BIOINFORMATICS

Vol. 21 Suppl. 1 2005, pages i204–i212
doi:10.1093/bioinformatics/bti1001

Enhanced position weight matrices using mixture models

Sridhar Hannenhalli[1,*] and Li-San Wang[2]

[1]Department of Genetics and [2]Department of Biology, University of Pennsylvania, Philadelphia, PA 19104, USA

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

ABSTRACT
Motivation: Positional weight matrix (PWM) is derived from a set of experimentally determined binding sites. Here we explore whether there exist subclasses of binding sites and if the mixture of these subclass-PWMs can improve the binding site prediction. Intuitively, the subclasses correspond to either distinct binding preference of the same transcription factor in different contexts or distinct subtypes of the transcription factor.
Result: We report an Expectation Maximization algorithm adapting the mixture model of Baily and Elkan. We assessed the relative merit of using two subclass-PWMs. The resulting PWMs were evaluated with respect to preferred conservation (relative to mouse) of potential sites in human promoters and expression coherence of the potential target genes. Based on 64 JASPAR vertebrate PWMs, 61–81% of the cases resulted in a higher conservation using the mixture model. Also in 98% of the cases the expression coherence was higher for the target genes of one of the subclass-PWMs. Our analysis of Reb1 sites is consistent with previously discovered subtypes using independent methods. Additionally application of our method to mutated sites for transcription factor LEU3 reveals subclasses that segregate into strongly binding and weakly binding sites with P-value of 0.008. This is the first study which attempts to quantify the subtly different binding specificities of a transcription factor on a large scale and suggests the use of a mixture of PWMs, instead of the current practice of using a single PWM, for a transcription factor.
Availability:
Contact: sridharh@pcbi.upenn.edu

1 INTRODUCTION

The protein complement in a cell is regulated at various levels including chromatin remodeling (Khorasanizadeh, 2004), transcription (Beckett, 2001), splicing (Singh, 2002), mRNA stability (Shim and Karin, 2002), export (Cullen, 2003) and translation (Kozak, 1992; Preiss and Hentze, 2003). While these mechanisms are poorly understood, transcriptional control is recognized as an important component of gene regulation. Gene transcription is controlled in a highly defined manner, both spatially and temporally (Davidson *et al*., 2002; Clyde *et al*., 2003; Kadonaga, 2004; Ptashne, 2004). This involves specific transcription factors (TFs) bound to their cognate *cis*-elements interacting with each other and the polymerase II complex. One of the first steps in the computational analysis of transcription is to model the DNA binding preference of a TF. The collection of binding sites obtained via different techniques like SELEX (Tuerk and Gold, 1990), footprinting (Guille and Kneale, 1997) or chromatin immunoprecipitation[1] (Horak and Snyder, 2002), can be used to represent the binding specificity of the TF. The various representations used in the literature are consensus, regular expression and the profiles or positional weight matrix (PWM) (Werner, 1999), aimed at concise representation capturing most of the known binding sites (high sensitivity) while minimizing ‘other’ sites captured by the representation (high specificity). PWM is a commonly used representation and is obtained by aligning examples of sequences bound by a TF and estimating the base preference at each position resulting in a $4 \times K$ matrix, where column i contains the base preferences for position i (Stormo, 2000). Often a TF may bind to degenerate sequences, resulting in non-specific PWMs and a high false-positive rate in binding site prediction. However, this varied binding specificity needs further inspection. It is possible that owing to a variety of reasons—dependence among positions, evolutionary mutation tendencies from base x to base y, compensatory mutation in other positions—the actual binding sites perhaps fall in subclasses. A global analysis of mutations among octamers using conservation among multiple yeasts species reveals that certain octamers fall in tightly linked clusters with functional interpretability (Tanay *et al*., 2004).

TFs operate as functional groups bound to their cognate binding sites, interacting with each other and ultimately with

[*]To whom correspondence should be addressed.

[1]In this approach one detects a set of genomic regions bound to a particular TF in the cell. The resulting regions are however large (few hundred bps). Using a combination of computational motif finding algorithms applied to bound regions one can detect the most likely motif or a PWM for the TF protein.

the polymerase II enzyme to initiate the transcription (Ptashne and Gann, 1997; Kadonaga, 2004). The set of functionally interacting TFs bound within relatively short regions are called transcriptional modules (Ludwig *et al*., 1998; Bolouri and Davidson, 2002; Thompson *et al*., 2004). This paradigm implies that the functionality of a bound factor depends on the other bound factors in the vicinity. For example, TF CREB, although constitutively bound to the cAMP response element, is more often functional in conjunction with a TATA-box (Conkright *et al*., 2003). There are also cases where the binding (as opposed to functionality) of a factor itself depends on the presence or absence of other *cis*-element or TFs (Hochschild and Ptashne, 1986; Lomvardas and Thanos, 2001). A particular TF can be part of many different modules interacting with a different set of TFs. For instance, TFs SWI4 and SWI6 interact with many different factors at different stages of the cell cycle to regulate specific genes (Lee *et al*., 2002). A slight change in the binding site sequence can change the exact binding of the TF to the DNA in way that is more or less suited for interacting with other factors (Ptashne, 2004). Thus it is conceivable that various binding sites for a factor have evolved to cope with these subtle context dependent interaction requirements.

In this paper, we investigate relative merits of representing the binding specificity of a TF using multiple PWMs instead of a single one, as is traditionally done. Given a set of binding sites that form the basis for a PWM, we search for significant subclasses of sites using mixture models (Bailey and Elkan, 1994). The determination of the most appropriate number of subclasses is a challenging issue in many different contexts. We have implemented two statistical criteria—Bayes and Akaike's Information Criteria (BIC and AIC) (Hastie *et al*., 2001)—to determine the number of subclasses. However, it is hard to interpret them biologically. Another issue with using a large number of subclasses is the lack of sufficient data (number of sites). These complications make it difficult to evaluate the results. Consequently we chose to specifically investigate merits of using two PWMs (corresponding to two subclasses) relative to the single original PWM. The rationale for this is that even if there are more than two 'true' subclasses, our two subclasses should be a generalization of those and should still show an improvement relative to the single cluster.

We evaluate the relative merits of mixture modeling using two independent criteria. The first one is based on the well-established correlation between evolutionary conservation and the functionality of a site (Thomas *et al*., 2003). The fraction of 'hits' of a PWM that are conserved can be used as a crude surrogate for how biologically relevant the PWM is. In fact a recent genome wide motif discovery approach uses preferential conservation as one of the criteria for motif detection (Kellis *et al*., 2003). Tanay *et al*. (2004) have discovered two distinct subclasses of Reb1 binding sites in yeast by analyzing the transition frequencies between different octamers based on whole genome multiple species conservation. The subtype they discovered also corresponds to differential binding energies (Wang and Warner, 1998). Application of our approach to a very limited number of experimentally validated Reb1 sites reveals the two subclasses. Furthermore, the two subtypes also show highly differential conservation relative to each other as well as relative to other potential Reb1 binding sites. Our second evaluation strategy is based on the expression coherence of the potential target genes using a PWM. If the genes targeted by a subclass-PWM have more similar expression profiles relative to the genes targeted by the original PWM, it may indicate a greater biological relevance of the subclass PWM (Pilpel *et al*., 2001; Banerjee and Zhang, 2003).

Based on 64 JASPAR vertebrate PWMs, 61% resulted in higher conservation using the subclass PWMs in human promoters relative to the original PWM. This fraction goes up to 81% when we only consider PWM whose subclass PWMs are 'very' dissimilar to each other. Also in 98% of the cases at least one of the subclass-PWM target genes showed higher expression coherence relative to genes that were the target of different subclass-PWMs. Furthermore, in 80% of the cases the average within-subclass expression coherence is higher than the across-subclass expression coherence. Additionally we applied the mixture modeling to binding sites for yeast TFs Reb1 and LEU3 and assessed the results relative to previous studies. Our analysis of Reb1 sites is consistent with previously discovered subclasses of Reb1 sites using independent methods (Tanay *et al*., 2004). There are 46 binding sites LEU3 with known DNA–protein binding energies. The two resulting subclasses of sites using our method segregate into strongly binding and weakly binding sites with P-value of 0.008. This is the first study which tries to quantify the subtly different binding specificities of TFs on a large scale. Our results suggest the use of a mixture of PWMs instead of a single PWM for some factors.

2 RESULTS

2.1 Reb1 analysis

Tanay and colleagues analyzed the transition patterns among conserved octamers in four-way alignment of yeast species (Tanay *et al*., 2004). The transition probability from one octamer to another was estimated using a large set of maximum parsimony trees, one for each octamer occurring in a gapless alignment between the four species. They discovered the 'multimodality' of the Reb1 TF. The network of octamers induced by all variants of Reb1 consensus reveals a two-family structure, where low-rate arcs separate a large Reb1 binding site family from a smaller one. The Reb1 promoter contains two autoregulatory sites, TTACCC$\underline{\text{G}}$ and TTACCC$\underline{\text{T}}$, each located in a different family, with distinct binding site affinities (Wang and Warner, 1998). The authors suggest the hypothesis of two distinct Reb1 modes of operation, each activating a different group of motifs in specific concentration. In order to detect these families using our mixture modeling, we need a fair number of experimentally verified binding sites, which does not exist for Reb1. Perhaps for this reason Reb1 is not represented in JASPAR. Simple observation of Reb1 sites in TRANSFAC revealed a strong preference for a 'T' in the last position and G is indeed the second most preferred position. When we applied our mixture modeling on the 15 experimentally known Reb1 sites extracted from TRANSFAC v7.2, the last position of the first subclass-PWM has G as the most probable base (probability = 0.85) and prob(T) = 0.15, whereas the last position of the second subclass-PWM has T as the most probable base with prob(T) = 0.5 and prob(G) = 0. Our mixture modeling segregated the sites among the preference at the last base into T or a G. We computed the occurrence of all four octamers (varying the base at the last position) in the yeast promoter regions for ~6000 genes in the 700 bp upstream regions using the *Saccharomyces* genome database (http://www.yeastgenome.org/). We also computed the fraction of these hits that fell in the four-way aligned region based on Kellis *et al*. (2003). The four octamers corresponding to four bases at the last position have percentage conservation of 13, 12, 52 and 19% for A, C, G and T, respectively. This analysis is consistent with previous studies and also provides further rationale for the use of preferential conservation as a means to reveal functionality.

2.2 Mixture PWMs revealed more conserved hits

There are numerous studies showing the correlation between evolutionary conservation and functionality of a region (Thomas *et al*., 2003). The whole genome binding site identification in yeast and human explicitly exploits this with significant success (Levy and Hannenhalli, 2002; Kellis *et al*., 2003; Xie *et al*., 2005). We use conservation statistics as a surrogate for functionality, to assess the relative advantage of using a mixture model. The following analysis was based on 64 vertebrate TFs in JASPAR (Sandelin *et al*., 2004). See the Methods section for details on how we arrived at this number. We applied the mixture modeling to these 64 PWMs resulting in two subclass PWMs $M_1$and M_2 for each original PWM M. To assess the relative merit of using mixture models with respect to conservation of potential binding sites in the human promoter region, we compared the following quantities, each measuring the fractional conservation in human promoters when using various PWMs:

- $ConsrFrac_{orig}$: using the original PWM M,
- $ConsrFrac_1$: using the first subclass PWM M_1,
- $ConsrFrac_2$: using the second subclass PWM M_2,
- $ConsrFrac_{1||2}$: using a 'mixture' of M_1 and M_2 according to their mixing probabilities (see Methods section).

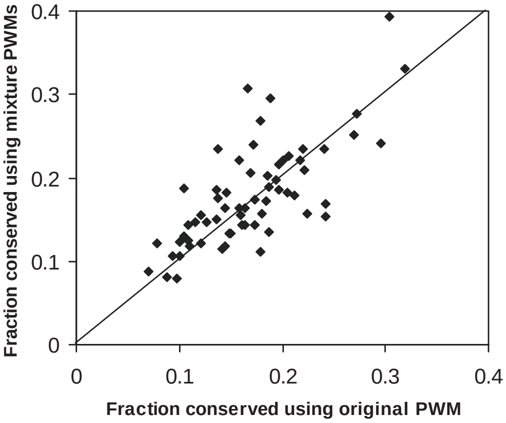

Fig. 1. For a majority (61%) of PWMs, the binding sites obtained using the mixture PWMs has better conservation in the human promoter regions. When we limit ourselves to the 25 original PWMs with fraction conservation ≤0.15, this number goes up to 86%.

Out of 64 JASPAR IDs, in 35 cases $ConsrFrac_{orig} < ConsrFrac_1$, in 35 cases $ConsrFrac_{orig} < ConsrFrac_2$, in 39 cases $ConsrFrac_{orig} < ConsrFrac_{1||2}$, in 25 cases all of the above were true. These 25 cases are listed in Table 1, and Figure 1 shows the distribution of $ConsrFrac_{orig}$ and $ConsrFrac_{1||2}$ for all 64 JASPAR IDs analyzed, 39 (61%) of which show an improvement. Also, among the 25 original PWMs for which the fractional conservation is <0.15, we observed an improvement with respect to fractional conservation in 19 (76%) using the mixture PWMs.

2.3 More dissimilar subclass PWMs offer greater improvements

Our mixture modeling gave us the best (in the sampled space) partitioning of the sites into subclasses. However, the associated likelihood score cannot be interpreted biologically as a measure of fitness in absolute terms. We used a measure of relative entropy between the subclass PWMs as a means to calibrate the fitness of the mixture models. Intuitively, the more dissimilar the two subclass PWMs are (and hence higher relative entropy), the more confidence we have in the existence of subclasses of binding specificities. The mean and standard deviation of the relative entropy per position (see the Methods section) for all pairs of the 64 TFs were 1.3 and 0.5, respectively. We found that we obtained improvements in fractional conservation for those JASPAR IDs with high relative entropy between the subclass PWMs (Fig. 2). The higher the threshold we set for relative entropy, the greater is the fraction of JASPAR IDs with increased fractional conservation observed: from 61% at the unrestricted case, the proportion of JASPAR IDs showing better conservation when the mixture PWMs were used went up to 81% when we only considered the PWMs for which the relative entropy between the subclass PWMs is at least 1.6.

Table 1. The 25 TFs that showed an improved conservation in the mixture models in $\mathit{ConsrFrac}_1$, $\mathit{ConsrFrac}_2$ and $\mathit{ConsrFrac}_{1||2}$ relative to $\mathit{ConsrFrac}_{orig}$

TF	JASPAR ID	$\mathit{Consr\ Frac}_{orig}$	$\mathit{Consr\ Frac}_1$	$\mathit{Consr\ Frac}_2$	$\mathit{Consr\ Frac}_{1\|\|2}$
Ahr-ARNT	MA0006	0.16	0.22	0.22	0.22
Chop-cEBP	MA0019	0.10	0.12	0.14	0.13
E2F	MA0024	0.20	0.20	0.21	0.22
FREAC-4	MA0031	0.14	0.16	0.16	0.16
GATA-3	MA0037	0.17	0.37	0.18	0.31
HLF	MA0043	0.10	0.17	0.18	0.19
Irf-2	MA0051	0.10	0.13	0.11	0.12
MEF2	MA0052	0.11	0.12	0.14	0.12
MZF_1–4	MA0056	0.21	0.22	0.22	0.23
MZF_5–13	MA0057	0.14	0.20	0.14	0.18
Myc-Max	MA0059	0.14	0.23	0.21	0.23
NRF-2	MA0062	0.22	0.22	0.23	0.23
Nkx	MA0063	0.14	0.18	0.16	0.15
PPARgamma	MA0066	0.13	0.14	0.15	0.15
Pax6	MA0069	0.11	0.15	0.13	0.14
RORalfa-1	MA0071	0.17	0.22	0.19	0.21
RORalfa-2	MA0072	0.08	0.12	0.10	0.12
SAP-1	MA0076	0.14	0.17	0.18	0.18
SPI-B	MA0081	0.18	0.22	0.24	0.27
SRY	MA0084	0.12	0.14	0.12	0.15
Staf	MA0088	0.11	0.14	0.11	0.12
USF	MA0093	0.15	0.19	0.18	0.18
Yin-Yang	MA0095	0.20	0.21	0.25	0.22
c-ETS	MA0098	0.30	0.33	0.44	0.39
n-MYC	MA0104	0.19	0.24	0.30	0.29

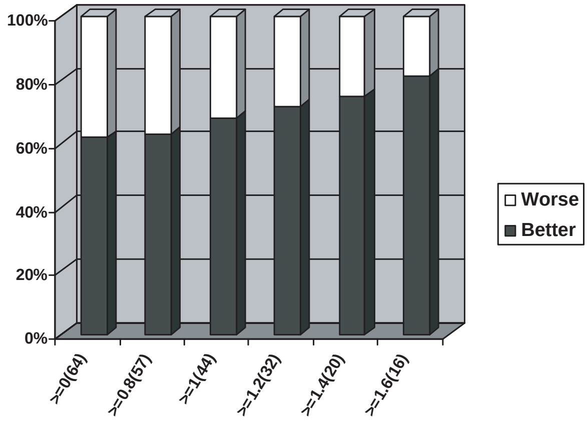

Fig. 2. The plot indicates that of the PWMs with higher relative entropy between subclass PWMs a larger fraction show improvement in fractional conservation. X-axis shows the threshold for the relative entropy, along with the number of PWMs qualifying that threshold in parentheses. The dark bar indicates the fraction of PWMs for which $\mathit{ConsrFrac}_{1||2} \geq \mathit{ConsrFrac}_{orig}$.

2.4 LEU3 analysis

To predict the DNA binding affinity of the yeast Leu3 TF, the *in vitro* equilibrium dissociation constants for 46 binding site variants was measured (Liu and Clarke, 2002). The authors also established that in this case the free energy of binding can be approximated as a sum of free energy contributions from each base. We applied our mixture modeling to the set of 46 sites and constructed two alternative PWMs. The WebLogo images (http://weblogo.berkeley.edu/) for the original and the two mixture PWMs are shown in Figure 3.

Recall that the two subtypes of Reb1 binding sites have differential binding energies (Wang and Warner, 1998). We checked whether the two clusters we detected have difference in binding energies. This is indeed the case. Out of 46 sites, cluster 1 had 22 sites and cluster 2 had 24 sites. At the energy cut-off of 27 nm K_d, 21 out of 22 (95%) subclass-1 sites fell above this threshold whereas 15 out 24 (62%) subclass-2 sites fell above this threshold. Given the 46 sites with their energy values and a given energy threshold if there were a total of 10 ‘low’ energy sites and 36 ‘high’ energy sites (based on a threshold of 27 nm K_d), and if we select 22 sites at random, the probability of choosing at least 21 (i.e. 21 or 22) ‘high’ energy sites is

$$\binom{10}{1}\cdot\binom{36}{21}\Big/\binom{46}{22}+\binom{36}{22}\Big/\binom{46}{22}=0.008.$$

Indeed the mixture models revealed binding site subtypes with different binding affinities similar to Reb1.

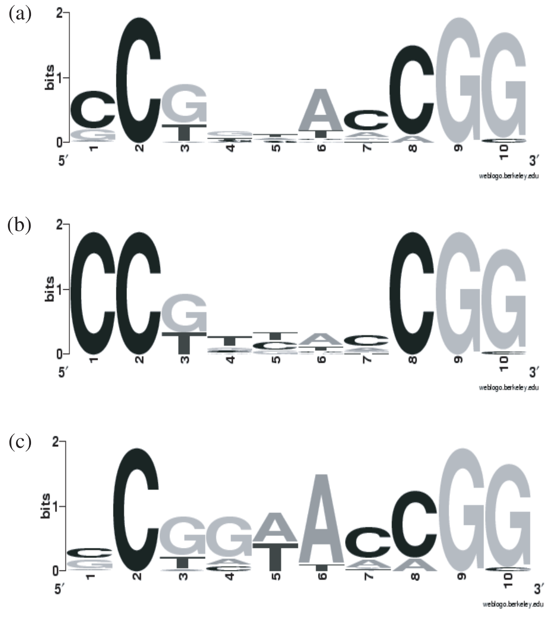

Fig. 3. The WebLogo image showing the positional weight matrices; (**a**) corresponds to the original set of Leu3 sites studies in (Liu and Clarke, 2002), and (**b**), (**c**) correspond to the two subclasses obtained from the mixture modeling.

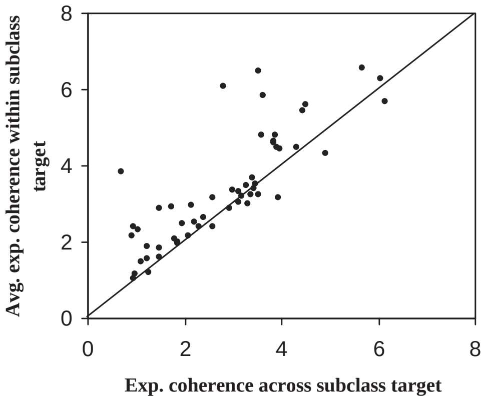

Fig. 4. The expression coherence across subclass-PWM targets against the weighted average of within-subclass target expression coherence for both subclasses.

2.5 Target genes of subclass PWMs revealed higher expression coherence

Out of 64 JASPAR IDs, in 9 cases one of the subclass PWMs yielded one or no unique target genes (see Methods section). For the remaining 55 cases, we computed the expression coherence within and across the unique target genes for each of the two subclass-PWMs (data not shown). Figure 4 compares the expression coherence across subclass-PWM targets against the weighted average of within subclass-PWM target expression coherence scores. In 44 of the 55 (80%) cases, the average expression coherence within subclass-PWM targets was higher than expression coherence of across subclass targets. Moreover, in all but one cases (98%) at least one of the two subclass PWMs had a coherence score higher than the cross coherence score. We regard these observations as independent evidence (in the sense that the improvements are not sequenced-based, as in the analysis of LEU3 dataset) supporting the notion that subclass PWMs based on our mixture model often provide better fidelity when used in transcription binding site prediction than the ‘use-all-observations’ single-PWM approach.

2.6 Cross-validation

An accurate prediction of independent, experimentally obtained set of binding sites provides the most direct test of the proposed approach. However, the scarcity of experimentally verified binding sites presents the obstacle for such an analysis on a large scale. Both TRANSFAC and JASPAR matrices are based on the handful of experimentally verified data available to date. An alterntaive way to perform such an evaluation is via cross-validation, requiring relatively large datasets. For a TF, the 10-fold cross validation works as follows. The known binding sites were randomly divided in 10 equal parts. Mixture models were generated using the nine parts and each site in the remaining part was scored using the original matrix and mixture matrices. The score obtained by original matrix was converted into P-value using the score distribution on a 1 Mb randomly selected human genomic sequence. The same was done for the mixture model score where the score by mixture of two PWMs was defined as maximum of the two scores. Thus we obtain two P-values for each site, one using original PWM and another using mixture. Comparison of these paired P-values was used to assess the relative advantage of mixture models.

As a proof of concept (owing to lack of data) we chose from among the matrices, which showed improvements based on conservation analysis, the three JASPAR matrices with at least 50 observed sites—MA0037 (67 sites), MA0052 (58 sites) and MA0080 (57 sites). For MA0037 and MA0080, the mixture model yielded better (lower) P-values for a majority of the sites. The one-sided Wilcoxon paired rank sum test showed a significance of 0.05 for MA0037 and 10^{-8} for MA0080.

3 METHODS

3.1 Positional weight matrix and binding sites

Among the two main resources for TFs binding sites and PWMs, namely TRANSFAC (Wingender *et al*., 1996) and JASPAR (Sandelin *et al*., 2004), we chose to use the latter, since it uses a higher stringency in choosing the published experimental data, and typically there are larger number of sites available for each PWM. JASPAR had 79 vertebrate TFs at the time of this study. Among these 79 vertebrate

TFs, the binding sites were available only for 66. The number of sites for each factor varies between 7 and 76 with an average of 29. Among these 66 factors, the binding site width for two of the factors is five bases and consequently the PWMs have very poor information content. Our search of potential binding sites for these factors at a specific P-value threshold (described below) does not yield any hits in the human promoter regions and are excluded from further analysis. Thus, the final set of PWMs that we analyzed is 64.

3.2 Human promoters and human–mouse conservation

Human promoters (1 kb upstream and 200 bp downstream of the $5'$ end of the gene) were extracted from July 2003 freeze of the UCSC genome database (http://www.genome.ucsc.edu/) for a total of 12 253 genes, corresponding to full length mRNAs from the DBTSS database (dbtss.hgc.jp). We also extracted the human–mouse conservation data from axtNET dataset at UCSC (Schwartz *et al.*, 2003). On average the fraction of promoter bases conserved between human and mouse is 0.32 with a minimum of 0.0 and a maximum of 0.93.

3.3 Estimating the mixture model using the EM algorithm

We adapted the probabilistic model in the MEME motif finding algorithm introduced in Bailey and Elkan (1995) to allow a mixture of an arbitrary number of base PWMs. Assume the input consists of an $m \times n$ data matrix X of m aligned DNA sequences of length n. The j-th symbol in the i-th sequence is denoted by X_{ij}. The underlying model is a mixture of k base PWMs. Each model j is associated with a $4 \times n$ PWM M_j; the probability of nucleotide x occurring at position i is $M_j[x, i]$. The relative probabilities of the mixture are specified by k non-negative numbers summing up to $1 : \lambda_1, \dots, \lambda_k$. The probability of observing sequence $X_i = (X_{i1}, \dots, X_{in})$ is

$$\Pr(X_i|M_1, \dots, M_k, \lambda_1, \dots, \lambda_k) = \sum_{j=1}^{k} \lambda_j \prod_{u=1}^{n} M_j[X_{iu}, u]$$

In the EM algorithm we introduce the matrix of indicator variables $Z = (Z_{ij})$ for the group memberships of input sequences: $Z_{ij} = 1$ indicates X_i is drawn from PWM M_j; Z and X form the complete data for the expectation maximization (EM) algorithm. The expected log likelihood of the complete data for any particular iteration in the EM algorithm is

$$\begin{aligned} &E_{Z|X,M_1,\dots,M_k,\lambda_1,\dots,\lambda_k}[l(M_1, \dots, M_k, \lambda_1, \dots, \lambda_k|Z, X)] \\ &\quad = \sum_{i=1}^{m} \sum_{j=1}^{k} E[Z_{ij}|X, M_1, \dots, M_k, \lambda_1, \dots, \lambda_k] \\ &\quad \times \log(\Pr(X_i|M_j)\lambda_j) \end{aligned}$$

Let $EZ_{ij} = E[Z_{ij}|X_i, M_1, \dots, M_k, \lambda_1, \dots, \lambda_k]$. This value can be computed using the Bayes' rule:

$$\begin{aligned} EZ_{ij} &= E[Z_{ij}|X_i, M_1, \dots, M_k, \lambda_1, \dots, \lambda_k] \\ &= \Pr(Z_{ij} = 1|X_i, M_i, \lambda_1, \dots, \lambda_k) \\ &= \frac{\Pr(X_i|M_j)\lambda_j}{\sum_{v=1}^{k} \Pr(X_i|M_v)\lambda_v} \end{aligned}$$

The M-step in the EM algorithm maximizes the expected complete log-likelihood over M and λ's. First, let $c_{juv} = \sum_{i=1}^{m} I(X_{iu} = v)EZ_{ij}$, the expected number of occurrences of symbol v at position u in PWM M_j. Using calculus one can show the following values maximize the expected complete log-likelihood (when $\beta = 0$):

$$\lambda_j^* = \frac{1}{m} \sum_{i=1}^{m} EZ_{ij},$$

$$M_j^*[v, u] = (c_{juv} + \beta) \Bigg/ \left(\sum_{p \in \{A,C,G,T\}} (c_{jup} + \beta) \right).$$

Here we introduce a non-negative, small constant $\beta = 10^{-6}$ so every parameter is positive even though $c_{juv} = 0$ (symbol v never occurred at position u in PWM M_j), since sometimes we set $c_{juv} = 0$ erroneously because the size of observations is small. We then update EZ_{ij} using the newly obtained parameters. The algorithm iterates until the likelihood converges.

3.4 Preferential conservation in promoters

For each PWM M, we computed the top 1000 scoring hits for M in the 12 253 promoter regions. This corresponds to a P-value of approximately one hit every 15 kb. This was done using an implementation of the tool *PWMSCAN* described in Levy and Hannenhalli (2002). Consider the two subclass PWMs M_1 and M_2 resulting from the matrix M. The mixture modeling described above not only compute the mixture PWMs, it also estimates the mixing probability or weights w_1 and w_2 of the two PWMs in the mixture. For each PWM M, we computed four sets of potential binding sites. Each set contained the 1000 top scoring hits in the 12 253 promoter regions using:

- Set_{orig}: the original PWM M.
- Set_1: the cluster PWM M_1.
- Set_2: the cluster PWM M_2.
- $Set_{1\|2}$: cluster PWM M_1 or M_2 using the mixing probabilities computed the mixture modeling. For a total of 1000 hits using the original matrix M, we chose top n_1 and n_2 hits for M_1 and M_2 such that $n_1 + n_2 = 1000$, and $n_1/n_2 = w_1/w_2$.

For each of these sets, we calculated the fraction of 1000 sites that were conserved between human and mouse (percentage identity ≥ 80%) using the human–mouse alignment from the UCSC axtNET dataset. Denote these by $ConsrFrac_{orig}$, $ConsrFrac_1$, $ConsrFrac_2$ and $ConsrFrac_{1\|2}$ respectively for the four sets mentioned above. To compare the relative merits of using the original PWM M versus using the two mixture PWMs M_1 and M_2, we compared $ConsrFrac_{orig}$ against $ConsrFrac_1$, $ConsrFrac_2$ and $ConsrFrac_{1\|2}$.

Given two PWMs M_1 and M_2, let p_{ijk} denote the probability of base k at position j of matrix M_i. Relative entropy is defined as

$$\mathrm{RE}(M_1, M_2) = \sum_{j=1}^{l} \sum_{k \in \{A,C,G,T\}} p_{1jk} \cdot \ln\left(\frac{p_{1jk}}{p_{2jk}}\right).$$

Clearly this is an asymmetric measure. We took the average of $\mathrm{RE}(M_1, M_2)$ and $\mathrm{RE}(M_2, M_1)$ as the relative entropy 'between' M_1 and M_2. We then normalized this by the PWM width l.

3.5 Expression coherence

For each pair of mixtures PWMs M_1 and M_2, let $G(M_1)$ and $G(M_2)$ be the sets of genes whose promoters had a 100% conserved hit by PWM M_1 or M_2, respectively. Let $UG(M_1) = G(M_1) - (G(M_1) \cap G(M_2))$ be the set of genes that are unique targets of M_1 (and, respectively, for M_2). We used the expression coherence (Pilpel *et al*., 2001; Banerjee and Zhang, 2003) to indicate the level of correlation in gene expression. Given a set of genes $UG(M)$, the expression coherence is the fraction of all pairwise correlations of $UG(M)$ exceeding the 99% percentile (the original paper used 95%) of pairwise correlations of 1000 randomly chosen genes. The cross-coherence score is defined similarly: we examined the fraction of all correlations across the two gene sets $UG(M_1)$ and $UG(M_2)$ exceeding the same threshold. To assess the improvement in expression coherence we take the weighted average of the within subclass PWM targets and compare against the expression coherence score of the across subclass targets. The weights are the number of gene-pairs in respective subclass-PWM target sets.

We used the Novartis U133A+GNF1B_101402 dataset (Su *et al*., 2004) to compute the coherence score. The dataset consists of 79 hybridization experiments on Affymetrix Human U133A array using distinct human tissues; each experiment has two replicates. Since some TFs may be ubiquitous for a substantial proportion of tissue types, we selected a subset of tissues for each TF such that the distribution of gene expression levels for genes in $G(M_1) \cup G(M_2)$ is significantly different from that for all genes using Fisher's randomization test with significance 0.05 (Conover, 1999). We then log-transformed the expression levels, and computed the coherence score.

4 DISCUSSION

Characterizing the binding specificity of TFs is a critical step in transcriptional regulation analysis. In the effort to improve the binding site prediction researcher have considered additional clues besides the binding specificity of a single TF. These include cross-species conservation (Levy and Hannenhalli, 2002; Kellis *et al*., 2003), clustering of binding sites (Berman *et al*., 2002), TF interactions (Pilpel *et al*., 2001; Hannenhalli and Levy, 2002; Lee *et al*., 2002; Banerjee and Zhang, 2003; Segal *et al*., 2003). In this work we have focused on a different aspect of the binding specificity of a TF. Various evolutionary pressures, compensatory mutations owing to dependence between positions within a binding site may lead to various subclasses of binding sites in different genetic and environmental contexts. We have specifically investigated the existence of subclasses of binding sites for a large set of TFs using the statistical mixture modeling technique, and whether employing the subclass PWMs can be used to predict binding sites with greater fidelity relative to using a single PWM without regard to the subclasses. We have shown that this is indeed the case for many TFs. Based on the genome wide chromatin immunoprecipitation assay, 70% of regions bound by CREB (see the Introduction section) do not have the standard CGTCA motif (Euskirchen *et al*., 2004). Possible reasons for this could be the existence of subclasses of CREB binding as well as dependencies on other factors for CREB binding. This work highlights an additional level of complexity in the accurate prediction of TF binding sites that have not been addressed so far.

Clearly multiple PWMs carry more information and can more specifically represent the known binding sites compared with a single PWM. There are other biological reasons why our mixture models yielded better results than the single-PWM approach. Owing to limited number of binding site observations, the resulting PWM may be strongly affected by outliers, and our mixture models were capable of detecting and rejecting them. Furthermore, different modes of operation for TF binding sites do exist (for instance Reb1), and these modes are determined by different distributions of binding site sequences; our mixture models detected this, and classified the binding sites into different subclasses.

Apart from the mixture PWMs, the EM approach also estimates the mixing probabilities of the PWMs. This provides a rational means to control for the overall false discovery rate while applying the mixture models for TFBS identification. The overall false positive rate is divided into the false positive rates for the individual matrices according to the mixing probabilities (see Section 2.2 for a specific example).

Finally, to make our method more applicable, means to estimate the optimal number of base-class PWMs is necessary; in our experience neither BIC nor AIC (see Introduction section) could return non-trivial estimates owing to lack of observations: whereas BIC always penalizes multiple clusters, AIC rewards multiple clusters excessively. As more experimental binding data become available [for instance under NIH's ENCODE initiative (http://www.genome.gov/10005107)], we will further assess the effectiveness of mixture modeling, as well as the means for controlling the number of base-class PWMs.

ACKNOWLEDGEMENTS

This research was supported by NIH 1-T32-HG000046-06 for L.-S.W. and Commonwealth of PA grant for S.H. We thank Junwen Wang for help with compiling promoter sequences.

REFERENCES

Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.

Bailey,T.L. and Elkan,C. (1995) The value of prior knowledge in discovering motifs with MEME. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **3**, 21–29.

Banerjee,N. and Zhang,M.Q. (2003) Identifying cooperativity among transcription factors controlling the cell cycle in yeast. *Nucleic Acids Res.*, **31**, 7024–7031.

Beckett,D. (2001) Regulated assembly of transcription factors and control of transcription initiation. *J. Mol. Biol.*, **314**, 335–352.

Berman,B.P., Nibu,Y., Pfeiffer,B.D., Tomancak,P., Celniker,S.E., Levine,M., Rubin,G.M. and Eisen,M.B. (2002) Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. *Proc. Natl Acad. Sci. USA*, **99**, 757–762.

Bolouri,H. and Davidson,E.H. (2002) Modeling DNA sequence-based cis-regulatory gene networks. *Dev. Biol.*, **246**, 2–13.

Clyde,D.E., Corado,M.S., Wu,X., Pare,A., Papatsenko,D. and Small,S. (2003) A self-organizing system of repressor gradients establishes segmental complexity in Drosophila. *Nature*, **426**, 849–853.

Conkright,M.D., Guzman,E., Flechner,L., Su,A.I., Hogenesch,J.B. and Montminy,M. (2003) Genome-wide analysis of CREB target genes reveals a core promoter requirement for cAMP responsiveness. *Mol. Cell*, **11**, 1101–1108.

Conover,W. (1999) *Practical Nonparametric Statistics*. John Wiley and Sons.

Cullen,B.R. (2003) Nuclear RNA export. *J. Cell. Sci.* **116**, 587–597.

Davidson,E.H., Rast,J.P., Oliveri,P., Ransick,A., Calestani,C., Yuh,C.H., Minokawa,T., Amore,G., Hinman,V., Arenas-Mena,C. *et al*. (2002) A genomic regulatory network for development. *Science*, **295**, 1669–1678.

Euskirchen,G. Royce,T.E., Bertone,P., Martone,R., Rinn,J.L., Nelson,F.K., Sayward,F., Luscombe,N.M., Miller,P., Gerstein,M. *et al*. (2004) CREB binds to multiple loci on human chromosome 22. *Mol. Cell. Biol.*, **24**, 3804–3814.

Guille,M.J. and Kneale,G.G. (1997) Methods for the analysis of DNA–protein interactions. *Mol. Biotechnol.*, **8**, 35–52.

Hannenhalli,S. and Levy,S. (2002) Predicting transcription factor synergism. *Nucleic Acids Res.*, **30**, 4278–4284.

Hastie,T., Tibshirani,R. and Friedman,J. (2001) *The Elements of Statistical Learning*. Berlin, Springer-Verlag.

Hochschild,A. and Ptashne,M. (1986) Cooperative binding of lambda repressors to sites separated by integral turns of the DNA helix. *Cell*, **44**, 681–687.

Horak,C.E. and Snyder,M. (2002) ChIP-chip: a genomic approach for identifying transcription factor binding sites. *Methods Enzymol.*, **350**, 469–483.

Kadonaga,J.T. (2004) Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors. *Cell*, **116**, 247–257.

Kellis,M., Patterson,N., Endrizzi,M., Birren,B. and Lander,E.S. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, **423**, 241–254.

Khorasanizadeh,S. (2004) The nucleosome. From genomic organization to genomic regulation. *Cell*, **116**, 259–272.

Kozak,M. (1992) Regulation of translation in eukaryotic systems. *Annu. Rev. Cell Biol.*, **8**, 197–225.

Lee,T.I., Rinaldi,N.J., Robert,F., Odom,D.T., Bar-Joseph,Z. Gerber,G.K., Hannett,N.M., Harbison,C.T., Thompson,C.M., Simon,I. *et al*. (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, **298**, 799–804.

Levy,S. and Hannenhalli,S. (2002) Identification of transcription factor binding sites in the human genome sequence. *Mamm. Genome*, **13**, 510–514.

Liu,X. and Clarke,N.D. (2002) Rationalization of gene regulation by a eukaryotic transcription factor: calculation of regulatory region occupancy from predicted binding affinities. *J. Mol. Biol.*, **323**, 1–8.

Lomvardas,S. and Thanos,D. (2001) Nucleosome sliding via TBP DNA binding *in vivo*. *Cell*, **106**, 685–696.

Ludwig,M.Z., Patel,N.H. and Kreitman,M. (1998) Functional analysis of eve stripe 2 enhancer evolution in Drosophila: rules governing conservation and change. *Development*, **125**, 949–958.

Pilpel,Y., Sudarsanam,P. and Church,G.M. (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.*, **29**, 153–159.

Preiss,T. and Hentze,M.W. (2003) Starting the protein synthesis machine: eukaryotic translation initiation. *Bioessays*, **25**, 1201–1211.

Ptashne,M. (2004) *A Genetic Switch*. Cold Spring Harbor Laboratory Press.

Ptashne,M. and Gann,A. (1997) Transcriptional activation by recruitment. *Nature*, **386**, 569–577.

Sandelin,A., Alkema,W., Engstrom,P., Wasserman,W.W. and Lenhard,B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, **32**, D91–D94.

Schwartz,S., Kent,W.J., Smit,A., Zhang,Z., Baertsch,R., Hardison,R.C., Haussler,D. and Miller,W. (2003) Human-mouse alignments with BLASTZ. *Genome Res.*, **13**, 103–107.

Segal,E., Yelensky,R. and Koller,D. (2003) Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. *Bioinformatics*, **19** (Suppl. 1) i273–282.

Shim,J. and Karin,M. (2002) The control of mRNA stability in response to extracellular stimuli. *Mol. Cells*, **14**, 323–331.

Singh,R. (2002) RNA–protein interactions that regulate pre-mRNA splicing. *Gene Expr.*, **10**, 79–92.

Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.

Su,A.I., Wiltshire,T., Batalov,S., Lapp,H., Ching,K.A., Block,D., Zhang,J., Soden,R., Hayakawa,M., Kreiman,G. *et al*. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA*, **101**, 6062–6067.

Tanay,A. Gat-Viks,I. and Shamir,R. (2004) A global view of the selection forces in the evolution of yeast cis-regulation. *Genome Res.*, **14**, 829–834.

Thomas,J.W., Touchman,J.W., Blakesley,R.W., Bouffard,G.G., Beckstrom-Sternberg,S.M., Margulies,E.H., Blanchette,M., Siepel,A.C., Thomas,P.J., McDowell,J.C. *et al*. (2003) Comparative analyses of multi-species sequences from targeted genomic regions. *Nature*, **424**, 788–793.

Thompson,W., Palumbo,M.J., Wasserman,W.W., Liu,J.S. and Lawrence,C.E. (2004) Decoding human regulatory circuits. *Genome Res.*, **14**, 1967–1974.

Tuerk,C. and Gold,L. (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, **249**, 505–510.

Wang,K.L. and Warner,J.R. (1998) Positive and negative autoregulation of Reb1 transcription in *saccharomyces cerevisiae*. *Mol. Cell Biol.*, **18**(7), 4368–4376.

Werner,T. (1999) Models for prediction and recognition of eukaryotic promoters. *Mamm. Genome*, **10**, 168–175.

Wingender,E., Dietze,P., Karas,H. and Knuppel,R. (1996) TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res.*, **24**, 238–241.

Xie,X., Lu,J., Kulbokas,E.J., Golub,T.R., Mootha,V., Lindblad-Toh,K., Lander,E.S. and Kellis,M. (2005) Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature*. **434**, 338–345.