DISCRIMINATIVE DETECTION OF CIS-ACTING REGULATORY VARIATION 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

The interaction between transcription factors and their DNA binding sites plays a key role for understanding gene regulation mechanisms. Recent studies revealed the presence of "functional polymorphism" in genes that is defined as regulatory variation measured in transcription levels due to the *cis*-acting sequence differences. These regulatory variants are assumed to contribute to modulating gene functions. However, computational identifications of such functional *cis*-regulatory variants is a much greater challenge than just identifying consensus sequences, because *cis*-regulatory variants differ by only a few bases from the main consensus sequences, while they have important consequences for organismal phenotype. None of the previous studies have directly addressed this problem. We propose a novel discriminative detection method for precisely identifying transcription factor binding sites and their functional variants from both positive and negative samples (sets of upstream sequences of both bound and unbound genes by a transcription factor) based on the genome-wide location data. Our goal is to find such discriminative substrings that best explain the location data in the sense that the substrings precisely discriminate the positive samples from the negative ones rather than finding the substrings that are simply over-represented among the positive ones. Our method consists of two steps: First, we apply a decision tree learning method to discover discriminative substrings and a hierarchical relationship among them. Second, we extract a main motif and further a second motif as a *cis*-regulatory variant by utilizing functional annotations. Our genome-wide experimental results on yeast *Saccharomyces cerevisiae* show that our method presented significantly better performances for detecting experimentally verified consensus sequences than current motif detecting methods. In addition, our method has successfully discovered second motifs of putative functional *cis*-regulatory variants which are associated with genes of different functional annotations, and the correctness of those variants have been verified by expression profile analyses.

1. Introduction

Transcription factors (TFs) are DNA-binding proteins at the terminals of signal transduction networks and, in genomic sequences, a TF binding site (motif) is a set of *cis*-regulatory elements that preserve a certain nucleotide composition, playing a key role in transcriptional regulations. Each transcription factor recognizes a specific binding site composed of similar substrings, referred to as *cis*-regulatory variants. Recently, such subtle variations were hypothesized to also play a key role in transcription control. [1,5] It is generally assumed that *cis*-regulatory variants are hard to be detected only by sequence analyses but rather require extensive experimental studies. [1]

While a number of methods have been proposed previously, computational identification of TF binding sites is still a challenging and unsolved problem. Most existing methods

for detecting motifs examine only the upstream sequences of clustered, and presumably co-regulated, groups of genes or bound genes by the same TF, and search for statistically over-represented motifs among them. Such well-known motif detecting algorithms include AlignACE, Multiple EM for Motif Elicitation (MEME), Yeast Motif Finder (YMF), and MDScan. [4] Since biological signals are subject to mutations and usually do not appear exactly, they typically use probability weight matrix (PWM) to represent motifs. On the other hand, genome-wide location analyses, referred to as chromatin immunoprecipitation (ChIP) microarray experiments, recently elucidated *in vivo* physical interactions between TFs and their chromosomal targets on the genome. [2,3] The ChIP microarray technique can be thought to provide reliable and useful information about direct binding of a specific protein complex to DNA. In other words, the ChIP data provide us the explicit interaction information about not only TF-DNA "binding" but also TF-DNA "unbinding".

Our fundamental idea for detecting 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 distinguish the upstream sequences between bound and unbound genes.

Compared with most existing methods, our new strategy has three distinct features. First, our method takes unbound upstream sequences into account as negative samples as well as bound sequences as positive ones. Several approaches using ChIP data have been proposed previously, [4] but they still focus on the positive samples alone. Second, we define motifs as "discriminative" substrings that correctly distinguish the upstream sequences of positive samples from those of negative ones instead of statistically over-represented patterns or well-conserved ones. Even if using statistical criteria, methods that only focus on over-represented patterns suffer from numerous spurious random similarities. Third, we use a discriminative machine learning technique for detecting motifs, and we search for motifs using an exact-match, which is the opposite of the current probabilistic search strategies. Existing methods try to represent a motif by one single model allowing biological noises (mismatches, insertions and deletions) to some extent. Yet their obtained model is characterized by one specific substring, referred to as *consensus*. If one single consensus sequence characterizes the positive samples, it must be more precisely detected by using an exact-match search when negative samples are taken into account. In addition, by allowing ambiguity, current methods can not distinguish between the consensus sequences and their functional variants. As a result, they fail to detect the subtle differences of motifs that lead to important consequences for organismal phenotype. In contrast with most existing methods, we search for main motifs and their functional variants by focusing on the subtle differences among substrings rather than allowing and unifying them.

To search for the most discriminative substrings, we employ the decision tree learning method. Decision trees are used for classification tasks whose concepts are defined in terms of a set of attribute-value pairs. A text-classification tree classifies an input text (sequence) into one category according to several tests whether the input sequence contains some specific substrings. The inductive learning problem of decision trees is to construct such a text-classification tree from already classified sequences. In this paper, we use the

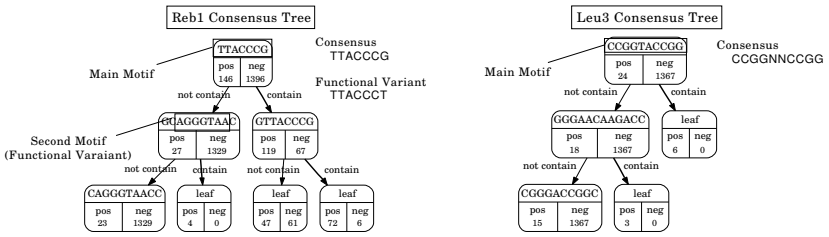


Figure 1. Motif detection by a decision tree learning method. These trees are constructed from both positive and negative samples of Reb1 and Leu3. The number of samples is shown in each node. The correctly identified consensus sequence and its previously inferred functional variant (only for Reb1) are shown inside the rectangles.

decision tree learning method for extracting sequence motifs given positive and negative samples. As a result of learning, substrings that are the most important and predictive for distinguishing the upstream sequences between positive and negative samples are extracted and are assigned to each internal node of the learned tree, which we call here a *consensus tree*. Figure 1 demonstrates the effectiveness of our method using the consensus tree. Our method correctly identified the consensus sequences for Reb1 and Leu3. As for the case of Reb1, a previous computational study [5] based on phylogenetic analysis only inferred the presence of Reb1 consensus variant. Our method succeeded to identify this variant and presented the relationships among them as a hierarchical tree structure. Further, our method inferred a number of *cis*-regulatory variants that have not previously been detected for many TFs through genome-wide experiments on S. cerevisiae.

2. Methods

Our method consists of two steps: (i) build a consensus tree by decision tree learning method, and (ii) search for highly functionally enriched motifs from the extracted substrings that are assigned in the internal nodes of the consensus tree.

In the preprocessing step, we select highly ChIP-array-enriched genes (binding P-value ≤ 0.001) as positive samples and low ChIP-array-enriched genes (binding P-value ≥ 0.80) as negative ones. The genome-wide location analyses assign P-value (confidence value) to each interaction between a TF and an intergenic region. It is reported that the empirical rate of false positives at a stringent P-value threshold ($P \leq 0.001$) is $6 - 10\%$ in the data of Ref. 3 and 4% in the data of Ref. 2. Since we assume that true motifs appear only in the upstream sequences of positive samples and not in those of negative ones, the use of a high confidence P-value threshold is required.

2.1. *Consensus Tree Construction*

We define motifs as informative substrings that can correctly classify genes into proper classes ('positive'/'negative') based on their upstream sequences. Thus, given a specific TF's positive and negative samples, our aim is to search for the most informative and hence discriminative substrings from the positive samples.

To accomplish this task, we use the decision tree learning method. We denote a se-

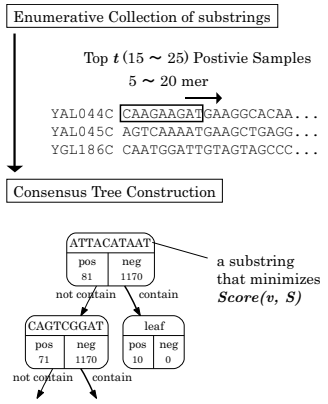


Figure 2. Consensus Tree Construction.

$BTLEARN(S, prnsrt, nsrt, keyl1, keyl2)$:

(1) Collect all the substrings.
$Keywords = \{v \mid keyl1 \leq |v| \leq keyl2\}$
(2) Output a consensus tree T.
T = *BTFIND*$(S,\ Keywords,\ prnsrt,\ nsrt)$

$BTFIND(S, Keywords, prnsrt, nsrt)$:

(1) If $(|S| - Occur(S, c_i))/|S| \leq nsrt$ is satisfied, return a subtree $T = c_i$.
(2) If $|S| \leq prnsrt$ is satisfied and the major class label with S is c_i, return a subtree $T = c_i$.
(3) If a substring v_g that minimizes $Score(v_g, S)$ is found from $v \in Keywords$, return v_g as an informative substring of the current node, a left-sided subtree T_0 and a right-sided subtree T_1.
T_0 = *BTFIND*$(S_0^v,\ Keywords - v,\ prnsrt,\ nsrt)$
T_1 = *BTFIND*$(S_1^v,\ Keywords - v,\ prnsrt,\ nsrt)$

Figure 3. Decision Tree Learning Algorithm.

quence by w, a substring by v, class labels ('positive'/'negative') by c and c_i, samples by S, and by S_0^v, S_1^v, $Occur$ as follows: $S_0^v = \{(w, c) \in S \mid w \text{ does not contain } v\}$, $S_1^v = \{(w, c) \in S \mid w \text{ contains } v\}$, $Occur(S, c_i) = |\{(w, c) \mid c = c_i\}|$. And v is "informative" if and only if $S_0^v \neq \emptyset$ and $S_1^v \neq \emptyset$.

If we have two classes ('positive' and 'negative') and denote their class labels by c_1 and c_2 respectively, the objective function is defined in Equation 1.

$$
\begin{aligned}
I(S) &= -\sum_{i=1}^{2} \frac{Occur(S, c_i)}{|S|} \log_2 \frac{Occur(S, c_i)}{|S|} \\
Loss(v, S) &= \frac{|S_0^v|}{|S|} I(S_0^v) + \frac{|S_1^v|}{|S|} I(S_1^v) \\
Score(v, S) &= Loss(v, S) + \tau \frac{1}{l} \log(p(v))
\end{aligned} \tag{1}
$$

where l is the length of v, $p(v)$ is the probability of generating v from a third-order Markov background model estimated from all the intergenic regions. τ is a free parameter and is chosen empirically. $Loss$ function indicates a weighted sum of the entropies of two sets that are divided by the presence of one specific substring. With the minimum entropy criterion, the most discriminative substring is the one that minimizes the $Score$ function.

The procedure of constructing a consensus tree by our decision tree learning method is shown in Figure 2. We begin by collecting every nonredundant w-mer in both strands of the top t $(15 - 25)$ positive samples, and then recursively search for the substring that minimizes the objective function with the current positive and negative samples from the collection of substrings, and divide both samples by the presence of it. The algorithm of decision tree learning is outlined in Figure 3. Given samples (S), two values for condition precedent ($prnsrt$ and $nsrt$) and lower and upper bounds of the length of the substring ($keyl1$ and $keyl2$), *BTLEARN* returns a learned consensus tree. By examining three TFs, we set $prnsrt = 10$, $nsrt = 0.01$, $keyl1 = 5$, and $keyl2 = 20$. We normalized the log likelihood of the background model, and set τ = 0.05

As a result of learning, substrings that are the most important and predictive for discrimination are extracted and are assigned to each internal node of the learned tree. Our decision tree learning method recursively split the search space, which is equivalent to clustering genes recursively by the presence of specific substrings. Therefore, we apply the following strategy for extracting the consensus sequence and their second variants: In a hierarchical structure of the learned tree, the main consensus sequence is extracted from the root node, and its significant second variants are extracted from the left children and the left descendants of the root node (Fig. 2). Since we assume that the number of significant functional variants is not large, we set the maximum depth of consensus trees to three.

2.2. *Extractions of cis-regulatory elements based on functional annotations*

After constructing the consensus tree, we search for a highly functionally enriched motif from an extracted substring in each internal node of the learned tree. Highly functionally enriched motif, which we call here a *functional consensus*, is the one whose target genes are highly associated with a same functional annotation. Target genes of a motif mean the genes which are included in the positive samples of the TF and whose upstream sequences contain a perfect-match to the motif. We assume that motifs are composed of several functional consensuses each of which regulates a specific set of genes. Since it is not usually possible to predict which nucleotide changes in motifs might affect expression, we search for main motifs and their variants by utilizing functional annotations.

We slide a window of length more than six along a discriminative substring in the node, and evaluate a motif in the window at each position by measuring its functional enrichment. For each window position, we calculate the hypergeometric P-value of independence between genes which are targets of the motif in the window and genes with the same GO biological process category, adjusted by Bonferroni correction for multiple testing. We collect the most functionally enriched motif as a functional consensus from every node.

The hypergeometric P-value is given by Equation 2.

$$P\text{-value} = \sum_{i=I}^{T} \frac{\binom{B}{i}\binom{G-B}{T-i}}{\binom{G}{T}} \tag{2}$$

where G is the total number of genes, B is the total number of genes in a particular biological process category, T is the number of target genes of the motif, and I is the number of genes which are targets of the motif and are in the particular biological process.

From the information-theoretic point of view, the most discriminative substrings are not necessarily be functionally enriched. Intuitively, they are too "informative" in the following sense. Since the ratio of nucleotide distribution in S. cerevisiae is approximately given by: A : T : G : C $=$ 32 : 32 : 18 : 18, the average information content of one nucleotide is: $I_{ave} = -\sum_{i\in\{A,T,G,C\}} p_i \log_2(p_i) \approx 1.94$ bits, where p_i is the frequency of occurrence of nucleotide i. The amount of information required to identify γ sites out of a possible Γ is given by: $I_\gamma = -\log_2 \frac{\gamma}{\Gamma}$ bits. Thus, if a motif is six base long and it occurs exactly once in every 1000 bases and may be placed in either of the two DNA strands in n sequences, the average information required to

identify a motif is then: $I_{actual} = -\log_2 \frac{n}{(n\times(1000-6+1)\times 2)} \approx 10.96$ bits. Therefore, $I_{actual}/I_{ave} \approx 5.64$ nucleotides are required to identify a motif from positive samples alone. However, in the discriminative framework, we search for a motif which appears only in the positive samples and must not appear in the negative ones. If we have p positive samples and q negative ones, the average information required to identify such a motif is: $I_{req} = -\log_2 \frac{p}{((p\times q)\times(1000-10+1)\times 2)}$ bits. In the case of $p = 50$ and $q = 1200$, $I_{req}/I_{ave} \approx 10.91$ nucleotides are required to identify such a discriminative motif.

The discussion stated above is just a rough approximation. In the discriminative framework, however, the required information tends to become high. Thus, to correctly identify functional consensuses, we need to "decompose" them by utilizing functional annotations. From the discussion stated above, we set the minimum length of a sliding window to six.

3. Experimental Results

3.1. *Data*

We collected the sequences of 1000 bp upstream of the translation start sites for 6270 genes on S.cerevisiae from SGD and SCPD, and two published genome-wide location data. [2,3] To search for functional consensuses and to assess the reliability of discovered *cis*-regulatory variants, we also collected various types of functional annotations, such as GO annotations for S.cerevisiae (process, component, and function), MIPS categories for S.cerevisiae (function, complex, motif, protein class, and phenotype), and a compendium of 827 gene expression profiles from 29 different publications. For evaluating obtained motifs, we collected all the 20 experimentally verified consensus sequences from TRANSFAC database and 25 from the literature that were reported in at least two papers. The average of the length of the collected motifs was 7.20 and the standard deviation of that was 2.27.

The total numbers of the location data that we used was 148. The number of positive samples ranged from 1 to 282 and that of negative ones ranged from 552 to 2084, with an average of 63 positive samples and 1177 negative ones per a TF. Due to the page limitation, we will only show typical experimental results for several TFs. The full results are available at our web site (`http://www.dna.bio.keio.ac.jp/reg_motifs`).

3.2. *Detection of Known Motifs*

We compared the motif detecting performance of our method with four other programs including AlignACE, MEME, YMF and MDScan. [4] AlignACE and MEME employ a heuristic local search approach, YMF employs an enumerative one, and MDScan employs a hybrid of enumerative and heuristic ones. Each program was run with default parameters. Note that, since the published consensus sequences are obtained empirically, they may not be the most functionally enriched and they are slightly different from literature to literature. Therefore, a discovered substring was considered to be consistent with the published consensus sequence if it contained at most one mismatch, insertion, or deletion.

When we only evaluated the top scoring motifs, that is, substrings that were assigned to the root nodes in the learned trees, our method correctly identified 38 out of 45 published

Table 2. Most Associated Functional Category.

Database		Motif	Associated Category	P value
GO	Process	TGACTC	amino acid metabolism	1.03 E-33
		GACTAA	nitrogen compound metabolism	5.782 E-19
	Function	TGACTC	molecular_function	5.89 E-10
		GACTAA	catalytic activity	1.07 E-04
	Component	TGACTC	cellular_component	2.57 E-10
		GACTAA	cellular_component	9.01 E-05
MIPS	function	TGACTC	amino acid metabolism	4.00 E-28
		GACTAA	amino acid metabolism	4.78 E-16
	complexes	TGACTC	Complexes by Systematic Analysis	2.31 E-16
		GACTAA	Complexes by Systematic Analysis	1.45 E-04
	protein class	TGACTC	Cys6 cysteine-zinc cluster	9.58 E-11
		GACTAA	Cys6 cysteine-zinc cluster	1.29 E-04
	phenotype	TGACTC	Auxotrophies, carbon and nitrogen utilization defects	1.43 E-05
		GACTAA	Methionine auxotrophy	1.41 E-03

Table 3. Differences of Expression Profiles.

Motif		Motif	t-test P value	Source
TGACTC	vs	GACTAA	1.27 E-07	Ploidy

Table 4. GO Terms for Gcn4.

GO Terms (assigned by YPD)
Amino acid biosynthesis
Cellular response to glucose starvation
Cellular response to nitrogen starvation
Cellular response to starvation
Response to stress
Nucleotide biosynthesis
Nucleobase, nucleoside, nucleotide and nucleic acid metabolism
Transcription from RNA polymerase II promoter
Regulation of transcription from RNA polymerase II promoter

Note: Terms that were associated with the main motif and the second motif are underlined.

consensus sequences. AlignACE identified 12, MEME identified 16, YMF identified 17, and MDScan identified 17 among 45 published consensus sequences. Within seven consensus sequences that our method failed to identify, four consensus sequences were discovered in other nodes of the learned trees. When we used random sequences generated from a third-order Markov background model as negative samples, our method identified 25. Table 1 shows 28 examples of 45 TFs used in our experiments, and shows discriminative substrings discovered by our method and discovered motifs with other four programs. In Table 1, motifs that were consistent with the published consensus sequences are underlined for our method and YMF and are shown with the mark of rectangles for AlignACE, MEME, and MDScan.

Our method clearly outperformed other programs, because all the existing methods only focus on the positive samples even if some of them were designed to utilize the location data. [4] In addition, our approach of using negative samples based on the location data was quite effective compared with using a random background model for negative samples. We assume that this result also contributed to the motif detecting performance of our method.

3.3. *Putative Cis-Regulatory Variants*

By performing the genome-wide search with our method on S. cerevisiae, we discovered putative functional variants for 17 TFs in total that were verified by both functional data analyses and expression profile analyses.

To assess the difference of expression profiles of two groups of targets, we used the paired t-test among all the Pearson correlations between every pair of genes within one group and those between every pair of genes each of which belongs to the different group. In other words, we assessed the difference between intra-cluster expression similarities and inter-cluster expression similarities. To select a meaningful threshold for both a hypergeometric P-value (functional enrichment) and a t-test P-value (expression difference), we calculated the average P-value of 1000 randomly selected motifs' targets for 10 times respectively, and we set a hypergeometric threshold to 0.1 and a t-test threshold to 0.01.

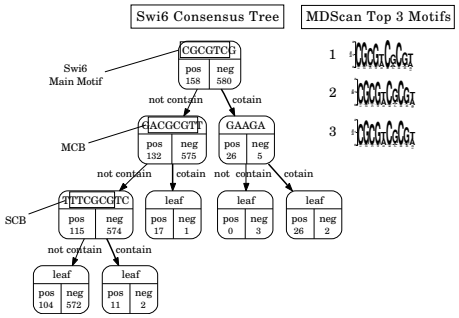


Figure 4. Relationship among different motifs induced by different complexes formed from the same non-DNA-binding cofactor, Swi6. (Swi6 forms two different complexes with different TFs, and each complex recognizes a specific motif)

Table 5. Most Associated GO Category.

Motif	Category	P value
CGCGTC	cell cycle	4.54 E-07
ACGCGT	cell cycle	1.10 E-06
TTCGCG	G1/S transition of mitotic cell cycle	6.82 E-07

Table 6. Top Two Associated TFs.

Motif	TFs	Overlaps	P value
ACGCGT	Swi6	46	1.58 E-56
	Mbp1	42	2.53 E-56
TTCGCG	Swi4	55	1.76 E-58
	Swi6	36	5.62 E-30

Table 7. Differences of Expression Profiles.

Motif		Motif	t-test P value	Source
CGCGTC	vs	ACGCGT	1.00 E-03	Stress Response
CGCGTC	vs	TTCGCG	1.32 E-10	Mitotic Cell Cycle
ACGCGT	vs	TTCGCG	7.81 E-06	Cell Cycle

Due to the page limitation, we only pick up Gcn4 as an example. Gcn4 regulates general control in response to amino acid or purine starvation. It involves in induction of genes required for utilization of poor nitrogen sources.

The discriminative substrings discovered in the root node and in the left children were TGACTCA (Table 1) and GATGACTAAC respectively. The discovered functional consensuses from them were TGACTC and GACTAA. Table 2-4 show the most associated functional categories, the difference of the expression profiles between those two motifs' targets, and the GO Terms for Gcn4 respectively. Table 2 and 4 indicate that targets of the main motif, TGACTC, primarily involve in the amino acid metabolism, while those of the second variant, GACTAA, involve in the nitrogen compound metabolism. Note that both target genes were predicted to be bound by the same TF from the location data, and targets of GACTAA had any significant overlaps with those of other TF's main motif. However, the expression profile analyses for them (Table 3) showed targets of GACTAA had a distinct biological property compared with those of the main motif of Gcn4 (TGACTC). Therefore, we concluded that GACTAA is a putative functional *cis*-regulatory variant of Gcn4.

3.4. *Detection of Multiple Motifs of Non-DNA-Binding Cofactors*

The representation of motifs as a hierarchical tree structure can be used for analyzing a relationship among multiple motifs induced by different complexes formed from the same cofactor. Our method correctly identified those relationships among motifs. To illustrate this, we pick up Swi6 as an example. (shown in Figure 4)

Swi6 is a non-DNA-binding cofactor of Mbp1 and Swi4. Swi6 and Mbp1 form MBF and Swi6 and Swi4 form SBF, both heterodimers are active during G1/S phase. Although Swi6 involves in both complexes, each complex recognizes a specific motif. MBF binds MCB (consensus:ACGCGT) and SBF binds SCB (consensus:CGCGAAA). Our method successfully identified both MCB and SCB from the positive and negative samples of Swi6, while MDScan failed to detect SCB. Further, our method presented the relationships be-

tween MCB and SCB as a hierarchical tree structure.

The functional consensuses of each internal node of the learned tree (Fig. 4) were CGCGTC, ACGCGT, and TTCGCG respectively. Table 5 shows the most associated GO biological process category for each motif's targets. Although these targets were predicted to be bound by Swi6 from the location data, targets of TTCGCG showed a distinct biological property. Their hypergeometric P-value associated with "cell cycle" category was just 0.00147. Table 6 shows the top two associated TFs with each motif. To determine the most associated TFs, we calculated the hypergeometric P-value of independence between targets of each motif and those of each TF's main motif, adjusted by Bonferroni correction (CGCGTC was excluded, since it was the main motif of Swi6). ACGCGT was highly associated with Mbp1, and TTCGCG was highly associated with Swi4. Table 7 shows the differences of expression profiles among each motif's targets. Targets of TTCGCG showed different expression profiles compared with others.

Table 5-7 clearly show the multimodality of Swi6. We assumed that the signal of MCB was stronger than that of SCB, since MCB-like motifs (CGCGTC, and ACGCGT) were discovered twice by our method and MDScan could only detect MCB. The consensus tree is thus able to reveal a relationship among multiple motifs of the same cofactor as a hierarchical tree structure.

4. Conclusion

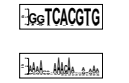
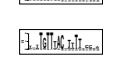
We present a novel discriminative motif detection method based on the location data. Our method significantly outperformed other motif detecting methods. Further, our method successfully detected putative functional *cis*-regulatory variants and also revealed the relationships among multiple motifs of the same cofactor for several TFs. Since our motifs obtained in this paper are just substrings, ongoing efforts for combining this method with methodologies of profile hidden Markov models will be published soon.

With the progress of genome-wide location analyses, we hope that our method can provide a useful platform for analyzing the regulatory functions of motifs including functional variants, and hence present more detailed analyses for transcriptional regulations.

References

1. C. Cowles, J. Hirschhorn, D. Altshuler, and E. Lander. Detection of regulatory variation in mouse genes. *Nature Genetics*, 32(3):432–437, 2002.
2. C. Harbison, D. Gordon, T. Lee, N. Rinaldi, K. Macisaac, N. Hannett 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. A. Tanay, I. Gat-Viks, and R. Shamir. A Global View of the Selection Forces in the Evolution of Yeast *Cis*-Regulation. *Genome Research*, 14(5):829–834, 2004.

Table 1. Comparison of Discovered Motifs.

TF name	Consensus	Our Method	AlignACE	MEME	YMF	MDScan
Abf1	TCAYTNTNNACG	TCACTATATACG			CACTWNAYACG	
Ace2	GCTGGT	GGGCGGGTG			CACACNCACAC	
Arg80	TTAAGTG	GCCGTTAAGT			CCGCGNCCGAC	
Bas1	TGACTC	CTGACTCCG			CACACNCACAC	
Cad1	TTASTAA	ATTAGTCAGC			CACACNCACAC	
Cbf1	TCACGTG	GTCACGTG			RTCACGTGAY	
Fkh1	GGTMAACAA	AAGGTAAACAA			AGGGGCGGGG	
Fkh2	GTAAACAA	TTGTTTACCTTT			CACACNCACAC	
Gcn4	TGACTCA	TGACTCA			AYATANATAYA	
Gln3	GATAAG	GATAAGATAAG			AYATANATAYA	
Hsf1	TTCNNGAA	TTCTAGAAG			CCCGTCTAGC	
Ino2	TTTCACATG	TTTTCACATGC			CGACCNCCGSG	
Ino4	TTTCACATG	TTTCACATG			ACGTANGTACG	
Leu3	CCGGNNCCGG	CCGGTACCGG			CCGGTNCCGGC	
Mbp1	ACGCGT	GACGCGTT			ACGCGWCGCG	
Mcm1	TCCYAATTNGG	CCAAATTAGG			CCGGGNCGTGC	
Msn2	MAGGGG	GCAGGGGCG			CTSCCNCATCC	
Msn4	CCCCT	CGAGAGCCCCA			CGAGGGCGCC	
Ndd1	TTGTTTAC	TTGTTTACCTTT			CACACNCACAC	
Pho4	CACGTG	CCACGTGC			CTACCCGGAG	
Rap1	ACCCATACA	ACCCATACA			CAYCCNTACAY	
Rcs1	TGCACCC	ACTGCACCC			ATATANRTATA	
Reb1	TTACCCG	TTACCCG			TSCGGGTAAY	
Ste12	TGAAACA	ATTTGAAACA			CCTGANTCAGG	
Sum1	GTGACNC	CTGACACCTG			ATATANATATA	
Swi4	CNCGAAA	GACGCGAAA			CACACNCACAC	
Swi6 (SCB)	CRCGAAA	TTTCGCGTC	none		none	none
Swi6 (MCB)	ACGCGT	CGCGTCG			ACGCGWCGCG	
Yap1	TTACTAA	TTAGTCAGCAT			CGACGNCGACG	