BIOINFORMATICS ORIGINAL PAPER
Vol. 21 no. 11 2005, pages 2636–2643
doi:10.1093/bioinformatics/bti402

Sequence analysis

A boosting approach for motif modeling using ChIP-chip data

Pengyu Hong[1], X. Shirley Liu[2], Qing Zhou[1], Xin Lu[2], Jun S. Liu[1,2] and Wing H. Wong[1,2,*]

[1]Department of Statistics, Harvard University, Cambridge, MA 02138, USA and [2]Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, USA

Received on July 30, 2004; revised on January 10, 2005; accepted on March 21, 2005
Advance Access publication April 7, 2005

ABSTRACT

Motivation: Building an accurate binding model for a transcription factor (TF) is essential to differentiate its true binding targets from those spurious ones. This is an important step toward understanding gene regulation.

Results: This paper describes a boosting approach to modeling TF–DNA binding. Different from the widely used weight matrix model, which predicts TF–DNA binding based on a linear combination of position-specific contributions, our approach builds a TF binding classifier by combining a set of weight matrix based classifiers, thus yielding a non-linear binding decision rule. The proposed approach was applied to the ChIP-chip data of *Saccharomyces cerevisiae*. When compared with the weight matrix method, our new approach showed significant improvements on the specificity in a majority of cases.

Contact: wwong@hsph.harvard.edu

Supplementary information: The software and the Supplementary data are available at http://biogibbs.stanford.edu/~hong2004/MotifBooster/.

1 INTRODUCTION

With the continuing explosive growth of sequenced genomes and genome-wide mRNA expression data, scientists are increasingly interested in modeling regulatory motifs and predicting binding targets of transcription factors (TFs). In this paper, we propose a discriminant approach that builds models to distinguish positive sequences (i.e. binding targets of a TF) from negative sequences (i.e. non-targets of a TF). Several approaches for this discriminant task have been proposed previously. DMotifs applies an enumerative search of the motif space and reports the best motif as a feature of the sequences that best differentiates positive from negative sequences (Sinha, 2002). Vilo *et al*. (2000) used a binomial formula for significance test to evaluate the occurrences of a motif in positive sequences against those in negative sequences. Similar to the approach of Vilo *et al*. (2000), the 'random selection null hypothesis' approach in Barash *et al*. (2001) tests the significance of a motif against negative sequences based on a hypergeometric distribution. Takusagawa and Gifford (2004) extended the works of Vilo *et al*. (2000) and Barash *et al*. (2001) to consider the effects of the lengths of sequences. The above approaches report motifs as consensus words, which are arguably less sensitive and precise than the corresponding weight matrix representations (Stormo *et al*., 1982).

Since the pioneering work of Stormo *et al*. (1982), the weight matrix model has become one of the most widely used models for representing motifs. A popular approach to estimating the parameters of a weight matrix *de novo* is to find a statistically enriched motif in positive sequences with respect to a background model (Stormo and Hartzell, 1989; Lawrence and Reilly, 1990; Lawrence *et al*., 1993; Liu *et al*., 1995; Barash *et al*., 2001). The background model, which usually is defined as an n-th order Markov model ($n = 0, 1, 2$ or 3), tries to capture all information in the non-binding sites that are much more heterogeneous than the binding sites. Such a background model is so general that the weight matrix model tends to have very low specificity. To better identify the non-binding sites that are very similar to the binding sites, Workman and Stormo (2000) proposed a discriminant method called ANN-Spec, which uses a Perceptron model and Gibbs sampling to train the weight matrix. They showed that the weight matrix models output by ANN-Spec have higher specificity than those built by non-discriminant approaches, such as MEME (Bailey and Elkan, 1994).

A motif reported as a weight matrix assumes that different positions of the motif are independent. Under this assumption, a weight matrix is essentially a linear classifier when used with a cutoff value to predict binding sites in sequences. Recent biological studies have demonstrated that individual positions of binding sites are not always independent (Bulyk *et al*., 2001, 2002; Man and Stormo, 2001), and suggested that some TFs recognize their targets in a non-linear fashion. Barash *et al*. (2003), adopted Bayesian networks to model dependencies in binding motifs as trees and mixtures of trees. The Bayesian tree model is similar to the one used in an early work by Agarwal and Bafna (1998) to model the dependency between bases. It is recently reported (Zhou and Liu, 2004) that a simpler pair-correlation model can largely account for all observed correlations among motif positions and using such a model in conjunction with the Gibbs sampling method suffers no overfitting problem. However, such a model still cannot accommodate some non-linear factors in discriminating positive and negative sequences.

It is widely accepted that a TF participates in controlling the mRNA levels of its target genes through its binding sites in the corresponding promoter regions. Hence, the REDUCE method (Bussemaker *et al*., 2001) and Motif Regressor (Conlon *et al*., 2003) were proposed to discover motifs by associating motif abundances with real-valued changes in genome-wide expression data. The REDUCE method enumerates all K-mers (DNA segments of length K) and checks whether the combinatorial effects of a set of K-mers can be used to explain changes of gene-expression data in a regression manner.

*To whom correspondence should be addressed.

Motif Regressor first uses MDSCAN (Liu *et al*., 2002) to generate a large set of matrix-based motif candidates that are enriched in the promoter regions of genes with the highest fold changes in gene expression data. Then it uses regression analyses to select motif candidates that are most relevant to the change of gene expressions. Nevertheless, neither approach exploits the potential of using negative sequences to change the parameters of a motif so as to increase the specificity of the model.

We propose a novel discriminant approach to enhance TF–DNA binding models using the boosting technique. First, we use the ChIP-chip data to select positive and negative sequences. In ChIP-chip experiments, DNA is crosslinked *in vivo* to proteins at sites of DNA–protein interaction and sheared to 500 bp–2 kb fragments. The DNA–protein complexes are precipitated by antibodies specific to the TF of interest. The precipitated protein-bound DNA fragments are PCR amplified, fluorescently labeled and hybridized to microarrays containing every promoter (sometimes also every ORF) in the genome. DNA fragments that are consistently enriched by ChIP-chip over repeated experiments are identified as positive sequences containing the protein–DNA interacting loci at ∼1 kb resolution. When compared with the gene-expression data, the ChIP-chip data provide much more accurate information about the genome-wide location of *in vivo* TF–DNA interactions, which enables us to assign definitive class labels to some promoter sequences with high confidence. Consequently, we can model the TF–DNA binding problem as a classification problem. We modify the confidence-rated boosting (CRB) algorithm (Schapire and Singer, 1999) to train a TF–DNA binding classifier as an ensemble model, which is a weighted combination of a set of base classifiers. The modified CRB algorithm automatically decides the number of base classifiers to be used so as to avoid overfitting. A key aspect of the boosting technique is that it forces some of the base classifiers to focus on the boundary between positive and negative samples, thus effectively reducing classification errors. We demonstrate the power of this approach by its performance on the ChIP-chip data of *Saccharomyces cerevisiae* (Lee *et al*., 2002).

2 METHODS

2.1 The ensemble model

We define a TF–DNA binding model as a weighted combination of a set of base classifiers $\{q_m(\bullet)\}$:

$$Q(S_i) = \sum_m \alpha_m q_m(S_i), \tag{1}$$

where α_m is the weight of $q_m(\bullet)$. The model weights can be normalized so that they sum up to 1. The class label of a DNA sequence S_i is decided by $\text{sign}(Q(S_i))$, with $+1$ denoting that S_i is a positive sequence. The base classifier has its root in the weight matrix method (Stormo *et al*., 1982). Let $f_m(\bullet)$ be the weight matrix model on which $q_m(\bullet)$ is based. And let the set $\{s_{ij}\}$ represent all K-mers in a DNA sequence S_i. The score of a K-mer s_{ij}, given $f_m(\bullet)$ is:

$$f_m(s_{ij}) = \sum_{k=1}^{K} \sum_{b\in\{\text{A,C,G,T}\}} w^m_{k,b}\, I_{k,b}\,(s_{ij}) - t, \tag{2}$$

where (1) $w^m_{k,b}$ is the parameter (in the logarithm scale) of the model $f_m(\bullet)$ for the nucleotide b at position k; (2) $I_{k,b}(s_{ij}) = 1$ if the k-th base of s_{ij} is b and $I_{k,b}\,(s_{ij}) = 0$, otherwise; (3) t is a threshold decided by some criteria (e.g. P-value). The higher the score, the more likely a site will be bound by the TF. The weight matrix model decides s_{ij} as a target of the TF if $f_m(s_{ij}) > 0$ and a non-target site, otherwise. We will show later that the threshold can be embedded into the parameter matrix $[w^m_{k,b}]_{k,b}$.

In many situations (e.g. ChIP-chip experiments), we only have information about whether a DNA sequence is bound by a TF, but do not know which sites in the sequence the TF binds to. Hence, given a weight matrix, we need to derive a scoring function to assess the likelihood of a DNA sequence as a target of a TF. This score should be affected by: (1) the number of matching sites in the sequence; and (2) the degree of the match for each matching site. The following function takes into account of the above factors and scores a sequence as:

$$h_m(S_i) = \log\left(\sum_{(r)} \mathrm{e}^{f_m(s_{ir})}\right), \tag{3}$$

where the sum is over the r best matching K-mers. This equation is similar to that proposed by Motif Regressor (Conlon *et al*., 2003). However, we limit it to the best r sites to avoid favoring very long sequences. Details for deciding the value of r are explained in Section 3.2.

The base classifier $q_m(\bullet)$ transforms the score of a sequence with a hyperbolic tangent function to a soft class prediction:

$$q_m(S_i) = \frac{1-\mathrm{e}^{-h_m(S_i)}}{1+\mathrm{e}^{-h_m(S_i)}} = \frac{\sum_{(r)} \mathrm{e}^{f_m(s_{ir})} - 1}{\sum_{(r)} \mathrm{e}^{f_m(s_{ir})} + 1}. \tag{4}$$

The hyperbolic tangent function is a scaled and biased *logistic* function, which has been used for motif site predictions (Barash *et al*., 2001; Segal *et al*., 2002).

2.2 Learn the ensemble model via boosting

We adopt the CRB algorithm (Schapire and Singer, 1999) to perform the following tasks in building an ensemble model $Q(\bullet)$: (1) deciding the number of linear classifiers $q_m(\bullet)$ in $Q(\bullet)$ and (2) learning the parameters of each $q_m(\bullet)$ and its weight α_m. Loosely speaking, in the first round, the CRB algorithm assigns equal weights to all samples and trains the first base classifier. In each of the rounds that follow, the boosting procedure gives higher weights to previously misclassified samples and learns a new base classifier with its weight using the reweighted samples. The final classifier is a linear assembly of weighted base classifiers from each round.

We made some modifications to the CRB algorithm to serve our purpose better. The modified CRB algorithm is outlined as Figure 1. Our first change tries to accommodate the unbalanced training set (the number of negative samples is much larger than that of positive ones) by assigning larger initial weights to the positive samples. Second, to prevent overfitting, we reserve some training sequences for internal test during training. The details of our implementations are explained in the next section.

3 IMPLEMENTATION

3.1 Initialize the weights of sequences

In our study, the number of negative sequences (usually in thousands) is often much larger than the positive ones (usually $<$100). Without proper adjustments, negative sequences would overwhelm a classifier and reduce its capability of recognizing positive sequences. As a remedy, we constrain the total weight of the positive sequences to be equal to that of the negative sequences (step b in Fig. 1). The sequences within each class have equal weights. This in effect imposes a higher penalty for misclassifying a positive sequence than misclassifying a negative one. Note that this heuristics is not equivalent to increasing the number of positive observations.

3.2 Learn base classifiers

The CRB algorithm (Schapire and Singer, 1999) is a Newton-like algorithm that constructs an ensemble model to minimize the upper bound on misclassification error

$$\text{Err} = \sum_i d_i^{(1)} \exp(-y_i\, Q(S_i)), \tag{5}$$

(a) Randomly reserve part of the training data for internal test. The remaining n training sequences and their class labels are denoted as $(S_1, y_1), \ldots, (S_n, y_n)$; $y_i \in \{-1, 1\}$.

(b) Initialize the weights of sequences $d_i^{(1)} (i = 1, \ldots, n)$.

(c) For $m = 1, \ldots, M$

(c.1) Train the parameters of $q_m(\bullet)$ and its weight α_m using the weighted sequences with the weights $\{d_i^{(m)}\}$.

(c.2) Update sequence weights: $d_i^{(m+1)} = \frac{d_i^{(m)} \exp(-\alpha_m y_i q_m(S_i))}{\sum_j d_j^{(m)} \exp(-\alpha_m y_j q_m(S_j))}$

(c.3) Use the reserved data to check if the overall model overfits the training data. Roll back ($m = m - 1$) and stop if it overfits.

(d) Output the final model $Q(\bullet) = \sum_m \alpha_m q_m(\bullet)$.

Fig. 1. The modified boosting algorithm.

where $d_i^{(1)}$ is the initial weight of S_i and y_i is the class label of S_i. Friedman *et al*. (2000) have detailed a discussions on the rationale of choosing the above criterion. In the m-th round, the CRB algorithm trains $q_m(\bullet)$ and its weight α_m to minimize the weighted error:

$$\varepsilon_m = \sum_i d_i^{(m)} \exp(-\alpha_m y_i q_m(S_i)), \tag{6}$$

where $d_i^{(m)}$ is the weight of S_i in the m-th round. In our case, the parameters to be estimated in each round include α_m, r and $\left[w_{k,b}^m\right]_{k,b}$. Basically, at step c.1 in Figure 1, we increase r from 1 to R (currently $R = 5$) by the step size 1. For each value of r, the parameters α_m and $\left[w_{k,b}^m\right]_{k,b}$ are initialized and refined to minimize the weighted error. Finally, the m-th round reports the values of r, α_m and $\left[w_{k,b}^m\right]_{k,b}$, which correspond to the minimum weighted error.

3.2.1 Initialization Since the motif must be an enriched pattern in the positive sequences, we take advantage of Motif Regressor (Conlon *et al*., 2003) to generate a good seed weight matrix for initializing $\left[w_{k,b}^m\right]_{k,b}$. The seed weight matrix, reported by Motif Regressor, has the best correlation between the logarithm of ChIP-chip P-value and motif-matching score of all training sequences. Let $\left[w_{k,b}^0\right]_{k,b}$ be the seed weight matrix. Given a value of r, we initialize α_m and $w_{k,b}^m$ as $\alpha_m(0) = 1$ and $w_{k,b}^m(0) = w_{k,b}^0 + (\sigma_{k,b} - t/K)$, respectively, where $\sigma_{k,b}$ is randomly generated in the range $[-0.2, 0.2]$ and t is the threshold as in Equation (2). The value of t is determined as the following. We first use the matrix $\left[w_{k,b}^0 + \sigma_{k,b}\right]_{k,b}$ to score all sites in the training sequences and obtain the minimum and maximum site scores as $t_{\min}$ and $t_{\max}$. Then, we increase t from $t_{\min}$ to $t_{\max}$ by the step size 0.1 and select the value that corresponds to the minimum weighted error under the current values of r and α_m.

3.2.2 Refinement The parameters $\left[w_{k,b}^m\right]_{k,b}$ and α_m are iteratively refined by a gradient-like method. In the n-th iteration ($n \geq 1$), use $\left[w_{k,b}^m(n-1)\right]_{k,b}$ to find the best r sites in each sequence as its representative sites, and update $\left[w_{k,b}^m(n)\right]_{k,b}$ and $\alpha_m(n)$ based on the corresponding gradients of the weighted error, i.e.:

$$\begin{aligned} w_{k,b}^m(n) &= w_{k,b}^m(n-1) - \frac{\eta_1}{(1+n/10)} \times \frac{\partial \varepsilon_m(n-1)}{\partial w_{k,b}^m(n-1)} \\ \alpha_m(n) &= \alpha_m(n-1) - \frac{\eta_2}{(1+n/10)} \times \frac{\partial \varepsilon_m(n-1)}{\partial \alpha_m(n-1)}, \end{aligned} \tag{7}$$

where the update rates are set as $\eta_1 = 0.05$ and $\eta_2 = 0.1$ based on our experience. The iteration stops if (1) the weighted error increases, (2) the improvement of error is <0.0001 or (3) the maximum number of iterations (currently 100) is reached. Note that a site s_{ij} is now scored as $\sum_{k=1}^{K} \sum_{b \in \{A,C,G,T\}} w_{k,b}^m(n) I_{k,b}(s_{ij})$, which is slightly different from Equation (2). The threshold t in Equation (2) is absorbed by $\left[w_{k,b}^m(n)\right]_{k,b}$ and is updated implicitly.

3.3 Prevent overfitting

A main challenge with the small number of positive samples is that one can easily overtrain the classifiers. Our strategy to alleviate this effect is to reserve a subset of the negative training sequences (5% in our current setting) and one positive training sequence for internal validation during training. The sequences are randomly selected. The weight of each reserved sequence is set as the initial weight of a training sequence with the same class label. Overfitting is checked using the reserved data at step c.3 in Figure 1. The boosting procedure will stop, if adding one more base classifier increases the error [as defined in Equation (5)] for the reserved sequence set.

Sometimes, the ensemble model may have only one base classifier, say $q_1(\bullet)$. We build a base classifier $q_\upsilon(\bullet)$ with its parameters as r_υ and $\left[w_{k,b}^0 - t_\upsilon/K\right]_{k,b}$, where r_υ and t_υ are decided by the initialization method (without $\sigma_{k,b}$) described in Section 3.2. The weight of $q_\upsilon(\bullet)$ is set as 1. We compare $q_\upsilon(\bullet)$ with $q_1(\bullet)$ and choose the one with a smaller weighted error as defined in Equation (5). The rationale for this step is that the current way for training base classifiers may not find the best one. This limitation can be amended by a weighted combination of multiple base classifiers. If the final model has only one base classifier, $q_\upsilon(\bullet)$ could be a better alternative.

4 RESULTS

4.1 Data

We used the ChIP-chip data reported in Lee *et al*. (2002). Positive sequences are selected using ChIP-chip P-value 0.001 as the cutoff. At this cutoff selection, the false positive rate is 6–10% and the false negative rate is ∼33% (Lee *et al*., 2002). Although the data are still noisy, they are the best genome-wide data of *in vivo* TF–DNA binding localization so far. To avoid having too few positive samples, we also required that each selected TF should have at least 25 positive sequences. Forty TFs (Lee *et al*., 2002) satisfy these criteria. Negative sequences were selected as those with ChIP-chip

Table 1. Data summary and cross-validation results for 31 ChIP-chip data

TF	Pos seq (no.)	Neg seq (no.)	Base classifiers (no.)	Average FP of weight matrix	Average FP of boosting	Improvement of boosting over weight matrix(%)
ABF1	176	3257	2	7.59	7.41	2.43
ACE2	46	2872	2	37.26	30.46	18.23
BAS1	31	3211	2	38.05	31.81	16.41
CAD1	27	3358	2	23.89	21.99	7.97
CBF1	28	2607	2	10.48	7.92	24.44
CIN5	116	3345	1	30.10	25.84	14.14
DAL81	32	3457	3	74.25	63.10	15.02
FHL1	124	3461	4	25.87	24.37	5.79
FKH1	40	3016	1	28.77	28.77	0.00
FKH2	72	3190	1	42.66	42.66	0.00
GCN4	56	2839	3	22.12	18.97	14.24
HAP4	42	3241	2	44.86	38.18	14.90
HSF1	34	2571	2	35.93	26.89	25.18
MBP1	74	3153	1	33.33	28.94	13.17
MCM1	59	3192	2	21.09	17.14	18.74
NRG1	59	3099	1	49.04	44.55	9.17
PDR1	45	3159	3	53.19	37.10	30.25
PHD1	70	3198	3	35.32	23.22	34.24
RAP1	127	3104	3	19.44	11.86	38.96
REB1	89	3021	2	10.30	9.92	3.70
RLM1	33	3164	1	39.16	39.16	0.00
SKN7	72	3089	1	45.10	37.87	16.02
SMP1	48	3276	2	51.15	43.78	14.41
STE12	54	3118	1	85.74	69.07	19.44
SUM1	41	3233	2	19.01	16.26	14.48
SWI4	90	3180	2	31.03	25.13	19.01
SWI5	72	3258	3	63.39	45.86	27.66
SWI6	65	3418	3	28.70	22.65	21.06
YAP1	35	2816	2	22.71	16.74	26.30
YAP5	55	3399	2	55.37	41.37	25.28
YAP6	65	3364	1	92.93	92.93	0.00

Columns 1, TF names; 2, number of positive sequences; 3, number of negative sequences; 4, number of base classifiers in the boosted classifier; 5, number of false positives FP_w using the weight matrix reported by Motif Regressor as a classifier; 6, number of false positives FP_b of the boosting method; 7, percentage of improvement of the boosting method over the weight matrix method, measured as $(FP_w - FP_b)/FP_w$.

ratio $\leq$1 and ChIP-chip P-value $\geq$0.05. Each selected TF has $\sim$3000 negative sequences. For each gene, we take its upstream sequence, up to 800 bp, not overlapping with the previous gene.

4.2 Boosting improves the specificity of motif models

To evaluate our method, we used the following cross-validation procedure. In each run, we leave one positive sequence and 5% of randomly selected negative sequences as the test data and train a classifier on the remaining data. This procedure is repeated 10 times for each positive sequence. The cross-validation error of each run is calculated as the number of false positives if the number of the false negatives is zero. The results are then averaged for all runs and compared. The detailed data, which include the sequence data, the ensemble models of the TFs, the logos of the ensemble models and all the test results, are available as the Supplementary data at http://biogibbs.stanford.edu/$\sim$hong2004/MotifBooster/.

We used Motif Regressor (Conlon *et al*., 2003) to find the seed weight matrix. For each TF, Motif Regressor called MDSCAN (Liu *et al*., 2002) to find candidate motifs of width 6–17 bases. At each width, MDSCAN reported the best 20 weight matrices enriched in the positive training sequences. Each weight matrix was used to score the training sequences. Motif Regressor then performed simple linear regression between the logarithm of ChIP-chip P-values and sequence scores. We chose the motif corresponding to the best regression P-value as our seed motif. We observed that Motif Regressor did not find significant enough motifs for nine TFs (DIG1, GAL4, GAT3, GCR2, IME4, IXR1, NND1, PHO4 and ROX1). It is possible that under the asynchronized growth condition, these TFs were not activated, or the modified tagged TFs have changed their binding characteristics. Table 1 summarizes the results for the remaining 31 TFs. Compared with the weight matrix reported by Motif Regressor, the ensemble models performed markedly better in 27 cases and evenly in 4 cases (FKH1, FKH2, RLM1 and YAP6). A closer examination on the four even cases reveals that each ensemble model only has one base classifier that is a direct conversion from the initial weight matrix.

The boosting approach also reported final models with single base classifier in 5 of 27 cases that performed better. These five TFs are CIN5, MBP1, NRG1, SKN7 and STE12. Since the base classifier is equivalent to a weight matrix model, these results indicate that

Table 2. Contributions of the base classifiers (BCs) in the "leave-one-out" cross validation tests

TF	BC no.	Average FP of WM	Average FP of BC 1	Average FP of BC (1 + 2)	Average FP of BC (1 + 2 + 3)	Average FP of BC (1 + 2 + 3 + 4)
ABF1	2	7.59	7.92	7.41	—	—
ACE2	2	37.26	62.74	30.46	—	—
BAS1	2	38.05	45.36	31.81	—	—
CAD1	2	23.89	24.97	21.99	—	—
CBF1	2	10.48	10.33	7.92	—	—
DAL81	3	74.25	75.81	65.91	63.10	—
FHL1	4	25.87	33.24	26.99	24.54	24.37
GCN4	3	22.12	25.92	20.15	18.97	—
HAP4	2	44.86	44.18	38.18	—	—
HSF1	2	35.93	27.29	20.71	—	—
MCM1	2	21.09	17.98	17.14	—	—
PDR1	3	53.19	49.38	40.31	37.10	—
PHD1	3	35.32	28.68	26.73	23.22	—
RAP1	3	19.44	19.31	13.59	11.86	—
REB1	2	10.30	11.30	9.92		
SMP1	2	51.15	49.79	43.78	—	—
SUM1	2	19.01	20.04	16.26	—	—
SWI4	2	31.03	37.10	25.13	—	—
SWI5	3	63.39	60.13	49.92	45.86	—
SWI6	3	28.70	38.22	26.43	22.65	—
YAP1	2	22.71	26.08	16.74	—	—
YAP5	2	55.37	45.88	41.37	—	—

Columns 1–7 are the TF names, number of BCs in the ensemble model, number of false positives of the weight matrix method and number of false positives of the ensemble model when its first 1, 2, 3 and 4 BCs are used, respectively. We order the base classifiers in each ensemble model so that their weights are in the descending order.

using negative information can help discover better weight matrices in many cases. This is consistent with the findings of Workman and Stormo (2000). However, the first base classifier does not always perform better than the initial weight matrix. Table 2 summarizes the contributions of the base classifiers for the cases where the boosting method selected more than one base classifier. The base classifiers in the final models are arranged in the descending order of their weights. The performances of 13 first base classifiers, i.e. the ones with the largest weights, are worse than those of the weight matrices reported by Motif Regressor. This may suggest that when the binding sites of a TF are 'heterogeneous' and maybe grouped into clusters, our boosting method finds base classifiers corresponding to different cluster profiles, whereas Motif Regressor reports an 'average' profile. Thus, a single base classifier may be too specific to a particular cluster and does not discriminate well globally.

5 DISCUSSION

For some cases, the ensemble model can reveal dependencies among motif positions. For example, Figure 2a displays the weight matrix found by Motif Regressor for RAP1, from which we can see that C and T dominate in position 5, and A and G dominate in position 8. But there is no further information on how these two positions might correlate with each other. In contrast, our boosting approach selected three base classifiers (Fig. 2b–d) to compose the final model. Two base classifiers favored C and A in positions 5 and 8, respectively, whereas the third one preferred T and G in those positions, respectively. This observation implies that positions 5 and 8 may cooperate in a certain way such that the change in one position correlates with the change in the other. As another example, we observe that positions 1, 10 and 13 of REB1 motif (Fig. 3) can be decomposed in a similar way. In its first base classifier, position 13 strongly prefers G; positions 1 and 10 are ambivalent about G and C, respectively. In the second base classifier, however, position 13 strongly disfavors G, and positions 1 and 10 strongly favor G and C, respectively. This suggests that the three positions may cooperate to facilitate the protein–DNA binding.

The boosting approach terminates with an ensemble of 2–3 base classifiers for most cases. This is atypical for applications using the boosting technique that usually can boost for hundreds to thousands of base classifiers. The small number of base classifiers could be due to three reasons. The first reason might be the unbalanced training data (∼100 positive versus ∼3000 negative sequences). We examined the sensitivity and specificity of each base classifier alone using the training samples (Fig. 4a). The sensitivity of base classifiers spreads out in the range of 40–90%, while their specificity concentrates in the range of 75–95%. This suggests that it is easier to train base classifiers to recognize negative samples in our case although the negative samples are more heterogeneous than the positive ones. We modify the boosting algorithm by adding more initial weights to the positive samples such that the initial total weights of two classes are equal. We note that although this method helps to bring out a less 'biased' classifier, it is not equivalent to increasing the number of positive observations. As shown in Figure 4b, base classifiers with higher sensitivity tend to have lower generalization errors. A similar trend can be observed for the specificity of base classifiers in Figure 4c. Figure 5a shows that it is more

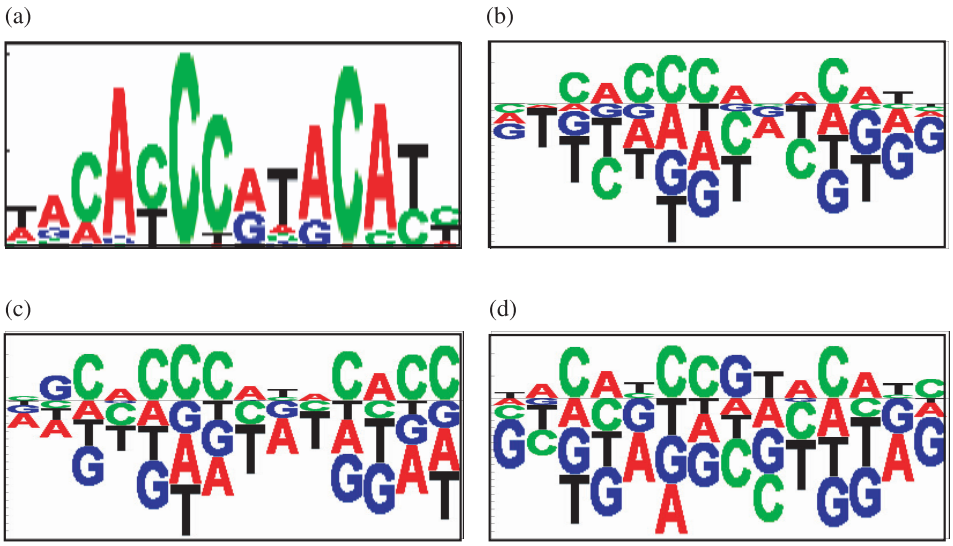

Fig. 2. Logos of the binding models of RAP1. (**a**) Position specific probability matrix. Logo of the weight matrix reported by Motif Regressor, drawn using the method of (Schneider and Stephens, 1990). (**b**), (**c**) and (**d**): Logos of the base classifiers 1, 2 and 3, respectively in the ensemble model reported by the boosting approach (weight of base classifier 1 = 0.31; weight of base classifier 2 = 0.30; weight of base classifier 3 = 0.39). Base classifiers have negative parameters and cannot be visualized in the same way. (b), (c) and (d) are drawn in the following way. The height of a letter corresponds to the absolute magnitude of its weight scaled by a factor k (For visualization purpose, $k = 3$ for positive weights and $k = 1$ for negative weights.) Letters are ordered by their weights. The black horizontal line represents zero. Letters above the zero line have positive weights, and those below the zero line have negative weights.

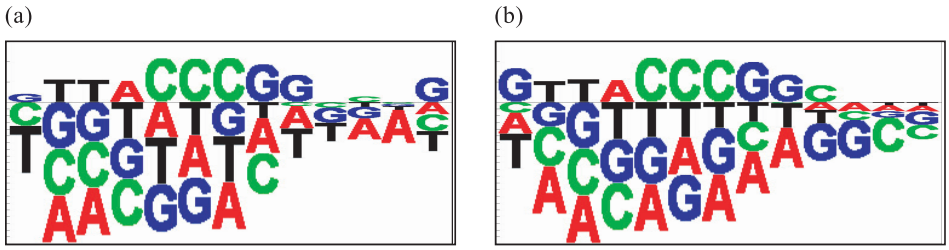

Fig. 3. Logos of the ensemble model of REB1. (**a**) The logo of base classifier 1 (Weight = 0.52). (**b**) The logo of base classifier 2 (Weight = 0.47).

likely to train base classifiers with relatively low training sensitivity and specificity when the size of positive sequences is small. Moreover, base classifiers trained with less positive samples are more likely to have higher generalization errors (Fig. 5b). Based on the above analyses, we reason that (1) base classifiers hardly overfit the training data in most cases and (2) the small size of positive samples does not provide enough information to boost for more base classifiers.

Second, the binding mechanisms of some TFs may indeed be almost linearly dependent of nucleotide types of the motif positions. For example, ABF1 has a much larger positive sample size (176) when compared with other TFs. Both the weight matrix and the ensemble model of ABF1 have low and comparable generalization errors (Table 1). The ensemble model has two base classifiers. The training sensitivity/specificity of the base classifiers are 93.18/94.66% and 90.34/95.58%. These results suggest that the binding mechanism of ABF1 may have little non-linearity because its samples can be well classified by linear decision rules including the weight matrix and the base classifiers. The base classifier becomes a 'strong' learner (i.e. it can explain most of the training data) in such a case. On the other hand, the mild performances of many other base classifiers suggest that the binding mechanisms of some other TFs could have relatively high non-linearity.

Finally, our approach initializes a base classifier using a seed matrix. The successive refining step may only explore a limited subspace around the seed matrix. The training of base classifiers can be improved by a sampling-based *de novo* motif finding algorithm that is capable of exploring a wider range of the solution space (e.g. by sampling at multiple temperature levels). Or we can replace the base learner with a simpler one, e.g. a simple decision tree that uses rules like whether a position should be C or not, etc. With the above modifications, the ensemble model could have more base classifier and capture more comprehensive features that lead to better classification performance. Nonetheless, the resultant base classifiers could be very diverse. Some base classifiers could represent highly degenerated motifs. One potential drawback of this alternative is the loss of biological interpretability of the ensemble model. Although it is still not perfectly understood why the number of base classifiers is small, our approach provides a good balance between the interpretability and the performances of the boosted models. Another choice for improving the boosted models is to train each base classifier only by a randomly selected subset of the full training set as suggested

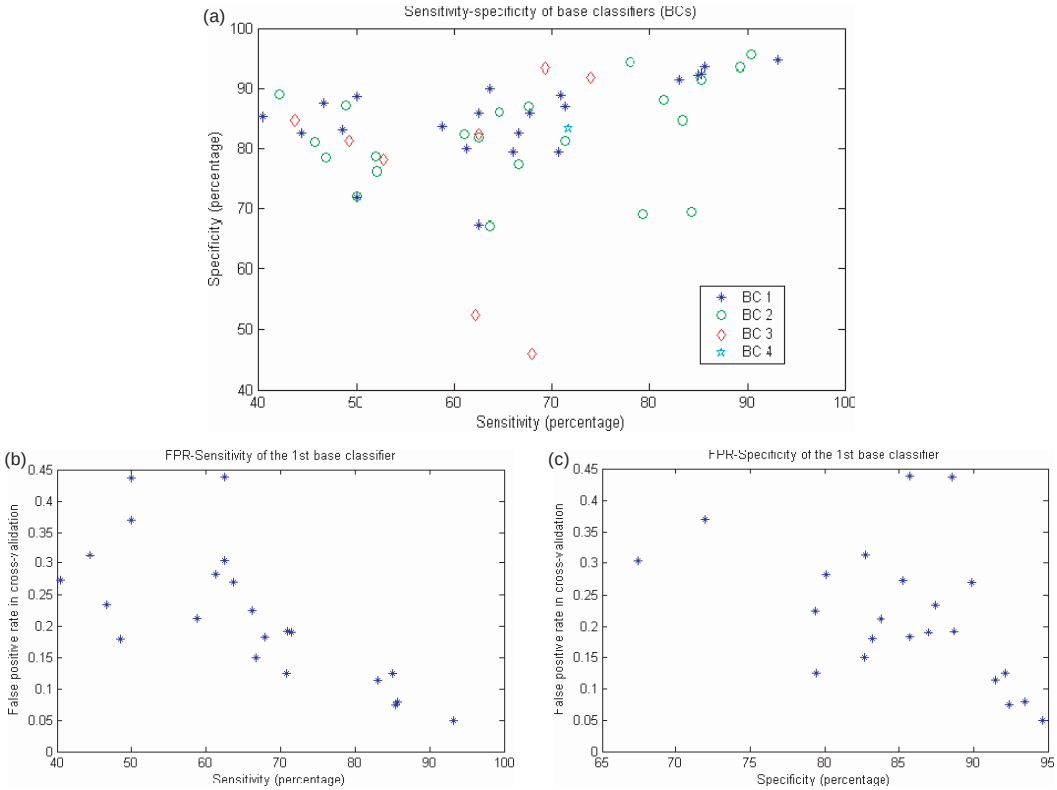

Fig. 4. (**a**) The training sensitivity (horizontal axis)–specificity (vertical axis) plot of the base classifiers. Star, circle, diamond and pentagram denote the sensitivity/specificity of the base classifiers, 1, 2, 3 and 4 respectively. (**b**) The cross-validation false positive rate (FPR)–training sensitivity (horizontal axis) plot of the base classifier 1. (**c**) The cross-validation FPR–training specificity (horizontal axis) plot of the base classifier 1.

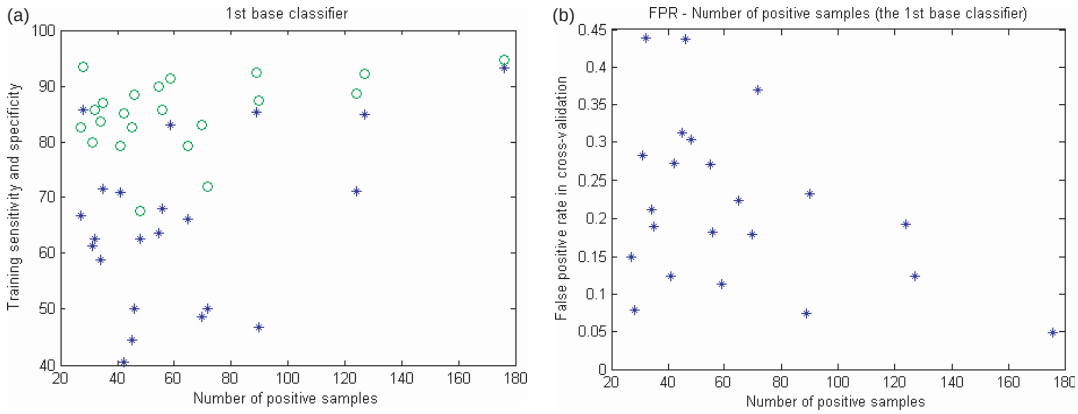

Fig. 5. The result plots of the first base classifiers. (**a**) Training sensitivity (star) and specificity (circle)–number of positive sequences (horizontal axis). (**b**) Cross-validation FPR–number of positive sequences (horizontal axis).

by Friedman (2002). It was reported that such kind of randomness has advantages in the situations of small samples and powerful weak learners.

6 CONCLUSION

We introduce a boosting-based method for modeling TF–DNA binding. By repeatedly fitting weight matrix based classifiers to weighted samples that focus on erroneous classifications, the boosting approach can build a more accurate TF–DNA binding model as a weighted combination of the base classifiers. The proposed approach was applied to the ChIP-chip data of *S.cerevisiae* and showed significant improvements on specificity in many cases. Like many recent studies that use mRNA microarray data to help refine regulatory binding motifs and infer combinatorial rules of transcription regulation (W. Wang *et al*., submitted for publication; Beer and

Tavazoie, 2004), we found that ChIP-chip data can be used to further refine motif models and reveal novel features of TF–DNA interactions. Currently, we use Motif Regressor to generate the seed motif for boosting. However, our algorithm is not limited to working with Motif Regressor and can be used to boost weight matrices reported by any motif finding algorithm.

ACKNOWLEDGEMENTS

The work of W.H.W. is supported by NIH-HG02341. The work of J.S.L. is supported by NIH-P20-CA96470 and NSF DMS-0244638. The work of P.H. is supported by NIH-GM67250. We thank the anonymous reviewers for constructive suggestions that helped us to unify the way to initialize and train base classifiers and inspired us to think hard on the overfitting issue of the ensemble models.

REFERENCES

Agarwal,P.K. and Bafna,V. (1998) Detecting non-adjoining correlations with signals in DNA. In *Proceedings of the Second Annual International Conference on Research in Computational Molecular Biology*, March 22–25, 1998, New York, USA. ACM Press, pp. 2–8.

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.

Barash,Y. *et al*. (2001) A simple hyper-geometric approach for discovering putative transcription factor binding sites. In *Algorithms in Bioinformatics: Proceedings of the 1st International Workshop*, LNCS 2149, pp. 278–293.

Barash,Y. *et al*. (2003) Modeling dependencies in protein–DNA binding sites. In *Prooceedings of the 7th Annual International Conference on Computational Molecular Biology (RECOMB 2003)*, Berlin, Germany, ACM Press, NY, pp. 28–37.

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

Bulyk,M.L. *et al*. (2001) Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. *Proc. Natl Acad. Sci. USA*, **98**, 7158–7163.

Bulyk,M.L. *et al*. (2002) Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res.*, **30**, 1255–1261.

Bussemaker,H.J. *et al*. (2001) Regulatory element detection using correlation with expression. *Nat. Genet.*, **27**, 167–171.

Conlon,E.M. *et al*. (2003) Integrating regulatory motif discovery and genomewide expression analysis. *Proc. Natl Acad. Sci. USA*, **100**, 3339–3344.

Friedman,J.H. (2002) Stochastic gradient boosting. *Comput. Stat. Data Anal.*, **38**, 367–378.

Friedman,J.H. *et al*. (2000) Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). *Ann. Statist.*, **28**, 337–407.

Lawrence,C.E. *et al*. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.

Lawrence,C.E. and Reilly,A.A. (1990) An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins*, **7**, 41–51.

Lee,T.I. *et al*. (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, **298**, 799–804.

Liu,J.S. *et al*. (1995) Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *J. Am. Stat. Assoc.*, **90**, 1150–1170.

Liu,X.S. *et al*. (2002) An algorithm for finding protein–DNA binding sites with applications to chromatin immunoprecipitation microarray experiments. *Nat. Biotechnol.*, **20**, 835–839.

Man,T.K. and Stormo,G.D. (2001) Non-independence of Mnt repressor–operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. *Nucleic Acids Res.*, **29**, 2471–2478.

Schapire,R. and Singer,Y. (1999) Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, **37**, 297–336.

Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.

Segal,E. *et al*. (2002) From promoter sequence to expression: A probabilistic framework. In *Proceedings of the 6th International Conference on Research in Computational Molecular Biology (RECOMB'02)*, Washington, DC, ACM Press, pp.263–272.

Sinha,S. (2002) Discriminative motifs. In *Proceedings of the 6th International Conference on Research in Computational Molecular Biology (RECOMB'02)*, Washington, DC, ACM Press, pp.291–298.

Stormo,G.D. and Hartzell,G.W.III (1989) Identifying protein-binding sites from unaligned DNA fragments. *Proc. Natl Acad. Sci. USA*, **86**, 1183–1187.

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.

Takusagawa,K. and Gifford,D. (2004) Negative information for motif discovery. *Pac. Symp. Biocomput.*, 360–371.

Vilo,J. *et al*. (2000) Mining for putative regulatory elements in the yeast genome using gene expression data. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **8**, 384–394.

Workman,C.T. and G.D. Stormo (2000) ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. *Pac. Symp. Biocomput.*, 467–478.

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