BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–15

W-AlignACE: An improved Gibbs sampling algorithm based on more accurate position weight matrices learned from sequence and gene expression/ChIP-chip data

Xin Chen [1,*], Lingqiong Guo [1], Zhaocheng Fan [2] and Tao Jiang [3]

[1]School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore
[2]Department of Computer Science and Technology, Tsinghua University, Beijing, China
[3]Department of Computer Science and Engineering, University of California at Riverside, USA

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: 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 also in discovering the regulatory targets of TFs.

Results: We study the question of how to learn a more accurate PWM from both binding sequences and gene expression (or ChIP-chip) data, and propose to find a PWM such that the likelihood of simultaneously observing both binding sequences and their associated gene expression (or ChIP-chip) data is maximized. To solve the above maximum likelihood problem, a sequence weighting scheme is thus introduced based on the observation that binding sites inducing drastic fold changes in mRNA expression (or showing strong binding ratios in ChIP experiments) are likely to represent a true motif. We have incorporated this new learning approach 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, mRNA expression and ChIP-chip data. These tests demonstrate that W-AlignACE is an effective tool for discovering TF binding motifs from gene expression (or ChIP-chip) data and, in particular, has the ability to find very weak motifs like DIG1 and GAL4.

Availability: http://www.ntu.edu.sg/home/ChenXin/Gibbs

Contact: chenxin@ntu.edu.sg

Supplementary materials: Available at *Bioinformatics* Online

1 INTRODUCTION

The discovery of regulatory motifs in DNA sequences is very important in systems biology as it is the first step towards understanding the mechanisms that regulate the expression of genes. With the advent of of high-throughput biotechnologies such as *cDNA microarray* and *chromatin immunoprecipitation* (ChIP), at least three computational strategies have been proposed to discover *de novo* binding motifs at very low costs. We summarize them briefly in Supplementary Figure 1. Although tremendous efforts have been made, motif finding remains a great challenge (Tompa *et al*., 2005).

Regulatory motifs are often modeled by *position weight matrices* (PWMs), which is a probabilistic model that characterizes the DNA binding preferences of a transcription factor (TF). Therefore, learning an accurate PWM plays a key role not only in modeling a TF's binding preferences but also in distinguishing its true binding sites from spurious sites. This is particularly critical for some motif discovery algorithms which rely heavily on position weight matrices, for instance, MEME and Gibbs sampler (Bailey and Elkan, 1994; Lawrence *et al*., 1993).

A PWM is generally learned from a collection of *aligned* DNA binding sites that are likely to be bound by a common TF. Theoretically, it is formulated as a maximum likelihood problem — finding a PWM such that the likelihood of the observed set of binding sites is maximized (Liu, 1994). To solve it, one may assume that the binding sites are independent random observations from a *product multinomial* distribution, from which it follows that each entry of the PWM will be proportional to the count of a nucleotide at the corresponding position. This is precisely the method commonly used to compute a PWM from a collection of binding sites (Stormo, 2000). However, learning from DNA binding sequences alone might not be sufficient to find a PWM that accurately models a TF's binding preferences. For example, an improvement could be made by taking the evolutionary history into account, as shown in PhyME (Sinha *et al*., 2004) and PhyloGibbs (Siddharthan *et al*., 2005).

In this paper, we study the question of how to learn an accurate PWM from both DNA binding sequences and expression data [1]. First, we extend the above maximum likelihood problem to find a PWM such that the likelihood of observing the combination of binding sequence and expression data is maximized. This extension is natural as expression data are direct observable results from the binding of TFs to DNA sequences. Then, a sequence weighting scheme is introduced to find a PWM, where every binding site is assigned a weight proportional to the logarithm fold change of mRNA expression of its downstream gene. Since binding sites

*To whom correspondence should be addressed

[1] In this paper, we use "gene expression" broadly to refer to not only standard mRNA microarray data, but also ChIP-chip data.

inducing drastic fold changes in expression (or showing strong binding ratios in ChIP experiments) are more likely to represent the true motif (Liu *et al*., 2002), the sequence weighting scheme could therefore offer an approximate while reasonably good solution to the new maximum likelihood problem at very low computational cost. Compared to the common learning approach, it allows to take advantage of gene expression variations explicitly so that a more accurate PWM is likely found. Third, we incorporate the sequence weighting scheme into the modern Gibbs sampling program AlignACE (Hughes *et al*., 2000; Roth *et al*., 1998), and the modified program is called W-AlignACE. Finally, we conduct large-scale tests on both simulated and real biological datasets, and compare the results of W-AlignACE with those obtained from well-known motif finding programs including AlignACE, MDscan (Liu *et al*., 2002), and MotifRegressor (Conlon *et al*., 2003). 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 all other three programs.

2 METHODS

2.1 Learning PWMs from sequences

As mentioned earlier, a PWM Θ is often used to characterize the nucleotide frequencies at each position of a binding site, where $\Theta = (\theta_1, \ldots, \theta_J)$ 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 jth position of a binding site, such that $\theta_{a,j} + \theta_{c,j} + \theta_{g,j} + \theta_{t,j} = 1$ for each j, $1 \le j \le J$. In general, Θ is assumed to follow a *product Dirichlet* distribution (Liu, 1994; Liu *et al*., 1995). 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 $\mathrm{Dir}(1, 1, 1, 1)$.

A PWM can be estimated from a collection of DNA sequences $\mathcal{R} = (R_1, \ldots, R_n)$ that correspond to *aligned* binding sites of a TF, where $R_i = (r_{i1} r_{i2} \cdots r_{iJ})$ represents the ith binding site, for each $i = 1, \ldots, n$, and r_{ij} is one of the nucleotides A, C, G, and T, for each $j = 1, \ldots, J$. These binding sites are assumed (Liu, 1994; Liu *et al*., 1995) to be independent random 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,

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

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

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

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 $\mathcal{R}$. 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 $B = (b_1 \ldots b_J)$ 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 sequences and expression

We propose a new approach to learning PWMs through the combination of both sequence and expression data. Let $\mathcal{E} = (E_1, \ldots, E_n)$ denote the fold changes of mRNA expression of downstream genes, where E_i is associated to the binding site R_i. [2] 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.

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 finding (Bussemaker *et al*., 2001; Conlon *et al*., 2003), most of which employ the third strategy that we discussed earlier in Supplementary 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,

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

or, for short, $\log \mathcal{E} \propto \pi(\mathcal{R}|\Theta)$. Therefore, we can reduce the maximum likelihood problem 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 (Leung *et al*., 2005; Hong *et al*., 2005). 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 itself (Liu, 1994).

2.3 A sequence weighting scheme

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,

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

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

$$\tilde{c}_{a,j} = \sum_{i=1}^{n} \delta(r_{ij}, \mathrm{A}) \cdot \log E_i, \quad \text{where} \quad \delta(r_{ij}, \mathrm{A}) = \begin{cases} 1, & \text{if } r_{ij} = \mathrm{A} \\ 0, & \text{otherwise} \end{cases}$$

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 motif (Liu *et al*., 2002). It follows that the desired PWM will be

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

Similarly, the conditional predictive distribution of a DNA sequence $B = (b_1 \ldots b_J)$ 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 learning PWMs is indeed done via a sequence weighting scheme; that is, every binding site contributes to the estimated PWM proportionally to the logarithm fold change of mRNA expression of its downstream gene. Note that $\pi(B|\Theta, \mathcal{E})$ would be completely equal to $\pi(B|\Theta)$ if every binding site induces the same fold change in expression. A simple example in Figure 1 clearly demonstrates the advantage of our approach to learning PWMs from both sequence and expression data.

The use of fold changes in expression as weights to estimate PWMs implicitly assumes that DNA sequences of motif elements exhibiting higher fold changes are more similar to the motif consensus pattern. This is plausible

[2] Note that, multiple binding sites may share the same downstream gene and thus its associated log fold change value.

(a)

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

(b)

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

(c)

	1	2	3	4	5
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. 1. 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 2.4 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. It should be noted that sequence weighting does not always lead to a PWM of higher information content. For example, if expression log ratios are (1, 2, 3, 4) instead of (4, 3, 2, 1) in the above example, then sequence weighting will find a PWM having lower information content (1.43 bits).

since such motif elements are more likely to represents a true motif. Moreover, since the binding energy of a TF protein to a site can be approximated as the sum of pairwise contact energy between the individual nucleotides and the protein (Djordjevic *et al*., 2003), different binding sites may indeed have different affinities for their cognate TFs. In evolution, there is not only selection force for TF binding sites to remain recognized by their TFs, but also selection force for preserving the binding strength of sites (Siddharthan *et al*., 2005), especially those inducing dramatic fold changes in expression.

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 (Liu, 1994; Liu *et al*., 1995). 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. In particular, we have incorporated this method of computing PWMs into a powerful Gibbs sampling program, AlignACE (Hughes *et al*., 2000; Roth *et al*., 1998). The modified program is called W-AlignACE, and has been implemented as a web server available to the public for free (see Availability). It is easy to see that the extra running time caused by sequence weighting is negligible.

2.4 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 as the relative entropy (*i.e.*, Kullback-Leibler distance) of binding sites with respect to the background base frequencies. However, it is well known that a highly conserved motif may not be statistically significant relative to the expectation for its random occurrences in the promoter sequences under consideration.

The MAP score is the metric for motif strength used by AlignACE to judge different motifs sampled during the course of the algorithm (Hughes *et al*., 2000). 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 REDUCE (Bussemaker *et al*., 2001) and also in MotifRegressor (Conlon *et al*., 2003) 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.

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

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 database (Matys *et al*., 2003). 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, it is very difficult to learn a PWM precisely since different learning methods usually produce different PWMs (see Figure 1). 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.*, ignore the weak region of the consensus pattern). It can be seen that such a comparison is not sensitive to either spurious 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

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 (Liu *et al*., 2002; Chen and Jiang, 2006), in the hope to explore in depth how a PWM learned from sequence and expression effects the performance of motif finding algorithms. Due to the page limit, the data generating procedure is outlined in the Supplementary Materials.

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 is assigned an expression value using a hyperbolic tangent function as in (Barash *et al*., 2001; Hong *et al*., 2005). 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.

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

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

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. MDscan and MotifRegressor, instead, are run only once for each dataset because they are deterministic algorithms (*i.e.*, no random seed required). Predicted motifs are sorted using their respective sorting schemes (*e.g.*, the MAP score for AlignACE), and only the top four are reported in each run 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 Database (SGD; Cherry *et al*., 1997) (see Supplementary Table 1 for the list of motif consensi), and compare it with the motifs found by MDscan, MotifRegressor, AlignACE, and W-AlignACE, respectively. [3] In our experiments, no prior knowledge on true motifs is assumed. Therefore, all the program parameters are set to their default values. [4] For instance, the default 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. Note that the evaluation method proposed in (Tompa *et al*., 2005) is not applicable here because W-AlignACE requires gene expression data in addition to promoter sequences.

3.2.1 mRNA expression data We have applied our algorithm to the publicly available dataset for yeast from the microarray experiments on environmental stress response (Gasch *et al*., 2000). 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. When this dataset is tested on on a workstation (3.2GHz CPU and 1GB RAM), both W-AlignACE and AlignACE take about 7 minutes each. In contrast, MDScan takes only 3 seconds and MotifRegressor 25 seconds.

YAP1 is a transcriptional activator required for oxidative stress tolerance, and is known to recognize the DNA sequence TTACTAA (Fernandes *et al*., 1997) or the sequence GCTTACTAA with higher binding specificity, as annotated [5] in SGD. 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 very weak pattern GATTAGTAAT ranked the 12th (MAP score: 10.09) in one run and GCTTAGTAAT ranked the 13th (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 difference 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 perform better than MDscan, but instead it reported GATTACTAAT as its second motif. These results give a solid example where W-AlignACE is more capable than AlignACE, MDscan, and MotifRegressor.

We also performed similar experiments on another two overexpression datasets concerning MSN2 and MSN4 (Gasch *et al*., 2000) and found that W-AlignACE was instead slightly outperformed by AlignACE in these two cases. This is mostly because the SGD-annotated motifs do not occur so frequently in the promoter regions of the most differently expressed genes. The detailed experimental results are discussed in the Supplementary Materials.

Source	Consensus	Rank
Fernandes *et al*., 1997	TTACTAA	-
SGD annotation	GCTTACTAA	-
W-AlignACE	GCTTACTAAT	2
AlignACE	GATTAGTAAT	12
	GCTTAGTAAT	13
MDscan	GATTACTAAT	1
MotifRegressor	GATTACTAAT	2

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.

3.2.2 ChIP-chip data We further apply our algorithm to the ChIP-chip data reported in (Lee *et al*., 2002). 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. 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 of (Hong *et al*., 2005), 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 gene, its promoter sequence is taken up to 800 bps upstream, but not overlapping with the previous gene.

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), because these three TFs were observed in (Hong *et al*., 2005) to be among the nine TFs (the other six are GAT3, GCR2, IME4, IXR1, PHO4, and ROX1) whose correct motifs are hard to find. Further 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-AlighACE are not evaluated here, and could still be true motifs.

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 nCGTnnnnAGTGAT for ABF1. Its MAP score is 351.866 as estimated by AlignACE, much lower than 436.877 by W-AlignACE (see Supplementary Table 2). In contrast, both program also reported an obviously spurious motif in the top four,

[3] Some 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.

[4] MotifRegressor 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.

[5] The annotated consensus is indeed GCTKACTAA using IUPAC ambiguity codes, where K represents the base G or T.

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
GCN4	56	AATGACTCAT	GATGAGTCAC	GGATGAGTCA	42.5719	GnATGAGTCAn	187.854
HAP4	42					CnnGnnnnTGATTGGnnC	62.6472
HSF1	34	TTTTCTAGAA		GAAnnTTCnAGAA	50.569	GAnnnTTCnAGAA	88.2247
MBP1	74	CGCGACGCGT		AAnAAACGCGT	36.9147	AnnAAACGCGTC	103.034
MCM1	59	CCTAATTAGG		TTnCCnnnTnnGGAAA	129.158	nTnCCnnAnnnGGAAA	179.82
NDD1	67	CCTAAATAGG		TTTCCnAAAnnGG	50.7552	CCnAAnnnGGnAAAnnnT	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			GCTTACTAAT	24.5596	ATTAGTAAGC	52.1866
YAP5	55						

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. Five of 40 TFs (GAT3, GCN4, IME4, IXR1, YAP6) do not have motif consensi annotated in the Saccharomyces Genome Database, and thus are not listed here.

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 finding correct motifs, but also enhances our confidence about the predicted motifs. This is further demonstrated by the following case studies.

ACE2 is a TF that activates the transcription of genes expressed in the G1 phase of the cell cycle (Dohrmann *et al*., 1992). 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 in two runs and among the top four in all runs. The highest MAP score estimated is 127.571 (see Supplementary Table 2). AlignACE did report the ACE2 motif in one of its runs but with a very low ranking of only 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 2 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 (Ren *et al*., 2000). In our previous study (Chen and Jiang, 2006), we incorporated the sequence weighting scheme into the basic Gibbs sampling algorithm from (Lawrence *et al*., 1993), 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 (Ren *et al*., 2000), which contains only 10 target genes. The current dataset from (Hong *et al*., 2005) contains 25 target genes. When run on this larger dataset, our previous algorithm (Chen and Jiang, 2006) 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 (Hong *et al*., 2005), partly because there is a 11-base gap (*i.e.*, degenerate region) in the middle of its consensus pattern, *i.e.* CGGnnnnnnnnnnnCCG. 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 2). Furthermore, these six genes are all among the top if we sort all genes in the dataset by

their ChIP-chip scores. [6] 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 pheromones (Ren *et al*., 2000). The ChIP-chip dataset from (Hong *et al*., 2005) 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 study (Chen and Jiang, 2006). 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 2. In conclusion, the sequence weighting scheme that learns PWMs from both sequence and ChIP 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 [7]. This could be due to several factors. First, the default cutoff used by MotifRegressor on the significance of linear regression might be too strict for these datasets. Second, the true motifs are too weak as evaluated by MotifRegressor based on the significance 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 parameter values that we set for MotifRegressor might not be optimal either, although their default values are preferred (see the first paragraph of Section 3.2).

4 DISCUSSION AND FUTURE RESEARCH

Learning an accurate PWM to characterize the binding sites of a TF is a fundamental problem in regulatory genomics, as it plays an important role in modeling regulatory motifs and also in discovering the regulatory targets of TFs. The commonly used learning approach relies on an implicit assumption that has never been questioned before. That is, all the putative binding sites contribute equally to the estimation of the motif PWM, regardless of the expression levels of downstream genes that they regulate. However, it is well known that expression levels (and fold changes) vary in a large range even among co-regulated genes, which perhaps suggests that the above assumption might not be fair, since the variations could be due to the sequence differences of their binding sites. Therefore, simply equating every binding site could result in an inaccurate PWM or a PWM with binding specificity so low that it would likely fail to differentiate true binding sites from spurious ones.

Another delicate problem concerns the selection of a set of (co-regulated) promoter sequences for motif finding. An ideal set shall contain promoter sequences that are (i) all bound by a common TF, and (ii) as many as possible. In practice, a pre-defined cutoff is often applied to expression fold changes (in particular from a single microarray experiment) or ChIP-chip scores to determine a set of promoter sequences that seem to be co-regulated. Note that a stringent cutoff will likely lead to a small set of promoter sequences and hence increases the chance of spurious motifs being statistically significant, while a relaxed cutoff might result in many promoter sequences that actually are not bounded by the TF of interest. Both cases might prevent true motifs from being successfully detected. As an example, if we increase the ChIP-chip score cutoff to reduce the size of the GAL4 dataset given in Table 3 by 50%, then AlignACE would be able to detect the correct motif as W-AlignACE is. Therefore, how to choose an appropriate cutoff value is also a nontrivial factor to the success of finding true motifs.

In this paper, to address the above two problems in the framework of learning PWMs, we formulate a maximum likelihood problem where the optimal PWM maximizes the likelihood of observing the combination of DNA binding sequence and expression data. A sequence weighting scheme is then proposed to offer an approximate while reasonably good solution to the maximum likelihood problem. The new learning approach via sequence weighting can be justified partially by the observation that binding sites inducing drastic fold changes in expression (or showing strong binding ratios in ChIP-chip experiments) are more likely to represent the true motif (Liu *et al*., 2002). Furthermore, it is easy to see that the new learning method reduces sensitivity to the fold change cutoff since sequences that have smaller fold changes have less impact on the construction of the PWMs.

It should be noticed that several computational methods have been developed to find motifs by taking gene expression variation into account (Bussemaker *et al*., 2001; Conlon *et al*., 2003; Keles *et al*., 2002; Liu *et al*., 2002; Segal *et al*., 2003; Wang *et al*., 2005), but all in a way different from ours. For example, MDscan (Liu *et al*., 2002) divides promoter sequences into two groups — a highly expressed group and a less expressed group, and treats the sequences from one group equally when estimating the motif. REDUCE (Bussemaker *et al*., 2001) and MotifRegressor (Conlon *et al*., 2003) use gene expression variation to select statistically significant PWMs via linear regression. Segal *et al*., 2002 and Wang *et al*., 2005 build a joint probabilistic model for promoter sequence and gene expression data, and then estimate PWMs via an expectation maximization algorithm. Here, we address a completely different and specific question; that is, how to learn a more accurate PWM from a set of aligned binding sequences in the presence of their associated expression data. More accurate PWMs can substantially enhance the capability of motif finding algorithms (*e.g.*, Gibbs sampling) to discover true but weak motifs, as demonstrated in our large scale tests of the program W-AlignACE on both simulated and real data. We conjecture that the idea can be applied to other programs based on PWMs.

We note in passing that there exist computational methods to find motifs in the promoter regions of genes that exhibit similar expression patterns across a variety of experimental conditions (Bussemaker *et al*., 2001). Here, our proposed method focuses on a single experimental condition (relative to a control condition). Previous studies (Keles *et al*., 2002) showed that focusing on a single experimental condition is crucial for identifying experiment-specific regulatory motifs. One reason for this is that averaging across experiments may destroy the significant relationship between the expression of genes and their regulatory motifs present only in a single experiment.

In conclusion, learning an accurate PWM from a collection of aligned binding sequences is a delicate problem that plays an important role in modeling a TF's binding preferences. In this paper, we tackled this problem by proposing a new approach to learning PWMs jointly from sequence and expression data. 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 gene expression or ChIP-chip data are available and correlated with the occurrences of the true motif in the promoter regions of the genes under study. On the other hand, the expression (or binding) information might misguide W-AlignACE when the true motif does not occur in the most differentially expressed or strongly bond genes, as discussed in the Supplementary Materials. 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 the discovery of cooperative motifs (or *cis*-regulatory modules).

[6] 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.

[7] More precisely, the current implementation of MotifRegressor uses MDmodule, instead of MDscan, as a feature extraction tool. MDmodule is a modified version of MDscan.

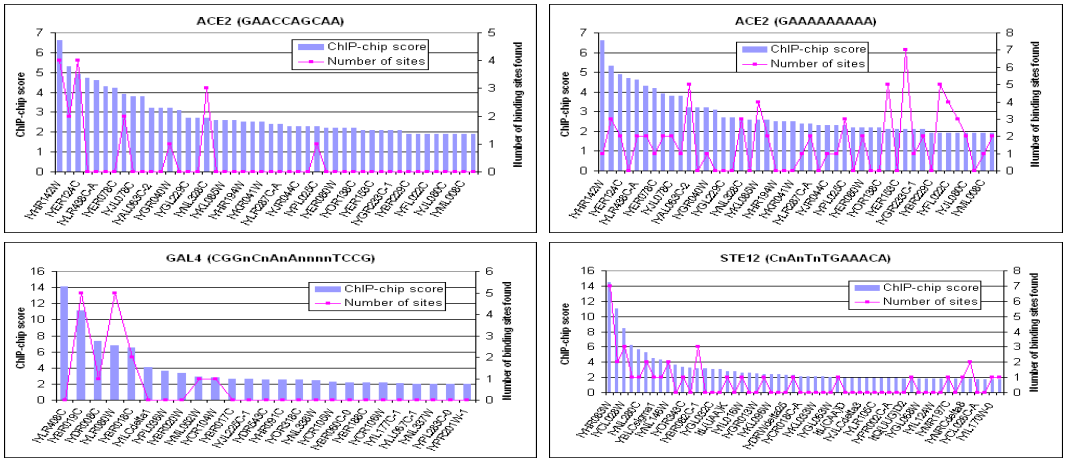

Fig. 2. The distributions of ChIP-chip scores and occurrences of binding sites of three TFs ACE2, GAL4 and STE12. The top right figure depicts the distribution for a spurious motif ranked the first by AlignACE with MAP score 104.81, and the other three figures correspond to three correct motifs all ranked the first by W-AlignACE with MAP scores, 127.571, 184.307, and 358.174, respectively. 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.

ACKNOWLEDGEMENT

XC's research is supported by an AcRF-SUG grant from Singapore Ministry of Education and TJ's research is supported by NSF grant IIS-0711129, NIH grant LM008991-01, NSFC grant 60528001, and a Changjiang Visiting Professorship at Tsinghua University. The authors would like to thank X. Shirley Liu for her valuable suggestions, and the anonymous reviewers for their helpful comments.

REFERENCES

Barash,Y. *et al*. (2001) A simple hyper-geometric approach for discovering putative transcription factor binding sites. *Algorithms in Bioinformatics: Proc. First International Workshop*, 278-293.

Bailey,T. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. ISMB*, **2**, 28-36.

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

Cherry,J. *et al*. (1997) Genetic and physical maps of Saccharomyces cerevisiae. *Nature*, **387**, 67-73.

Chen,X. and Jiang,T. An improved Gibbs sampling method for motif discovery via sequence weighting. *Proc. of Comput. Syst. Bioinfo.*, 239-247.

Conlon,E. *et al*. (2003) Integrating regulatory motif discovery and genome-wide expression analysis. *PNAS*, **100**, 3339-3344.

Dohrmann,P. *et al*. (1992) Parallel pathways of gene regulation: homologous regulators SWI5 and ACE2 differentially control transcription of HO and chitinase. Genes Dev. **6**, 93-104.

Djordjevic,M. *et al*. (2003) A biophysical approach to TF binding site discovery. *Genome Res.*, **13**, 2381-2390, 2003.

Fernandes,L. *et al*. (1997) Yap, a novel family of eight bZIP proteins in Saccharomyces cerevisiae with distinct biological functions. *Mol. Cell. Biol.*, **17**, 6982-6993.

Gasch,A. *et al*. (2000) Genomic expression programs in the Response of Yeast cells to environmental changes. *Mol. Biol. Cell*, **11**, 4241-4257.

Hughes,J. *et al*. (2000) Computational identification of *Cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.*, **296**, 1205-1214.

Hong,P. *et al*. (2005) A boosting approach for motif modeling using ChIP-chip data. *Bioinformatics*, **21**, 2636-2643.

Keles,S. *et al*. (2002) Identification of regulatory elements using a feature selection method. *Bioinformatics*, **18**, 1167-1175.

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

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

Leung,H. *et al*. (2005) Finding motifs with insufficient number of strong binding sites. *Journal of Computational Biology*, **12**, 686-701.

Liu,J. (1994) The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*, **89**, 958-966.

Liu,J. *et al*. (1995) Bayesian models for multiple local sequence alignment and Gibbs sampling strategies, *Journal of the American statistical Association*, 90, 1156-1170.

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

Matys,V. *et al*. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res*, **31**, 374-378.

Neuwald,A. *et al*. (1995) Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci.*, **4**, 1618-1632.

Roth,F. *et al*. (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotech.*, **16**, 939-945.

Ren,B. *et al.* (2000) Genome-wide location and function of DNA binding proteins. *Science*, **290**, 2306-2309.

Segal,E. *et al*. (2002) From sequence to expression: a probabilistic framework. *RECOMB*, 263-272.

Sinha,S. *et al*. (2004) PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinformatics*, **5**, 170, 2004.

Stormo,G. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16-23, 2000.

Siddharthan,R. *et al*. (2005) PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput. Biol.*, **1**, e67, 0534-0555.

Segal,E. *et al*. (2003) Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. *Bioinformatics*, **19**, i273-i282.

Tompa,M. *et al*. (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotechnology*, **23**, 2005.

Wang,W. *et al*. (2005) Inference of combinatorial regulation in yeast transcriptional networks: a case study of sporulation. *PNAS*, **102**, 1998-2003.