using the forward algorithm.[16] Each such component of the Fisher score vector measures how a given parameter in the model contributes to the total posterior probability. For each positive and negative promoter region sequence, we create a Fisher score vector that consists of Fisher scores for all transitions in the model except those within the motif sequences, which are fixed at a probability of 1, and for all emission probabilities.

These vectors are then used to train a support vector machine to discriminate between the positive and negative promoter region sequences. We use a particularly simple SVM optimization algorithm that was introduced by Jaakkola *et al.*[12] Our implementation is described in Brown *et al.*[13] and is freely available on the web (www.cs.columbia.edu/compbio/svm). Briefly, the SVM learning algorithm constructs a maximum-margin hyperplane that separates the negative and positive examples in a training set. The margin is soft, in the sense that it permits some misclassifications, as might be expected in a noisy data set. The hyperplane is calculated in an implicit feature space, whose dimensionality depends upon the choice of kernel function used. We use the radial basis kernel function $K(\vec{\mathbf{X}}, \vec{\mathbf{Y}}) = exp(-||\vec{\mathbf{X}} - \vec{\mathbf{Y}}||^2/2\sigma^2)$, which has previously been shown to provide good recognition performance for protein sequence and gene expression classification.[12,13] Once the separating hyperplane is constructed, the SVM can be used to predict the classifications of previously unseen examples. For a more complete explanation of our SVM methods, see [13,17] and the accompanying web page (www.cse.ucsc.edu/research/compbio/genex).

3 Results

To test our method, we analyze two classes of genes in the budding yeast *S. cerevisiae*, cytoplasmic ribosomes and nucleosomal complex proteins (core histones), as defined by the MIPS Yeast Genome Database.[18] These classes were selected because they contain patterns of motif occurences that we believed could be learned using our method. The majority of cytoplasmic ribosomes are regulated by the Rap1p transcription factor, and this factor usually binds in pairs in close proximity.[19] The histone genes are coregulated during the cell cycle and share a distinctive promoter structure.[20] The coordinate expression of the genes in these classes has also been demonstrated in recent DNA microarray experiments.[21,17]

For training, we use the 2465 annotated genes [a] originally used by Eisen *et al.*[21] Predictions are then made on the remaining 3807 yeast genes, many of which are unannotated. Upstream sequences consisting of 1000 base pairs for all of the genes were obtained from the Stanford Genome web site (genome-www.stanford.edu). Complete data and results from these experiments are available at www.cs.columbia.edu/compbio/prom-svm.

Only one motif discovered by Improbizer for these two classes clearly corresponds to a known TFBS, that for Rap1p. Our motif has a consensus TWWACAYCCRTACATYWY. While Rap1p binding sites are found in many genes, the

[a] This is the updated version of the orginal 2467 gene dataset.

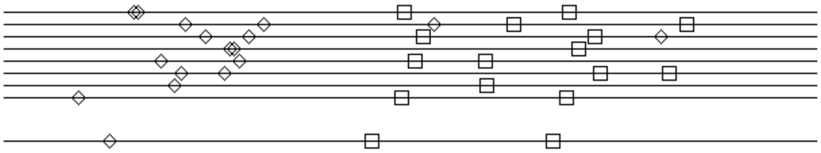

Figure 1: **Patterns of motif occurences in the nucleosomal promoters.** Each line corresponds to a single 1000-base pair promoter region. Squares and diamonds indicate occurrences of motifs 1 and 2, respectively. The top eight lines are promoters of known nucleosomal genes. The final line is a promoter from a gene (YOR084W) identified by the Meta-MEME + SVM method.

one we discovered may have features specific to the ribosomal class, which will aid in classification performance.[19]

Three of the four motifs found by Improbizer are statistically surprising. Improbizer reports motif scores that are scaled relative to the given data set. We therefore assess the quality of a given motif by running Improbizer five times on shuffled versions of the given dataset and comparing the relative motif scores of the original motifs to the scores generated by the control runs. For the ribosomal genes, the Rap1p motif has a score 41.8% greater than the corresponding average control run score. The score of the second motif, however, exceeds the average score of the second control run motif by only 6.0%. For the nucleosomal genes, both motif scores, with consensus sequences TTACCACCK and YHCGGGCGM, exceed the control run scores by more than 20%.[b] We therefore expect the first ribosomal motif and both nucleosomal motifs to provide useful classification.

The observed patterns of motif occurrences support the hypothesis that relative motif positions are conserved throughout each class. An illustration of this conservation in the nucleosomal complex genes is shown in Figure 1. The obvious pattern of occurences suggests that using the number and locations of the motifs will be useful in recognizing members of the class.

Figure 2 presents a visualization of the Fisher score vectors created from the cytoplasmic ribosomal proteins. The vectors of class members are clearly similar to one another and different from the vectors of genes not in the class. The vectors were created using both Improbizer motifs, but the figure illustrates that the second motif does not provide a sharp contrast between ribosomal and non-ribosomal genes.

In order to quantify the relative discriminative power of parameters in the

[b] Improbizer found one slightly more surprising motif with consensus ATATATAAA, but we elected not to use this motif.

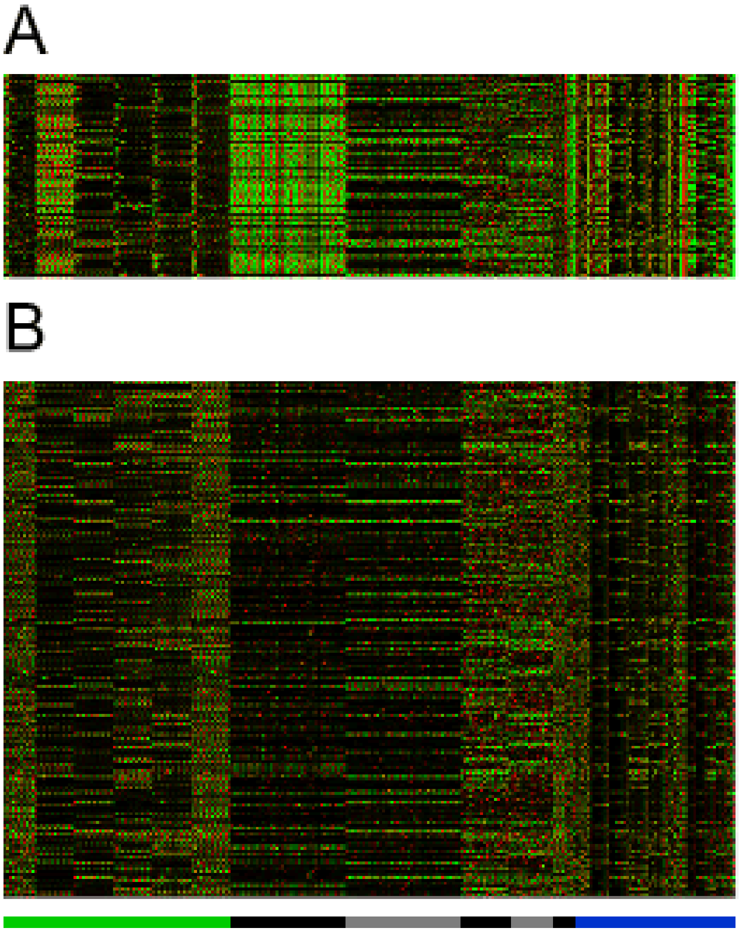

Figure 2: **Fisher score vectors for a representative class.** **A**. Fisher score vectors for the 121 cytoplasmic ribosomal protein gene 5' UTRs. Green indicates negative values of the Fisher score, red positive, black represents zero. **B**. Fisher score vectors for a representative subset of the remaining genes in the dataset. The colored bars at the bottom apply to both **A** and **B** and indicate the (arbitrary) organization of the scores in this image. Green bar: Spacer states. Black bars: Forward-strand versions of motifs. Grey: Reverse complements of motifs. Blue: State transitions. This image was generated by TreeView (rana.stanford.edu).

model, we use the Fisher criterion score[22] to identify components of the Fisher score vectors most relevant to the given classification.[c] For a given feature j, we compute the mean and standard deviation of that feature across the positive examples (μ_j^+ and σ_j^+, respectively) and across the negative examples (μ_j^- and σ_j^-). The Fisher criterion score, $(\mu_j^+ - \mu_j^-)^2/((\sigma_j^+)^2 + (\sigma_j^-)^2)$, gives higher values to features whose means differ greatly between the two classes, relative to their variances.

This analysis shows that many of the non-motif features of the HMM are strongly discriminative. In both classes, the transitions between motifs, which are at the right in Figure 2, are among the most significant features. In particular, transitions leading to the null motif are consistently emphasized, showing that the presence or absence of the specific motifs generated from the class is a key feature, as expected. In the cytoplasmic ribosomal class model, self-transitions for several spacer states are discriminative, suggesting that differences in the location or spacing of motifs within the promoter are informative. Among these states is the spacer state on the self-transition to the Rap1p binding site motif, which is often found in closely-spaced pairs within the 5' UTRs of the ribosomal protein genes. In addition, some motif states have high Fisher criterion scores. In general, these scores correspond to states in which either one base is highly conserved, or in which one base is seen with a very small probability. This tendency reflects the preference of a given transcription factor for bases in certain positions. In the ribosomal class, none of the motif positions in the less-conserved second motif is strongly discriminative. Thus, given a set of negative training examples, the SVM must learn to focus on the highly discriminative parameters and ignore less useful parameters.

Classification of the unannotated genes produces 36 predictions in the ribosomal class and 1 prediction in the nucleosomal complex class. The single nucleosomal gene prediction, YOR084W, is a putative lipase and is shown in Figure 1. The relative locations of the TFBS motifs in this promoter closely matches those of other genes in the family, suggesting that this gene belongs in the class. Table 1 lists the ribosomal gene predictions. As can be seen, eight of the 36 have been annotated elsewhere as cytoplasmic ribosomal proteins. Another five are identified as transporters.

The availability of DNA microarray gene expression data for yeast provides an additional means of validating the prediction results. A set of coregulated genes should exhibit strongly correlated expression profiles. We use published gene expression data[21] to compute the average Pearson correlation coefficient

[c] These two names both refer to R. A. Fisher, but the criterion score and score vectors are otherwise unrelated.

Table 1: **Cytoplasmic ribosomes predictions** Listed are the genes from the unannotated class predicted to be regulated similar to the cytoplasmic ribosomes. The discriminant value calculated by the SVM is shown, with higher discriminants indicating a stronger prediction. Annotations for these genes are given when available. Genes in boldface have correlated gene expression profiles, as described in the text.

Gene	Disc	Annotation
YMR194W	0.546	RPL43B; cytoplasmic ribosomal protein L36A
YBR190W	0.495	
YBR220C	0.391	acetyl-coenzyme A transporter (AcCoAT) family
YBL049W	0.303	
YHR002W	0.300	Mitochondrial carrier family member
YJL049W	0.275	
YNL101W	0.239	amino acid/auxin permease (AAAP) family
YDL133C-A	0.222	RPL41B; cytoplasmic ribosomal protein L41B
YMR230W	0.200	RPS10B; cytoplasmic ribosomal protein S10B
YMR144W	0.186	
YOR292C	0.182	
YPR068C	0.165	HOS1; putative histone deacetylase
YGR260W	0.155	TNA1; nicotinic acid transporter
YPL034W	0.117	
YML010C-B	0.108	
YDR281C	0.097	PHM6; involved in phosphate metabolism
YJR094W-A	0.095	RPL43B; cytoplasmic ribosomal protein L43B
YNR062C	0.092	
YKL151C	0.090	
YML073C	0.083	RPL6A; cytoplasmic ribosomal protein L6A
YDR467C	0.082	
YPL189W	0.080	GUP2; putative active glycerol transporter
YPL249C-A	0.079	RPL36B; cytoplasmic ribosomal protein L36B
YLR362W	0.068	STE11; kinase involved in mating signalling
YML026C	0.057	RPS8B; cytoplasmic ribosomal protein S8B
YOR249C	0.053	APC5; subunit of the Anaphase Promoting Complex
YDR479C	0.050	
YLR287C-A	0.047	
YGR283C	0.042	
YIL063C	0.042	YRB2; Ran-GTPase-binding protein
YCL036W	0.029	
YFR031C-A	0.029	RPL2A; cytoplasmic ribosomal protein L2A
YMR014W	0.025	
YGL136C	0.013	Possible S-adenosylmeth-dependent methyltransferase
YDR325W	0.009	YCG1; involved in chromatin structure
YBR180W	0.001	DTR1; dityrosine transporter

between the expression profile of each predicted gene with the profiles of other genes in the class. We then compare this average similarity with similar values derived from 1000 randomly selected genes. This analysis shows that the average correlation for the predicted nucleosomal gene (0.308) is more than one standard deviation away from the mean (mean = 0.026, stdev = 0.170). The analysis also identifies eight predicted ribosomal genes (indicated in bold in Table 1) with strongly correlated expression profiles.

4 Discussion

We have demonstrated a method for extracting from promoter sequences higher-order features that capture information about the occurences of transcription factor binding site motifs. We use these features for the supervised learning of classes of coregulated genes. The resulting classifier combines generative and discriminative models using the Fisher kernel method,[11] which has previously been shown to produce excellent protein family classification performance.[12] The HMM provides an understandable probabilistic model capable of handling variable-length sequences and missing data. The SVM provides improved classification performance relative to the HMM via the SVM's ability to learn from both positive and negative examples.

While the method we describe is promising, it is clear that the first step to obtaining good discrimination is generating highly conserved, specific motifs. Thus the success of the method is dependent on the motif-discovery phase. For example, the promoter regions of nucleosomal complex proteins are known to contain multiple occurrences of an activating TFBS, consensus GCGAAAAANTNNGAAC, and six of the eight regions also contain a negative site, consensus TNNACGCTNAANGNC.[20] Some ribosomal genes are known to be regulated by abf1, consensus RTMRYBNNNNACG, instead of or in addition to Rap1p.[19] These motifs were not found using our current settings of Improbizer, most likely due to our preference for shorter motifs. In the future, we plan to explore finding motifs using multiple parameter settings and possibly other methods in order to obtain better motifs. We expect that this will improve the classification performance of our method.

In addition to motif content and relative motif locations, other kinds of information may be relevant to the classification of coregulated genes. For example, the base composition of the spacer regions between motifs may relate to DNA bendability, which in turn influences the ability of transcription factors to bind to the DNA.[23] Thus, future promoter models should explicitly include information about bendability. Further improvements might come from including general promoter sequence elements such as the TATA box. While not specific to a particular class of genes, they would assist in the accurate modeling of the overall promoter structure.

One of the greatest challenges in developing a classifier of coregulated genes is identifying trustworthy training and test sets. The MIPS database is useful in this regard because the annotations therein are largely based upon wet lab experiments. However, MIPS does not provide classifications based upon coregulation. In this paper, we use two protein complexes, under the hypothesis that such complexes tend to share regulatory elements. This hypothesis does not always hold true. Consequently, apparent errors in the predictions made by the classifier may result from the inclusion of non-coregulated genes in the class, or from the exclusion of coregulated genes from the class. An alternative to MIPS would be to use classes of genes that are known to be coregulated, such as the ones identified by Van Helden *et al.*,[24] but such classes rarely are accompanied by large sets of "negative example" genes that are known not to be coregulated. A large-scale comparison of the method presented here with other promoter classification techniques must await the availability of a suitable gold standard database of coregulated genes. This database may be available soon from microarray analyses of transcription factor mutants.

The correlation analysis of gene expression data with promoter classification predictions presented above is only the first step toward combining the information available in gene expression and promoter sequence data. In future work, we will develop Fisher kernel models that are capable of learning from heterogeneous data sets that include combine promoter sequence and microarray expression data.

Acknowledgments

The authors wish to thank Jim Kent for providing help with the Improbizer software. WNG is funded by an Award in Bioinformatics from the PhRMA Foundation.

References

1. YJ Hu, S Sandmeyer, C McLaughlin, and D Kibler. Combinatorial motif analysis and hypothesis generation on a genomic scale. *Bioinformatics*, 16:222–232, 2000.
2. J van Helden, AF Rios, and J Collado-Vides. Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *NAR*, 28:1808–1818, 2000.

3. LJ Jensen and S Knudsen. Automatic discovery of regulatory patterns in promoter regions based on whole cell expression data and functional annotation. *Bioinformatics*, 16:326–333, 2000.
4. S Sinha and M Tompa. A statistical method for finding transcription factor binding sites. In *ISMB*, 2000.
5. AM McGuire, JD Hughes, and GM Church. Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. *Genome Research*, 10:744–757, 2000.
6. KA Koch and DJ Thiele. Functional. *JBC*, 274:23572–23760, 1999.
7. KT Arndt, C Styles, and GR Fink. Multiple global regulators control HIS4 transcription in yeast. *Science*, 237:874, 1987.
8. AG Hinnebusch, G Lucchini, and GR Fink. A synthetic his4 regulatory element confers general amino acid control on the cytochrome c gene (CYC1) of yeast. *PNAS*, 82:498–503, 1985.
9. VN Vapnik. *Statistical Learning Theory.* Wiley, 1998.
10. N Cristianini and J Shawe-Taylor. *An Introduction to Support Vector Machines.* Cambridge UP, 2000.
11. T Jaakkola and D Haussler. Exploiting generative models in discriminative classifiers. In *NIPS 11.* Morgan Kauffmann, 1998.
12. T Jaakkola, M Diekhans, and D Haussler. Using the Fisher kernel method to detect remote protein homologies. pages 149–158, Menlo Park, CA, 1999. AAAI Press.
13. MPS Brown, WN Grundy, D Lin, N Cristianini, C Sugnet, TS Furey, M Ares, and D Haussler. Knowledge-based analysis of microarray gene expression data using support vector machines. *PNAS*, 97:262–267, 2000.
14. TL Bailey and CP Elkan. Fitting a mixture model by expectation-maximization to discover motifs in biopolymers. In *ISMB*, pages 28–36. AAAI Press, 1994.
15. WN Grundy, TL Bailey, CP Elkan, and ME Baker. Meta-MEME: Motif-based hidden Markov models of protein families. *CABIOS*, 13:397–406, 1997.
16. LR Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, 77:257–286, 1995.
17. MPS Brown, WN Grundy, D Lin, N Cristianini, C Sugnet, M Ares, and D Haussler. Support vector machine classification of microarray gene expression data. Technical Report UCSC-CRL-99-09, UC Santa Cruz, 1999.
18. HW Mewes, D Frishman, C Gruber, B Geier, D Haase, A Kaps, K Lemcke, G Mannhaupt, F Pfeiffer, C Schüller, S Stocker, and B Weil. MIPS: a database for genomes and protein sequences. *NAR*, 28:37–40, 2000.

19. RF Lascaris, WH Mager, and RJ Planta. DNA-binding requirements of the yeast protein Rap1p as selected in silico from ribosomal protein gene promoter sequences. *Bioinformatics*, 15:267–277, 1999.

20. MA Osley, J Gould, S Kim, M Kane, and L Hereford. Identification of sequences in a yeast histone promoter involved in periodic transcription. *Cell*, 45:537–544, 1986.

21. M Eisen, P Spellman, PO Brown, and D Botstein. Cluster analysis and display of genome-wide expression patterns. *PNAS*, 95:14863–14868, 1998.

22. C Bishop. *Neural Networks for Pattern Recognition.* Oxford UP, 1995.

23. RF Lascaris, E de Groot, PB Hoen, WH Mager, and RJ Planta. Different roles for Abf1p and a T-rich promoter element in nucleosome organization of the yeast RPS28A gene. *NAR*, 28:1390–1396, 2000.

24. J van Helden, B Andre, and J Collado-Vides. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *JMB*, 281:827–842, 1998.