BIOINFORMATICS

Vol. 21 Suppl. 1 2005, pages i387–i393
doi:10.1093/bioinformatics/bti1002

A motif-based framework for recognizing sequence families

Roded Sharan[1,]* and Eugene W. Myers[2]

[1]School of Computer Science, Tel-Aviv University, Tel-Aviv 69978, Israel and
[2]Computer Science Division, University of California at Berkeley, 387 Soda Hall,
Berkeley, CA 94720, USA

Received on January 15, 2005; accepted on March 27, 2005

ABSTRACT

Motivation: Many signals in biological sequences are based on the presence or absence of base signals and their spatial combinations. One of the best known examples of this is the signal identifying a core promoter—the site at which the basal transcription machinery starts the transcription of a gene. Our goal is a fully automatic pattern recognition system for a family of sequences, which simultaneously discovers the base signals, their spatial relationships and a classifier based upon them.

Results: In this paper we present a general method for characterizing a set of sequences by their recurrent motifs. Our approach relies on novel probabilistic models for DNA binding sites and modules of binding sites, on algorithms to study them from the data and on a support vector machine that uses the models studied to classify a set of sequences. We demonstrate the applicability of our approach to diverse instances, ranging from families of promoter sequences to a dataset of intronic sequences flanking alternatively spliced exons. On a core promoter dataset our results are comparable with the state-of-the-art McPromoter. On a dataset of alternatively spliced exons we outperform a previous approach. We also achieve high success rates in recognizing cell cycle regulated genes. These results demonstrate that a fully automatic pattern recognition algorithm can meet or exceed the performance of hand-crafted approaches.

Availability: The software and datasets are available from the authors upon request.

Contact: roded@tau.ac.il

1 INTRODUCTION

We initially began this work with the problem of trying to identify core promoters—the sites at which the basal transcription machinery starts the transcription of a gene. This is one of the earliest sequence signals studied in computational biology dating back to Gary Stormo's PhD thesis on using perceptrons in 1981. It remains one of the hardest signals to find, harder even than the now trendy *cis*-regulatory signals, also known as distal and proximal promoters, which serve as the binding sites of complexes that interact with and modulate the activity of core promoters. Solving the problem of finding core promoters is very important as most gene prediction programs routinely miss the 5′ exon because they are geared to recognize coding sequence. Among other implications, this has great impact on the accuracy of the upstream region in which one looks for *cis*-regulatory control.

We wished to study a classifier of the signal by the examination of a collection of positive and negative examples. The basic idea is to first recognize potentially distinguishing attributes or patterns and then study which combinations of these attributes discriminate positive from negative examples. The idea is quite natural and there have been several other attempts along these lines (Pavlidis *et al*., 2001; Ben-Hur and Brutlag, 2003), mainly focusing on the classification task. The specific problem of recognizing eukaryotic core promoters has been studied by several authors and various approaches have been reported for it, including neural networks (Reese, 2001), linear discriminant analysis (Hannenhalli and Levy, 2001) and hidden Markov models (Ohler *et al*., 2002). The last method, called McPromoter, is the best in-class and hand-crafted classifier for Drosophila core promoters based on a great deal of human analysis and insight.

Here, we present a unified framework for the task of recognizing sequence families. The framework consists of two components: (1) algorithms that recognize unusual patterns or attributes of a number of types within the training dataset and (2) a support vector machine (SVM) that uses the attributes studied for the classification. Specifically, we study sequence motifs that discriminate positive from negative examples. We also study discriminative sequence modules, consist of spatial combinations of motifs. The studying relies on novel probabilistic models for these signals. In addition, we made our system extensible by permitting the introduction of hand-crafted attributes if desired. These particular choices of base attribute classes to search over the result in a classifier whose performance on the core promoter problem is comparable with the state-of-the-art McPromoter (Ohler *et al*., 2002) and

*To whom correspondence should be addressed.

exceeds that of NNPP (Reese, 2001). Moreover, we came to realize that the method extends well beyond our original goal and to illustrate this we apply it here to the problems of classifying alternatively spliced exons, in human, and recognizing genes that are under cell-cycle control in yeast. In comparison with a previous approach for detecting alternatively spliced exons, we are able to show increased sensitivity of the predictions.

2 METHODS

We tackle the following classification problem: the input consists of a training set of sequences with positive and negative examples, and a test set; the goal is to devise a classifier for the positive examples that will best discriminate between positives and negatives on the test set. We devise a two-phase scheme for this problem: in the first phase we use the training data to study attributes (features) that are prevalent in the positive sequences compared with background (negative) sequences. The attribute vector of each sequence consists of three types of attributes: (1) discriminative motifs, (2) discriminative modules of motifs and (3) external attributes that are unique to the specific application. In the second phase we train a SVM for the classification problem using the attributes studied as sequence features. The two phases are described in detail in the following sections.

2.1 Studying discriminative motifs

We model each motif using the standard position weight matrix (PWM) representation (Bailey and Elkan, 1994; Roth *et al.*, 1998), which assumes independence between positions in a binding site. This model assigns a weight to each position in the motif and each nucleotide $n \in \{A, C, G, T\}$, representing the extent to which the nucleotide's presence in this position is associated with the motif.

For studying PWMs we adapt the discriminative motif model by Segal *et al*. (2002). This model is specified using a logistic function with p position-specific weights $w_i[n]$, one for each position i and each nucleotide $n \in \{A, C, G, T\}$, and a threshold w_0. For a sequence example s, denote its nucleotide sequence by $s.S = s.S_1, \dots, s.S_L$. For a motif m, denote by $s.m$ the location of occurrence of m in s, with the convention that if m does not occur in s then $s.m = -1$. The model assumes that a positive sequence contains exactly one occurrence of the motif (we extend and refine this model below) and a negative sequence does not contain the motif. The location distribution of a motif's occurrence within a positive sequence is assumed to be uniform, i.e. the motif occurs with equal probability at each of the $L - p + 1$ possible positions in the sequence. Positions in which the motif does not occur (within positive and negative sequences alike) are modeled using a 0-order Markov model. Under this model, the probability of a motif occurrence given the sequence is:

$$P(s.m \geq 0 \mid s.S_1, \dots, s.S_L, \theta_m) = \text{logit}\left(w_0 + \log\left(\sum_{j=1}^{L-p+1} p_m(j) \exp\left\{\sum_{i=1}^{p} w_i[s.S_{i+j-1}]\right\}\right)\right)$$

where θ_m is the set of parameters for the motif, $p_m(j) = 1/(L - p + 1)$ and $\text{logit}(x) = 1/(1 + e^{-x})$ is the logistic function. [The reader is referred to Segal *et al*. (2002) for more details on the model and the likelihood derivation.]

We extend the above model to take into account the possible bias in the location of certain motifs along the input sequences. Such bias was observed previously for promoter regions [see e.g. Tanay and Shamir (2003) and Beer and Tavazoie (2004)]. We use a simple model for the location preference, in which the sequence is equally partitioned to k parts ($k = 10$), each having a certain probability of containing the motif, and within each part the probability of occurrence is assumed to be uniform. For a given motif, we empirically estimate the distribution of the locations of its occurrences along the positive sequences (see below). We redefine $p_m(j)$ based on the estimated distribution.

A complicating factor in applying this model to study the motif parameters from the data is that we do not expect the motif m to occur in every core promoter sequence, but only in a fraction of the sequences. Thus, we treat the positive training data as noisy. Precisely, let T be a set of labels for the training sequences, specifying for each sequence s whether it is a positive or a negative example. We further denote, T^+ as the set of positive examples, T^- as the set of negative examples and S as the set of all nucleotide sequences $\{s.S | s \in T\}$. Define $q_m \equiv P(s \in T^+ | s.m = -1)$ to be the probability that a sequence is a core promoter given that motif m does not occur in it. This probability reflects the fraction r_m of positive sequences containing the motif m: $q_m = \left(1 + \frac{a}{1 - r_m}\right)^{-1}$, where a is the ratio of negative to positive examples. The likelihood of the data under this extended model is:

$$\begin{aligned} P(T | S, \theta_m, q_m) = \prod_{s \in T^+} &\{P(s.m \geq 0 | s.S, \theta_m) \\ &+ q_m(1 - P(s.m \geq 0 | s.S, \theta_m))\} \\ &\times \prod_{s \in T^-} \{(1 - q_m)(1 - P(s.m \geq 0 | s.S, \theta_m))\} \end{aligned}$$

The model parameters that need to be studied are the fraction of positive sequences containing the motif and the position specific weights for the motif. We seek parameters that optimize the likelihood of the data. Since this optimization problem has no closed form solution, we use a conjugate gradient ascent to find a local optimum in the parameter space. The starting point for the gradient ascent process has a large impact on the quality of the result; it is found using a method similar to Barash *et al*. (2001), which we describe next.

2.2 Initialization of the motif model

The initialization of PWMs that correspond to putative binding sites is done using a three-stage process: First, discriminative 'consensus' sequence patterns are identified; second, these consensus sequences are scored to quantify their enrichment in the positive sequences versus the negative ones; third, occurrences of these patterns along the positive sequences are used to compute an initial PWM for the corresponding motifs.

The pattern search is done in an exhaustive manner, scoring all sequences of length 6–8 bp, which are called seeds. To score a seed we count the number of its occurrences up to one mismatch in the positive and negative examples. We compute a hypergeometric P-value for these counts, and retain only seeds that have an adjusted P-value <0.01 (we use a Bonferroni correction to adjust the P-values for multiple testing). We also compute an enrichment P-value against a first-order Markov model of the positive sequences, and filter seeds that do not pass the 0.01 significance level. The surviving seeds are further filtered in a greedy fashion to ensure that no two seeds are similar in sequence or significantly overlap in their occurrences.

For each remaining seed, the initial position specific weights are computed by averaging over all occurrences (up to one mismatch) of this seed. We use the seed occurrences also to possibly extend the PWM at each end by positions whose information content exceeds a threshold. Once the initial weights are determined, the parameters of the location distribution are estimated by considering, for each positive sequence, only the highest-scoring match of the pattern to the sequence.

2.3 Studying discriminative modules

In addition to the motif-based features, we also study more complex patterns, namely, spatial combinations of motifs, or modules. We seek modules that are abundant in the positive sequences relative to the negative ones. Studying modules allows us to identify signals that are too weak at the motif level, and also to associate motifs whose co-occurrence has a functional significance.

To this end, we generalized the above motif model to ordered combinations of two or more motifs. For simplicity, we describe the model for the case in which the module consists of two ordered motifs. The assumption is that two factors (e.g. transcription factors) can interact if and only if the occurrences of their motifs are at least d_L and at most d_U bases apart. We further assume a prior uniform distribution on the distance between the two motifs, although more complex distributions can be incorporated into the model. The probability that a module M, consisting of a pair of motifs (m_1, m_2), occurs at a given position i of a sequence s is therefore:

$$P(s.M = i|s.S) = \frac{1}{d_U - d_L + 1} P(s.m_1 = i|s.S) \sum_{j=i+d_L}^{i+d_U} P(s.m_2 = j|s.S)$$

where

$$P(s.m_k = l|s.S) = \text{logit}\left(w_0^{(m_k)} + \sum_{t=1}^{p} w_t^{(m_k)}[s.S_{t+l-1}]\right)$$

One can study this model using the same gradient ascent approach used for the single motif model. The initialization of the model is done by enumerating pairs of seeds (consensus sequences) that occur up to one mismatch within a window of size $w (w = 50)$. These putative modules are scored by computing their enrichment in the positive set, using a hypergeometric test. Significant pairs are then initialized in a way similar to the initialization of seeds for the motif model.

2.4 Adding external attributes

Up till now we have described a general framework for studying discriminative attributes from sequence data. However, depending on the specific problem, there may be properties that are important for the classification task and cannot be expressed as sequence motifs. For instance, Sorek *et al*. (2004) show that exons whose length is divisible by three are less likely to be constitutive. Thus, in each of the applications described below we also add to our attribute vectors those attributes that were found to be discriminative for that specific classification problem.

In addition, we add one more features to the attribute vectors, representing the fit of a sequence to a probabilistic model of the positive sequences versus the negative sequences. Specifically, we compute a first-order Markov model for the positive and negative sequences and define this feature to be the log odds of being a positive versus being a negative example.

2.5 Training the SVM

SVM is a classification method based on finding a separating hyperplane between positive and negative samples that maximizes the distance (margin) between the samples and the hyperplane (in case the samples are not separable, mis-classification errors are combined into the optimization criterion).

SVMs allow an implicit mapping of the sample vectors (sequences in our case) into a high-dimensional, non-linear feature space, in which the samples may be better separated through the use of a similarity function between pairs of samples, called kernel. In our framework the sequence vectors are mapped onto attribute vectors and a linear SVM is applied to those. To optimize internal parameters of the SVM we use a cross-validation approach. The trained SVM is used

Table 1. Motifs and modules in Drosophila core promoters

Motif/module	Name	Consensus	Length	P-value	MEME-short	MEME-long
1	DRE	ATCGATAG	8	1E−33	+	+
2	—	GGTCACACT	9	3E−23	+	+
3	DPE	CGGTCG	6	2E−19	+	−
4	—	CAGCACTG	8	4E−14	+	−
5	—	CAGCTGGT	8	4E−13	+	−
6	—	CCGATAAC	8	8E−13	−	−
7	—	CGACGACG	8	1E−12	−	−
8	—	TCGCCGCG	8	4E−11	−	−
9	TATA	CTATAAAA	8	6E−9	+	−
10	—	CGAGCGGC	8	7E−9	+	−
11	INR	CTCAGTCG	8	3E−7	+	−
12	—	GGTATTTT	8	5E−5	+	−
13	—	TCGGCAGC	8	6E−5	−	−
1	12 + 2	GGTATTTT:GGTCACAC	≤50	9E−16	−	−
2	DRE + 6	ATCGATAG:CCGATAAC	≤50	6E−11	−	−
3	INR+DPE	CTCAGTCG:CGGTCG	≤50	7E−4	−	−

Top: the thirteen top-scoring motifs. For each motif, indicated are its common name (if such is known), its consensus sequence, its P-value (Bonferroni corrected) and whether it was identified by MEME, as reported in Ohler *et al*. (2002). MEME was applied both to the original 300 bp sequences (long) and to shorter segments from −60 to +40 bp (short). We consider a motif to match a MEME motif if their consensus sequences are identical up to one mismatch. Bottom: the three significant modules. The name of each module refers to the motifs that comprise it.

to classify new sequences. In order to measure our confidence in each prediction, we compute a confidence score based on the approach by Platt (1999). This is done by fitting a logistic function to the output of the SVM.

2.6 Evaluation measures

Given a test set, denote by TP, FP, TN and FN the numbers of true positive, false positive, true negative and false negative predictions. The sensitivity of a set of predictions is defined as the percentage of positives that are correctly predicted, i.e. $\text{sens} = \text{TP}/(\text{TP} + \text{FN})$. The specificity is defined as the percentage of negatives that are correctly predicted, i.e. $\text{spec} = \text{TN}/(\text{TN} + \text{FP})$. The FP rate equals $1 - \text{spec} = \text{FP}/(\text{TN} + \text{FP})$. For some applications (e.g. core promoter identification—see below) the number of TN in the test set far exceeds the number of TPs. In such cases, we replace the specificity measure with an adjusted specificity, defined as $\text{aspec} = \text{TP}/(\text{TP} + \text{FP})$.

It is convenient to visualize a range of sensitivities and specificities obtained by an algorithm using a receiver operating characteristic (ROC) curve, which depicts the sensitivity of the predictions as a function of the FP rate, or a ROC-like curve that depicts the sensitivity as a function of $(1 - \text{aspec})$.

3 RESULTS

3.1 Drosophila core promoters

As a first test, we applied our method to predict core promoters in the Drosophila genome. A core promoter is a short sequence region (about 100 bp), flanking the transcription start site, that serves as a recognition site for the basal transcription apparatus. Common core promoter elements include the TATA box at −31 to −26 bp, its extension, BRE, at −37 to −32 bp, the initiator, INR, at −2 to +4 bp and a downstream element, DPE, at +28 to +32 bp. A fifth element, DRE, was implicated to be abundant in core promoters in Ohler *et al*. (2002).

The training dataset that we used was prepared by Ohler *et al*. (2002) and includes a set of 1842 core promoters, 1799 intronic sequences and 2859 coding sequences. These sequences are 300 bp long, where for core promoters they extend from −250 to +50 bp. In order to take advantage of this partition of the sequences, we trained our model twice: first, to discriminate between core promoters and intronic sequences; and second, to discriminate between core promoters and coding sequences. Since coding sequences are very different from core promoter sequences in their nucleotide content, we used only external attributes for the second classification task. We restricted the program to identify the 15 top-scoring motifs or modules, and retained only significant motifs and modules whose frequency in the positive set was estimated to be ≥10%. In total, the algorithm identified 13 significant motifs and three significant modules, which are summarized in Table 1.

The 13 motifs that we have studied include four known core promoter elements: TATA, INR, DRE and DPE. Five of the other nine motifs that the algorithm studied match motifs that were studied by MEME (Bailey and Elkan, 1994) on the same dataset (Table 1). Specifically, Ohler *et al*. (2002) applied MEME to the core promoter sequences, where in one application MEME was applied to the entire 300 bp segments and in the other MEME was applied to shorter segments from

−60 to +40 bp. While the first application failed to recover most of the known core promoter elements, the 10 top-scoring motifs of the second application included nine of the motifs that our algorithm identified. We note that both our method and MEME did not recover the BRE motif, which could imply that it is underrepresented in the data.

In addition, we studied three significant modules on this dataset, which are shown in Table 1. The first module consists of motifs 12 and 2. These two motifs were reported to have a high frequency of co-occurrence in core promoter sequences (Ohler *et al*., 2002). The second module consists of the DRE element and motif 6. The third module consists of the INR and DPE motifs. This module structure is one of the most common core promoter structures reported in the literature (Butler and Kadonaga, 2002).

Following Ohler *et al*. (2002) we also used 14 external attributes that capture the physical properties of DNA sequences, and were shown to discriminate between core promoters and other sequences. Specifically, the computation of these properties uses experimentally derived tables on physical properties of di- or tri-nucleotides, such as bendability, GC-content, conformation, etc. Full details on these properties and their computation can be found in Ohler *et al*. (2001). We used the average value of each property along the core promoter segment from −60 to +40 bp as a feature. Note that more complex features can be computed based on the external attributes, but this was not the focus of our analysis.

To test the performance of our algorithm we applied it after training to identify core promoters in the well annotated Adh region (Ohler *et al*., 2002). This region is 2.9 Mb long and contains 92 annotated open reading frames (not included in our training data). The core promoter predictions were computed by sliding a window across each of the strands, calculating its confidence score, and choosing local maxima of these confidence scores as the predictions. To evaluate the results we used the same quality measures employed in Ohler *et al*. (2002): sensitivity and adjusted specificity. ROC-like curves of the results are presented in Figure 1; a comparison with existing methods is given in Table 2. These results (Fig. 1) also demonstrate the utility of using both discriminative motifs and modules for the classification task. We further examined the utility of modeling the location preference of motifs by comparing our results with a variant of the algorithm that assumes a uniform distribution for the location of motif occurrences. The comparison shows a mild improvement in accuracy when using the extended model (i.e. modeling location preference; data not shown).

3.2 Alternative splicing in human

As a second test, we applied our method to classify alternative and constitutive exons. Recently, Sorek *et al*. (2004) have reported on a method to identify alternatively spliced exons based on their sequence characteristics rather than on EST data. Specifically, they have shown that alternative exons tend to have length divisible by three and tend to be conserved along with their flanking sequence between human and mouse.

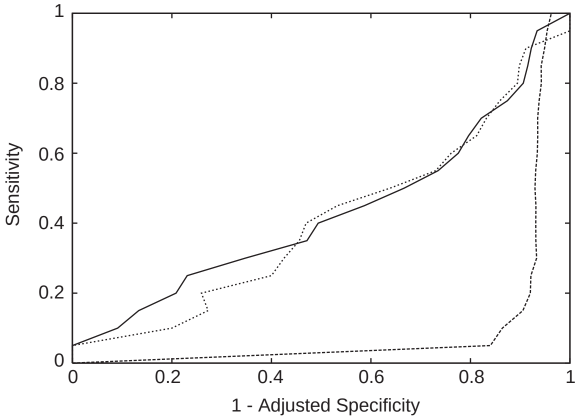

Fig. 1. Performance on the Adh region, shown as ROC-like curves, where the *x*-axis is (1 − aspec) and the *y*-axis is the sensitivity of the predictions. The solid, dotted and dashed curves describe the performance of the algorithm when using both discriminative motifs and modules, motifs only and no motifs or modules (i.e. using only external attributes), respectively.

Table 2. Comparison of classification results on the Adh region

Sensitivity (%)	Adjusted specificity MotifBased (%)	McPromoter (%)	NNPP (%)
20	79	69	14
35	53	51	10
50	33	40	6
65	20	29	—

For each sensitivity level, the adjusted specificity of each method is indicated. The results of McPromoter are adapted from Ohler *et al*. (2002). The results of NNPP are adapted from Reese (2001), and were based on a smaller training set.

We tested our method on the training data reported by Sorek *et al*. (2004), which consists of flanking sequences for 243 alternative exons and 1753 constitutive ones. Following Sorek *et al*. (2004), we evaluated our results using 5-fold cross-validation. The algorithm studied two to three significant motifs in each cross-validation iteration, with two motifs consistently studied in the majority of the iterations: TCCTTTTT and TTGTCTGT. No modules were studied for this dataset.

A ROC curve of the results is presented in Figure 2, where the specificities and sensitivities represent averages over the cross validation iterations. In particular, as highlighted in Table 3, our method compares favorably with that reported by Sorek *et al*. (2004).

We note that a recent paper by Dror *et al*. (2005) analyzed this data using an SVM-based classification procedure and

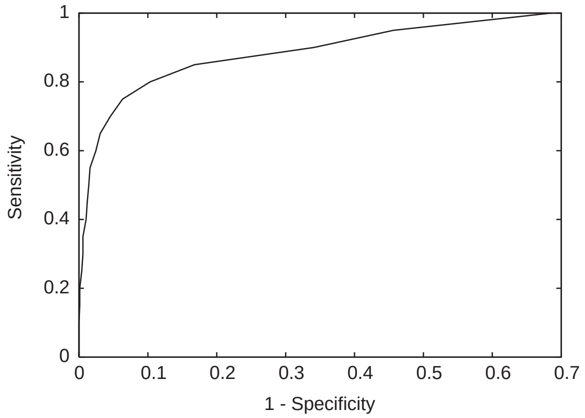

Fig. 2. ROC curve for the classification of alternatively spliced exons.

Table 3. Classification results on the exon dataset of Sorek *et al*. (2004)

Method	Sensitivity (%)	Specificity (%)
MotifBased	40.3	99.4
(Sorek *et al*., 2004)	32.3	99.7

The specificity and sensitivity percentages represent averages over five cross-validation iterations.

reported on a sensitivity rate of 50%. However, the results here and in Sorek *et al*. (2004) are not directly comparable with those of Dror *et al*. (2005), since the latter study used a different validation method (the data was partitioned into a training and a test set) and took advantage of additional external attributes that were not part of the original data of Sorek *et al*. (2004).

3.3 Cell-cycle regulation in yeast

As a third test of our method, we applied it to recognize cell cycle regulated genes in yeast according to their promoter sequences. The assumption underlying this experiment was that cell cycle regulated genes carry in their promoter sequences unique signals, corresponding to the binding sites of cell cycle regulators. To compile a training dataset we downloaded 500 bp promoter sequences for all yeast genes. We classified a gene as cell cycle regulated if it was reported to have a periodic expression pattern in Spellman *et al*. (1998). In total, we had 6323 genes with promoter information, 799 of which were determined to be cell cycle regulated. As a negative set we arbitrarily selected 800 other genes. We tested our success rate using 5-fold cross validation. A ROC curve of the results is depicted in Figure 3. We further examined the motifs that were studied by the algorithm. Throughout the cross validation iterations the algorithm studied three to four significant motifs, consistently identifying motifs whose consensus sequences matched those of the known cell cycle regulators MBP1, SWI4 and SWI6.

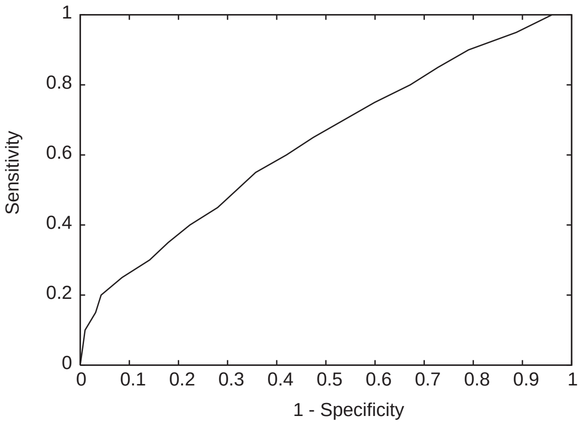

Fig. 3. ROC curve for the classification of cell cycle regulated genes.

4 CONCLUSIONS

We have presented a general framework for the characterization and classification of a family of related sequences based on recurrent sequence motifs and modules of motifs. We demonstrated several applications of our framework to identifying core promoters, alternatively spliced exons and cell cycle regulated genes. There are many possible extensions to our work, including (1) more refined modeling of the position preference of a motif; (2) modeling the distance distribution among motifs in a module; (3) design of kernel functions for the classification task based on the approach by Lanckriet *et al*. (2004) to provide explicit treatment of the problem of combining features of different types; and (4) application of our method to classify other sequence families, such as core promoters in other species, promoter regions of tissue-specific genes and promoter regions of genes with specific expression patterns.

ACKNOWLEDGEMENT

Part of the work of R.S. was done while doing his post-doc at the University of California, Berkeley.

REFERENCES

Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of ISMB*, Stanford, CA, pp. 28–36.

Barash,Y., Bejerano,G. and Friedman,N. (2001) A simple hypergeometric approach for discovering putative transcription factor binding sites. *Proceedings of WABI*, Bergen, Norway, pp. 278–293.

Beer,M.A. and Tavazoie,S. (2004) Predicting gene expression from sequence. *Cell*, **117**, 185–198.

Ben-Hur,A. and Brutlag,D. (2003) Remote homology detection: a motif based approach. *Bioinformatics*, **19** (Suppl. 1), I26–I33.

Butler,J.E.F. and Kadonaga,J.T. (2002) The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes Dev.*, **16**, 2583–2592.

Dror,G., Sorek,R. and Shamir,R. (2005) Accurate identification of alternatively spliced exons using support vector machine. *Bioinformatics*, **21**, 897–901.

Hannenhalli,S. and Levy,S. (2001) Promoter prediction in the human genome. *Bioinformatics*, **17**, S90–S96.

Lanckriet,G.R.G., De Bie,T., Cristianini,N., Jordan,M.I. and Noble,W.S. (2004) A statistical framework for genomic data fusion. *Bioinformatics*, **20**, 2626–2635.

Ohler,U., Niemann,H., Liao,G.C. and Rubin,G.M. (2001) Joint modeling of DNA sequence and physical properties to improve eukaryotic promoter recognition. *Bioinformatics*, **17**, S199–S206.

Ohler, U., Liao, G.-C., Niemann,H. and Rubin,G.M. (2002) Computational analysis of core promoters in the Drosophila genome. *Genome Biol.*, **3**, 1–12.

Pavlidis,P. Furey,T.S., Liberto,M., Haussler,D. and Grundy,W.N. (2001) Promoter region-based classification of genes. *Pac. Symp. Biocomput.*, 2001, 151–163.

Platt,J.C. (1999) Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Technical report. Microsoft Research.

Reese,M.G. (2001) Application of a time-delay neural network to the annotation of the *Drosophile melanogaster* genome. *Comput. Chem.*, **26**, 51–56.

Roth,F.P., Hughes,J.D., Estep,P.W. and Church,G.M. (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.*, **16**, 939–945.

Segal,E., Barash,Y., Simon,I., Friedman,N. and Koller,D. (2002) From sequence to expression: a probabilistic framework. *Proceedings of RECOMB*, Washington, DC, pp. 263–272.

Sorek,R., Shemesh,R., Cohen,Y., Basechess,O., Ast,G. and Shamir,R. (2004) A non-EST-based method for exon-skipping prediction. *Genome Res.*, **14**, 1617–1623.

Spellman,P. T., Sherlock,G., Zhang,M.Q., Iyer,V.R., Anders,K., Eisen,M.B., Brown,P.O., Botstein,D. and Futcher,B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.

Tanay,A. and Shamir,R. (2003) Modeling transcription programs: inferring binding site activity and dose-response model optimization. *Proceedings of RECOMB*, Berlin, Germany, pp. 301–310.