BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 21 2007, pages 2823–2828
doi:10.1093/bioinformatics/btm473

Sequence analysis

OSCAR: One-class SVM for accurate recognition of *cis*-elements

Bo Jiang[1], Michael Q. Zhang[1,2] and Xuegong Zhang[1,*]

[1]MOE Key Laboratory of Bioinformatics and Bioinformatics Division, TNLIST/Department of Automation, Tsinghua University, Beijing 100084, China and [2]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11274, USA

Received on April 25, 2007; revised on August 29, 2007; accepted on September 11, 2007

Advance Access publication October 5, 2007

Associate Editor: Alex Bateman

ABSTRACT

Motivation: Traditional methods to identify potential binding sites of known transcription factors still suffer from large number of false predictions. They mostly use sequence information in a position-specific manner and neglect other types of information hidden in the proximal promoter regions. Recent biological and computational researches, however, suggest that there exist not only locational preferences of binding, but also correlations between transcription factors.

Results: In this article, we propose a novel approach, OSCAR, which utilizes one-class SVM algorithms, and incorporates multiple factors to aid the recognition of transcription factor binding sites. Using both synthetic and real data, we find that our method outperforms existing algorithms, especially in the high sensitivity region. The performance of our method can be further improved by taking into account locational preference of binding events. By testing on experimentally-verified binding sites of GATA and HNF transcription factor families, we show that our algorithm can infer the true co-occurring motif pairs accurately, and by considering the co-occurrences of correlated motifs, we not only filter out false predictions, but also increase the sensitivity.

Availability: An online server based on OSCAR is available at http://bioinfo.au.tsinghua.edu.cn/oscar.

Contact: zhangxg@tsinghua.edu.cn

1 INTRODUCTION

Identification of transcription factor binding sites (TFBSs) and the corresponding motifs plays a pivotal role in the understanding of the transcriptional regulation mechanism. Recently, the advent of DNA microarray techniques and ChIP-chip experiments has significantly improved our ability to identify those important *cis*-regulatory elements. However, since these experimental procedures are expensive and time-consuming, computational methods are still needed for predicting TFBSs.

The task of computational identification of TFBSs can be divided into two main categories. One is motif discovery, i.e. to discover a motif as well as its putative sites in a collection of genomic sequences that are expected to be bound by the same factor. Many state-of-the-art algorithms, such as MEME (Bailey and Elkan, 1994), Bioprospector (Liu *et al.*, 2001), YMF (Sinha and Tompa, 2002), etc. (for a recent review, see Tompa *et al.*, 2005), were designed to this end. The other task is motif search, i.e. to scan potential binding sites of a known transcription factor, often, on a genomic scale. Several tools like MatInspector (Quandt *et al.*, 1995) and MATCH™ (Kel *et al.*, 2003) were developed to search for putative TFBSs in DNA sequences by using position weight matrices (PWMs; Stormo *et al.*, 1982) in TRANSFAC® (Wingender *et al.*, 2000) or JASPAR (Sandelin *et al.*, 2004) databases. In this article, we focus on the latter task, which is to recognize the putative binding sites of known transcription factors.

Previous works to predict new putative TFBSs based on known binding sites still suffer from large number of false predictions when applied in a genomic scale. Most commonly used tools ignore the extra information in and beyond the TFBSs, such as the conservation of the binding site locations and the co-occurrences of other motifs in the promoter sequences. There is ample evidence that many *cis*-regulatory elements show preferred locations (FitzGerald *et al.*, 2004; Frith *et al.*, 2003; Xie *et al.*, 2005) and precise organizations (Boyer *et al.*, 2005; Odom *et al.*, 2006) within promoter sequences. A search of the binding sites on the entire genome without such information usually returns a large number of sites, many of which are not functional *in vivo*, although they would probably bind to the transcription factor if they were in proper genomic contexts.

Recently, many classification algorithms, such as support vector machines (SVMs), have been applied to discriminate false predictions from the true ones in TFBS recognition (Holloway, 2005; Sun *et al.*, 2006) and other applications. Jaakkola *et al.* (2000) proposed a discriminative framework with a variant of SVMs to detect protein homology, which can potentially be applied to the detection of TFBSs. Special string kernels in SVMs were also designed to process the regulatory regions of genes, in order to recognize a given class of promoter region, and simultaneously identify a collection of relevant, discriminative sequence motifs (Leslie *et al.*, 2004; Rätsch *et al.*, 2005; Sharan and Myers, 2005; Sonnenburg *et al.*, 2005a,b; Vert *et al.*, 2005). However, as with other classification techniques, a set of known positive and negative samples must be supplied to the SVMs. With limited instances of experimentally-verified binding sites, it is hard to determine a 'negative' set of sequences, to which a transcription factor will certainly

*To whom correspondence should be addressed.

not bind (Hong *et al.*, 2005). In fact, there exists no site (separated from its context) that will never be bound by a transcription factor, but only sites that are unlikely to be bound. In this sense, recognition of TFBSs may not be well characterized as a standard two-class classification problem.

In this article, we exploit the one-class SVM, rather than two-class approaches, to estimate the support of probability distribution of known TFBSs, i.e. the region where most known (positive) samples live in the feature space. Our method, named OSCAR (One-class Support vector machine for *Cis*-element Accurate Recognition), simultaneously considers the nucleotide composition of all the positions within a binding site, and further incorporates locational preferences and co-occurrences of DNA motifs. We first demonstrate that our novel approach outperforms existing algorithms on synthetic data. Applying our method to promoter regions in the Eukaryotic Promoter Database (EPD; Praz *et al.*, 2002) with annotated TFBSs from TRANSFAC®, we find that the performance of our algorithm can be further improved by considering the locational preferences of binding events. We also show that the mutual interactions of transcription factors can be identified and exploited by using our method. Furthermore, the application of our method to recognize the binding sites of GATA and HNF transcription factor families indicates that the integration of co-occurrence information not only decrease false positives, but also increase the sensitivity of the prediction.

2 MATERIALS AND METHODS

2.1 Databases for identification of TFBSs

In this study we used two databases: (1) TRANSFAC® (release 9.4; Wingender *et al.*, 2000) is a database that provides data on transcription factors and their binding sites in promoters of eukaryotic genes as well as a library of PWMs. It also offers an additional profile, called 'nonredundant profile for vertebrates', by categorizing PWMs into groups on the basis of the linked transcription factors and selecting just one 'best' matrix from each group. (2) The EPD (Praz *et al.*, 2002) is an annotated nonredundant collection of eukaryotic promoters, for which the transcription start site (TSS) has been determined experimentally. The annotation part of an entry includes description of the initiation site mapping data, cross-references to other databases including TRANSFAC®.

2.2 One-class SVM and basic OSCAR algorithm

One-class SVM (Schölkopf *et al.*, 2001) was proposed to estimate the support of a high-dimensional distribution, i.e. the region where most of the data live in the high-dimensional feature space. It returns a prediction function f that takes the value $+1$ in a 'small' region capturing most of the data points, and -1 elsewhere.

Given known binding sites of a particular transcription factor, we first encode each site s with length L into a binary string $\mathbf{x}$ with length $4L$. The nucleotide at each position of the binding site is mapped into a 4-dimension vector in $\{0,1\}$-space with following rule: A↦0001, C↦0010, G↦0100 and T↦1000. To determine whether a DNA sequence s_0 with length L is a putative binding site of the transcription factor, s_0 is transformed into a binary string $\mathbf{x}_0$ in the same way, and the prediction function is given as:

$$f = \text{sign}((w \cdot x_0) - \rho), \tag{1}$$

where the weight w and parameter ρ are determined by the one-class SVM training algorithm given below. Note that w can be denoted in a matrix form: $\mathbf{M}_{i,j} = w_{4\times i+j}$, where $i = 1,\ldots,L$, and $j = 1,\ldots,4$ correspond to four nucleotides. Regarding the parameter ρ in prediction function (1) as a threshold, we come to a scoring scheme similar to PWM-based methods, yet in our method nucleotide preferences at different positions within a binding site are considered simultaneously in SVM model.

To be more specific, consider the training data $\mathbf{x}_1,\ldots,\mathbf{x}_l \in \Phi$, where $\mathbf{x}_i$ is a binary string encoded by a known binding site, Φ is a $\{0, 1\}$-vector space with $4L$ dimension, and l is the number of known binding sites. One-class SVM solves the following quadratic programming problem:

$$\min_{w\in\Phi,\,\xi\in\mathbf{R}^l,\,\rho\in\mathbf{R}} \frac{1}{2}\|w\|^2 + \frac{1}{vl}\sum_i \xi_i - \rho, \tag{2}$$

$$\text{subject to } (w \cdot \mathbf{x}_i) \geq \rho - \xi_i, \xi_i \geq 0. \tag{3}$$

Here, $v \in (0,1)$ is a user-defined parameter.

For the $\mathbf{x}_j$'s that the equalities in (2) hold at the solutions of w and ρ, we have $\rho = (w \cdot \mathbf{x}_j)$ and $\mathbf{x}_j$ is named a 'support vector'. Since nonzero slack variables ξ_i are penalized in the objective function, we can expect that the prediction function (1) will be positive for most samples contained in the training set, while the regularization term $\|w\|$ will still be small. The actual trade-off between these two goals is controlled by v. In fact, Schölkopf *et al.* (2001) have proved that v is an upper bound on the fraction of outliers and a lower bound on the fraction of support vectors.

To solve w and ρ from problem (2) and (3), we use the LIBSVM package (Chang and Lin, 2001) for the implementation of a one-class SVM. In our applications, one-class SVM is used to construct prediction functions with training sets containing only experimentally-verified binding sites of known transcription factors. The algorithmic framework of a one-class SVM not only enables us to exploit the nucleotide composition of known binding sites, but also allows us to consider their locational preferences and effects of TFBS co-occurrences.

2.3 Incorporating locational preference

Binding sites of many transcription factors show preferred locations within promoter regions, which are influenced by nucleosome occupancy (Segal *et al.*, 2006), chromatin structure and other factors. The locational preferences of binding events can be conducive to the recognition of functional binding sites in genome. A one-class SVM can serve as a good algorithmic frame, which allows us to exploit this type of information when the relative locations to the TSS of known binding sites are available. To encode such information, we first divide the promoter regions that contain the binding sites in training data into K consecutive intervals relative to TSS, where K and the length of each interval are user-defined parameters. We recommend choosing the interval length from 150 bp to 300 bp, which is approximately the total length of a nucleosome unit and the linker region. The location of a known binding site, if available, is described by a binary string with length K, in which a nonzero value indicates the occurrence of binding sites in the corresponding interval (see examples in Fig. 1). This additional string is concatenated to the original string that depicts the nucleotide composition of a binding site, as the training vector of the one-class SVM. Note that if the location of binding site is not known, the additional string will be all zero, and the corresponding method will become the basic OSCAR algorithm.

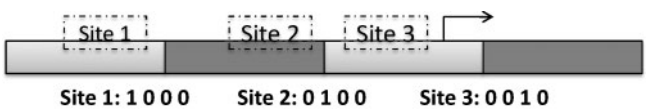

Fig. 1. Illustration for encoding locational preferences in OSCAR-L. The arrow indicates TSS; colored intervals delineate the division of the promoter region; boxes denote the locations of binding sites, and the locational coding of each site is given below.

To search for putative binding sites of the transcription factor in a promoter region, the region is divided into intervals according to its TSS in the same way as described above. The sites in different intervals are encoded accordingly, and determined by the prediction function in the one-class SVM. To distinguish the algorithm that incorporates locational preferences from the basic OSCAR algorithm, we call this one OSCAR-L.

2.4 Exploiting co-occurrence information

Gene regulation is often achieved by a precise organization of multiple transcription factors. Some transcription factors are by nature moderately or poorly specific in their DNA binding and achieve higher specificity only in the context of other binding partners. In this subsection, we further extend OSCAR to recognize putative binding sites with a proximal binding of other co-factors. We call this method OSCAR-C.

First, a set of M candidate motifs that may co-occur with a specific transcription factor (named 'primary TF' here) is obtained from the database or predefined by the user. To search for the binding sites of the primary TF in a promoter region, M additional bits are added to the previous binary string, each of which indicates whether the corresponding candidate motif appears in this region. In the training process, the occurrences of a candidate motif in a promoter region can be determined from the annotations in the database. While in the predictions when we do not have the knowledge about the binding of candidate motifs, the prediction function obtained by the OSCAR-L algorithm described in the above subsection can be used to identify the occurrences of candidate motifs.

The candidate motif set may contain irrelevant motifs, and lead to over-fitting of the algorithm. Thus, the accuracy of prediction might be improved through the exclusion of candidate motifs with little relevance to the binding of the primary TF. Besides, knowing which candidate motifs are relevant can give insight to the interactions between transcription factors. Therefore, a criterion similar to SVM-RFE (Guyon *et al.*, 2002) is proposed to perform the selection of co-occurring candidate motifs according to their contributions in predicting the binding of the primary TF. Assume that w_m is the weight corresponding to the occurrence of the mth candidate motif in prediction function, we use w_m^2 as the ranking score to select relevant candidate motifs. Let J denote the objective function in (2), which becomes $(1/2)\ \|w\|^2$ when data points are separable from the origin. Our ranking criterion is then explained by the OBD algorithm (LeCun *et al.*, 1990), which approximates the change in objective function due to the removal of the i-th feature by expanding objective function J in the Taylor series to the second order:

$$\Delta J(m) = \frac{\partial J}{\partial w_m}\Delta w_m + \frac{\partial^2 J}{\partial w_m^2}(\Delta w_m)^2, \tag{4}$$

where $m = 1,\ldots,M$. At the minimum of J, the first order term can be neglected. With $J = (1/2)\ \|w\|^2$, Equation (4) becomes $\Delta J(m) = (\Delta w_m)^2$, where $\Delta w_m = w_m$ corresponds to removing the m-th feature.

Given a cut-off on the ranking list, we can obtain candidate motifs relevant to the binding of the primary TF. The indicators for the binding of selected motifs are used in the one-class SVM algorithm together with other features, and an integrated prediction function is finally obtained.

3 EXPERIMENT RESULTS

3.1 Results on synthetic data

We begin by evaluating our methods on synthetic data. To this end, we retrieve the binding sites of transcription factors from TRANSFAC® (release 9.4) nonredundant profiles, and 63 vertebrate transcription factors with more than 10 annotated binding sites are selected. To avoid over-fitting effects in evaluating algorithms, Leave-One-Out Cross Validation (LOOCV) is performed together with the synthetic procedure: for a set with N known binding sites of a transcription factor, we first randomly implant a site into an artificial sequence that is generated according to uniform nucleotide compositions, and then we apply the basic OSCAR algorithm, which uses the remaining $N-1$ sites as training data, to predict putative binding sites in the synthetic sequence. This step is repeated N times with each site implanted into a random sequence. To make comparisons, we test the performance of three other methods: simple PWM based on log-odds ratio (LOPWM), MATCH™ (Kel *et al.*, 2003) and P-Match (Chekmenev *et al.*, 2005), which are closely interconnected with TRANSFAC® database. MATCH™ is a representative PWM-based tool for searching putative transcription factor binding sites in DNA sequences, while P-Match is a newly developed tool that combines pattern matching and weight matrix approaches.

In synthetic data, the locations of binding events are known with certainty, which enable us to evaluate the predicting ability of different tools according to known binding sites in simulated sequences. A predicted position is a true-positive (TP) prediction if it coincides with a known binding site, and it is a false-positive (FP) prediction otherwise. To assess the performance of algorithms, we calculate the true positive rate (TPR, also known as 'sensitivity'), which is the ratio of the number of true positives to the total number of known binding sites, and false positive rate (FPR), which is the ratio of the number of false positives to the total number of nonsite positions. A curve of FPR versus TPR can be plotted while a threshold parameter is varied. This curve, which is called ROC (receiver operating characteristics) curve, is a comprehensive and objective way to compare the performance of methods as a tradeoff between specificity and sensitivity. ROC curves are obtained by setting v of one-class SVM in a range from 0.1 to 0.9. To evaluate MATCH and P-Match, we vary the two cut-offs of each algorithm in a complete range of values, and report the maximum TPR achieved at a given FPR.

In Figure 2, the area under the ROC curve indicates that OSCAR is able to pick out more TP sites given fixed number of false predictions in the region of high FPR (with a relatively small v). In the low FPR region, the performance of OSCAR is comparable with LOPWM and MATCH algorithms. This can be explained by the fact that the number of support vectors in the training samples increases as v becomes larger, and as a result, the weight vectors of the basic OSCAR algorithm become similar to the entries of traditional PWMs.

3.2 Results on EPD promoter regions

To verify our methods on real biological data, we determine the locations of TFBSs in TRANSFAC® database according to the annotations in EPD (release 85). Among all the transcription factors in TRANSFAC® nonredundant profiles, binding sites of 45 vertebrate transcription factors can be identified, corresponding to totally 124 promoters in EPD85 (data available from our website). Promoter regions from upstream

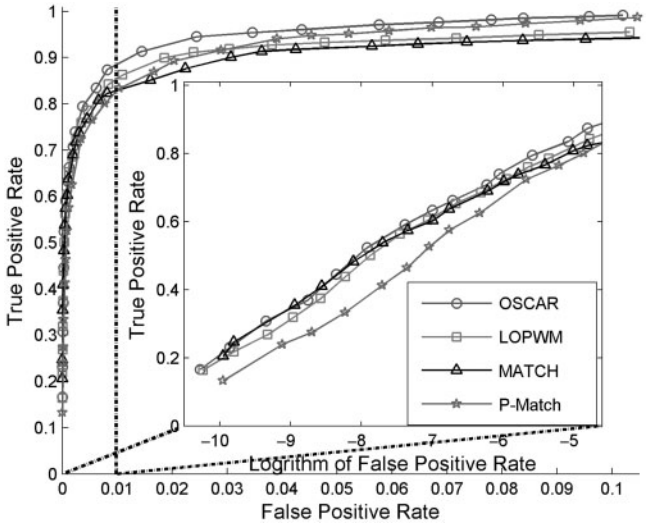

Fig. 2. ROC curves show the performances of different methods on synthetic data. Note that a logarithmic scale transformation on the *X*-axis is used to show the low FPR performances of different methods.

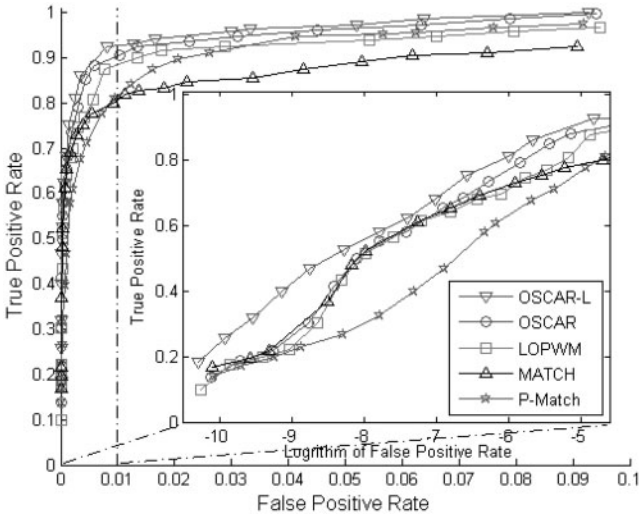

Fig. 3. ROC curves show the performances of different methods on EPD promoter regions, and confirm the advantage of incorporating locational preference of binding events. Note that a logarithmic scale transformation on the *X*-axis is used to show the low FPR performances of different methods. TPR is obtained by LOOCV, and FPR is estimated on randomly shuffled EPD promoter regions.

1000 bp to downstream 300 bp are retrieved. To test the sensitivity of algorithms, we use the LOOCV procedure as follows: for a set with N known binding sites of a transcription factor, we obtain a prediction function on $N-1$ sites and test the remaining site based on the score. This step is repeated N times with each site serving as testing sample. TPR is equal to the number of predicted binding sites divided by total number of known sites. For the purpose of test, we construct a background set, which presumably contains few true TFBSs, by randomly permuting the sequences of promoter regions in the training set with conserved di-nucleotides (Coward, 1999). Predictions were made on the background set by moving a scanning window of the motif length at the step of 1 bp. The FPR is calculated as the ratio of the number of predictions to the total number of base-pairs on background sequences. The evaluations of LOPWM, MATCH, P-Match and the basic OSCAR algorithms are performed by using the same procedure. Besides, we also apply the OSCAR-L algorithm to incorporate the locational preference, by dividing each promoter region into five consecutive intervals with equal length of 260 bp. ROC curves in Figure 3 show the overall performance of different methods on EPD promoters for all 45 transcription factors.

As we can see, OSCAR and P-Match are generally more accurate than LOPWM and MATCH in the area of high FPR, while OSCAR, LOPWM and MATCH are comparable in the low FPR region, which are consistent with observations in literature (Chekmenev *et al.*, 2005) and the results on our synthetic data. As expected, the overall performance can be improved by incorporating locational preferences into the basic OSCAR algorithm (OSCAR-L). At a given sensitivity level, OSCAR-L will result in a smaller FPR than the basic OSCAR algorithm when averaged over all 45 transcription factors (see Supplementary Table 1).

3.3 Results on GATA and HNF family

In order to check the ability of our method to reveal known TFBSs through identifying and utilizing co-occurring motifs (OSCAR-C), we test OSCAR-C and other methods on two additional data sets involving GATA factor and HNF factor family, which have relatively large numbers of binding sites that can be identified on EPD promoters. To apply the OSCAR-C algorithm, 141 motifs in TRANSFAC® database for vertebrates constitute the candidate motif set. Their co-occurrences with the binding sites of the primary TF are input as additional features to obtain the prediction function. To further evaluate the statistical significance of co-occurring motifs selected by the OSCAR-C algorithm, we randomly permute the bases in the promoter sequences, while preserving the confirmed sites of the primary factor. The permutation is repeated for 100 times, and the OSCAR-C algorithm is applied to each permutated set. The distribution of the largest and second largest absolute values of weights are used to access the significance of the top two motifs on the ranking list, respectively.

3.3.1 GATA factor binding sites Transcription factors in the GATA family are so-called because they bind to the consensus DNA sequence '(A/T)GATA(A/G)'. They are shown to play critical roles in development, including in cell-fate specification, regulation of differentiation and control of cell proliferation and movement (Patient and Meghee, 2002). We extract 13 EPD promoter regions, which contain 25 TRANSFAC annotated binding sites of a GATA transcription factor (V$GATA_Q6). The two most relevant motifs that co-occur with the binding of GATA factor is identified by the OSCAR-C algorithm and given in the first row of Table 1, together with the *P*-values evaluated by the procedure described above. The co-occurrences of GATA with NF-Y and Sp1 factors are

Table 1. Motifs co-occurring with the binding of GATA factor and HNF family identified by the OSCAR-C algorithm

TF family	Primary motif	Co-occurring motifs	*P*-value
GATA	V$GATA_Q6	V$NFY_Q6_01	<0.01
		V$SP1_Q2_01	0.02
HNF	V$HNF1_Q6_01	V$PIT1_Q6	0.01
		V$AP4_Q6_01	0.02
	V$HNF3_Q6_01	V$CEBP_Q3	0.01
		V$NF1_Q6_01	0.01
	V$HNF4_Q6_01	V$SRF_Q5_02	<0.01
		V$USF_Q6_01	0.03

Table 2. TPR (%) of different method to identify GATA factor binding sites given fixed levels of FPR

FPR (%)	OSCAR-C	OSCAR-L	OSCAR	P-Match	MATCH
1.0	100.0	96.0	92.0	68.0	48.0
0.5	96.0	84.0	84.0	48.0	40.0
0.1	84.0	64.0	48.0	NA[a]	20.0
0.05	68.0	48.0	44.0	NA[a]	20.0

[a]The lowest FPR that P-Match could achieve on this data set is 0.23%, corresponding to a TPR of 36.0%.

consistent with biological evidence in the literature (e.g. Huang *et al*., 2004; Furusawa, *et al*., 2003).

We incorporate co-occurrences of two motifs to predict binding sites of GATA factor. Again, we use LOOCV procedure to test the TPRs of MATCH, P-Match, the basic OSCAR, OSCAR-L and OSCAR-C algorithms. FPRs are also evaluated by randomly shuffling the promoter regions in training set. Table 2 shows the TPRs of different methods given fixed number of false predictions. A complete plot of ROC curves is given in Supplementary Figure 1. We can see that there is a clear advantage from utilizing co-occurrence information, indicating that the integration of co-occurring motifs not only decreases false predictions, but also increases sensitivity by enhancing the signal strength.

3.3.2 HNF family binding sites Hepatocyte nuclear factors 1, 3 and 4 (HNF-1, -3 and -4) are liver-enriched transcription factors that function in the regulation of several liver-specific genes (Ktistaki and Talianidis, 1997). Totally 22 promoter regions are extracted from EPD85 with annotated binding sites for transcription factor HNF-1 (on 7 promoters), HNF-3 (on 7 promoters) and HNF-4 (on 13 promoters) in TRANSFAC®. Motifs co-occurring with different factors in HNF family are identified and evaluated separately, and their TRANSFAC IDs and corresponding *P*-values are shown in Table 1. These results are supported by evidence from the biological experimental literature (e.g. Antes and Levy-Wilson, 2001; Elholm *et al*., 1996; Groupp *et al*., 1994; Hiesberger *et al*., 2004; Kahn, 1997).

Table 3. TPRs (%) of different method to identify TFBSs of transcription factors in HNF family given fixed levels of FPR

FPR (%)	OSCAR-C	OSCAR-L	OSCAR	P-Match	MATCH
1.0	96.7	93.3	93.3	43.3	53.3
0.5	93.3	90.0	90.0	40.0	26.7
0.1	86.7	80.0	76.7	33.3	10.0
0.05	76.7	73.3	66.7	30.0	3.3

The co-occurrence information is exploited by OSCAR-C. Predictions are made separately for different factors in HNF family, and the results are summed up in Table 3 and Supplementary Figure 2. From the results in Table 3 and ROC curves in Supplementary Figure 2, we can find that they are consistent with what we have observed in the other studies, and that the performance of OSCAR-C is superior over other methods.

4 DISCUSSION

In this article, we use the one-class SVM algorithm to identify *cis*-regulatory elements in promoter regions when only limited, 'positive' samples are available. To improve the accuracy of prediction, we further consider the preferences of binding site locations, and the co-occurrences of other motifs in promoter regions. Applying our method to synthetic as well as real data, we have demonstrated the advantage of the proposed strategy, and confirmed the conclusion that incorporation of locational preference can improve the performance of prediction. We also illustrate how the method can be used to identify and utilize co-occurring motif pairs in predicting binding sites of two transcription factor families.

Note that, in the OSCAR algorithm, we construct a linear prediction function, which assumes an additive contribution from each position towards the score. Different from PWM-based methods, our method allows for encoding each binding site as a unitary vector. Some higher-order models, such as Bayesian tree structure (Barash *et al*., 2003) or di-nucleotide interactions (Zhou and Liu, 2004), were exploited to improve the prediction accuracy. However, even in cases where intra-site interactions exist, the additive model has been suggested to be a good approximation (Benos *et al*., 2002). In our applications, we limit to construct linear prediction rules. The 'kernel trick' in SVM, which allows for nonlinearly mapping into a high-dimensional feature space, can be used as well. We have tried both polynomial and RBF kernels in our experiments, but little improvement has been observed. Since high-dimensional feature spaces may suffer from low generalization ability, we recommend using the linear kernel in this context.

An extensive investigation of the correlated motifs in prediction can help us to better understand gene regulations mechanisms. In this article, we focus on how to recognize the binding sites of a particular transcription factor more accurately by exploring the related information more efficiently, but the method presented in this article can also be used

to infer the co-occurrences of motifs in the future. In addition, we use discrete interval indicators for the division of promoter regions in order to utilize the locational information. Given larger amount of available data, the algorithm may be further improved by using the exact positions of binding events directly. Moreover, other types of information, such as the occupancy of nucleosome (an algorithm that can use nucleosome occupancy information in TFBS prediction is in preparation), evolutionary conservation between related species, and the tissue specificity of the downstream genes, may be further exploited by incorporating them into our method.

ACKNOWLEDGEMENTS

We thank Dr Andrew Smith for his help with the Shufflet program, and Dr Xiaoyue Zhao for helpful discussions.
This work is partially supported by NSFC grants 60540420569 (X.Z. and M.Q.Z.), 30625012 (X.Z.), the National Basic Research Program 2004CB518605 (X.Z.), the Changjiang Professorship Award (M.Q.Z.) of China and NIH grant HG001696 (M.Q.Z.) of USA.

Conflict of Interest: none declared.

REFERENCES

Antes,T.J. and Levy-Wilson,B. (2001) HNF-3 beta, C/EBP beta, and HNF-4 act in synergy to enhance transcription of the human apolipoprotein B gene in intestinal cells. *DNA Cell Biol.*, **20**, 67–74.

Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, Menlo Park, California, pp. 28–36.

Barash,Y. *et al.* (2003) Modeling dependence in protein-DBA binding sites. In *RECOMB'03*. Berline, Germany.

Benos,P.V. *et al.* (2002) Additive in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Res.*, **30**, 4442–4451.

Boyer,L.A. *et al.* (2005) Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell*, **122**, 947–956.

Chang,C.C. and Lin,C. (2001) LIBSVM: a library for support vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm, accessed date: 28 November 2006 (version 2.83).

Chekmenev,D.S. *et al.* (2005) P-Match: transcription factor binding site search by combining patterns and weight matrices. *Nucleic Acids Res.*, **33**, W432–W437.

Coward,E. (1999) Shufflet: shuffling sequences while conserving the k-let counts. *Bioinformatics*, **15**, 1058–1059.

Elholm,M. *et al.* (1996) Regulatory elements in the promoter region of the rat gene encoding the acyl-CoA-binding protein. *Gene*, **173**, 233–238.

FitzGerald,P.C. *et al.* (2004) Clustering of DNA sequences in human promoters. *Genome Res.*, **14**, 1562–1574.

Frith,M.C. *et al.* (2003) Cluster-Buster: finding dense clusters of motifs in DNA sequences. *Nucleic Acids Res.*, **31**, 3666–3668.

Furusawa,M. *et al.* (2003) Molecular cloning of the mouse AMY-1 gene and identification of the synergistic activation of the AMY-1 promoter by GATA-1 and Sp1. *Genomics*, **81**, 221–233.

Groupp,E.R. *et al.* (1994) Characterization of the distal alpha-fetoprotein enhancer, a strong, long distance, liver-specific activator. *J. Biol. Chem.*, **269**, 22178–22218.

Guyon,I. *et al.* (2002) Gene selection for cancer classification using support vector machines. *Mach. Learn.*, **46**, 389–422.

Hiesberger,T. *et al.* (2004) Mutation of hepatocyte nuclear factor-1beta inhibits Pkhd1 gene expression and produces renal cysts in mice. *J. Clin. Invest.*, **113**, 814–825.

Holloway,D.T. *et al.* (2005) Integrating genomic data to predict transcription factor binding. *Genome Inform.*, **16**, 83–94.

Hong,P. *et al.* (2005) A boosting approach for motif modeling using ChIP-chip data. *Bioinformatics*, **21**, 2636–2643.

Huang,D.Y. *et al.* (2004) GATA-1 and NF-Y cooperate to mediate erythroid-specific transcription of Gfi-1B gene. *Nucleic Acids Res.*, **32**, 3935–3946.

Jaakkola,T. *et al.* (2000) A discriminative framework for detecting remote protein homologies. *J. Comput. Biol.*, **7**, 95–114.

Kahn,A. (1997) Transcriptional regulation by glucose in the liver. *Biochimie*, **79**, 113–118.

Kel,A.E. *et al.* (2003) MATCH^{TM}: a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.*, **31**, 3576–3579.

Ktistaki,E. and Talianidis,I. (1997) Modulation of hepatic gene expression by hepatocyte nuclear factor 1. *Science*, **277**, 109–112.

LeCun,Y. *et al.* (1990) Optimum brain damage. *Advances in Neural Information Processing Systems 2*. Morgan Kaufmann, San Mateo, CA. 598–605.

Leslie,C.S. *et al.* (2004) Mismatch string kernels for discriminative protein classification. *Bioinformatics*, **20**, 467–476.

Liu,X. *et al.* (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. In Altman,R.B. (ed.) *Proceedings of the 6th Pacific Symposium on Biocomputing*. Word Scientific Publish Company, Hawaii, USA, pp. 127–138.

Odom,D.T. *et al.* (2006) Core transcriptional regulatory circuitry in human hepatocytes. *Mol. Syst. Biol.*, **2**, 2006.0017. doi:10.1038/msb4100059.

Patient,R.K. and Mcghee,J.D. (2002) The GATA family (vertebrates and invertebrates). *Curr. Opin. Genet. Dev.*, **12**, 416–422.

Praz,V. *et al.* (2002) The Eukaryotic Promoter Database, EPD: new entry types and links to gene expression data. *Nucleic Acids Res.*, **30**, 322–324.

Quandt,K. *et al.* (1995) MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res.*, **23**, 4878–4884.

Rätsch,G. *et al.* (2005) RASE: recognition of alternatively spliced exons in C.elegans. *Bioinformatics*, **21**, i369–i377.

Segal,E. *et al.* (2006) A genomic code for nucleosome positioning. *Nature*, **442**, 772–778.

Sandelin,A. *et al.* (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, **32**, 91–94.

Schölkopf,B. *et al.* (2001) Estimating the support of a high-dimensional distribution. *Neural Comput.*, **13**, 1443–1471.

Sharan,R. and Myers,E.W. (2005) A motif-based framework for recognizing sequence families. *Bioinformatics*, **21**, i387–i393.

Sinha,S. and Tompa,M. (2002) Discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res.*, **30**, 5549–5560.

Sonnenburg,S. *et al.* (2005a) Large Scale Genomic Sequence SVM Classifiers. In *Proceedings of the 22nd International Conference on Machine Learning*. ACM Press, Bonn, Germany, pp. 849–856.

Sonnenburg,S. *et al.* (2005b) Learning interpretable SVMs for biological sequence classification. *RECOMB 2005, LNBI 3500*, pp. 389–407.

Stormo,G.D. *et al.* (1982) Use of the 'Perceptron' algorithm to distinguish translational initiation sites in *E.coli*. *Nucleic Acids Res.*, **10**, 2997–3011.

Sun,Y. *et al.* (2006) Using feature selection filtering methods for binding site prediction. In *Proceedings of the 5th IEEE International Conference on Cognitive Informatics*. IEEE CS Press, Beijing, China, pp. 566–571.

Tompa,M. *et al.* (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.*, **23**, 137–144.

Vert,J.P. *et al.* (2005) Kernels for gene regulatory regions. In Weiss,Y. *et al.* (eds.), *Advances in Neural Information Processing Systems 18*. MIT Press, Cambridge, MA, pp. 1401–1408.

Wingender,E. *et al.* (2000) TRANSFAC®: an integral system for gene expression regulation. *Nucleic Acids Res.*, **28**, 316–319.

Xie,X. *et al.* (2005) Systematic discovery of regulatory motifs in human promoters and 30 UTRs by comparison of several mammals. *Nature*, **434**, 338–345.

Zhou,Q. and Liu,J.S. (2004) Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics*, **20**, 909–916.