Incorporating Nucleosome Occupancy to Predict Transcription Factor Binding Sites

Bo Jiang

April 29, 2007

ABSTRACT

Traditional methods to identify potential binding sites of known transcription factors usually neglect abundant information hidden in a large genomic scale. Recent researches, however, reveal that the intrinsic nucleosome organization in genome can be predicted and may help in directing transcription factors towards their target sites. In this paper, we present a biophysical approach, Nu-OSCAR, which incorporates predicted nucleosome occupancy to aid the recognition of transcription factor binding sites. Testing on experimentally verified binding sites of yeast transcription factors, we demonstrate that our method outperforms existing algorithms that uses only site-specific binding information of transcription factors. An online server based on Nu-OSCAR is available at http://bioinfo.au.tsinghua.edu.cn/nu_oscar .

INTRODUCTION

Identification of transcription factor binding sites (TFBSs), which are also known as DNA motifs, plays a pivotal role in the understanding of the mechanism of transcriptional regulation. Recently, with the advents of high-throughput microarray techniques and ChIP-chip experiments, the availability of large amount of data on eukaryotic transcription factors (TFs) and their binding sites puts forward the computational task of predicting potential binding sites of those important regulatory elements on a genomic scale.

To study the site specificity of transcription factors and to search putative TFBSs in DNA sequences, positional weight matrices (PWMs) are constructed from local alignments of known binding sites, and adopted by several tools like MatInspector (Quandt *et al*., 1995) and MATCH^{TM} (Kel *et al*., 2003). However, a search of the binding sites on the entire genome with only local site-specific information usually returns a large number of sites, many of which are thought to occur at random and biologically non-functional *in vivo*, Is there any mechanism determining the binding of transcription factors in genomic contexts? And how the accuracy of TFBS prediction can be improved by incorporating such information?

It is well known that eukaryotic genomic DNA exists not as naked DNA, but rather as highly compacted chromatin. A fundamental component in the hierarchical levels of chromatin structure is called “nucleosome”, which contains a ~147-base-pair (bp) stretch of DNA tightly wrapped around the octamer of the histone proteins, and is separated from each other by unwrapped linker DNA that is typically 10-50 bp in length. Although the histone tail domains in nucleosomes potentially interact with some particular proteins (Jenuwein and Allis, 2001), access to DNA occupied by nucleosomes is generally inhibited for transcription factors (Richmond and Davey, 2003). An attractive hypothesis is that genomes could target the binding of transcription factors towards appropriate sites by encoding stable nucleosome over those non-functional sites. This idea suggests that the prediction of TFBSs could benefit from viewing binding events in a broader perspective. Recent studies also revealed that nucleosome organizations are partially explained by their intrinsic sequence preferences, which have significant regulatory roles (Segal et al., 2006; Ioshikhes et al., 2007). In their works, the researchers illustrated that some transcription factors show significant lower nucleosome occupancy at their functional sites.

In this paper, we derived a novel algorithm, called Nu-OSCAR, which incorporates nucleosome occupancy information to identify TFBSs in promoter regions. Our approach is based on a biophysical view of equilibrium interactions between transcription factor binding and nucleosome occupancy. By applying our method on experimentally verified binding sites of 61 yeast transcription factors, we demonstrate that the performance of TFBS prediction can be improved over site-specific approaches by considering information from a large genomic scale. Our results confirm the existence of a more refined and integrate genomic code for transcriptional regulation, and give an example of how this code can be deciphered to elucidate the regulatory process of gene expression.

MATERIALS AND METHODS

Biophysical model of interaction between TF and nucleosome DNA

First, consider a short piece of DNA *S*, which has two states: state $"FS"$ in which *S* is free of nucleosome *in vivo*, and the other state $"OS"$ in which it is occupied by nucleosome. The occupancy preferences of *S*, which can be determined from a larger genomic context around the DNA segment, are characterized by the probability that *S* is covered by a nucleosome:

$$p_o = p_o(S) = \frac{[OS]}{[FS]+[OS]},$$

where $[FS]$ and $[OS]$ denotes the concentration of *S* that is free of or occupied by nucleosomes, respectively. The overall binding process of a TF to a DNA segment *S* with two states can be represented as

$$OS + FS + TF \leftrightarrow TF - FS + TF - OS,$$

which can be broken down into two competing equilibrium reactions:

$$FS + TF \leftrightarrow TF - FS \qquad k_f = \frac{[TF - FS]}{[TF][FS]},$$

and

$$OS + TF \leftrightarrow TF - OS \qquad k_o = \frac{[TF - OS]}{[TF][OS]},$$

where $[TF]$ denotes the concentration of the transcription factor, $[TF - FS]$ and $[TF - OS]$ denote concentrations of protein-DNA complexes, k_o and k_f are the equilibrium constants of the reactions. Assume $\alpha = k_f / k_o$, and the binding energy of the corresponding TF to a DNA segment *S* that is free of nucleosome is $E(S)$, then

$$k_f = K \exp(-\beta E(S)), \tag{1}$$

where $\beta = 1/k_B T$, K is a constant, T is absolute temperature and k_B is Boltzmann constant. The equilibrium probability of *S* being bound to the corresponding TF is given by

$$p(S) = \frac{[TF - FS]+[TF - OS]}{[FS]+[OS]+[TF - FS]+[TF - OS]} = \frac{(k_f(1-p_o)+k_o p_o)[TF]}{(k_f(1-p_o)+k_o p_o)[TF]+1}.$$

According to formula (1), the above formula can be written as

$$p(S) = \frac{1}{C_{\alpha,p_o} \exp\{\beta(E(S)-\mu)\}+1}, \tag{2}$$

where $C_{\alpha,p_o} = 1/(\alpha(1-p_o)+p_o)$, and $\mu = k_B T(\ln(K[TF]) - \ln\alpha)$ is the chemical potential. When $(1-p_o)$ is relatively small, we have

$$C_{\alpha,p_o} = \exp\{-\log(1+(\alpha-1)(1-p_o))\} \approx \exp\{(1-\alpha)(1-p_o)\},$$

and thus, formula (2) can be approximated by :

$$p(S) \approx \frac{1}{\exp\{\beta(E(S)+\varepsilon_0(1-p_o)-\mu)\}+1}$$

with $\varepsilon_0 = (1-\alpha)/\beta = (1-k_f/k_o)k_B T$.

To proceed further, we need an expression of the binding energy $E(S)$. Here, we use the simplest model of protein-DNA binding, which assumes that the interaction of a given base with the factor does not depend on the neighboring bases, and expresses $E(S)$ as

$$E(S) \approx \sum_{i=1}^{L}\sum_{j=1}^{4} \varepsilon_{i,j} s_{i,j},$$

where *L* is the length of DNA segment S, and $s_{i,j}$ characterizes the sequence by setting $s_{i,j} = 1$ if the *i*-th base in S is *j*, and $s_{i,j} = 0$ otherwise. $\varepsilon_{i,j}$ is the interaction energy with the nucleotide *j* at position $i = 1,2,...,L$ of S. Furthermore, we can write the equilibrium binding probability of *S* into a Fermi-Dirac form:

$$p(S) \approx \frac{1}{\exp\{\beta(E_1(S)-\mu)\}+1}$$

with $E_1(S) = \sum_{i=1}^{L}\sum_{j=1}^{4} \varepsilon_{i,j} s_{i,j} + \varepsilon_0(1-p_o(S))$ being generalized binding energy of *S*.

Let set $\mathbf{S}_N = \{S^{(1)}, S^{(2)}, ..., S^{(N)}\}$ contain N binding sites that has been observed to be bound by a specific TF in experiments. Under the assumption that the distribution of binding energies of random DNA segments can be approximated by Gaussian distribution, it follows form an argument in Dejordjevic et al. (2003) that the maximum likelihood inference of observing the sequences comprising set $\mathbf{S}$ can be equivalent to the minimizing the variance of energies from the distribution. By further assuming a uniform background nucleotide distribution (if this is not the case, we may use $(s_{i,j} - b_j)$ instead of $s_{i,j}$, with b_j equal to the background frequency of nucleotide j), and shifting energies so that the overall average binding energy is zero, we can make inferences about binding energy and other parameters by solving the following problem:

$$\min_{\varepsilon_{i,j},\varepsilon_0} \quad \frac{1}{2}\left(\sum_{i=1}^{L}\sum_{j=1}^{4} \varepsilon_{i,j}^2 + \varepsilon_0^2\right), \tag{3}$$

$$\text{subject to} \quad E_1(S^{(k)}) = \sum_{i=1}^{L}\sum_{j=1}^{4} \varepsilon_{i,j} s_{i,j}^{(k)} + \varepsilon_0(1-p_o(S^{(k)})) \le \mu, \quad k = 1,2,...,N. \tag{4}$$

One-class support vector machine and the Nu-OSCAR algorithm

Note that we are free to set the value of chemical potential μ in (4), and Dejordjevic et al. (2003) suggested that we can fix the value of $\mu = 1$. However, it might be unnecessary, or even impossible to impose such constraint on all observed binding sites. Here, we introduce a slack variable ξ to penalize binding sites that violate the constraint in (4), and minimize the chemical potential at the same time. Then, problem (3) and (4) can be written as:

$$\min_{\varepsilon_{i,j},\varepsilon_0,\ ,\mu} \frac{1}{2}\left(\sum_{i=1}^{L}\sum_{j=1}^{4}\varepsilon_{i,j}^2 + \varepsilon_0^2\right) + \frac{1}{\nu N}\sum_{k=1}^{N}\xi_k + \mu\,, \tag{5}$$

$$\text{subject to}\quad E_1(S^{(k)}) = \sum_{i=1}^{L}\sum_{j=1}^{4}\varepsilon_{i,j}s_{i,j}^{(k)} + \varepsilon_0(1 - p_o(S^{(k)})) \le \mu + \xi_k,\ \xi_k \ge 0\,,\quad k = 1,2,...,N\,, \tag{6}$$

where $\nu \in (0,1)$ is a user-defined parameter.

At this point, we encounter a quadratic programming problem similar to one-class support vector machine (one-class SVM; Schölkopf *et al*., 2001) , which is proposed to estimate the support of a high-dimensional distribution. Note that parameter ν actually controls the trade-off between variance of energies and fraction of binding sites with energy below μ. It can be shown that ν is an upper bound on the fraction of binding sites with energy above chemical potential μ.

In our applications, we use the LIBSVM package (Chang and Lin, 2001) for the implementation of one-class SVM to solve $\varepsilon_{i,j}, \varepsilon_0$ and μ from problem (5) and (6). To predict whether an arbitrary DNA segments S with length L is bound by the TF, we can use the following prediction function:

$$f(S) = \text{sign}\left[\mu - \sum_{i=1}^{L}\sum_{j=1}^{4}\varepsilon_{i,j}s_{i,j} - \varepsilon_0(1 - p_o(S))\right],$$

which is positive when S is a putative binding site of the TF. By regarding the μ in the prediction function as threshold, we come to a scheme similar to PWM-based methods, yet in our method nucleotide preferences at different positions within a binding site are considered simultaneously in the a biophysical model, and the nucleosome occupancy information around each site is further exploited. We named our algorithm Nu-OSCAR (Nucleosome-Occupancy Study for *Cis*-elements Accurate Recognition). To make comparisons, we also implement a simpler algorithm that only use the sequential composition of TFBSs, and call it OSCAR (for more details about the OSCAR algorithm, please see Jiang et al., 2006).

Nucleosome occupancy prediction and TFBS datasets

To determine the probability of a DNA segment *S* being covered by a nucleosome, we need to predict the nucleosome occupancy status within promoter regions. To this end, the probabilistic nucleosome-DNA interaction model derived in Segal *et al*. (2006) is exploited. Let $p_o(S_i)$ be the probability that *i*-th base pair in a sequence *S* with length *L* is occupied by a nucleosome, which can be calculated by using a dynamic programming algorithm (prediction software available at http://genie.weizmann.ac.il/pubs/nucleosomes06). The calculation of p_o is straightforward,

$$p_o(S) = \frac{1}{L}\sum_{i=1}^{L} p_o(S_i)\,.$$

In our applications, the TFBS data is obtained from Young's group (Harbison *et al*. 2004). They have determined the binding sites on all the yeast promoters for 102 transcription factors with high-confidence conservation criteria. We locate binding sites of these transcription factors to 1273 intergenic sequences, and 61 transcription factors with more than 10 identified binding sites constitute our experimental datasets.

RESULTS

Nucleosome occupancies at functional and spurious sites

Predicting functional TFBSs is a very challenging task originating from the nature of binding sites that are very short in length (usually ranging from 6 to 12 bp). Therefore, putative TFBSs occur at a high frequency across a genome and result in an overabundance of false-positive predictions. A possible explanation for those spurious (very similar in pattern but presumed non-functional) sites is that nucleosome occupancy tends to inhabit non-functional sites for regulatory factors. We test this hypothesis by examining whether the functional and conserved binding sites of transcription factors had higher or lower predicted nucleosome occupancy compared to their spurious sites. To this end, we apply the basic OSCAR algorithm to scan the binding sites for the 61 transcription factors on yeast promoter regions, and define spurious sites as those prediction results that do not coincide with the experimentally reported locations of functional binding sites (Harbison *et al*., 2004). Then, nucleosome occupancies on the promoter regions are determined by using the method given in Segal *et al*. (2006). At the sensitivity level of 95% (*i.e.* 95% of functional sites are recognized by OSCAR), we perform a student t-test to compare the distribution of average predicted occupancies at the functional sites and at the spurious sites, and take $p<0.05$ to be significant. The results are shown in Figure 1 and Supplementary Table 1.

From Figure 1 and Supplementary Table 1, we can find that for 13 (21%) transcription factors the predicted nucleosome occupancy at their functional binding sites is significantly lower compared with their spurious sites. Only one (0.02%) factor show significantly higher predicted occupancy at its functional site. Those regulatory factors with significant lower predicted occupancy obtained in our experiment are highly consistent with the results in previous studies (Bernstein *et al.*, 2004; Segal *et al*., 2006). Moreover, if we lowered the threshold in determining binding sites, the effect of nucleosome occupancy will become more evident. For instance, at 99% sensitivity level, about 46% (28 out of 61) factors show significantly lower predicted nucleosome occupancy at their functional sites (see Supplementary Table 2). These results confirm the previous observation that the binding of transcription factors are influenced by nucleosomes competing for DNA occupancy, and suggest that the accuracy of TFBS prediction could be improved by considering such effect.

Overall performance of the Nu-OSCAR algorithm

In **Materials and Methods**, we developed a new algorithm, Nu-OSCAR, which combines the sequence specificities of regulatory factors and predicted nucleosome occupancies of genomic sequences in TFBS prediction. Nu-OSCAR simultaneously considers the sequential composition individual binding site and the nucleosome positioning information around each site from a larger genomic scale, the resulting prediction function depicts the thermodynamic equilibrium between the nucleosomes and the site-specific transcription factors that compete with nucleosomes for occupancy along the promoter regions. To evaluate the predictive ability of Nu-OSCAR, we

perform a Leave-One-Out Cross Validation (LOOCV) on the yeast TFBS datasets as follows: for a set with *N* known binding sites of a transcription factor, we obtain a prediction function based on *N*-1 sites and test the remaining site based on the function. This step is repeated *N* times with each site serving as testing sample. A predicted site is a true-positive (TP) prediction if it coincides with experimentally reported binding site (Harbison *et al*., 2004), 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 number of true positives to the total number of known binding sites, and false positive rate (FPR), which is the number of false positives to the total number of non-site positions. A curve of FPR versus TPR can be plotted while a threshold parameter is varied. This curve, called ROC (Receiver Operating Characteristics) curve, is a comprehensive and objective way to compare the performance of methods as a trade off between specificity and sensitivity. To make comparisons, we also use the above procedure to test the predictive ability of the basic OSCAR algorithm and P-Match (Chekmenev *et al*., 2005), which is a newly developed tool that combines pattern matching and weight matrix approaches.

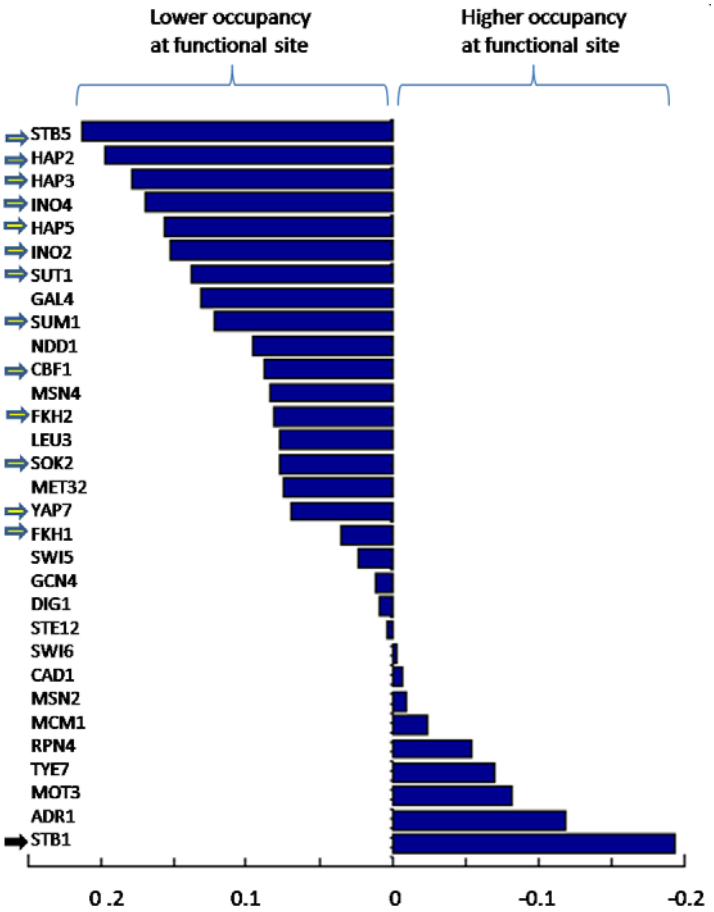

Figure 1. The histogram shows the differences in predicted nucleosome occupancy between functional and spurious sites predicted by the basic OSCAR algorithm (only results for 31 out of 61 transcription factors are shown here, and complete and detailed results are given in Table 1). Light arrows indicate 13 transcription factors showing significantly lower occupancy at functional sites compared with spurious sites; and a black arrow indicates the only factor having significantly higher occupancy at functional sites.

The results of the different methods are plotted as ROC curves in Figure 2A. As we can see,

OSCAR outperforms P-Match, which is consistent with our previous observations (Jiang *et al.*, unpublished). Notably, the performance of Nu-OSCAR is superior over the other two algorithms, especially in the area of high sensitivity, which confirms the advantage of incorporating nucleosome occupancy information. At 95% sensitivity level, Nu-OSCAR results in a smaller false positive rate than the basic OSCAR algorithm for 75% transcription factors (45 out of 61), and equal or a slightly larger false positive rate for the remaining transcription factors. To further investigate the effect of using occupancy information, we perform a student t-test to compare the distribution of average predicted occupancies at the functional sites and at the spurious sites predicted by Nu-OSCAR. Given a significant level of 0.05, only two (0.03%) factor show significantly lower predicted occupancy at its functional site (see Supplementary Table 3). The result of the test indicates that the information associated with nucleosome occupancy in regulating locations of TFBSs has been sufficiently exploited by Nu-OSCAR.

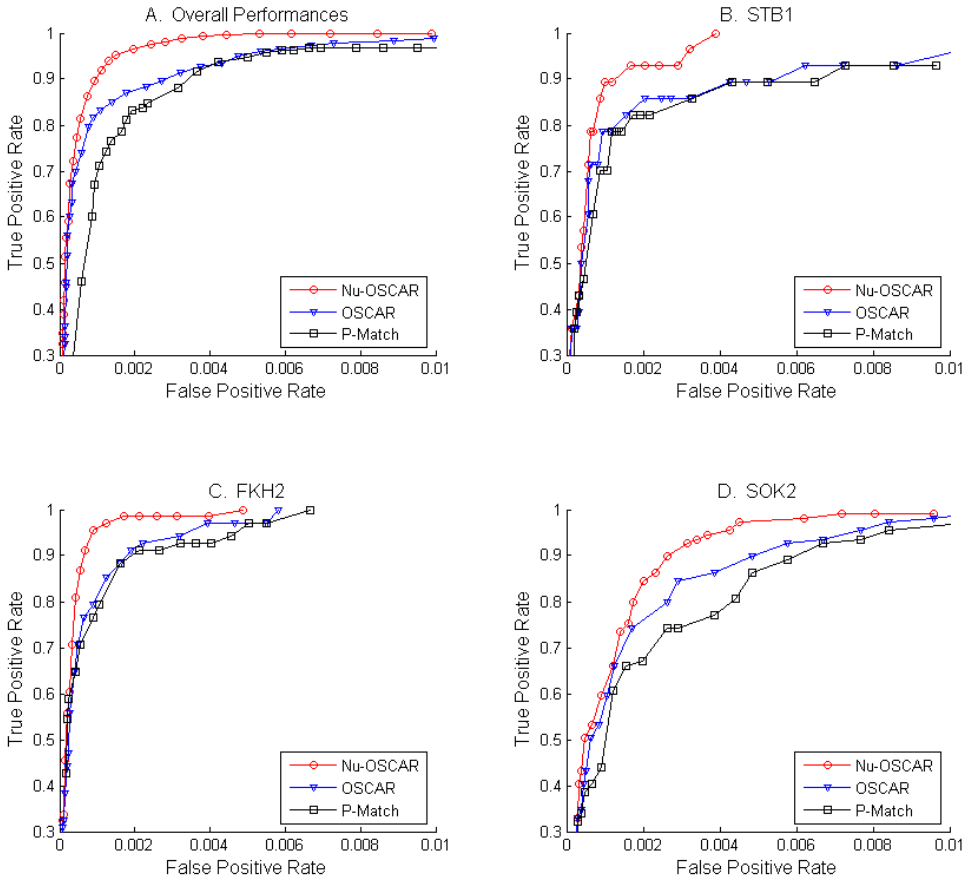

Figure 2. ROC curves indicate the high performance of the Nu-OSCAR algorithm on overall dataset with 61 transcription factors (A), STB1 factor with significant higher occupancy at functional sites (B), FKH2 factor (C) and SOK2 factor (D) with significant lower occupancy at functional sites. These results confirm the advantage of incorporating nucleosome occupancy information.

TFs with significant nucleosome occupancy preferences

Among all the transcription factors, STB1 factor shows significantly higher predicted occupancy

at its functional binding sites with a t-test p-value of 0.006. This result may be explained by the biological function of STB1, as a component of SIN3 histone deacetylase complex (Kasten and Stillman, 1997). From the ROC curves in Figure 2B, we can find that there is a clear advantage of utilizing occupancy information to predict binding sites of STB1 factor.

FKH2 and SOK2 are two transcription factors that have most significant lower occupancy at their functional sites, with t-test p-values of 0.0004 and 0.0005, respectively. Notably, FKH2 was identified as a component of SWI5 (a chromatin-remodeling complex; Pic *et al*., 2000), and the inclusion of FKH2 in the complex is facilitated by DNA bending induced by MCM1 (Kumar *et al*., 2000), which usually occurs at linker DNA. ROC curves of different methods applied in recognizing binding sites of FKH2 and SOK2 factors are given in Figure 2C and 2D. We can find that the results are consistent with what we have obtained in the previous studies, and especially that Nu-OSCAR provides superior recognition accuracies.

DISCUSSION

In this paper, we propose Nu-OSCAR, a novel approach exploiting predicted nucleosome occupancy to improve the accuracy of TFBS prediction based on a biophysical model. Applying our method to scan TFBSs on yeast promoter regions, we have demonstrated the advantage of incorporating information from a larger genomic scale, and confirmed the model in which nucleosomes mask spurious sites that occur randomly in large genome, thereby directly contributing to regulation of transcription at the chromatin level.

Note that our approach grounds on the sequence-based prediction of nucleosome occupancy, which still requires further research. Besides, DNA sequence is not the sole determinant of nucleosome occupancy. Specific mechanisms to alternate the positions of some nucleosomes regulated by chromatin remodeling enzymes under different conditions remain unresolved. An integration of biophysical and structural properties of nucleosomes and DNA, or the availability of abundant and elaborate experimental results, which allow for more accurate positioning of nucleosomes, will lead to further improvements in TFBS recognition as well. Most recently, Ozsolak *et al*. (2007) described a high-resolution approach to examine nucleosome positioning in human promoters, providing the possibility of applying our method to more complex genomes.

We just take a first step towards the understanding of dynamic interactions between the site-specific transcription factors and other elements contributing to gene regulation, and only focus on the influence of nucleosome occupancy in the basic level of complicate chromatin architecture. An extensive investigation of interactions among synergistic regulatory factors, and other epigenetic effects like histone modification code or DNA methylation, may help us better understand the regulatory mechanisms *in vivo*, and further enhance our ability to predict the regulatory sites of DNA binding proteins. In this sense, our method provides a useful algorithmic frame to incorporate other genomic encoded features to identify *cis*-regulatory elements.

References

Bernstein, B. E., Liu, C. L., Humphrey, E. L., Perlstein, E. O. and Schreiber, S. L. (2004) Global nucleosome occupancy in yeast. *Genome Biology*, 5: R62.

Chang, C.C., and Lin, C. (2001) LIBSVM : a library for support vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Chekmenev, D.S., Haid, C. and A.E. Kel. (2005) P-Match: transcription factor binding site search by combining patterns and weight matrices. *Nucleic Acids Research*, 33: W432-W437.

Djordjevic, M. and Sengupta, A.M. and Shraiman, B.I. (2003) A biophysical approach to transcription factor binding site discovery. *Genome Research*, 13: 2381-2390.

Harbison, C. T., Gordon, D. B., Lee, T. I., Rinaldi, N. J., Macisaac, K. D., Danford, T. W,, Hannett, N. M., Tagne, J. B., Reynolds, D. B., Yoo, J., Jennings, E. G., Zeitlinger, J., Pokholok, D. K., Kellis, M., Rolfe, P. A., Takusagawa, K. T., Lander, E. S., Gifford, D. K., Fraenkel E. and Young R. A. (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431, 99-104.

Ioshikhes, I. P., Albert, I., Zanton, S. J. and Pugh, B. F. (2006) Nucleosome positions predicted through comparative genomics. *Nature Genetics*, 38: 1210-1215.

Jenuwein, T. and Allis, C. D. (2001) Translating the histone code. *Science*, 293: 1074-1080.

Jiang, B., Zhang, M.Q., Zhang, X. (2007) OSCAR: One-class SVM for Accurate Recognition of *Cis*-elements. *Unpublished*. http://bioinfo.au.tsinghua.edu.cn/oscar/oscar.pdf.

Kasten, M. M. and Stillman, D. J. (1997) Identification of the Saccharomyces cerevisiae genes STB1-STB5 encoding Sin3p binding proteins. *Molecular General Genetics*, 256: 376-86.

Kel , A.E., Gößling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O.V. and Wingerder, E. (2003) MATCHTM: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Research*, 31: 3576-3579.

Kumar, R., Reynolds, D. M,, Shevchenko, A., Shevchenko, A., Goldstone, S. D., Dalton, S. (2000) Forkhead transcription factors, Fkh1p and Fkh2p, collaborate with Mcm1p to control transcription required for M-phase. *Current Biology*, 10: 896-906.

Ozsolak, F., Song, J. S., Liu, X. S. and Fisher, D. E. (2007) High-throughput mapping of the chromatin structure of human promoters. *Nature Biotechnology*, 25: 244-248.

Pic, A., Lim, F. L., Ross, S. J., Veal, E. A., Johnson, A. L., Sultan, M. R., West, A. G., Johnston, L. H., Sharrocks, A. D., Morgan, B. A. (2000) The forkhead protein Fkh2 is a component of the yeast cell cycle transcription factor SFF. *EMBO Journal*, 19:3750-61

Quandt, K., Frech, K., Karas, H., Wingender, E. and Werner, T. (1995) MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Research*, 23: 4878-4884.

Richmond, T. J. and Davey, C. V. (2003) The structure of DNA in the nucleosome core, *Nature*, 423: 145-150.

Schölkopf, B., Platt, J., Shawe-Taylor, J., Smola, A.J. and Williamson, R.C. (2001) Estimating the support of a high-dimensional distribution. *Neural Computation*, 13: 1443-1471.

Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thåström, A., Field, Y., Moore, I.K., Wang, J.Z. and Widorn J. (2006) A genomic code for nucleosome positioning. *Nature*, 442, 772-778.