FORMATICS

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

any signals in biological sequences are based
ce or absence of base signals and their spatial
One of the best known examples of this is the
ng a core promoter—the site at which the basal
achinery starts the transcription of a gene. Our
utomatic pattern recognition system for a family
which simultaneously discovers the base sig-
tial relationships and a classifier based upon

his paper we present a general method for
a set of sequences by their recurrent motifs.
relies on novel probabilistic models for DNA
nd modules of binding sites, on algorithms to
m the data and on a support vector machine
models studied to classify a set of sequences.
te the applicability of our approach to diverse
ging from families of promoter sequences to
tronic sequences flanking alternatively spliced
re promoter dataset our results are comparable
f-the-art McPromoter. On a dataset of altern-
exons we outperform a previous approach.
ve high success rates in recognizing cell cycle
es. These results demonstrate that a fully auto-
recognition algorithm can meet or exceed the
f hand-crafted approaches.
he software and datasets are available from the
equest.
d@tau.ac.il

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

of NNPP (Reese, 2001). Moreover, we came
the method extends well beyond our original
ustrate this we apply it here to the problems of
ernatively spliced exons, in human, and recog-
that are under cell-cycle control in yeast. In
ith a previous approach for detecting alternat-
exons, we are able to show increased sensitivity
ons.

DS

e following classification problem: the input
training set of sequences with positive and
nples, and a test set; the goal is to devise a
the positive examples that will best discrim-
positives and negatives on the test set. We
-phase scheme for this problem: in the first
the training data to study attributes (features)
lent in the positive sequences compared with
negative) sequences. The attribute vector of
e consists of three types of attributes: (1) dis-
notifs, (2) discriminative modules of motifs
nal attributes that are unique to the specific
the second phase we train a SVM for the clas-
blem using the attributes studied as sequence
e two phases are described in detail in the
ions.

ng discriminative motifs

h motif using the standard position weight mat-
presentation (Bailey and Elkan, 1994; Roth *et al.*,
assumes independence between positions in a
This model assigns a weight to each position in
each nucleotide $n \in \{A, C, G, T\}$, representing
which the nucleotide's presence in this position
with the motif.

g PWMs we adapt the discriminative motif
al *et al.* (2002). This model is specified using a
on with p position-specific weights $w_i[n]$, one
on i and each nucleotide $n \in \{A, C, G, T\}$, and
$_0$. For a sequence example s, denote its nucle-
e by $s.S = s.S_1, \ldots, s.S_L$. For a motif m, denote
ation of occurrence of m in s, with the conven-

a motif occurrence given the sequence is:

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

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:

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

od similar to Barash *et al*. (2001), which we

zation of the motif model

ion of PWMs that correspond to putative bind-
ne using a three-stage process: First, discrim-
nsus' sequence patterns are identified; second,
s sequences are scored to quantify their enrich-
sitive sequences versus the negative ones; third,
f these patterns along the positive sequences
mpute an initial PWM for the corresponding

search is done in an exhaustive manner, scoring
of length 6–8 bp, which are called seeds. To
e count the number of its occurrences up to one
he positive and negative examples. We com-
eometric P-value for these counts, and retain
at have an adjusted P-value <0.01 (we use
correction to adjust the P-values for multiple
lso compute an enrichment P-value against a
rkov model of the positive sequences, and filter
ot pass the 0.01 significance level. The surviv-
urther filtered in a greedy fashion to ensure that
re similar in sequence or significantly overlap
ences.
aining seed, the initial position specific weights
by averaging over all occurrences (up to one
this seed. We use the seed occurrences also to
d the PWM at each end by positions whose
ontent exceeds a threshold. Once the initial
etermined, the parameters of the location dis-
estimated by considering, for each positive
y the highest-scoring match of the pattern to

ng discriminative modules

the motif-based features, we also study more
rns, namely, spatial combinations of motifs, or
seek modules that are abundant in the positive
tive to the negative ones. Studying modules
dentify signals that are too weak at the motif
to associate motifs whose co-occurrence has a
nificance.
we generalized the above motif model to

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,

and modules in Drosophila core promoters

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

p-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 EME, 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). f 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 mprise it.

v sequences. In order to measure our confidence ction, we compute a confidence score based on by Platt (1999). This is done by fitting a logistic e output of the SVM.

ation measures

set, denote by TP, FP, TN and FN the num- positive, false positive, true negative and false ictions. The sensitivity of a set of predictions the percentage of positives that are correctly . sens = TP/(TP + FN). The specificity is percentage of negatives that are correctly pre- pec = TN/(TN + FP). The FP rate equals P/(TN + FP). For some applications (e.g. core tification—see below) the number of TN in the xceeds the number of TPs. In such cases, we ecificity measure with an adjusted specificity, ec = TP/(TP + FP).

ent to visualize a range of sensitivities and spe- ned by an algorithm using a receiver operating

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%.

p. While the first application failed to recover own core promoter elements, the 10 top-scoring econd application included nine of the motifs thm identified. We note that both our method d not recover the BRE motif, which could imply represented in the data.

we studied three significant modules on this are shown in Table 1. The first module con- s 12 and 2. These two motifs were reported frequency of co-occurrence in core promoter hler *et al*., 2002). The second module consists ement and motif 6. The third module consists d DPE motifs. This module structure is one of mon core promoter structures reported in the ler and Kadonaga, 2002).

hler *et al*. (2002) we also used 14 external attrib- re the physical properties of DNA sequences, wn to discriminate between core promoters uences. Specifically, the computation of these s experimentally derived tables on physical di- or tri-nucleotides, such as bendability, onformation, etc. Full details on these prop- ir computation can be found in Ohler *et al*. sed the average value of each property along oter segment from -60 to $+40$ bp as a fea- more complex features can be computed based l attributes, but this was not the focus of our

erformance of our algorithm we applied it after ntify core promoters in the well annotated Adh *t al*., 2002). This region is 2.9 Mb long and con- ated open reading frames (not included in our The core promoter predictions were computed ndow across each of the strands, calculating its re, and choosing local maxima of these confid- the predictions. To evaluate the results we used ty measures employed in Ohler *et al*. (2002): adjusted specificity. ROC-like curves of the sented in Figure 1; a comparison with exist- s given in Table 2. These results (Fig. 1) also e utility of using both discriminative motifs and e classification task. We further examined the eling the location preference of motifs by com- lts with a variant of the algorithm that assumes

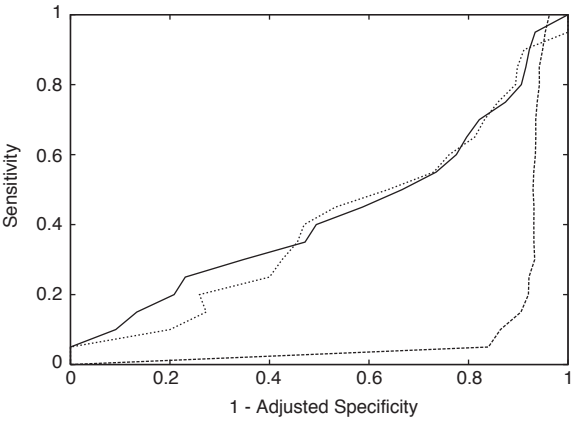


Fig. 1. Performance on the Adh region, shown as ROC-like curves, where the x-axis is $(1 - \text{aspec})$ and the y-axis is the sensitivity of the predictions. The solid, dotted and dashed curves describe the performance of the algorithm when using both discriminative motifs and modules, motifs only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

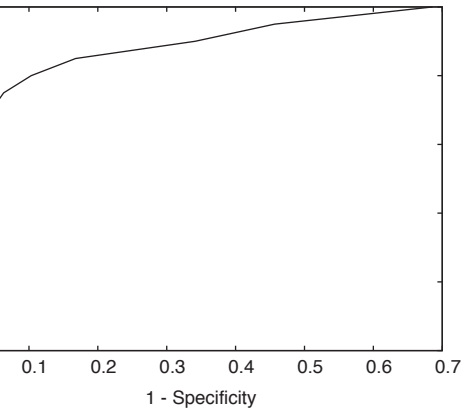
Table 2. Comparison of classification results on the Adh region

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.

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.

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 con-



urve for the classification of alternatively spliced

cation results on the exon dataset of Sorek *et al*. (2004)

	Sensitivity (%)	Specificity (%)
	40.3	99.4
4)	32.3	99.7

sensitivity percentages represent averages over five cross-validation

sensitivity rate of 50%. However, the results rek *et al*. (2004) are not directly comparable with *et al*. (2005), since the latter study used a differ- method (the data was partitioned into a training and took advantage of additional external attrib- e not part of the original data of Sorek *et al*.

ycle regulation in yeast

st of our method, we applied it to recognize ulated genes in yeast according to their pro- ces. The assumption underlying this experiment cycle regulated genes carry in their promoter que signals, corresponding to the binding sites regulators. To compile a training dataset we 00 bp promoter sequences for all yeast genes.

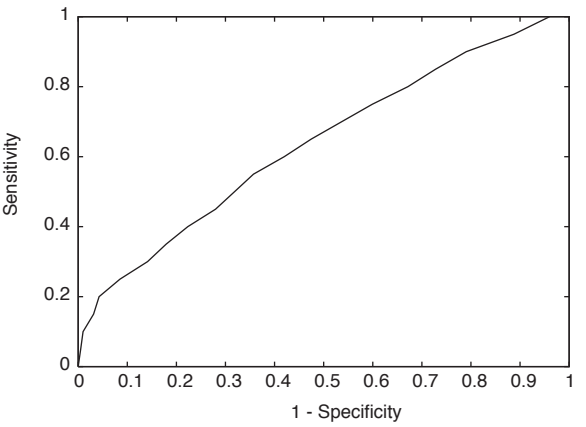


Fig. 3. ROC curve for the classification of cell cycle regulated genes.

consensus sequences matched those of the known cell cycle regulators MBP1, SWI4 and SWI6.

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.

d Brutlag,D. (2003) Remote homology detec-
f based approach. *Bioinformatics*, **19** (Suppl. 1),

l Kadonaga,J.T. (2002) The RNA polymerase II core
ey component in the regulation of gene expression.
6, 2583–2592.
R. and Shamir,R. (2005) Accurate identification
ely spliced exons using support vector machine.
cs, **21**, 897–901.
and Levy,S. (2001) Promoter prediction in the human
informatics, **17**, S90–S96.
G., De Bie,T., Cristianini,N., Jordan,M.I. and
2004) A statistical framework for genomic data
formatics, **20**, 2626–2635.
ann,H., Liao,G.C. and Rubin,G.M. (2001) Joint
f DNA sequence and physical properties to
aryotic promoter recognition. *Bioinformatics*, **17**,

G.-C., Niemann,H. and Rubin,G.M. (2002) Com-
alysis of core promoters in the Drosophila genome.
., **3**, 1–12.
y,T.S., Liberto,M., Haussler,D. and Grundy,W.N.
oter region-based classification of genes. *Pac. Symp.*
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.