

# Discriminative Discovery of Transcription Factor Binding Sites from Location Data

Yuji Kawada and Yasubumi Sakakibara  
Department of Biosciences and Informatics, Keio University  
3-14-1 Hiyoshi, Kohoku-ku, Yokohama, 223-8522, Japan  
yuji@dna.bio.keio.ac.jp, yasu@bio.keio.ac.jp

## Abstract

**Motivation:** *The availability of genome-wide location analyses based on chromatin immunoprecipitation (ChIP) data gives a new insight for in silico analysis of transcriptional regulations.*

**Results:** *We propose a novel discriminative discovery framework for precisely identifying transcriptional regulatory motifs from both positive and negative samples (sets of upstream sequences of both bound and unbound genes by a transcription factor (TF)) based on the genome-wide location data. In this framework, our goal is to find such discriminative motifs that best explain the location data in the sense that the motifs precisely discriminate the positive samples from the negative ones. First, in order to discover an initial set of discriminative substrings between positive and negative samples, we apply a decision tree learning method which produces a text-classification tree. We extract several clusters consisting of similar substrings from the internal nodes of the learned tree. Second, we start with initial profile-HMMs constructed from each cluster for representing putative motifs and iteratively refine the profile-HMMs to improve the discrimination accuracies. Our genome-wide experimental results on yeast show that our method successfully identifies the consensus sequences for known TFs in the literature and further presents significant performances for discriminating between positive and negative samples in all the TFs, while most other motif detecting methods show very poor performances on the problem of discriminations. Our learned profile-HMMs also improve false negative predictions of ChIP data.*

## 1. Introduction

In genomic sequences, a motif is a set of *cis*-regulatory elements that preserve a certain nucleotide composition, playing a key role in transcriptional regulations. A large

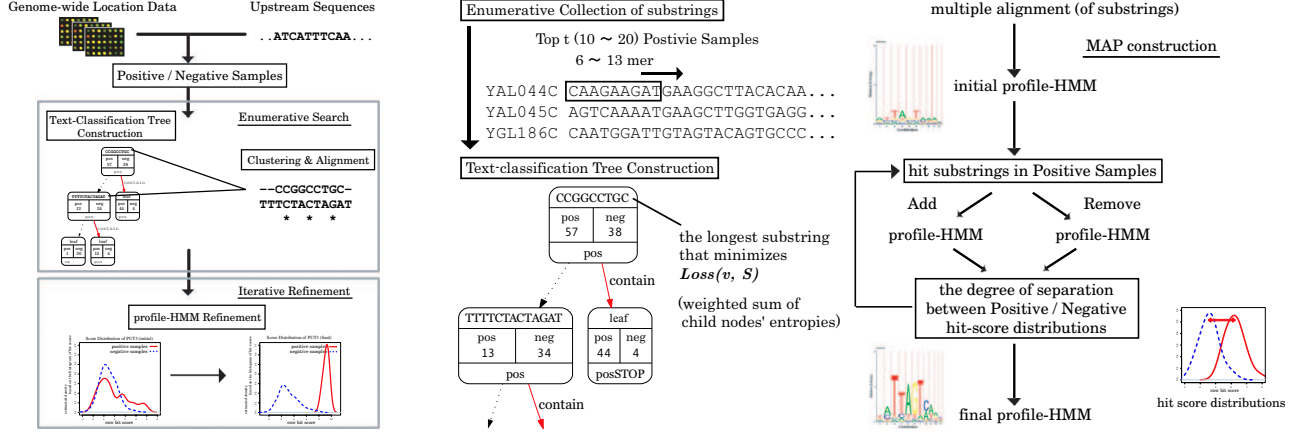
number of algorithms for finding motifs have been proposed previously [4], and most of them search for statistically over-represented patterns in the upstream sequences of co-regulated genes. On the other hand, genome-wide location data recently elucidated the *in vivo* physical interactions between transcription factors and their chromosomal targets on the genome [2, 3]. The ChIP data provide us the explicit and reliable interaction information about not only TF-DNA "binding" but also TF-DNA "unbinding".

Our fundamental idea for motifs is that the true motif appears only in the upstream sequences of the target genes controlled and bound by the TF and does NOT appear in those of the unbound ones. This idea leads us to a discriminative approach to find true motifs that correctly distinguish the upstream sequences between bound and unbound genes from positive and negative samples. We employ two machine learning techniques, decision tree learning for extracting most discriminative substrings and iterative reconstructions of hidden Markov models (HMMs) for improving the discrimination accuracies of motifs, where we use profile-HMMs to represent motifs.

Our genome-wide experimental results on yeast show that the discriminatively learned profile-HMMs agree with almost all the consensus sequences for well-known TFs in the literature and further present significant performances for discriminating upstream sequences between positive and negative samples, while most of the motifs discovered by other existing methods perform poorly on the problem of discriminations. The learned profile-HMMs also improve the false negative predictions of ChIP data that could not be predicted as bound genes by the location data in spite of the biological evidences in the literature.

## 2. Methods

Our algorithm consists of two steps: (i) build a text-classification tree by decision tree learning, extract relevant substrings, and create initial profile-HMMs; and (ii) itera-



**Figure 1. A schematic flow diagram of our proposed method (left), an illustration of text-classification tree construction (middle), and a flow diagram of iterative refinement of profile-HMMs (right).**

tively refine the profile-HMMs to improve the discrimination accuracies (Figure 1. left). Our method takes both location and genome data as input, and outputs a motif in the form of a profile-HMM for each TF.

In the preprocessing step, we select highly ChIP-array-enriched genes ( $p\text{-value} \leq 0.001$ ) as positive samples and least ChIP-array-enriched genes ( $p\text{-value} \geq 0.99$ ) as negative ones. Since we assume that true motifs only appear in the upstream sequences of positive samples and do not appear in those of negative ones, the use of a high confidence  $p\text{-value}$  threshold assures our assumption.

To build a text-classification tree, we then begin by collecting every nonredundant  $w\text{-mer}$  ( $6 \leq w \leq 13$ ) in both strands of the top  $t$  ( $10 \sim 20$ ) positive samples and recursively splits both positive and negative samples by the presence of a specific substring (Figure 1. middle). By using the minimum entropy criterion, we search for the longest substring that best minimizes the  $Loss$  function defined in Equation (1) from the collection of substrings. We denote a sequence by  $w$ , a substring by  $v$ , a class label ('positive' or 'negative') by  $l$ , samples by  $S$ , and by  $S_0, S_1, Occur()$  as follows;  $S_0^v = \{(w, l) \in S \mid w \text{ doesn't contain } v\}$ ,  $S_1^v = \{(w, l) \in S \mid w \text{ contains } v\}$ ,  $Occur(S, l_i) = |\{(w, l) \mid l = l_i\}|$ .

$$I(S) = - \sum_{i=1}^2 \frac{Occur(S, l_i)}{|S|} \log_2 \frac{Occur(S, l_i)}{|S|}$$


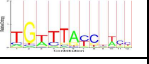
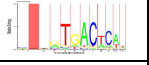
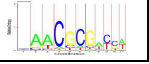
$$Loss(v, S) = \frac{|S_0^v|}{|S|} I(S_0^v) + \frac{|S_1^v|}{|S|} I(S_1^v) \quad (1)$$

$Loss$  function indicates a weighted sum of the entropies of two sets that are divided by the presence of some specific substring and we use pair-HMMs for local alignment to decide whether an upstream sequence contains each substring.

As a result of learning, substrings that are important for discrimination are extracted and are assigned to each internal node of the learned tree. Extracted substrings and their reverse complements are then clustered via  $k\text{-medoids}$  algorithms using each normalized pairwise similarity, a global-alignment score divided by the length of the longer substring, as the distance metric. Note that, in this case, the distance indicates not dissimilarity but similarity between substrings. The  $k\text{-medoids}$  clustering was performed 500 times to find a clustering with a maximal sum of intra-cluster distances. To find the optimal number of clusters, this process was performed with different number of clusters (from half of the total number of substrings to two), and the number with the minimal inter-cluster distance is adopted. Members of each cluster are multiple aligned by ClustalW, and a profile-HMM representing a putative motif is created from each multiple alignment.

The iterative strategy for refining initial profile-HMMs is similar to the one adopted by PSI-BLAST. It runs a local search for finding similar substrings with the current profile-HMM on the positive samples, selects the "best" hit substring in the sense that a new profile-HMM constructed by adding this substring to the current training set most significantly improves the discrimination accuracy, and it runs a next search. Similarly, the "best" substring constituting the current training set is removed. The discrimination accuracy of a profile-HMM means the capability of separating two hit-score distributions for positive and negative samples respectively assigned by the profile-HMM. These steps are iterated until convergence is achieved, that is, when there exists no such substring that improves the discrimination accuracy, the iteration stops (Figure 1. right). Each "iteration" refines the profile-HMM and, finally, we adopt the most discriminative one as a motif.

**Table 1. Examples of the discovered motifs.**

TF name	discriminative substrings	learned profile-HMMs	published consensus
CBF1	-ATCACGTGACAC GGTCACCCAA--- *****		TCACRTGA
FKH1	TTGTTTACCTTTC -TGATTGTGG--- ***		TTGTTTAC
GCN4	GTCGTGACTCA- ATCATGACAATT ** *****		ARTGACTCW
MBP1	AAACGCGTCCT		ACGCGT

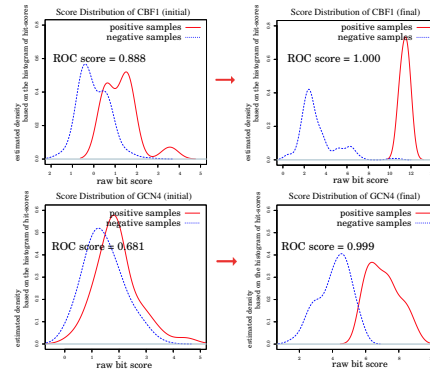
To quantify the separation capabilities of profile-HMMs, we use two metrics, ROC AUC [1] and the Fisher criterion. These metrics indicate the degree of separation between two distributions and a higher value means a greater separation.

The main reason to use different metrics is to obtain a better (local) optimum since no general methods exist for obtaining the global optimum. Two metrics are used in turn, and only when the convergence is achieved with both metrics, the iteration stops. We also try using six more metrics (e.g. entropy, MNCP [1], ICV, discriminant function), but the combination of those two metrics mentioned above turned out to perform best.

### 3. Results and discussion

We collect the sequences of 1000 bp upstream of the translation start sites of 6270 genes on yeast from SGD and SCPD, and two published genome-wide location data [2, 3]. To compare the discrimination accuracies of motifs obtained by our method with others, we also collect all the experimentally verified binding sites that are included in TRANSFAC (release 8.1) [5] and SCPD, and computationally discovered motifs of Harbison *et al* [2].

Since the number of overlapping TFs that are included in both TRANSFAC and two location data is 62, we applied our method to those 62 TFs. The total numbers of positive and negative samples of them are 3447 and 6906 respectively. As the transcriptional regulatory network is known to form a scale-free network, the number of positive samples ranges from 5 to 267 and that of negative ones ranges from 18 to 262, with an average of 56 positive samples and 105 negative ones per TF. Due to the page limitation, we will only show some typical results for several TFs using the location data of Harbison *et al* [2]. The full results for all the 62 TFs are available at our web site ([http://www.dna.bio.keio.ac.jp/disc\\_motifs](http://www.dna.bio.keio.ac.jp/disc_motifs)).

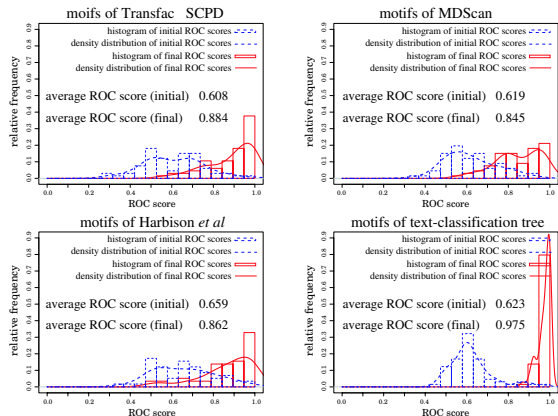


**Figure 2. Improvements of discrimination accuracies.**

To compare the consensus sequences with our learned profile-HMMs, we collect all the positional weight matrices (PWMs) from TRANSFAC. Since these PWMs have been constructed from experimental data taken from different binding studies of TFs and their cognate DNA sequences, the consensus sequences of them are thought to be quite reliable. There are 20 PWMs whose TFs are included in our study and 18 consensus sequences of them agree with our learned profile-HMMs. We also collect the consensus sequences of other 25 TFs that are described in at least two papers, and 21 of them agree with our discovered motifs. Table 1 shows the discriminative substrings extracted from the text-classification trees, learned profile-HMMs, and the published consensus for four TFs. The learned profile-HMMs are shown in HMM motif Logo. Our method correctly identifies the *cis*-regulatory patterns that agree with the published consensus, although they tend to be longer than the published consensus since our decision tree learning method searches for the longest substrings that best minimize the objective function.

Figure 2 shows the improvements of discrimination accuracies, that is, the hit-score distributions for positive and negative samples assigned by initial and final profile-HMMs for CBF1 and GCN4. As shown in Figure 2, discrimination accuracies of profile-HMMs are significantly improved by our iterative refinement method.

Due to the inherent variability of consensus sequences, it is difficult to evaluate the obtained motifs quantitatively. Thus, we use the discrimination accuracies of profile-HMMs measured by ROC AUC to assess the reliabilities of obtained motifs in addition to the similarities between the consensus sequences and our discovered motifs. To compare the motif-detecting performance of our method with those of other existing methods, we use three different initial profile-HMMs. They are constructed from the binding sites of TRANSFAC and SCPD, discovered motifs of



**Figure 3. Histograms of ROC scores of 62 TFs with different initial profile-HMMs.**

Harbison *et al* [2], and those of MDScan [4] respectively. Figure 3 shows the histograms of ROC scores ( $x$ -axis denotes ROC scores and  $y$ -axis denote percentages against the whole) with different initial profile-HMMs and the subsequent final profile-HMMs of all the 62 TFs.

From Figure 3, while our decision tree learning method itself shows a comparable performance to others, our iterative refinement method shows the significant improvement of discrimination accuracies. Hence, our text-classification tree is thought to identify a good starting point for our greedy iterative refinement. In other words, we think that by our decision tree learning method we can correctly identify the discriminative properties that are underlaid in the sequences of positive samples instead of numerous spurious similarities among them.

To compare the robustness of our method with other well-known methods, AlignACE and MDScan [4], we run 10-fold cross validation tests for nine TFs. We classify genes into bound or unbound ones depending on the score threshold at which the misclassification rate is the lowest. The results of four TFs shown in Table 2 demonstrate the effectiveness of our method. Our proposed method, the decision tree learning combined with the iterative refinement, is thought to be less dependent on the training data and discover more reliable motifs.

Although there exists no gold standard for evaluating the sensitivity of motif detecting methods, we assessed the sensitivity of our method by evaluating how our learned profile-HMMs improve the false negative predictions of location data. We collect from TRANSFAC all the genes that are experimentally verified to be bound but cannot be predicted by the location data [2]. There are 95 false negative genes for 32 TFs in total. Table 3 shows the number of false negative genes of the location data [2] as well as that of correctly identified ones by our method for four TFs. Our

**Table 2. 10-fold cross validation results.**

	Our method	MDScan	AlignACE
TF name	Test set	Test set	Test set
ABF1	0.967	0.841	0.827
CAD1	0.924	0.774	0.738
FKH2	0.891	0.826	0.819
SWI6	0.896	0.835	0.829

**Table 3. Improvement of false negative predictions of location data.**

TF name	False Negatives	Correctly Identified
ABF1	13	10
GCN4	6	3
MCM1	7	7
RAP1	11	10

learned profile-HMMs correctly identify 66 genes among them (69.5% improvements).

In conclusion, we present a novel discriminative motif discovery method based on the location data. The results indicate that our decision tree learning method correctly identifies the published consensus sequences for known TFs and the discrimination accuracies of obtained motifs in the form of profile HMMs are significantly improved by our iterative refinement method. Moreover, our iterative refinement method can be combined with any other motif detecting methods. With the progress of genome-wide location analyses, we hope that our method can provide more detailed view of motifs and hence present more reliable relationships between TFs and their target genes.

## References

- [1] N. Clarke and G. Joshua. Rank order metrics for quantifying the association of sequence features with gene regulation. *Bioinformatics*, 19(2):212–218, 2003.
- [2] C. Harbison, D. Gordon, T. Lee, N. Rinaldi, K. Macisaac, N. H. T. Danford, J. Tagne, D. Reynolds, J. Yoo, E. Jennings, et al. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431:99–104, 2004.
- [3] T. Lee, N. Rinaldi, F. Robert, D. Odom, Z. Bar-Joseph, G. Gerber, N. Hannett, C. Harbison, C. Thompson, I. Simon, et al. Transcriptional regulatory networks in *saccharomyces cerevisiae*. *Science*, 298:799–804, 2002.
- [4] X. Liu, D. Brutlag, and J. Liu. An algorithm for finding protein-dna binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nature Biotechnology*, 20(8):835–839, 2002.
- [5] E. Wingender, X. Chen, E. Fricke, R. Geffers, R. Hehl, I. Liebich, M. Krull, V. Matys, H. Michael, R. Ohnhauser, et al. The transfac system on gene expression regulation. *Nucleic Acids Research*, 29(1):281–283, 2001.