Ensemble Machine Methods for DNA Binding

Yue Fan and Mark A. Kon*
Dept. of Mathematics and Statistics
Boston University
111 Cummington St, Boston, MA 02215, U.S.A

Charles DeLisi
Bioinformatics and Systems Biology
Boston University, Boston, MA 02215, U.S.A.

Abstract

We introduce three ensemble machine learning methods for analysis of biological DNA binding by transcription factors (TFs). The goal is to identify both TF target genes and their binding motifs. Subspace-valued weak learners (formed from an ensemble of different motif finding algorithms) combine candidate motifs as probability weight matrices (PWM), which are then translated into subspaces of a DNA k-mer (string) feature space. Assessing and then integrating highly informative subspaces by machine methods gives more reliable target classification and motif prediction. We compare these target identification methods with probability weight matrix (PWM) rescanning and use of support vector machines on the full k-mer space of the yeast S. cerevisiae*. This method, SVMotif-PWM, can significantly improve accuracy in computational identification of TF targets. The software is publicly available at http://cagt10.bu.edu/SVMotif*

1 Introduction

Transcription factor binding sites (TFBSs) within DNA regulatory regions are among the most important functional elements in the genome. An improved understanding of how transcription factors (TFs) interact with DNA gives a better overall picture of regulatory pathways within the cell. For a fixed transcription factor t, the study of its binding preference can be divided into three problems: (a) identification of target genes of t, (b) identification of binding motifs (characteristic DNA sequences of length approximately 10 to which t binds) and (c) identification of functional binding sites (DNA binding positions).

At this point hundreds of genomes have been sequenced, and recent large scale computational methods have provided many significant results toward solving these problems [8, 22, 18, 1, 2, 19, 16]. In this paper, we are only interested in how to apply ensemble machine methods to solve the first two problems.

Computational methods typically begin analysis with motif identification by searching for common patterns of DNA subsequences in a collection of suspected target regions (positive set). A binding motif for t is typically represented as a position weight matrix (PWM) whose j^{th} column consists of the four probabilities of the DNA bases A, C, G, and T in position j of the motif. The motif can then be used to detect new binding target genes and corresponding binding sites by rescanning its PWM through the promoter regions of candidate genes [5]. A promoter position with a rescanning score higher than a given threshold is reported as a new binding site and suggests a new target.

More recently machine learning has been used in these types of analysis [9, 15]. A common approach is based on first solving the (classification) problem in (a), of identifying targets/nontargets of t. Feature maps into a high dimensional feature space $\mathcal{F}$ based on counts of k-mers (promoter substrings of length k) have been used to represent gene promoters [9, 14] as well as amino acid sequences [15] using the spectrum map. Such a feature map $\phi(g) = \mathbf{x}$ takes a promoter region g and maps it into $\mathbf{x} \in \mathcal{F}$, whose i^{th} component counts the number of appearances in g of the i^{th} k-mer of an indexed list. A classifier can be trained on this feature space and used to classify new gene promoter regions as targets or non-targets [9]. This direct framework for target identification does not need prior estimates of the motif and achieves state-of-the-art performance. The motif can also be constructed from using the most important classification features (promoter subsequences) [14].

Analysis of 13 motif discovery algorithms [23] suggests that no single motif identification system can perform consistently well over all transcription factors. To take advantage of strengths of different algorithms, some ensemble methods have been developed recently which combine results of different algorithms. They usually perform a strategy known as motif ranking, which starts with a collection of potential binding motifs (typically as PWMs) predicted by one or more component discovery algorithms, and aims

This work was partially supported by the NIH under grant number GM080625-01

to select the most reliable ones. Clustering methods can also be applied to merge similar candidates before assessments. For comparing predictions from a single algorithm, a simple solution aggregates motif scores from multiple runs as a final motif score [10]. However, a lack of algorithm-independent motif scores limits this method in comparisons of motifs from different algorithms. WebMOTIFS (the web interface of TAMO) [21, 6] computes p-values of hypergeometric enrichment scores [7] for different motifs as a cross-algorithmic measure. A more sophisticated measure is a Bayesian scoring function [11, 3]. This measures how well a proposed PWM fits known targets in the training data by computing the posterior probability of observing them. The reliability measure usually consists of two main components, enrichment and degeneracy. Enrichment measures whether the motif is abundant in positives, and is often based on the number of rescanning hits. Degeneracy is usually measured by the entropy distance between a motif and the genome background. Machine methods introduce discriminatory power (between targets and non-targets in the training set) as another piece of information.

Subspace-valued weak learners: We will introduce an ensemble machine learning method of independent interest, which will be used in the conversion of candidate PWMs to subspaces of the k-mer feature space $\mathcal{F}$. In this method, the component weak learners output subspaces of the feature space $\mathcal{F}$ (as candidate dimensional reductions), rather than scalar answers (as in, e.g., bagging and boosting). It is a new way to integrate and augment information from individual PWMs.

For the problem of identifying regulatory targets, subspace methods show a large improvement against conventional PWM rescanning methods. Each input algorithm produces PWMs, each of which produces a different subspace of $\mathcal{F}$. We hence term these algorithms *subspace-valued weak learners*. In addition, a biologically validated PWM can also be converted to a subspace as well in this framework. As compared to machine methods using the entire k-mer space, these ensemble methods achieve similar or better results by using spaces of greatly reduced dimension.

For the problem of finding binding motifs, ensemble subspace methods are generally preferable over full feature space methods since they provide more reliable and complete predictions. They allow introduction of orthogonal information types obtained by weak learners, which can be chosen as non-machine learning methods. The success of this alternative means of target prediction for a TF based on known motifs provides us with a new motif ranking measure based on discriminatory power. Combining this with enrichment and degeneracy measures, ensemble methods perform very well predicting targets of yeast TFs. Though we have mainly focused on target and motif identification, our future work includes developing new methods for identification of functional binding sites with these approaches.

Independence of bases in PWM: We believe that a significant part of the improvement in the use of our machine TF target classification methods over PWM methods occurs because gene promoter scanning by machine methods effectively removes sometimes inappropriate base independence assumptions which are implicit in PWM rescanning methods. For example, suppose an SVM is used to scan a genome to find gene targets for a TF t, and the importances of the most significant features (k-mers) are measured by their weight $\mathbf{w}_i$ in the SVM function $f(x) = \mathbf{w} \cdot \mathbf{x} + b$ ($\mathbf{w}, \mathbf{x} \in \mathcal{F}$). This collection of most significant k-mers gives an empirical approximation of the true joint distribution of k bases in the motif, in particular including the joint variation of the bases.

2 Machine Learning Tools

2.1 K-mer Feature Space

Denoting the set of DNA nucleotides as $\Sigma = \{A, C, G, T\}$, any given gene promoter region is a finite sequence g of letters in Σ. Let Σ^k be the set of all sequences of length k. A simple and useful method of mapping g into the k-mer feature space $\mathcal{F}$ is the feature map

$$\phi_{sp}(g) = \{n_{a_i}(g)\}_{i=1}^{4^k},$$

where a_i ranges over all of Σ^k and n_{a_i} is the number of exact matches for a_i in g. Thus the feature vector $\phi_{sp}(g)$ will be a 4^k dimensional vector containing frequency counts of all possible k-mers in Σ^k. The map ϕ_{sp} is hence called the *spectrum map* [15]. Since transcription factors do not distinguish sequences and their reverse complements, exact matches above also include reverse complement matches.

2.2 Position Weight Matrix (PWM) Subspaces

In DNA sequence analysis, a PWM is a $4 \times l$ matrix, $M = (\theta_{ij})$ whose j^{th} column θ_j defines the distribution of $\{A, C, G, T\}$ appearing at position j in the corresponding motif. Assuming independence among motif positions, it is simple to compute the probability that a given k-mer of length l is a binding site (with the PWM's motif) by multiplying corresponding probabilities in each column. For a k-mer shorter than l, letters 'N' representing any nucleotide can be added to the end, allowing the product to ignore the corresponding positions. If we assume additional columns beyond the original PWM size whose frequencies represent the genome background, the probability for longer k-mers can also be defined.

A PWM is generated empirically from a large number of likely binding sites. We can also perform the reverse process (PWM → k-mers) by randomly generating a set of k-mers with fixed length k following the distribution defined by the PWM. If an infinite number of k-mers were generated in this way, they would cover all k-mers which have positive probabilities. To reduce noise effects in the PWM, we only collect those k-mers with PWM probabilities greater than a given cut-off. We call this subset as *profile set* or *profile subspace* (if seen as a subspace of the k-mer feature space $\mathcal{F}$) of the corresponding PWM. Although these k-mers are defined analytically, the set can be reasonably approximated by generating a small random sample $\mathcal{S}$ (See Algorithm 1). Notice that these n_0 k-mers have a large number of repeats and (S) will have a much smaller size than n_0.

```
Input: $4 \times l$ matrix $M = (\theta_{ij})$, fixed length $k$,
       number of $k$-mers $n_0$
Output: $\mathcal{S}$, the profile set
foreach $i = 1 : n_0$ do
    $s_0 \sim$ Unif$(1, l - k + 1)$;
    for $j = 1 : k$ do
        Generate $\alpha_j \sim$ Multinomial $(\theta_{s_0+j})$;
    end
    $\mathcal{S} = \mathcal{S} \cup \alpha_1, \ldots, \alpha_k$;
end
```

Algorithm 1: Generating the PWM Profile Subspace

The dimension of the profile subspace strongly depends on the degeneracy of the PWM. A PWM with uniform distribution over all columns effectively has the entire space $\mathcal{F}$ as its profile space. Meanwhile, the profile space of a PWM of size l with a point mass distribution over each column has only $l - k + 1$ dimensions.

An interesting property of profile subspaces relates to the fact that the reliability of the the corresponding motif can be measured by its discriminatory power. That is, if the genes of interest are projected onto the profile subspace of a true binding motif, they should be more easily separable than when they are projected onto the profile subspace of a false binding motif (Fig. 1). This property provides a measure of reliability which is used to construct ensemble motif discovery methods later in Section 4.

2.3 Use of Subspace-Valued Weak Learners

Let $\mathcal{F}$ again denote the entire k-mer feature space. In the language of machine learning ensembles, a collection of motif discovery algorithms $\{A_i\}_{i=1}^{m}$ which predict groups of PWMs by optimizing different objective functions can be viewed as 'weak learners', whose individual outputs are feature subspaces $\mathcal{F}_i$, each arising as above from a single PWM. Such algorithms A_i now with subspace outputs will be denoted as *subspace-valued weak learners*.

The combination of subspace-valued outputs of an ensemble of algorithms (in this case different motif finding algorithms) to form a larger search subspace of the feature space $\mathcal{F}$ seems to not to have been used previously in machine learning. Based on the results of the pilot study mentioned below, this method can be useful in solving the transcription regulation problem.

Each subspace corresponds to a discovered pattern by which the positives (targets of TF t) might be differentiated from negatives. For identification of target genes of t (goal(a)), the final goal will be to build a meta-classifier over the span $\oplus\mathcal{F}_i$ of these subspaces, instead of the larger full k-mer space $\mathcal{F}$, in order to reduce the redundancy of too many parameters. Note that each profile subspace $\mathcal{F}_i$ is built through sampling out of a single input PWM/binding motif. For identification of binding motifs (goal(b)), the final goal will be to select or construct a new profile subspace which has the best cross-validational discriminatory power. A better predicted PWM/binding motif can be deduced from the identified profile subspace. We will discuss only the above two goals here.

3 Binding/Nonbinding Gene Classification

3.1 Spanning Space Method for Subspace-valued Ensembles

The most intuitive way to combine several informative subspaces $\mathcal{F}_i$ of the feature space $\mathcal{F}$ should arise from forming their span $\oplus\mathcal{F}_i$, the smallest subspace containing all of them. Then a classifier (e.g. an SVM using cross validation) can be run on examples binding/nonbinding genes, using only features in $\oplus\mathcal{F}_i$. This gives a way to filter information in a very high dimensional $\mathcal{F}$ through dimensional reduction. It is equivalent to using the union of all *profile sets* (see above) as features to train the classifier. To save computational cost, we have restricted all generated k-mers to have the same length.

3.2 Combined Kernel Method

The above simple direct sum method ignores the fact that each individual *profile set* is assessed separately based on cross-validation in order to rank the quality of the PWM which generates it. This is assumed to perform as a whole when detecting positives. Since some basis elements (profile k-mers) of different $\mathcal{F}_i$ may intersect, one has to be careful not to mix them in direct sum $\oplus\mathcal{F}_i$. As an alternative method, we compute the individual kernel matrix K_i for each profile subspace $\mathcal{F}_i$ of PWM M_i by restricting the inner products to $\mathcal{F}_i$ only. i.e.

$$K_i(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle_{\mathcal{F}_i} = \sum_{k \in \mathcal{F}_i} x_{ik} x_{jk}$$

where x_i and x_j are $\mathcal{F}$-feature vectors for gene i and j in the training sample.

Then a combined kernel matrix is defined as

$$K = \sum_{i=1}^{m} \beta_i K_i$$

Note using K is equivalent to using a kernel (on the full space $\mathcal{F}$) which is weighted on each k-mer s proportionally to β_i, the number of appearances of s in different $\mathcal{F}_i$. In order to standardize the scale when adding kernel matrices, each kernel is normalized by $K(x,y) = K(x,y)/\sqrt{K(x,x)K(y,y)}$.

Finally, the classifier is trained on K. In the simplest case in which $\beta_i = 1$ for all i, this suggests a direct combination by concatenating all profile sets without eliminating duplicates. Such a choice of feature space geometry should perform similarly to the first method; however, the weights β_i can be very helpful when a large number of incorrect or non-informative PWMs exist in the pilot predictions.

3.3 Synopsis Feature Space Method

The above two methods are ways of combining subspace-valued weak learners. A third is the *synopsis* method. A synopsis vector is an unadjusted continuous value produced by a learning algorithm A_i. For example, in the case of SVM, the synopsis vector is:

$$f_i(\mathbf{x}_k) = \langle \mathbf{w}, \mathbf{x}_k \rangle = \sum_{j=1}^{n} \alpha_j K(\mathbf{x}_k, \mathbf{x}_j) y_j.$$

The label of a novel feature vector $\mathbf{x}$ is predicted by choosing a threshold b_i and taking the sign of $f_i + b_i$. If the machine is trained on an informative subspace $\mathcal{F}_i$ of $\mathcal{F}$, the resulting synopsis feature should be a good predictor of the class of x. In the case of SVM, this can be done with the proper choice of a threshold b_i. If the synopsis feature of a new SVM classifier trained on another subspace $\mathcal{F}_j$ provides a useful separation of samples based on different information, combining these two synopsis features into a vector in a (2 dimensional) synopsis feature space $\mathcal{F}_s$ can improve resolution through the two types of information. Continuing this way to combine synopsis feature information from m different subspaces, we have trained a meta-classifier over the full synopsis feature space $\mathcal{F}$. This method can also adaptively assess the quality of information and assign small coefficients to bad synopsis features.

Generally, merging similar motifs in the candidate pool results in a small synopsis feature space, with dimension between 10 to 20. Such low dimension allows the use of classical statistical methods, such as logistic regression or linear discriminant analysis. An alternative is the use of non-linear SVM rather than a linear one to improve classification accuracy.

3.4 Comparison between Subspace Method and Rescanning

Rescanning a gene promoter sequence g is widely used to computationally determine whether g is a target of (bound by) the TF t, if the binding motif (PWM) is known. The $4 \times l$ PWM is first transformed into a log ratio matrix in order to compute a *log ratio score*:

$$LM_{ij} = log\left(\frac{M_{ij}}{B_i}\right), i = 1,2,3,4 \text{ and } j = 1,2,\ldots,l$$

where LM_{ij} and M_{ij} are the $(i,j)^{th}$ elements in each matrix, and B_i is the background frequency of $i \in \Sigma = \{A,C,G,T\}$. The log ratio score of l-mer $\alpha_1\alpha_2\cdots\alpha_l$, $\alpha_i \in \Sigma$, is

$$LR\left(\alpha_1\alpha_2\ldots\alpha_k\right) = \sum_{j=1}^{l} LM_{\alpha_j,j}$$

The matrix LM is scanned through all the possible l-mer positions in a probe sequence g. A hit represents a position where the LR is above a given threshold c. All the probes g having a hit are then classified as binding targets (positives). Efforts have been made [13] to find realistic background frequencies B_i, which would improve the accuracy of functional binding site identification. However, finding a good threshold c is always a problem. Since PWM hit scoring implicitly assumes independence among motif positions, the LR under this assumption may not be able to discriminate among promoters versus false positives.

The above k-mer feature subspace methods count exact matches of k-mers in promoter g, and classify g a target if it contains sufficiently many exact matches from the true binding motif. Thresholds for such counts can then be set, as is done effectively in SVM classification, where a weighted count $\mathbf{w} \cdot \mathbf{x}$ is thresholded for selection of a promoter g with feature vector $x = \phi(g)$ as a target. As compared to rescanning methods which report all matches to these as hits, a machine learner (e.g. SVM [25, 24, 4, 27]) can learn which k-mers are more informative than others (e.g. through their weight in $\mathbf{w}$). The ROC curve (Fig. 1) also illustrates the improvement.

4 Binding Motif Discovery

At this point we have a set of motifs predicted by each weak learner A_i. The true binding motif may be one of them, may be partially covered by two or more motifs, or may not be detected at all. The goal of the ensemble method is to identify the most reliable motif among them, rather than to detect new binding motifs. Thus, if none of the algorithms A_i has detected the true binding motif, we will not expect the ensemble method detect it.

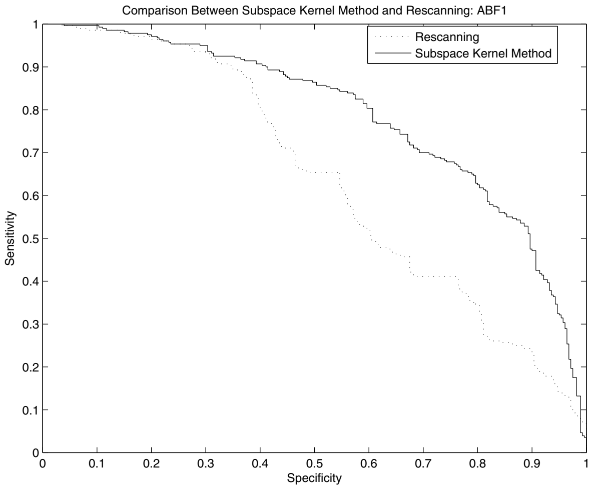

Figure 1. Comparison between Subspace Method and Rescanning

4.1 Spanning Space Method

Motivated by our previous work [14], the spanning space method is again the first approach to doing ensemble prediction. The method is described in more detail in [14], but basically involves agglomeration of k-mers which as features in $\oplus\mathcal{F}_i$ are most successful in predicting targets g of the TF t. These agglomerated k-mers then are used to form one or more candidate PWMs for binding sites of t. However this points us back to constructing new motifs based on discriminating k-mers in feature vectors $\mathbf{x} = \phi(g)$ of positive targets. We are then limited by the common drawback of word enumeration methods, namely limited width of input k-mers and of the predicted PWM. More specifically, if $k = 6$ the width of the predicted PWM is typically around 10. Use of longer k-mers can help in this case, though such k-mers may bring more variance into the dataset due to very low occurrence frequencies, and thus sparse feature vectors.

4.2 Kernel Method: Individual Assessment

To avoid dealing with very long k-mer features in this algorithm, we aim to select the most informative PWMs obtained by other algorithms and ultimately integrate them through agglomeration. Based on biological knowledge, most transcription factors recognize targets by identifying specific sequence pattern. In the pool of candidate PWMs, some of these contain information (true pattern) which can be used to separate positives from negatives, while others cannot provide such information. Hence strength in classifying promoters is consistent with a PWM which represents a true motif. Therefore those PWMs which through a learned algorithm have the strongest classification ability should be reported as the most likely ones to represent true motifs.

The kernel method proceeds as follows. An SVM is trained on each subspace kernel K_i (Section 3.2), each arising from a single PWM. The cross validated prediction strengths are then computed. Varying the machine classification threshold over different values, an ROC curve and AUC can be computed as a measure of discriminatory power. The kernels and thus motifs are ranked by their corresponding AUC. Since an AUC under 0.5 indicates the kernel performs worse than randomly, we eliminate those motifs and their PWM from the final pool used in agglomeration.

4.3 Synopsis Feature Space Method: Joint Information

The above *kernel method* still scores motifs individually. The *synopsis feature space method* reduces the generally large feature space $\oplus\mathcal{F}_i$ used in PWM evaluation to a low dimensional one $\mathcal{F}_s$, with dimension m equal to the number of input motifs. Each feature in $\mathcal{F}_s$ arises from a single input PWM. This reduction allows, for example, a simple combination of motifs in case where two predicted PWMs can recover mutually exclusive parts of a true binding motif. Their individual discriminatory power may not significantly higher than that of a non-functional pattern or a cofactor binding motif. In such a case, to assess the PWMs jointly in this way can provide more reliable results.

Running feature selection over the *synopsis feature space* gives us a way to do this very easily. In a synopsis mapping each PWM becomes a single axis in a new reduced feature space $\mathcal{F}_s$. Each synopsis feature is the best single scalar (by a given criterion) for separating targets and non-targets in its corresponding subspace $\mathcal{F}_i$. For example, if we define each synopsis feature to be the value $\mathbf{w}_i \cdot \mathbf{x}$ of a linear SVM classifier in each subspace, combining synopsis vectors in $\mathcal{F}_s$ is equivalent to analysis in the subspace $\text{span}\{\mathbf{w}_i\}_i$.

To avoid over-fitting, the feature selection in $\mathcal{F}_s$ should proceed on cross-validational synopsis features. That is, the entire dataset D is first divided into n mutually exclusive parts $\{D_k\}_{k=1}^n$ and the cross-validational synopsis feature on $\mathcal{F}_i$ for $x \in D_k$ is

$$f_i(\mathbf{x}) = \mathbf{w}_i^{(-k)} \cdot \mathbf{x} + b_i^{(-k)}$$

where $-k$ indicates the classifier is trained with D_k excluded.

5 Results

5.1 Ensemble Classification of Target Genes

In general, large experimentally curated sets of binding targets for a given TF t are not available. Therefore we employ a widely used ChIP-chip [17] dataset from Young's lab. We have selected probe sequences with p-value< 0.001 as positives and those with p-value > 0.8 as negatives to train our classifiers. Five-fold cross-validation is used for all experiments. We use only AlignACE and BioProspector as ensemble weak learners, and no machine learning method is involved in the pilot study of ensemble learners in order to avoid possible over-fitting. We choose the top 30 motifs from AlignACE and the top 3 form BioProspector with specified widths running from 6 to 15. This yields 60 motifs in the candidate pool, some of which are very similar.

Fig. 2 shows the receiver operating characteristic (ROC) curve for four methods on the *S. cerevisiae* TF $GCN4$, the 8-mer space method and three methods in Section 3. We choose an 8-mer space $\mathcal{F}$ for all subspace methods for purposes of comparison. 1000 8-mers are randomly generated from each of 60 PWMs. In the figure, we see a large improvement with the ensemble subspace method in terms of area under curve (AUC). The dimension of $\mathcal{F}$ is $4^8 = 65535$, but profile subsets yield around only 1500 unique 8-mers, giving better classification. This greatly reduced dimension implies biologically that the true motif for $GCN4$ should have a highly conserved (minimally varying) short core. For the TF $ABF1$ (Fig. 3), however, the analogous procedure gives similar performance instead of a large improvement. This suggests biologically that the true binding motif for $ABF1$ should be longer and should have a broader probability distribution in some columns.

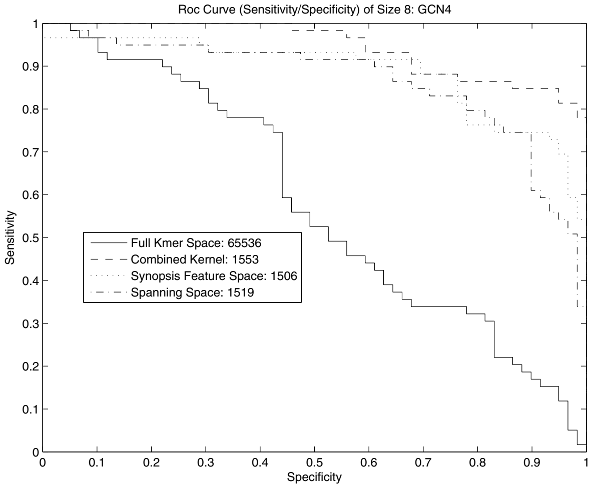

Figure 2. ROC curve for GCN4

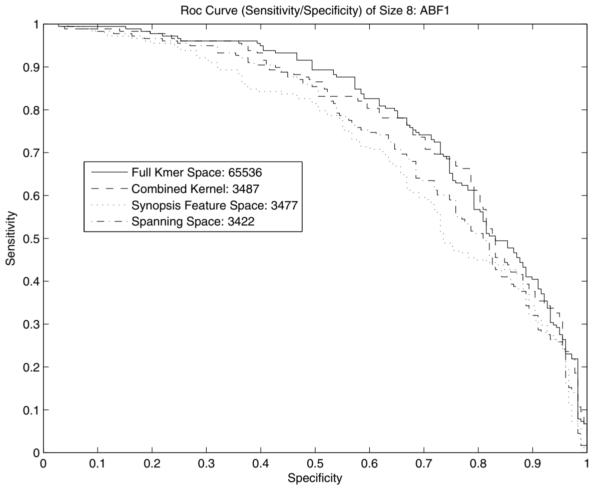

Figure 3. ROC curve for ABF1

We are also interested in how classification accuracy changes as we generate more and longer k-mers from each motif. Figure 2 shows the trends of sensitivity, specificity, positive predictive value and AUC for $GCN4$. The number of k-mers varies from 100 to 3000 and width k also varies from 5 to 8. Though no significant improvement occurs from varying the number of random k-mers, we see a clear improvement from increasing k-mer width. This again suggests that the true motif is highly conserved since a small number of random k-mers cover the profile set. The core part of the motif may have 7 bases, since we cannot improve with 8-mers. The trend for $ABF1$ (Fig. 5) can suggest a conserved two block motif based on the evidence that 5-mer is enough but AUC is low, where each block is short and highly conserved.

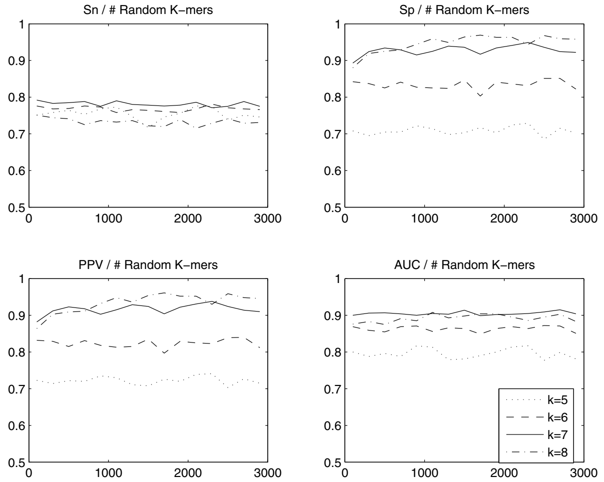

Figure 4. AUC trend for GCN4

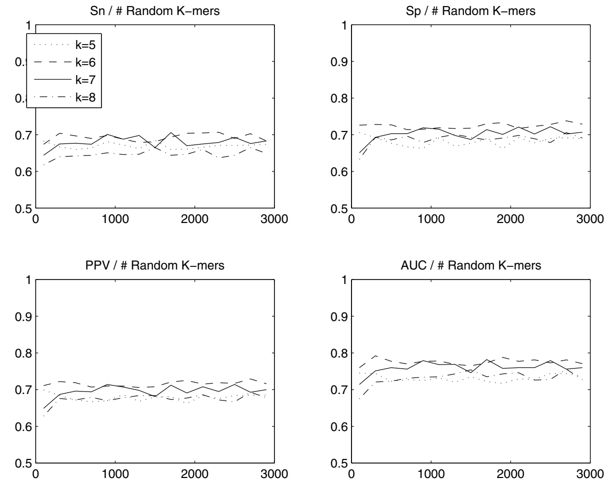

Figure 5. AUC trend for ABF1

AlignACE	BP	SVMotif	Union	Kernel	Synopsis
17	15	17	21	16	18

Table 1. Number of Motifs Detected

5.2 Ensemble Motif Discovery Method

For prediction of binding motifs, we use results from BioProspector [18], AlignACE [22], and SVMotif [14]. 26 TFs are selected from the list in Supplemental File 1 of MacIsaac's 2006 paper [20]. Only TFs with reported binding motifs in Transfac [26] are tested. For each TF, positives are selected from ChIP-chip data with p-value $p < 0.005$ and negatives with $p > 0.8$. The top 50 motifs from AlignACE, top 20 from SVMotif, and top 3 each from BioProspector with widths from 6 to 15 are used as the initial candidate pool. Some PWMs are very similar and are merged first in order to reduce correlations in predictor. As a result roughly 50 PWMs are left to be evaluated.

True (curated) motifs for 21 TFs are actually in this pool and the ensemble methods can predict 16 and 18 of these. However, this experiment aims to show the ability of SVMotif-PWM in integrating all information from different sub-learners. The ideal case is that one of the sub-learners reports the true motif, and the combined method takes advantage of it, also predicting this motif. Experiment shows that this is true in over 75 percent of the cases. Detailed output is available at http://cagt10.bu.edu/SVMotif

6 Method

6.1 Greedy Agglomeration

After the assessment of PWM quality (section 4), there are still non-informative PWMs (false patterns). An agglomeration procedure is helpful for merging similar PWMs and revealing a representative motif cluster. A greedy method with two thresholds has been used. Analogous to the clustering method for k-mers used in [14], a set of PWM clusters is formed recursively through addition of new PWM in a list by matching similarity to existing clusters. If the similarity score between a PWM M and the PWM of an existing cluster $\mathcal{C}$ is larger than the threshold **IN**, then M is added into $\mathcal{C}$. If it is smaller than another threshold (the **NEW** threshold), M starts a new cluster. If the score is between the two thresholds, M is temporarily skipped in the list. The list is ordered by decreasing PWM ranking. If at the end of the list no more PWMs can either be added into clusters or start new ones, a new random seed is selected from the remaining un-clustered PWMs. This continues until all PWMs are clustered. Once clusters are completed, a motif score combining entropy, feature importance and binding ratio is computed to rank them.

6.2 Motif Similarity

The above computation of motif similarity has two components, direct similarity and divergence from a random background. Each column of a PWM defines a distribution over Σ. The *Kullback-Leibler* (KL) *divergence* measures the distance between two of these as $D_{KL}(X\|Y) = \sum_{\alpha_i} X_i \log(X_i/Y_i)$. Here α_i runs through the alphabet in Σ and X_i and Y_i are two defined distributions. KL divergence measures the distance of distribution X from the reference distribution Y. The full similarity score between two PWM columns is defined as

$$D(P\|Q) = D_{KL}(P\|B) - D_{KL}(P\|Q) = \sum_{\alpha_i} P_i \log \frac{Q_i}{B_i}$$

where P and Q are columns from two motifs, and B is the background distribution. The first term measures divergence of P from the background and the second measures similarity between P and Q. In general, $D(P\|Q) \neq D(Q\|P)$, but the roles of P or Q will be clear from the context.

We can now define similarity between PWMs N and M, with M as reference, to be

$$A(N,M) = \max_{l} \left\{ \sum_{i} D(N_{i+l}, M_i) \right\}$$

with l the position lag in the alignment; the index i runs only through overlapping positions. In the agglomeration procedure, the reference PWM M has is that of a cluster $\mathcal{C}$.

6.3 Adding M into $\mathcal{C}$

Each cluster is represented by a *weight PWM* (WPWM) different from a conventional *frequency PWM*. Since the ranked PWMs in the list have different importance scores assigned by AUC (Section 4.2) or R-SVM (Section 4.3), a k-mer with higher importance is thus counted more than a k-mer with lower importance. Thus each column in a WPWM includes importance weights of nucleotides in addition to frequencies.

6.4 PWM Sources

This ensemble method integrates results from a collection of weak learners which compute PWMs in different ways from sets of known targets of a given TF. The components can be learning methods such as BioProspector [18], AlignACE [22], SVMotif [14] and so on. A good choice combines optimization methods which focus on different aspects of binding specificity and output PWMs with highly varied information. An additional source of candidate PWMs can be known (curated) PWMs of transcription factors, e.g., from the UCSC genome browser [12], which has binding motifs for 94 *S. cerevisiae* TFs. A third possible source of PWMs is k-mers or PWMs which are conserved across the genome and across different species, since functional sites tend to be conserved. As mentioned above, this framework provides a general method for integrating different information sources.

References

[1] T. Bailey and C. Elkan. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning*, 21:51–80, 1995.

[2] T. L. Bailey, N. Williams, C. Misleh, and W. W. Li. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucl. Acids Res.*, 34(suppl_2):W369–373, 2006.

[3] D. Che, S. T. Jensen, L. Cai, and J. S. Liu. Best: Binding-site estimation suite of tools. *Bioinformatics*, 21(12):2909–2911, 2005.

[4] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, Cambridge, 2000.

[5] M. C. Frith, Y. Fu, L. Yu, J.-F. Chen, U. Hansen, and Z. Weng. Detection of functional DNA motifs via statistical over-representation. *Nucl. Acids Res.*, 32(4):1372–1381, 2004.

[6] D. B. Gordon, L. Nekludova, S. McCallum, and E. Fraenkel. TAMO: a flexible, object-oriented framework for analyzing transcriptional regulation using DNA-sequence motifs. *Bioinformatics*, 21(14):3164–3165, 2005.

[7] C. Harbison et al. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431:99–104, 2004.

[8] G. Hertz, G. Hartzell, and G. Stormo. Identification of consensus patterns in unaligned dna sequences known to be functionally related. *Comput Appl Biosci*, 6:81–92, 1990.

[9] D. T. Holloway, M. A. Kon, and C. DeLisi. Machine learning methods for transcription data integration. *IBM Journal of Research and Development*, 50(6):631–644, 2006.

[10] J. Hu, B. Li, and D. Kihara. Limitations and potentials of current motif discovery algorithms. *Nucleic Acids Res*, 33:4899–4913, 2005.

[11] S. T. Jensen and J. S. Liu. BioOptimizer: a Bayesian scoring function approach to motif discovery. *Bioinformatics*, 20(10):1557–1564, 2004.

[12] D. Karolchik and R. Baertsch et al. The UCSC Genome Browser Database. *Nucl. Acids Res.*, 31(1):51–54, 2003.

[13] N.-K. Kim, K. Tharakaraman, and J. L. Spouge. Adding sequence context to a Markov background model improves the identification of regulatory elements. *Bioinformatics*, 22(23):2870–2875, 2006.

[14] M. A. Kon, Y. Fan, D. Holloway, and C. DeLisi. Svmotif: A machine learning motif algorithm. In *ICMLA '07: Proceedings of the Sixth International Conference on Machine Learning and Applications*, pages 573–580, Washington, DC, USA, 2007. IEEE Computer Society.

[15] R. Kuang, E. Ie, K. Wang, K. Wang, M. Siddiqi, Y. Freund, and C. Leslie. Profile-based string kernels for remote homology detection and motif extraction. In *CSB '04: Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference*, pages 152–160, Washington, DC, USA, 2004. IEEE Computer Society.

[16] C. Lawrence, S. Altschul, M. Boguski, J. Liu, A. Neuwald, and J. Wootton. Detecting subtle sequence signals: a gibbs sampling strategy for multiple alignment. *Science*, 262:208–214, 1993.

[17] T. Lee and R. Young et al. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, 298(5594):799–804, 2002.

[18] X. Liu, D. Brutlag, and J. Liu. Bioprospector: discovering conserved dna motifs in upstream regulatory regions of co-expressed genes. *Proceedings of the Sixth Pacific Symposium on Biocomputing*, pages 127–138, 2001.

[19] X. S. Liu, D. L. Brutlag, and J. S. Liu. An algorithm for finding protein-dna binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat Biotech*, 20:835–839, 2002.

[20] K. MacIsaac et al. An improved map of conserved regulatory sites for saccharomyces cerevisiae. *BMC Bioinformatics*, 7(1):113, 2006.

[21] K. A. Romer, G.-R. Kayombya, and E. Fraenkel. WebMOTIFS: automated discovery, filtering and scoring of DNA sequence motifs using multiple programs and Bayesian approaches. *Nucl. Acids Res.*, 35(suppl_2):W217–220, 2007.

[22] F. Roth, J. Hughes, P. Estep, and G. Church. Finding dna regulatory motifs within unaligned noncoding sequences clustered by whole-genome mrna quantitation. *Nature Biotechnology*, 16:939–945, 1998.

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

[24] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1998.

[25] V. Vapnik. *Statistical Learning Theory*. Wiley, New York, 2000.

[26] E. Wingender, P. Dietze, H. Karas, and R. Knüppel. Transfac: a database on transcription factors and their dna binding sites. *Nucleic Acids Res*, 24(1):238–241, January 1996.

[27] X. Zhang et al. Recursive svm feature selection and sample classification for mass-spectrometry and microarray data. *BMC Bioinformatics*, 7(1):197, 2006.