

LEARNING POSITION WEIGHT MATRICES FROM SEQUENCE AND EXPRESSION DATA

Xin Chen* and Lingqiong Guo

*School of Physical and Mathematical Sciences
Nanyang Technological University, Singapore*

*Email: chenxin@ntu.edu.sg, guol0005@ntu.edu.sg

Zhaocheng Fan

*Department of Computer Science and Technology
Tsinghua University, Beijing, China*

Email: fanzhaocheng@tsinghua.org.cn

Tao Jiang

*Department of Computer Science and Engineering
University of California at Riverside, USA*

Email: jiang@cs.ucr.edu

Position weight matrices (PWMs) are widely used to depict the DNA binding preferences of transcription factors (TFs) in computational molecular biology and regulatory genomics. Thus, learning an accurate PWM to characterize the binding sites of a specific TF is a fundamental problem that plays an important role in modeling regulatory motifs and discovering the binding targets of TFs. Given a set of binding sites bound by a TF, the learning problem can be formulated as a straightforward maximum likelihood problem, namely, finding a PWM such that the likelihood of the observed binding sites is maximized, and is usually solved by counting the base frequencies at each position of the aligned binding sequences. In this paper, we study the question of accurately learning a PWM from both binding site sequences and gene expression (or ChIP-chip) data. We revise the above maximum likelihood framework by taking into account the given gene expression or ChIP-chip data. More specifically, we attempt to find a PWM such that the likelihood of simultaneously observing both the binding sequences and the associated gene expression (or ChIP-chip) values is maximized, by using the sequence weighting scheme introduced in our recent work. We have incorporated this new approach for estimating PWMs into the popular motif finding program AlignACE. The modified program, called W-AlignACE, is compared with three other programs (AlignACE, MDscan, and MotifRegressor) on a variety of datasets, including simulated data, publicly available mRNA expression data, and ChIP-chip data. These large-scale tests demonstrate that W-AlignACE is an effective tool for discovering TF binding sites from gene expression or ChIP-chip data and, in particular, has the ability to find very weak motifs.

1. INTRODUCTION

The discovery of regulatory motifs in DNA sequences is very important in systems biology as it is the first and important step towards understanding the mechanisms that regulate the expression of genes. It is well-known that direct experimental determination of transcription factor (TF) binding motifs is not practical or efficient in many biological systems⁷. However, recent advances in high-throughput biotechnology such as *cDNA microarray* and *chromatin immunoprecipitation* (ChIP) offer a chance to discover *de novo* binding motifs at very low costs. Taking advantage of these new technologies,

at least three computational strategies for motif discovery have been proposed in the literature. We summarize them briefly in Figure 1. Although tremendous efforts have been made in the past decade, motif finding still remains a great challenge³⁰.

Regulatory motifs (or TF binding sites) are often modeled by *position weight matrices* (also known as *position specific scoring matrices*), which is a probabilistic model that characterizes the DNA binding preferences of a TF. Therefore, learning an accurate position weight matrix plays a key role not only in modeling a TF but also in distinguishing its true binding sites from spurious sites. This is particularly valuable for some motif discovery algorithms

*Corresponding author.

rately models a TF. An improvement could be made by taking the evolutionary history into account, as shown in PhyME²⁷ and PhyloGibbs²⁸. More recently, with the advent of DNA microarray and ChIP technologies, gene expression (or ChIP-chip) data has proven to be particularly valuable for motif discovery, as it represents an observable effect resulting directly from the binding of TFs. As illustrated in Figure 1, regression-based methods find motifs by correlating putative binding sites with expression data^{3, 5, 7, 18, 14}. However, none of them take advantage of expression/ChIP data explicitly to estimate an accurate PWM.

In our recent study, a *sequence weighting* scheme was proposed to estimate a PWM by explicitly taking gene expression variations (or binding ratios) into account⁶. We then incorporated it into the basic Gibbs sampling algorithm for motif discovery¹⁶, but with the limitation that it could only run in the site sampling mode (*i.e.*, it assumes exactly one binding site per sequence). Our preliminary experiments showed that sequence weighting was a quite effective approach to the estimation of PWMs, since it helped find weak motifs in two datasets (for TFs GAL4 and STE12) that were missed by the original (basic) Gibbs sampler. However, our recent tests on 40 ChIP-chip datasets from¹⁴ indicate that the approach still has a large room for improvement since the sequence-weighted Gibbs sampler would miss many of the motifs found by AlignACE, which is more modern Gibbs sampling algorithm^{13, 25}.

In this paper, we continue the development of the sequence weighting approach, and present several further improved motif discovery results. First, we extend the maximum likelihood problem naturally to find a PWM such that the likelihood of observing the combination of binding sites and expression data is maximized. The extension provides a theoretical foundation of the sequence weighting scheme, which is missing in⁶. Since binding sites inducing dramatic fold changes in expression (or showing strong binding ratios in ChIP experiments) are more likely to represent the true motif¹⁷, the sequence weighting scheme could therefore offer an approximate while reasonably good solution to the new maximum likelihood problem at very low computational cost. Second, we incorporate the sequence weighting scheme into the modern Gibbs sampling program, AlignACE^{13, 25}. The modified program is called W-AlignACE. Dif-

ferent from our previous implementation of sequence weighting in⁶, W-AlignACE is able to run in the motif sampling mode. In other words, it allows zero or multiple binding sites to occur in a promoter sequence. Third, we conduct large-scale tests on two high-throughput datasets including gene expression and ChIP-chip data, and compare the results of W-AlignACE with those obtained from AlignACE, MDscan¹⁷, and MotifRegressor⁷. Our results demonstrate that W-AlignACE performed the best in all tests, and was able to find very weak motifs such as those for DIG1 and GAL4, which were missed by the other three program.

The remainder of the paper is organized as follows. We first formulate a maximum likelihood problem for learning PWMs jointly from sequence and expression data in Section 2. Our experiments on simulated data will be presented in Section 3.1, and experiments on large real biological datasets, including mRNA expression and ChIP-chip data, will be presented in Section 3.2. Finally, some concluding remarks are given in Section 4.

2. LEARNING POSITION WEIGHT MATRICES

We consider how to estimate a PWM from binding sequences alone and from both binding sequences and expression/ChIP-chip data separately.

2.1. Learning PWMs from sequences

As mentioned before, a PWM Θ is often used to characterize the nucleotide frequencies at each position of a binding site, where

$$\Theta = (\theta_1, \dots, \theta_J)$$

and $\theta_j = (\theta_{a,j}, \theta_{c,j}, \theta_{g,j}, \theta_{t,j})^T$ represents the probability of observing the four nucleotides A, C, G, and T at the j th position of a binding site, such that $\theta_{a,j} + \theta_{c,j} + \theta_{g,j} + \theta_{t,j} = 1$ for each j , $1 \leq j \leq J$. In general, Θ is assumed to follow a *product Dirichlet* distribution^{19, 20}. Hence, the prior distribution on Θ is

$$\pi(\Theta) = \pi_1(\theta_1) \cdots \pi_J(\theta_J),$$

where $\pi_j(\theta_j)$ is a Dirichlet distribution $\text{Dir}(1, 1, 1, 1)$.

A PWM can be estimated from a collection of DNA sequences $\mathcal{R} = (R_1, \dots, R_n)$ that correspond to *aligned* binding sites of a TF, where $R_i =$

$(r_{i1}r_{i2}\cdots r_{iJ})$ represents the i th binding site, for each $i = 1, \dots, n$. These binding sites are assumed^{19, 20} to be randomly independent observations from a *product multinomial* distribution with parameter Θ ; that is, r_{ij} 's are mutually independent, and with probability $\theta_{a,j}$ take the nucleotide A, for example. It thus follows that the posterior distribution of Θ is also a product of independent Dirichlet distributions,

$$\pi(\Theta|\mathcal{R}) = \prod_{j=1}^J \text{Dir}(c_{a,j} + 1, c_{c,j} + 1, c_{g,j} + 1, c_{t,j} + 1),$$

where $c_{a,j}$, for example, is the count of nucleotide A among all the j th bases of the binding sites in \mathcal{R} . Further, by maximizing the likelihood of Θ , *i.e.*, $\pi(\mathcal{R}|\Theta)$, we have

$$\theta_{a,j} \propto c_{a,j} + 1, \quad \theta_{c,j} \propto c_{c,j} + 1,$$

$$\theta_{g,j} \propto c_{g,j} + 1, \quad \theta_{t,j} \propto c_{t,j} + 1.$$

That is, the probability of observing the nucleotide A (C, G, or T) at position j of a binding site is proportional to the count of nucleotide A (C, G, or T) among all the j -th position of the binding sites in \mathcal{R} .^a Indeed, this is exactly the method commonly used to estimate a PWM Θ for a TF, given a collection of its binding sites. Consequently, the conditional predictive distribution of a DNA sequence $B = (b_1 \dots b_J)$ will be

$$\pi(B|\Theta) \propto \prod_{j=1}^J \theta_{b_j,j} \propto \prod_{j=1}^J (c_{b_j,j} + 1).$$

2.2. Learning PWMs from both sequences and expression

We propose a new approach to learning PWMs through the combination of both sequence and expression data. The method can be easily extended to ChIP-chip data. Let $\mathcal{E} = (E_1, \dots, E_n)$ denote the fold changes of mRNA expression of downstream genes, where E_i is associated to the binding site R_i .^b We want to find a PWM Θ such that its likelihood $\pi(\mathcal{R}, \mathcal{E}|\Theta)$ is maximized; that is, Θ can best “explain” both the sequence and expression data simultaneously. The hope is that such a newly formulated problem will result in a PWM with significantly improved discriminative power. Finding the maximum

likelihood $\pi(\Theta|\mathcal{R}, \mathcal{E})$, however, is expected to be very hard, as it is conditioned on two disparate types of data whose exact quantitative correlation is not completely clear yet. Note that, expression fold changes are assumed to be induced as a result of the binding between DNA sequences and a TF.

Linear correlation between sequence and expression, *i.e.*, assuming additivity of binding sites' contributions to expression, has been used in several existing methods for motif site predictions^{3, 7}, most of which employ the third strategy that we discussed earlier in Figure 1. For the sake of a simple argument, the expression (log fold change) is assumed to be correlated proportionally to the conditional predictive distribution of its corresponding sequence, that is,

$$\log E_i \propto \pi(R_i|\Theta), \quad \text{for each } i, 1 \leq i \leq n,$$

or, for short, $\log \mathcal{E} \propto \pi(\mathcal{R}|\Theta)$. Therefore, we can reduce the maximum likelihood problem $\pi(\mathcal{R}, \mathcal{E}|\Theta)$ to the problem of finding a PWM Θ such that sequence \mathcal{R} fits expression $\log \mathcal{E}$ the best by linear correlation. A natural method to solve such a fitting problem is via an EM-like iteration, *i.e.*, starting with an initial PWM and then refining it iteratively^{18, 14}. However, such an iterative process is generally very time consuming. Moreover, it is clearly infeasible to incorporate such a process into a Gibbs sampling algorithm, which is an iterative algorithm by itself¹⁹.

In order to approximate Θ with an effective algorithm, we assume that the posterior distribution $\pi(\Theta|\mathcal{R}, \mathcal{E})$ is a product of independent Dirichlet distributions as $\pi(\Theta|\mathcal{R})$ but with different parameters; that is,

$$\pi(\Theta|\mathcal{R}, \mathcal{E}) = \prod_{j=1}^J \text{Dir}(\tilde{c}_{a,j} + 1, \tilde{c}_{c,j} + 1, \tilde{c}_{g,j} + 1, \tilde{c}_{t,j} + 1),$$

where $\tilde{c}_{a,j}$, for example, is the count of nucleotide A *weighted* by $\log \mathcal{E}$ among all the j th bases of the binding sites in \mathcal{R} . In other words,

$$\tilde{c}_{a,j} = \sum_{i=1}^n \delta(r_{ij}, A) \cdot \log E_i,$$

$$\text{where } \delta(r_{ij}, A) = \begin{cases} 1, & \text{if } r_{ij} = A \\ 0, & \text{otherwise} \end{cases}$$

^aThe additive term of 1 in the above formula is due to the prior distribution of $\pi_j(\theta_j)$.

^bNote that, multiple binding sites may share the same downstream gene and thus its associated log fold change value.

We can see that the above setting of parameters can be justified partially by the biological observation that binding sites inducing big fold changes in expression are more likely to represent a true motif¹⁷. It follows that the desired PWM will be

$$\theta_{a,j} \propto \tilde{c}_{a,j} + 1, \quad \theta_{c,j} \propto \tilde{c}_{c,j} + 1,$$

$$\theta_{g,j} \propto \tilde{c}_{g,j} + 1, \quad \theta_{t,j} \propto \tilde{c}_{t,j} + 1.$$

Similarly, the conditional predictive distribution of a DNA sequence $B = (b_1 \dots b_J)$ will be

$$\pi(B|\Theta, \mathcal{E}) \propto \prod_{j=1}^J \theta_{b_j,j} \propto \prod_{j=1}^J (\tilde{c}_{b_j,j} + 1).$$

Consequently, the new approach to the learning of PWMs is indeed done via the sequence weighting scheme recently proposed in⁶. Note that $\pi(B|\Theta, \mathcal{E})$ is completely equal to $\pi(B|\Theta)$ if every binding site induces the same fold change in gene expression. Figure 2 illustrates a simple example that clearly demonstrates the advantage of our new approach for learning PWMs from both sequence and expression data.

Gibbs sampling is known to be a very effective strategy for motif discovery. Its basic idea is to construct a Markov chain of a random variable X with $\pi(X)$ as its equilibrium distribution. For details on Gibbs sampling algorithms, the reader is referred to^{19, 20}. The above new predictive distribution $\pi(B|\Theta, \mathcal{E})$ can be used, in place of $\pi(B|\Theta)$, to implement a collapsed Gibbs sampling algorithm^{19, 20}. In particular, we have incorporated this method of computing PWMs into a powerful Gibbs sampling program, AlignACE (for *Aligns Nucleic Acid Conserved Elements*^{13, 25}). The modified program is called W-AlignACE, and available at <http://www.ntu.edu.sg/home/ChenXin/Gibbs>.

2.3. Quality measures of putative motifs

Putative motifs are generally scored and ranked before they are reported, because only the top few motifs undergo further investigations in practice. Therefore, a metric is needed to measure the goodness of putative motifs. Indeed, the metric to be chosen plays an important role in the success of motif discovery. An inappropriate metric might lower the rank of a *bona fide* motif so that it is unlikely to be discovered.

Information content is often used to measure the degree of nucleotide conservation in a motif given a probabilistic model Θ . It is defined as¹⁵

$$\text{IC} = \frac{1}{J} \sum_{j=1}^J \sum_{b \in \{A,C,G,T\}} \theta_{b,j} \log \frac{\theta_{b,j}}{\theta_{b,0}},$$

where $\theta_0 = (\theta_{A,0}, \theta_{C,0}, \theta_{G,0}, \theta_{T,0})^T$ is the nucleotide frequencies in the background sequence such that they sum up to one. The logarithm is often taken with base two to express the information content in bits. If each residue is equally probable in the background sequence then the information content can be as large as 2, representing the most conserved motif. Note that, however, a highly conserved motif may not be statistically significant relative to the expectation for its random occurrences in the promoter sequences under consideration. Figure 2 shows an example where sequence weighting might improve the information content of a PWM, although this is not necessarily always the case.

The MAP score is the metric for motif strength used by AlignACE to judge different motifs sampled during the course of the algorithm¹³. It is calculated for a motif by taking into account factors such as the number of aligned binding sites, the number of promoter sequences, the degree of nucleotide conservation, and the distribution of information-rich positions. Therefore, it is believed to be a more sensitive measure for assessing different motifs, in particular, those having different widths and/or different numbers of aligned binding sites.

Another alternative is to measure the statistical significance of correlation between putative motifs and gene expression. For example, the p -values from multiple linear regression are employed in REDUCE³ and also in MotifRegressor⁷ to rank putative motifs. Such a metric takes into account the variation of gene expression data, and is thus more plausible from the biological perspective. Note that, however, the presence of a few spurious binding sites may reduce the significance value dramatically. Therefore, it is not a robust metric.

2.4. Performance evaluation of putative motifs

To show the predictive ability of a motif discovery approach, we need an accurate yet feasible method to evaluate putative motifs. The most accurate method

	$\log \mathcal{E}$	1	2	3	4	5
(a)	4	A	C	T	G	A
	3	A	G	T	G	A
	2	A	G	T	C	A
	1	A	C	A	C	A

	1	2	3	4	5	
(b)	A	1	0	.25	0	1
	C	0	.5	0	.5	0
	G	0	.5	0	.5	0
	T	0	0	.75	0	0

	1	2	3	4	5	
(c)	A	1	0	.1	0	1
	C	0	.5	0	.3	0
	G	0	.5	0	.7	0
	T	0	0	.9	0	0

Fig. 2. Estimating PWMs. (a) A collection of four aligned DNA sequences bound by a TF, and the logarithmic fold changes in expression of their corresponding downstream genes listed in the first column. (b) The PWM learned from sequences alone. Its information content (see section 2.3 for definition) is 1.44 bits. (c) The PWM learned from both sequences and expression. Its information content improves to 1.53 bits, indicating the higher binding specificity of the motif. For instance, the TF is shown to bind to nucleotide G more preferentially than C at the fourth position, although both have the same counts observed in the sequences. Indeed, it can be justified by the fact that the nucleotide G occurs at the fourth position of the sequences that induce large fold changes in expression.

is clearly to directly verify if putative binding sites are true or not. This requires that the *bona fide* binding sites are already known before the evaluation, which, however, is not the case for most biological datasets. Therefore, the use of this method is limited to simulation experiments²⁴.

The second method is to compare the PWM of a putative motif with that of the true one. The true PWMs used for evaluation should be able to correctly reflect the binding preference of TFs. However, not many true motif PWMs have been found and are available in the public databases. For instance, of the 40 motifs that we study below, only 9 have PWMs in the TRANSFAC database²². Furthermore, these PWMs might not be considered true due to at least two reasons. First, they are derived from as few as eight binding sequences. Second, the computational method for learning a true PWM from binding sequences is questionable (see Figure 2). These reasons discourage us from using PWMs as benchmark for reliable performance evaluation, in particular at a large scale.

The third choice is to consider the consensus pattern of a putative motif. The consensus pattern is generally described using IUPAC-ambiguity codes, and hence a more rough (but robust) representation of TF binding preference than its corresponding PWM. In the IUPAC code of a motif, $\{A, C, G, T\}$ indicate the most conserved region of a consensus pattern, which we refer to as the *core* of a consensus pattern. Note that the core is the most informative part of a consensus. To compare motifs, a putative motif is usually considered true if its consensus core matches that of the true motif (*i.e.*, the weak region of the consensus pattern are ignored). It can be seen that such a comparison is not sensitive to either spu-

rious binding sites or the scarcity of binding sites, as is the previous method using PWMs.

Based on these observations, we will compare consensus cores in the performance evaluation of our predicted motifs in this study.

3. EXPERIMENTAL RESULTS

In this section, we present our test results of W-AlignACE on both simulated and real datasets. Note that the evaluation method proposed in³⁰ is not applicable here because W-AlignACE requires ChIP-chip or expression data in addition to promoter sequences.

3.1. Simulated data

We first perform tests on randomly generated sequence data, with artificially planted motif instances, to get an insight into the algorithm’s idealized performance under controlled conditions. Here, we generate more complicated simulated data than those used in many other studies^{6, 17}, in the hope to explore in depth how a PWM learned from sequence and expression effects the performance of motif finding algorithms. The data generating procedure is summarized as follows.

- (1) Manually create a motif consensus sequence consisting of a specified number of nucleotides. In our experiments, we consider three motif widths, $J = 6, 8, 10$, reflecting different levels of difficulty for motif finding.
- (2) Randomly generate 100 promoter sequences of 600 bases each.
- (3) A PWM Θ of size 4×10 is randomly generated according to the motif consensus and a given value for its information content.

- (4) Randomly generate 60 motif occurrences (*i.e.*, binding sites) according to the motif probabilistic model given by the PWM Θ .
- (5) Among 100 promoter sequences, we will not plant any binding sites in the bottom 50. That is, the 60 binding sites are planted in the top 50 promoter sequences at random positions by replacing segments of the same width. Because the planted positions are randomly selected, some of the top 50 promoter sequences may not contain any binding sites, while the others may contain multiple sites. Therefore, the total number of promoter sequences without any planted binding sites may exceed 50.
- (6) The hyperbolic tangent function, which is a scaled and biased logistic function, has been used in several studies to model the relation between sequence and expression^{1, 14}. Similarly, we use it to estimate expression values hypothetically induced by the planted binding sites. For example, for a promoter sequence S planted with m binding sites $\mathcal{R} = (R_1, \dots, R_m)$, where $R_i = (r_{i1}r_{i2} \dots r_{iJ})$ is the i th binding site in S , its expression value can be set using the following sequence of formulae,

$$f_{\Theta}(R_i) = \sum_{j=1}^J \log \frac{\theta_{r_{ij},j}}{\theta_{r_{ij},0}},$$

$$h_{\Theta}(S) = \log \left(\sum_{i=1}^m e^{f_{\Theta}(R_i)} \right),$$

$$q_{\Theta}(S) = \frac{1 - e^{-h_{\Theta}(S)}}{1 + e^{-h_{\Theta}(S)}},$$

$$E_{\Theta}(S) = 2^{1 + \frac{q_{\Theta}(S)}{W}},$$

which starts with computing the log-odds between the posterior probability of binding sites and a background probability of nucleotides. Note that the maximum expression value assigned this way could be close to 4. For those having no binding sites planted, the expression values are set to be randomly in the interval $[1, 2)$, simulating a commonly occurred situation in microarray experiments where some genes

may not have any binding site of the TF under investigation in their promoter regions, but are more or less expressed (possibly due to the binding of other TFs). Compared to equal expression values to be assigned, random expression values impose more difficulties on W-AlignACE to finding a correct motif, but apparently has no effect on AlignACE.

For each motif width, ten test datasets are generated with varying degrees of conservation, giving rise to a total of 30 datasets. Each dataset has 100 promoter sequences, each of which assigned an expression value as described above. We run both program AlignACE and W-AlignACE on the data, and then compare their predicted motifs with the planted motif. A predicted motif is considered true if it has the same consensus core as the planted motif. The results are summarized in Table 1. We can see that W-AlignACE is able to find more true motifs than AlignACE, and in most cases, the true motif is ranked the first among the list of reported motifs if sorted by their MAP scores.

3.2. Real data

Due to the stochastic nature of Gibbs sampling, we run for each dataset both programs AlignACE and W-AlignACE five times with different random seeds. MDscan¹⁷ and MotifRegressor⁷, instead, are run only once for each dataset because they are deterministic algorithms. Predicted motifs are sorted using their respective sorting schemes (*e.g.*, the MAP score for AlignACE), and only the top four are reported since the remaining motifs (ranked after the fourth) are generally too insignificant to be considered as true. In order to evaluate our method, we retrieve the consensus pattern for each motif from the Saccharomyces Genome Database⁴ (see <http://www.yeastgenome.org/>), and compare it with the motifs found by MDscan, MotifRegressor, AlignACE, and W-AlignACE, respectively.^c In our experiments, no prior knowledge on true motifs is assumed. Therefore, all the program parameters are set to their default values.^d For instance, the de-

^cSome motif consensi in the Saccharomyces Genome Database were obtained from putative binding sites, which have not been verified experimentally. Therefore, caution must be taken when using them as benchmark data.

^dMotifRegressor requires as many as 17 input parameters, for which we chose a typical setting (*i.e.*, their default values are generally preferred). The specific command line thus used to run MotifRegressor is “MotifRegressor MRexpression.txt MRsequences.txt yeast.int 1 1 2 1 1 2.0 1.5 5.0 5.0 10 10 50 30 MRoutput.txt”. For its detailed explanation, please refer to the documentation of MotifRegressor.

Table 1. Test results on 30 simulated datasets. For each motif width, we performed the test on ten PWMs with varying information contents.

Motif width	Information content	AlignACE Rank if found	W-AlignACE Rank if found
$J = 6$	0.65, 0.74, 0.77, 0.81, 0.88	-, -, -, -, -	-, -, 3, -, -
	0.91, 0.98, 1.01, 1.01, 1.18	-, -, -, -, -	-, -, 1, -, 1
$J = 8$	0.61, 0.71, 0.72, 0.88, 0.91	-, -, -, -, -	-, -, -, -, 1
	0.96, 1.02, 1.04, 1.08, 1.17	-, -, -, -, 1	2, -, 1, 1, 1
$J = 10$	0.63, 0.74, 0.79, 0.82, 0.93	-, -, -, -, 1	-, -, -, 2, 1
	0.98, 1.01, 1.03, 1.03, 1.03	1, -, 1, -, 1	1, 1, 1, -, 1

fault number of columns to align is set to 10. Working with default values is indeed a common practice, especially when the discovery of *novel* motifs is intended.

3.2.1. mRNA expression data

We have applied our algorithm to the publicly available dataset for yeast from microarray experiments on environmental stress response¹¹. A sample of 100 most induced genes by YAP1 overexpression is used here to demonstrate the advantage of the new learning approach in motif discovery. The log fold changes of these genes in mRNA expression range from 1.04 to 3.55.

YAP1 is a transcriptional activator required for oxidative stress tolerance, and is known to recognize the DNA sequence TTACTAA¹⁰ or the sequence GCTTACTAA with higher binding specificity, as annotated^e in the Saccharomyces Genome Database (<http://db.yeastgenome.org/cgi-bin/locus.pl?locus=YAP1>). Our experimental results show that, AlignACE failed to report any motifs containing the consensus pattern TTACTAA of the YAP1 motif among the top four motifs in each run. Instead, W-AlignACE successfully found the known YAP1 motif GCTTACTAAT and ranked it the second (MAP score: 126.68). A closer examination on all the putative motifs revealed that, AlignACE reported a weak pattern GATTAGTAAT ranked 12 (MAP score: 10.09) in one run and GCTTAGTAAT ranked 13 (MAP score: 9.41) in another run. Although both contain the complementary inverse of TTACTAA, neither exactly matches GCTTACTAA, the YAP1 motif annotated in the Saccharomyces Genome Database. Note that the second weak pattern above differs from the YAP1 motif by only one base at the sixth position, if we ignore the differ-

ence in motif width. MDscan reported the pattern GATTACTAAT as its top ranked motif, which differs from the YAP1 motif by one base at the second position. MotifRegressor did not performed better than MDscan, but instead it reported GATTACTAAT as its second motif. These results give a solid example where W-AlignACE is more accurate than AlignACE, MDscan, and MotifRegressor.

Table 2. Test results on the publicly available datasets from the yeast environmental stress response microarray experiment. Note that, only W-AlignACE discovered the YAP1 motif consensus in the Saccharomyces Genome Database without any mismatching.

Source	Consensus	Rank
Fernandes <i>et al.</i> ¹⁰	TTACTAA	-
SGD annotation	GCTTACTAA	-
W-AlignACE	GCTTACTAAT	2
AlignACE	GATTAGTAAT	12
	GCTTAGTAAT	13
MDscan	GATTACTAAT	1
MotifRegressor	GATTACTAAT	2

3.2.2. ChIP-chip data

We further apply our algorithm to the ChIP-chip data reported in²¹. Recall that a ChIP-chip experiment uses chromatin immunoprecipitation (ChIP), followed by the detection of enriched fragments using DNA microarray hybridization, to determine the genomic-binding location of TFs (TFs). Although the data are still noisy, they are the best genome-wide data of *in vivo* TF-DNA binding localization so far¹⁴. Forty datasets, each containing genes targeted by one TF, have been obtained using ChIP-chip *p*-value 0.001 as the cutoff in the study of¹⁴, and are publicly available at <http://biogibbs.stanford.edu/~hong2004/MotifBooster/>. The sizes of these datasets range from 25 up to 176 genes. For each

^eThe annotated consensus is indeed GCTKACTAA using IUPAC ambiguity codes, for which K represents the base G or T.

gene, its promoter sequence is taken up to 800 bps upstream, but not overlapping with the previous gene.

As mentioned earlier, we use consensus cores annotated in the *Saccharomyces* Genome Database as benchmark, and compare them with the putative motifs reported by MDscan, MotifRegressor, AlignACE, and W-AlignACE. To evaluate our method, we search for the putative motifs with consensus cores matching the annotated ones (while ignoring the difference in motif width), and consider them as being correct. Table 3 summarizes all the true motifs found for the forty TFs under investigation. At a first glance, it is already very encouraging to see that W-AlignACE successfully found the correct motifs for three TFs, DIG1, GAL4, and NDD1. This is especially interesting since these three TFs were observed in ¹⁴ to be among the nine TFs (the other six are GAT3, GCR2, IME4, IXR1, PHO4, and ROX1) whose correct motifs are hard to find. ^f

Compared to the other three program (MDscan, MotifRegressor, and AlignACE), W-AlignACE in general performed strongly. It found correct motif patterns for all the datasets that AlignACE was able to solve, and also for six additional datasets (ACE2, DIG1, GAL4, HAP4, STE12, SWI5). We further notice that in most cases, W-AlignACE reported a PWM with a much higher MAP score than AlignACE when a correct motif was found by both. When a spurious motif was reported, however, the MAP scores estimated by both program are comparable. For instance, both AlignACE and W-AlignACE found the correct consensus pattern nCGTmnnnAGTGAT for ABF1. Its MAP score is 351.866 as estimated by AlignACE, much lower than 436.877 estimated by W-AlignACE. In contrast, both program also reported an obviously spurious motif in the top four, GAAAAAAAAA. Its MAP scores are 176.129 and 165.632 given by AlignACE and W-AlignACE, respectively. All the above show that the new PWM learning approach via sequence weighting could increase the signal-to-noise ratio of a correct motif, but not of a spurious motif. Therefore, it may have a profound impact on the success of computational motif discovery, because it not only

increases the chance of find correct motifs, but also enhances our confidence about the predicted motifs. This is further demonstrated by the following case studies. Note that, the full test results are available at <http://www.ntu.edu.sg/home/ChenXin/Gibbs>, and so is the program W-AlignACE.

ACE2 is a TF that activates the transcription of genes expressed in the G1 phase of the cell cycle ⁸. Its ChIP-chip data in our study consists of 46 target genes. W-AlignACE successfully discovered the correct ACE2 motif, and ranked it the first with the highest MAP score 127.571. AlignACE did report the ACE2 motif in one of its runs but with a very low ranking of 9 (MAP score: 22.2304). In contrast, GAAAAAAAAA is the top motif found by AlignACE, having the MAP score as high as 104.081. Figure 3 depicts the distributions of some motifs in the promoter sequences, from which we can see that functional binding sites are more likely to occur in the promoter sequences having higher ChIP-chip scores. This observation is precisely the basis of W-AlignACE and why it performs better than AlignACE. Also note that, both MDscan and MotifRegressor failed to report any motifs resembling the correct ACE2 motif.

GAL4 is among the most characterized transcriptional activators, which activates genes necessary for galactose metabolism ²⁶. In our previous study ⁶, we incorporated the sequence weighting scheme into the basic Gibbs sampling algorithm from ¹⁶, which was only allowed to run in the site sampling mode (*i.e.*, assuming that exactly one binding motif occurs in each input promoter sequence), and tested it successfully on a small ChIP-chip data from the genome-wide location analysis ²⁶, which contains only 10 target genes. The current dataset from ¹⁴ contains 25 target genes. When run on this larger dataset, our previous algorithm ⁶ failed to find any motifs resembling the correct GAL4 motif (mostly likely because it was limited to the site sampling mode and could not properly handle multiple/zero occurrences of the correct motif). Indeed, GAL4 is a well-known motif that is too weak to be easily detected ¹⁴, partly because there is a 11-base gap (*i.e.*, degenerate region) in the middle of

^fFurther notice that, four of the above mentioned six TFs, GAT3, GCR2, IME4, and IXR1, do not have motif consensi annotated in the *Saccharomyces* Genome Database. Therefore, their motifs found by W-AlignACE are not evaluated here, and could still be true motifs.

Table 3. Experimental results on 40 CHIP-chip datasets. The highlighted rows indicate TFs for which W-AlignACE was able to find the correct motifs but AlignACE failed. The TFs with asterisks do not have motif consensus patterns annotated in the Saccharomyces Genome Database.

TF	#seq	MDscan	MotifRegressor	AlignACE		W-AlignACE	
		Consensus	Consensus	Consensus	MAP	Consensus	MAP
ABF1	176	CGTATATAAT		nCGTnnnnAGTGAT	351.866	nCGTnnnnAGTGAT	436.877
ACE2	46					GAACCAGCAA	127.571
BAS1	31	TGACTCCTTT		nnnAGGAGTCA	26.242	TGACTCCGnnnnnGA	164.367
CAD1	27	GATTACTAAT		GCTGACTAAT	22.3769	TGCTTAnTAAT	55.0084
CBF1	28	TCACGTGACC		nGGTCACGTG	91.5147	nGGTCACGTG	112.272
CIN5	116			ATTACATAAnC	25.7981	GnTTAnGTAAGC	162.825
DAL81	32						
DIG1	35					CnTnTGAAACAn	246.198
FHL1	124	TGTATGGGTG	TGTATGGGTG	ATGTnCGGGTG	241.916	ATGTnCGGGTG	370.814
FKH1	40						
FKH2	72	TGTTTACAAT		AAnGTAAACAA	40.8666	AAAnnGTAAACA	185.944
GAL4	25					CGGnCnAnAnnnnTCCG	184.307
GAT3*	31						
GCN4	56	AATGACTCAT	GATGAGTCAC	GGATGAGTCA	42.5719	GnATGAGTCAn	187.854
GCR2*	27						
HAP4	42					CnnGnnnnTGATTGGnnC	62.6472
HSF1	34	TTTTCTAGAA		GAAnnTTCnAGAA	50.569	GAnnnTTCnAGAA	88.2247
IME4*	27						
IXR1*	28						
MBP1	74	CGCGACGCGT		AAnAAACGCGT	36.9147	AnnAAACGCGTC	103.034
MCM1	59	CCTAATTAGG		TTnCCnnnTnnGAAAA	129.158	nTnCCnnAnnnGAAAA	179.82
NDD1	67	CCTAAATAGG		TTTCnAAAnGG	50.7552	CCnAAnnnGnAAAnnnT	222.986
NRG1	59		CCCTAGGCGC				
PDR1	45						
PHD1	70						
PHO4	41						
RAP1	127	TGTATGGATT		ATGTnTGGGTG	204.493	ATGTnTGGGTG	255.127
REB1	89	TCCGGGTAAC		nCCGGGTAAC	216.424	nCCGGGTAAC	262.57
RLM1	33						
ROX1	28						
SKN7	72						
SMP1	48						
STE12	54	TGAAACACAT				CnAnTnTGAAACA	358.174
SUM1	41	TGTGACAGTA		GTGnCAGnAAA	50.0198	GTGnCAGnAAA	69.7947
SWI4	90	AACGCGAAAA		GnnnCGCGAAAA	66.0847	GnGnCGCGAAAA	247.458
SWI5	72					AAnnnnnAGAnnGCTGG	109.432
SWI6	65			GnGnCGCGAAAA	48.4327	GnGnCGCGAAAA	49.8036
YAP1	35			GTTACTAAT	24.5596	ATTAGTAAGC	52.1866
YAP5	55						
YAP6*	65						

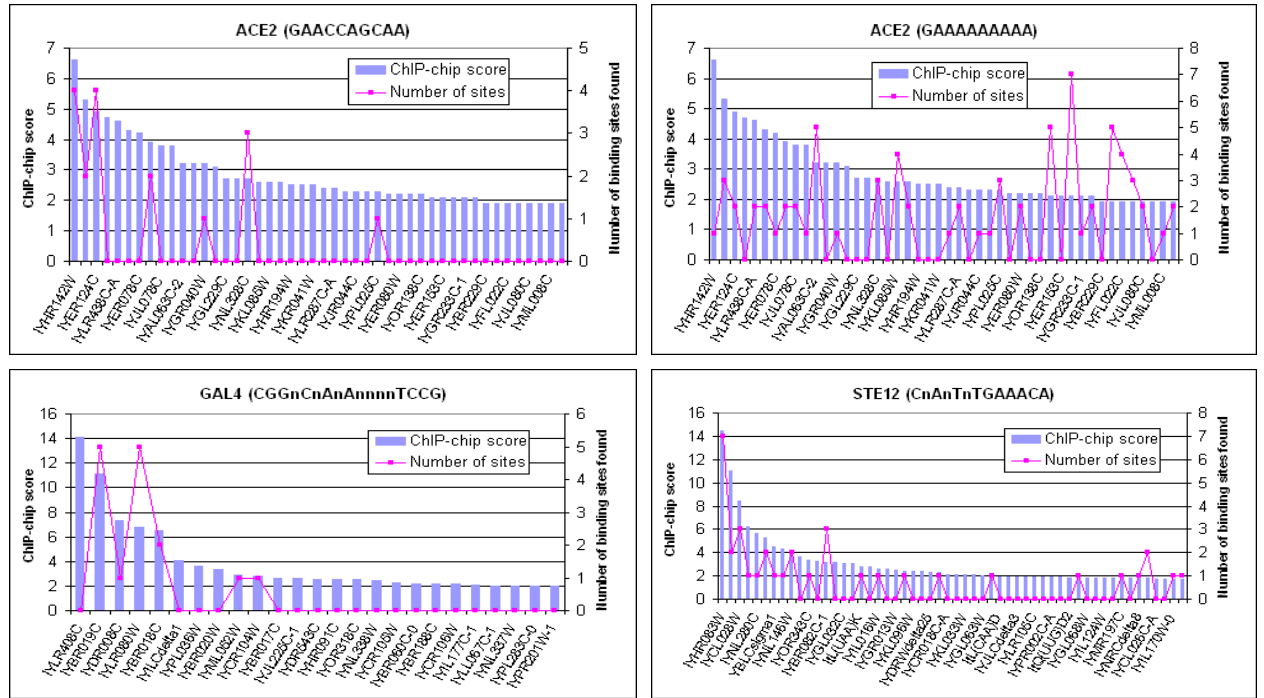


Fig. 3. The distributions of ChIP-chip scores and occurrences of the binding sites of three TFs ACE2, GAL4 and STE12. The top right figure depicts the distribution for a spurious motif ranked the first by AlignACE with MAP score 104.81, and the other three figures correspond to three correct motifs all ranked the first by W-AlignACE with MAP scores, 127.571, 184.307, and 358.174, resp. We can see that the correct motifs occur in promoter sequences with high scores more frequently than in those of low scores. This property generally does not hold for spurious motifs, whose occurrences are not expected to have any correlation with ChIP-chip scores or expression values.

its consensus pattern, *i.e.* CGGnnnnnnnnnnCCG. Therefore, the new dataset for GAL4 presents a new challenge for computational motif discovery methods. W-AlignACE once again performed remarkably better than AlignACE. It ranked the correct GAL4 motif the first with MAP score 184.307. In contrast, AlignACE failed to find the correct GAL4 motif, and neither did MDscan or MotifRegressor. A closer examination on the GAL4 dataset reveals that there are only 6 of the 25 genes whose promoter sequences contain the exact consensus pattern (see Figure 3). Furthermore, these six genes are all among the top if we sort all genes in the dataset by their ChIP-chip scores.[§] This might explain the failure of AlignACE and the success of W-AlignACE in the GAL4 dataset. MDscan failed perhaps because it was not optimized for finding gapped motifs.

STE12 is a DNA-bound protein that directly controls the expression of genes in response of haploid yeast to mating pheromones²⁶. The ChIP-chip dataset from¹⁴ consists of 54 pheromone-induced genes in yeast likely to be directly regulated by STE12. This data is also much larger than the dataset consisting of 29 genes used in our previous study⁶. W-AlignACE once again found the correct motif and ranked it the first with MAP score 358.174. On the contrary, AlignACE ranked the correct motif only the fourteenth with a much lower MAP score of 49.0173. This is not surprising, because once again most of the occurrences of the correct motif are located in the promoter regions of genes having high ChIP-chip scores, as shown in Figure 3. In conclusion, the sequence weighting scheme that learns PWMs from both sequence and expression data could indeed boost AlignACE's ability to pick correct motifs from sequences with noisy background.

It is interesting to note that MotifRegressor performed much worse than MDscan in this test, although the former uses the latter as a feature extraction tool to find candidate motifs^h. This could be due to several factors. First, the cutoff used by MotifRegressor on the significance of linear regression might be strict. Second, the true motifs are too weak as evaluated by MotifRegressor based on the signifi-

cance of linear regression (*e.g.*, due to the presence of spurious binding sequences). Third, the parameter setting that MotifRegressor applied to MDscan did not work as well as the default one, which we used to test MDscan. Last, the parameters that we set for MotifRegressor might not be optimal either.

4. DISCUSSION AND FUTURE RESEARCH

Learning an accurate PWM from a collection of aligned binding site sequences is a delicate problem that plays an important role in modeling a TF. In this paper, we tackled this problem by proposing a new approach to learning PWMs jointly from sequence and expression. We believe that this approach could be a very useful enhancement to many of the motif discovery programs that are based on PWMs, such as Gibbs sampling and MEME. Our preliminary experiments on Gibbs sampling support this belief, and demonstrate that W-AlignACE is a very effective tool for biologists to computationally discover TF binding motifs when the gene expression or ChIP-chip data are given. The web W-AlignACE service is provided at <http://www1.spms.ntu.edu.sg/~chenxin/W-AlignACE>. Our future work includes more delicate/theoretical treatment of multiple motif occurrences, and treatment of multiple experiment expression data (which are usually time series data) and cooperative motifs (or *cis*-regulatory modules).

ACKNOWLEDGMENTS

XC's research is supported by a start-up fund from NTU. TJ's research is supported by NSF grant CCF-0309902, NIH grant LM008991-01, NSFC grant 60528001, National Key Project for Basic Research (973) grant 2002CB512801, and a Changjiang Visiting Professorship at Tsinghua University.

References

1. Y. Barash, G. Bejerano, and N. Friedman. A simple hyper-geometric approach for discovering putative transcription factor binding sites. *Algorithms in Bioinformatics: Proc. First International Workshop*, number 2149 in LNCS, pp. 278-293, 2001.

[§]Unfortunately, there is no GAL4 binding site at the upstream of the top gene, which actually presents more challenge to W-AlignACE than to AlignACE for discovering the correct motif.

^hMore precisely, the current implementation of MotifRegressor uses MDmodule, instead of MDscan, as a feature extraction tool. MDmodule is a modified version of MDscan.

2. T. Bailey and C. Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28-36, 1994.
3. H. Bussemaker, H. Li, and E.D. Siggia. Regulatory element detection using correlation with expression. *Nat. Genet.*, **27**, 167-171, 2001.
4. J. Cherry, C. Ball, *et al.* Genetic and physical maps of *Saccharomyces cerevisiae*. *Nature*, **387**, 67-73, 1997.
5. D. Chiang, P. Brown, and M. Eisen. Visualizing associations between genome sequences and gene expression data using genome-mean expression profiles. *Bioinformatics*, **17**, S49-S55, 2001.
6. X. Chen and T. Jiang. An improved Gibbs sampling method for motif discovery via sequence weighting. *Proc. of Computational System Bioinformatics*, 239-247, 2006.
7. E. Conlon, X. Liu, J. Lieb, and J. Liu. Integrating regulatory motif discovery and genome-wide expression analysis. *PNAS*, **100**, 3339-3344, 2003.
8. P. Dohrmann, G. Butler, K. Tamai, S. Dorland, J. Greene, D. Thiele, and D. Stillman. Parallel pathways of gene regulation: homologous regulators SWI5 and ACE2 differentially control transcription of HO and chitinase. *Genes Dev.* **6**, 93-104, 1992.
9. J. Dolan, C. Kirkman, and S. Fields. The yeast STE12 protein binds to the DNA sequence mediating pheromone induction. *Proc. Natl. Acad. Sci. USA*, **86**, 5703-5707, 1989.
10. L. Fernandes, C. Rodrigues-Pousada, and K. Struhl. Yap, a novel family of eight bZIP proteins in *Saccharomyces cerevisiae* with distinct biological functions. *Mol. Cell. Biol.*, **17**, 6982-6993, 1997.
11. A. Gasch, P. Spellman, C. Kao, O. Carmel-Harel, M. Eisen, G. Storz, D. Botstein, P. Brown. Genomic expression programs in the Response of Yeast cells to environmental changes. *Mol. Biol. Cell*, **11**, 4241-4257, 2000.
12. I. Holmes and W. Bruno. Finding regulatory elements using joint likelihoods for sequence and expression profile data. *Proc. of ISMB*, 202-210, 2000.
13. J. Hughes, P. Estep, S. Tavazoie, and G. Church. Computational identification of *Cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.*, **296**, 1205-1214, 2000.
14. P. Hong, X. Liu, Q. Zhou, X. Lu, J. Liu, and W. Wong. A boosting approach for motif modeling using ChIP-chip data. *Bioinformatics*, **21**, 2636-2643, 2005.
15. G. Hertz and G. Stormo. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563-577, 1999.
16. C. Lawrence, S. Altschul, M. Boguski, J. Liu, A. Neuwald, J. Wootton. Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208-214, 1993.
17. 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**, 835-839, 2002.
18. H. Leung, F. Chin, S. Yiu, R. Rosenfeld, and W. Tsang. Finding motifs with insufficient number of strong binding sites. *Journal of Computational Biology*, **12**, 686-701, 2005.
19. J. Liu. The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*, **89**, 958-966, 1994.
20. J. Liu, A. Neuwald, and C. Lawrence. Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *Journal of the American statistical Association*, **90**, 1156-1170, 1995.
21. T. Lee, *et al.* Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, **298**, 799-804, 2002.
22. V. Matys, E. Fricke, *et al.* TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res*, **31**, 374-378, 2003.
23. A. Neuwald, J. Liu, and C. Lawrence. Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci.*, **4**, 1618-1632, 1995.
24. P. Pevzner and S. Sze. Combinatorial approaches to finding subtle signals in DNA sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **8**, 269-278, 2000.
25. F. Roth, J. Hughes, P. Estep, and G. Church. Finding DNA regulatory motifs within unaligned non-coding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotech.*, **16**, 939-945, 1998.
26. B. Ren, *et al.* Genome-wide location and function of DNA binding proteins. *Science*, **290**, 2306-2309, 2000.
27. S. Sinha, M. Blanchette, and M. Tompa. PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinformatics*, **5**: 170, 2004.
28. R. Siddharthan, E. Siggia, E. Nimwegen. PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS Computational Biology*, **1**, e67, 0534-0555, 2005.
29. E. Segal, R. Yelensky, and D. Koller. Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. *Bioinformatics*, **19**: i273-i282, 2003.
30. M. Tompa, N. Li, *et al.* Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotechnology*, **23**, 2005.