AN IMPROVED GIBBS SAMPLING METHOD FOR MOTIF DISCOVERY VIA SEQUENCE WEIGHTING

Xin Chen*
School of Physical and Mathematical Sciences
Nanyang Technological University, Singapore
**Email: chenxin@ntu.edu.sg*

Tao Jiang
Department of Computer Science and Engineering
University of California at Riverside, USA
Currently visiting at Tsinghua University, Beijing, China
Email: jiang@cs.ucr.edu

The discovery of motifs in DNA sequences remains a fundamental and challenging problem in computational molecular biology and regulatory genomics, although a large number of computational methods have been proposed in the past decade. Among these methods, the Gibbs sampling strategy has shown great promise and is routinely used for finding regulatory motif elements in the promoter regions of co-expressed genes. In this paper, we present an enhancement to the Gibbs sampling method when the expression data of the concerned genes is given. A sequence weighting scheme is proposed by explicitly taking gene expression variation into account in Gibbs sampling. That is, every putative motif element is assigned a weight proportional to the fold change in the expression level of its downstream gene under a single experimental condition, and a position specific scoring matrix (PSSM) is estimated from these weighted putative motif elements. Such an estimated PSSM might represent a more accurate motif model since motif elements with dramatic fold changes in gene expression are more likely to represent true motifs. This weighted Gibbs sampling method has been implemented and successfully tested on both simulated and biological sequence data. Our experimental results demonstrate that the use of sequence weighting has a profound impact on the performance of a Gibbs motif sampling algorithm.

1. INTRODUCTION

Discovering motifs in DNA sequences remains a fundamental and challenging problem in computational molecular biology and regulatory genomics [19], although a large number of computational methods have been proposed in the past decade. The motif finding problem can be simply formalized as the problem of looking for short segments that are over-represented among a set of long DNA sequences. Previously proposed methods for finding motifs broadly fall into two categories: (a) deterministic combinatorial approaches based on word statistics [15, 14, 2, 6], and (b) probabilistic approaches based on local multiple sequence alignment [10, 1, 8, 11]. Typical methods in the first category search the promoter sequences of co-regulated genes for various sized motifs exhaustively, and then evaluate their significance by a statistical method, whereas methods in the second category rely on local search techniques such as expectation maximization and Gibbs sampling. The latter methods also usually represent a motif as a *position specific scoring matrix* (PSSM) (which is also commonly referred to as a position weight matrix).

Gibbs sampling has shown to be a very promising strategy for motif discovery. The original implementation of Gibbs sampling was done in the site sampling mode, which assumes that there is exactly one motif element (notably a transcript factor binding site) located in each (promoter) sequence. Since its first application to find conserved DNA motifs in the early 90's [10], quite a few improvements have been made in the literature to improve its effectiveness. These improvements include: (a) motif sampling allowing zero or multiple motif elements located in each sequence [13]; (b) incorporation of a higher-order Markov background model [18, 11]; (c) column sampling allowing gaps within a motif [13]; (d) incorporation of phylogeny information [17]; and so on. Be-

[*] Corresponding author.

sides these enhancements, we observe below two important aspects common to all previous implementations of Gibbs sampling (and also common to most other motif finding algorithms).

First, the promoter DNA sequences upstream of a collection of co-expressed genes are often taken as the input to a motif finding algorithm. This is because that co-expression is usually taken as an evidence of co-regulation, which is in turn assumed to be controlled by a common motif. Molecular biology has been revolutionized by *cDNA microarray* and *chromatin immunoprecipitation* (ChIP) techniques, which allow us to simultaneously monitor the mRNA expression levels of many genes under various conditions. With gene expression data in hand, one generally applies a selected threshold on fold changes in the expression level under a single experimental condition relative to some control condition in order to retrieve a set of co-expressed genes. Here, it naturally gives rise to a question: how to select a threshold properly such that the motif would be more easily and reliably found? Notice that motif elements with large fold changes in expression are in general more likely to represent a true motif [12]. However, the statistical significance (*e.g.*, p-values) of these motif elements may not increase as the threshold increases, because an increase in threshold may also simultaneously reduce the number of co-expressed genes. On the other hand, lowering the threshold may cluster more genes that are less likely co-regulated, and also decrease the statistical significance as well. This is a dilemma that was not addressed by any previous Gibbs sampling algorithm.

Second, a PSSM is commonly used to represent a probabilistic motif model by taking into account the base variation of motif elements at different positions. Specifically, given a PSSM of a motif, $\mathcal{Q}$, each component $q_{i,j}$ describes the probability of observing base j at position i in the motif. In all previous implementations of Gibbs sampling, a PSSM is estimated from a set of putative motif elements, which are sampled from the input promoter sequences, with components proportional to the observed frequencies of each base at different positions. This relies on an implicit assumption that has never been questioned before. That is, all motif elements, regardless of the downstream genes that they regulate, should contribute equally to the components of the motif PSSM. However, we know that expression levels (and fold changes) vary in a large range even among co-regulated genes, which could perhaps suggest that the above assumption might not be fair. In other words, equating every motif element could result in an inaccurate PSSM. Note, however, that an accurate motif model for a transcription factor is essential to differentiate its true binding sites from spurious ones.

In this paper, we address the above two problems together by one scheme referred to as *sequence weighting*. It is natural to assume that motif elements with dramatic fold changes in expression are more likely to represent a true motif. Therefore, we want to estimate a PSSM such that it can explicitly reflect such a (nonuniform) likelihood distribution over motif elements. One way to achieve this is to assign each motif element a weight, *e.g.*, proportional to the fold change in expression, and then to estimate each component q_{ij} of the PSSM as the *weighted* frequencies of base j at position i among all motif elements. One can see that a *weighted* PSSM favors putative motif elements showing large fold changes in expression. On the other hand, a putative motif element with small fold changes, which is less likely to represent the true motif, will not affect a weighted PSSM as much.

The use of fold changes in expression as weights to estimate PSSMs implicitly assumes that the DNA sequences of motif elements exhibiting higher fold changes are more similar to the motif consensus pattern. This is plausible since such motif elements are more likely to represent the true motif. Moreover, since the binding energy of a transcription factor (TF) protein to a DNA site can be approximated as the sum of pairwise contact energy between the individual nucleotides and the protein [20], different binding sites may indeed have different affinities for their cognate transcription factors. In evolution, there is not only selection force for TF binding sites to remain recognized by their TFs, but that also selection force for preserving the strength of binding sites [17], especially those showing dramatic fold changes in expression.

We have incorporated the sequence weighting scheme into the Gibbs sampling algorithm originally developed in [10, 7]. In real applications on a set of co-expressed genes, we can assign each input promoter sequence a weight proportional to the fold change in gene expression obtained from a cDNA microarray experiment, or proportional to the so-called *binding ratio* determined by a *genome-wide location analysis*, which is a popular approach that combines a modified chromatin immunoprecipitation procedure with DNA microarray analysis for studying genome-wide protein-DNA interactions and transcription regulation [16]. In a genome-wide location analysis, a binding ratio is calculated by taking the average of fold changes in expression over three independent microarray experiments.

Our implementation of Gibbs sampling via sequence weighting has been successfully tested on both simulated and real biological sequence data. We considered two sets of genes regulated by the transcriptional activator Gal4 and Ste12 respectively, and their expression levels and binding ratios were determined by the genome-wide location analysis [16]. The test results show that the use of sequence weighting has a profound impact on the performance of the Gibbs motif sampling algorithm.

The rest of the paper is organized as follows. The next section introduces the basic Gibbs sampling algorithm and the proposed sequence weighting scheme. Preliminary experiments on simulated data and real data are presented in Section 3. Section 4 gives some concluding remarks.

2. GIBBS SAMPLING THROUGH SEQUENCE WEIGHTING

In this section, we start with the description of the basic Gibbs sampling algorithm, and then introduce the new method of estimating position specific scoring matrices (PSSMs) via sequence weighting.

2.1. The Motif Model

A DNA motif is usually represented by a set of short sequences that are all binding sites of some transcription factor protein. Due to base variation at binding sites, the pattern of a DNA motif is conveniently described by a probabilistic model of base frequencies at each position, for which a common mathematical representation is a so-called position specific scoring matrix (PSSM). A PSSM $\mathcal{Q}$ consists of entries $q_{i,j}$, which give the probabilities of observing base j at position i of a binding site. The main assumption underlying this motif model is that the bases occurring at different positions of a DNA motif are probabilistically independent.

Assume that we are given a set of N binding sites, $s_1, s_2, \ldots, s_N$, of width W each. Let $J = 4$ be the number of bases in the alphabet $\{A, C, G, T\}$, and $c_{i,j}$ be the observed count/frequency of base j at position i. A widely used method to estimate a PSSM from these binding sites is simply given by [a]

$$q_{i,j} = \frac{c_{i,j}}{c_i}$$

where c_i is the sum of $c_{i,j}$ over the alphabet; that is, $c_i = \sum_{j=1}^{J} c_{i,j}$.

With the PSSM $\mathcal{Q}$, we are able to estimate the probability $P(s|\mathcal{Q})$ of an arbitrary sequence s being generated by the motif model as

$$P(s|\mathcal{Q}) = \prod_{i=1}^{W} q_{i,s_i}$$

where s_i is the base of s at position i. On the other hand, a background sequence model $\mathcal{P}$ is estimated to depict the probabilities p_j of base j occurring in the background sequence. The probability of the sequence s being generated by $\mathcal{P}$ is given by

$$P(s|\mathcal{P}) = \prod_{i=1}^{W} p_{s_i}$$

Therefore, the likelihood that s is a true binding motif of interest under the motif mode $\mathcal{Q}$ versus the background model $\mathcal{P}$ is given by the formula

$$L(s \mid \mathcal{P}, \mathcal{Q}) = \frac{P(s|\mathcal{Q})}{P(s|\mathcal{P})}$$

The most useful quantity characterizing the quality of a PSSM $\mathcal{Q}$ is its information content I, defined as

$$I = \frac{1}{W} \times \sum_{i=1}^{W} \sum_{j=1}^{J} q_{ij} \log \frac{q_{ij}}{p_j}$$

[a] In an actual implementation, a "pseudocount" should be added to each $c_{i,j}$ in order to avoid a zero frequency for any base not actually observed.

where the logarithm is often taken with base 2 to express the information content in bits. The information content thus ranges from 0 to 2, reflecting the weakest to the strongest motifs.

2.2. Basic Gibbs Sampling Algorithm

The basic motif finding problem is, given a set of DNA sequences, $S_1, S_2, \ldots, S_N$, to look for short sequence segments of specified width W that are over-represented among the input sequences. In real biological applications, the input sequences (of a typical size of 800 bps) are usually taken from the upstream regions of co-expressed genes, W ranges from 5 bp to 16 bp, and the output segments are putative binding sites.

Gibbs sampling has proven to be a powerful strategy for finding weak DNA motifs. The most basic implementation of Gibbs sampling, known as a site sampler, assumes that there is exactly one binding site located in each input sequence. The details of this implementation are described in [10]. In Figure 1, we briefly summarize it in order to show how sequence weighting is incorporated into Gibbs sampling. Note that $S_k[i, i+W-1]$ denotes the substring of width W starting at i and ending at $i+W-1$ in sequence S_k.

2.3. Estimating the PSSM via Sequence Weighting

To start, we introduce two more notations in order to incorporate sequence weights into the computation of a PSSM. Given a set of N binding sites, s_1, s_2, …, s_N, of width W each, let w_k be the weight associated with the input sequence S_k, reflecting in some way the contribution of the sequence S_k to the PSSM as discussed above. The sequence weights can be normalized so that they sum up to N. We define a binary function $\delta(i, j, k)$ as

$$\delta(i,j,k) = \begin{cases} 1, \text{ if } s_k(i) = j \\ 0, \text{ otherwise} \end{cases}$$

where $s_k(i)$ is the base at position i of sequence k.

In order to incorporate sequence weights into the Gibbs Sampling algorithm, we propose to compute $c_{i,j}$ as the *weighted* count of base j at position i of the binding motif, *i.e.*

$$c_{i,j} = \sum_{k=1}^{N} w_k \delta(i,j,k)$$

Then, we estimate $q_{i,j}$ as before, but using the weighted counts $c_{i,j}$. That is,

$$c_i = \sum_{j=1}^{J} c_{i,j}$$

and

$$q_{i,j} = \frac{c_{i,j}}{c_i}, \quad 1 \le i \le W,\ 1 \le j \le J$$

One can easily see that the above is a natural extension of the original construction of PSSM where the weights for all sequences involved were assumed to be equal.

We have implemented the above sequence weighting scheme into the Gibbs motif sampler software developed in [10, 7]. Only the necessary parts of the source code have been modified so that we could make a fair comparison between the original Gibbs sampler and this modified version. It is easy to see that the extra running time caused by sequence weighting is negligible.

3. EXPERIMENTAL RESULTS

In order to test the performance of the above weighted Gibbs sampler, we have applied it to both simulated and real sequence data, and compared its results with the original Gibbs sampler [10, 7]. The simulated data sets allow us to compare the performance of the algorithms in an idealized situation that does not involve the complexities of real data. For our tests on real data, we use two sets of genes in *Saccharomyces cerevisiae* (yeast) that were determined by ChIP-array experiments [16] to be co-regulated by two proteins Ste12 and Gal4, respectively.

3.1. Simulated Data

In our simulation studies, a motif model was created as follows. First, 20 short DNA sequences (of width W) were randomly generated for binding sites of a common transcription factor with varying degrees of conservation. The seed transcription factor binding site is described by a consensus pattern.

```
Input: A set of DNA sequences $S_1$, $S_2$,..., $S_N$ and the motif width $W$
Output: The starting position $a_k$ of the motif in each sequence $S_k$;
    A PSSM $\mathcal{Q} = [q_{i,j}]$ for the putative motif model
begin
  Initialization:
    Randomly select a position $a_k$ for the motif in each sequence $S_k$
    Estimate the background base frequencies $p_j$, for $j$ from 1 to $J$, to obtain $\mathcal{P}$
  Repeat until convergence:
    Predictive update step:
      Randomly select a sequence $S_z$ from the input sequences
      Take the set of putative binding sites $\{S_k[a_k, a_k + W - 1] \mid 1 \leq k \leq N, k \neq z\}$
      Estimate the PSSM $\mathcal{Q}$ from $\{S_k[a_k, a_k + W - 1] \mid 1 \leq k \leq N, k \neq z\}$
    Sampling step:
      Estimate $P(S_z[n, n + W - 1] \mid \mathcal{Q})$ for every position $n$ in sequence $S_z$
      Estimate $P(S_z[n, n + W - 1] \mid \mathcal{P})$ for every position $n$ in sequence $S_z$
      Randomly select a new position $a_z$ in $S_z$ according to $L(S_z[n, n + W - 1] \mid \mathcal{P}, \mathcal{Q})$
end
```

Fig. 1. The basic Gibbs sampling algorithm.

The degree of conservation is measured by the Hamming distances to the consensus pattern, and the weakest binding sites have one half of their bases different from those at the corresponding positions of the consensus. Second, a set of 20 promoter sequences of 800 bases long were randomly generated, each with a binding site implanted at a randomly selected position. Finally, each promoter sequence was assigned a weight as the degree of conservation of the implanted binding site, based on the observation that binding sites with dramatic fold changes in gene expression are more likely to represent true motifs [12]. In the test, we chose five different motif widths ($W = 8, 10, 12, 14, 16$), reflecting different levels of difficulty for motif finding. For each motif width, 100 test data sets were generated, giving rise to a total of 500 data sets.

Both the original Gibbs sampler and the weighted version were applied to search for motifs in each data set, and the top three motifs were reported from each program. A found motif is considered correct if its consensus sequence differs from the the planted motif consensus pattern by at most two bases. We are interested in the number of times each program successfully detects the motif inserted in the 100 tests for each motif width, and the average rank of the correct motif if it comes up in the top three. The results are summarized in Table 1. Each program was run twice on each test data set with the option of column sampling [13] turned on or turned off, respectively.

We can see that the weighted Gibbs sampling method was able to find more correct motifs than the original Gibbs motif sampler in all the tests. It is particularly promising in the discovery of weak (*i.e.*, short) motifs, as twice as many correct motifs were found by the weighted method when $W = 8$ (without column sampling) than by the original Gibbs sampling method.

3.2. Real Biological Data

3.2.1. *Ste12*

The transcription activator Ste12 is a DNA-bound protein that directly controls the expression of genes in response of haploid yeast to mating pheromones [16]. We will use it to demonstrate how the sequence weighting scheme could boost the prediction accuracy of the Gibbs sampling method.

The genome-wide location analysis is a promising approach to monitor protein-DNA interactions across a whole genome [16]. It combines a modified chromatin immunoprecipitation (ChIP) procedure with DNA microarray analysis in order to provide a relative binding of the protein of interest to a DNA sequence. Such an analysis on epitope-tagged

Table 1. Simulation results on 20 sequences of 800 bases

Tests	The original Gibbs motif sampler [7, 10]				Gibbs sampling via sequence weighting			
	with column sampling		without column sampling		with column sampling		without column sampling	
	Times found	Avg. rank	Times found	Avg. rank	Times found	Avg. rank	Times found	Avg. rank
$W = 8$	22	2.50	23	1.91	31	1.61	49	1.90
$W = 10$	35	2.03	50	2.00	58	1.97	73	1.97
$W = 12$	53	1.91	73	1.97	77	1.92	88	2.08
$W = 14$	74	1.89	84	1.96	80	1.86	95	1.96
$W = 16$	70	2.00	92	1.85	88	1.76	95	2.00

Ste12 has determined that 29 pheromone-induced genes in yeast are likely to be directly regulated by Ste12 [16]. Figure 2 lists these genes and their binding ratios extracted from [16]. Note that, in the figure, names with all capital letters, such as STE12, are used to represent genes, and names that begin with a capital letter, such as Ste12, represent DNA binding motifs.

Of great interest is to find the sites bound by Ste12 in the promoter sequences of these 29 genes. For this purpose, we extracted up to 800 bp upstream regions of each gene from the *Saccharomyces* genome database, and assigned each sequence a weight as the relative binding ratio obtained from the genome-wide location analysis. Both the original Gibbs sampling algorithm and our weighted version were run on all 29 sequences, and their experimental results were then compared.

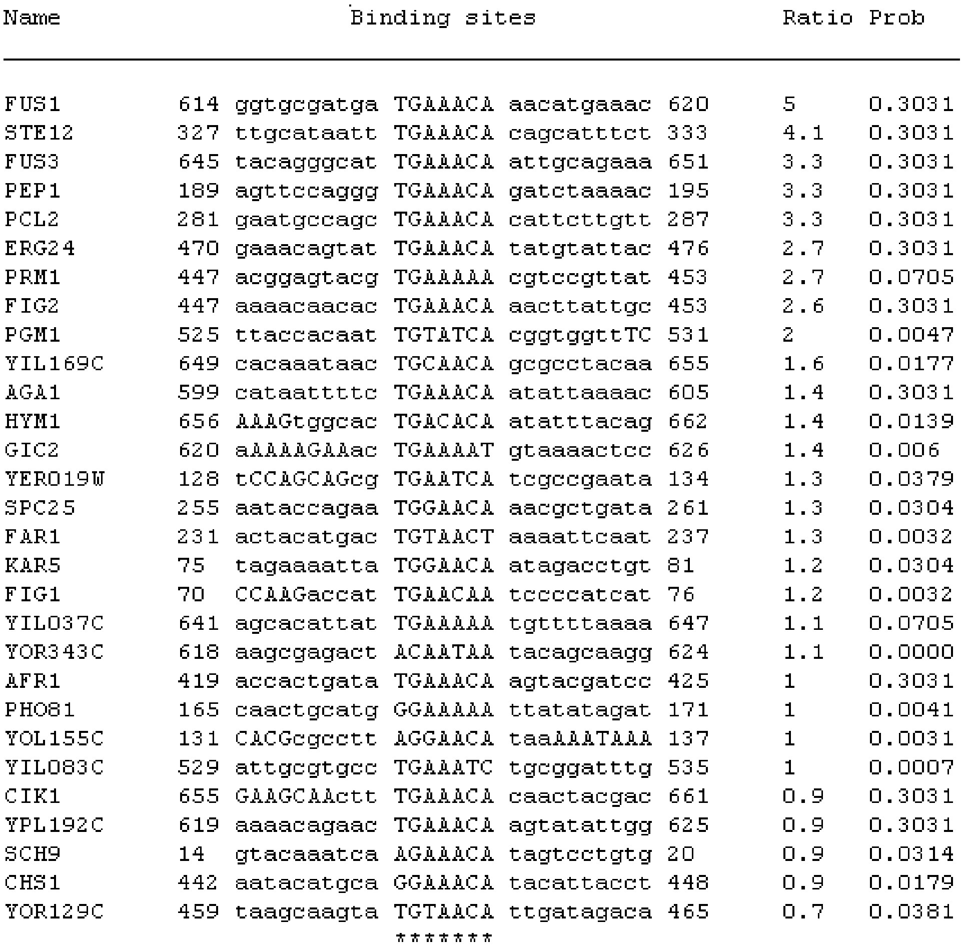
Due to the stochastic nature of Gibbs sampling, we ran both programs 10 times with different random seeds, and each time the top ten putative motifs were reported. One can see that the same motif might be reported in different runs. Of the 100 putative motifs, the original Gibbs sampling algorithm did not find any motif resembling the known Ste12 consensus pattern TGAAACA [5]. Our algorithm, however, found the correct Ste12 motif six times in 10 runs, and ranked the Ste12 motif the second among all the putative motifs in terms of information content. Figure 2 lists the putative binding sites found by the weighted Gibbs sampling method upstream of the 29 genes regulated by Ste12, and Figure 3 shows the corresponding weighted PSSM. The information content of this PSSM is 1.09, indicating a very strong motif that has been detected by our algorithm.

The above experimental results are very encouraging, but not surprising to us. One can see from Figure 2 that, roughly speaking, the higher a relative binding ratio is, the closer the concerned binding site is to the known motif consensus pattern (in terms of sequence Hamming distance). In particular, each of the top six sequences in the table contains a binding site exactly matching the Ste12 motif consensus pattern. Once some of these binding sites have been selected by chance, they are strongly favored in the construction of the (weighted) PSSM due to their large weights. This process tends to recruit more correct sites, which in turn further improve the specificity of the PSSM.

3.2.2. *Gal4*

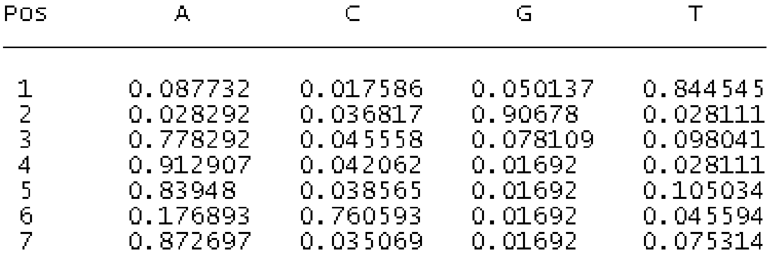
Gal4 is among the most characterized transcriptional activators, which activates genes necessary for galactose metabolism [16]. It provides another test showing that sequence weighting really improves the performance of the original Gibbs sampling algorithm. The genome-wide location analysis [16] found 10 genes to be regulated by Gal4 and induced in Galactose, with varying relative binding ratios (see Figure 4).

We performed the same experiment on Gal4 as we did on Ste12. Of the 100 putative motifs, the original Gibbs sampling algorithm once again failed to find any motif similar to the known Gal4 consensus pattern $\mathrm{CGGN}_{11}\mathrm{CCG}$. This result was unexpected by us because the binding sites of Gal4 are actually well conserved among the input sequences (shown below). With sequence weighting, however, our algorithm successfully discovered the exact Gal4 motif with the highest information content (1.80) among 100 putative motifs. These putative binding sites are listed in Figure 4, while the weighted PSSM is given in Figure 5. This clearly shows the advantage of sequence weighting that we have implemented in the Gibbs sampling algorithm, although we suspect that the algorithms might have found or missed the Gal4 motif by chance due to its very low statistical significance and the stochastic nature of Gibbs sampling.



Name	Binding sites	Ratio	Prob
FUS1	614 ggtgcgatga TGAAACA aacatgaaac 620	5	0.3031
STE12	327 ttgcataatt TGAAACA cagcatttct 333	4.1	0.3031
FUS3	645 tacagggcat TGAAACA attgcagaaa 651	3.3	0.3031
PEP1	189 agttccaggg TGAAACA gatctaaaac 195	3.3	0.3031
PCL2	281 gaatgccagc TGAAACA cattcttgtt 287	3.3	0.3031
ERG24	470 gaaacagtat TGAAACA tatgtattac 476	2.7	0.3031
PRM1	447 acggagtacg TGAAAAA cgtccgttat 453	2.7	0.0705
FIG2	447 aaaacaacac TGAAACA aacttattgc 453	2.6	0.3031
PGM1	525 ttaccacaat TGTATCA cggtggttTC 531	2	0.0047
YIL169C	649 cacaaataac TGCAACA gcgcctacaa 655	1.6	0.0177
AGA1	599 cataattttc TGAAACA atattaaaac 605	1.4	0.3031
HYM1	656 AAAGtggcac TGACACA atatttacag 662	1.4	0.0139
GIC2	620 aAAAAGAAac TGAAAAT gtaaaactcc 626	1.4	0.006
YER019W	128 tCCAGCAGcg TGAATCA tcgccgaata 134	1.3	0.0379
SPC25	255 aataccagaa TGGAACA aacgctgata 261	1.3	0.0304
FAR1	231 actacatgac TGTAACT aaaattcaat 237	1.3	0.0032
KAR5	75 tagaaaatta TGGAACA atagacctgt 81	1.2	0.0304
FIG1	70 CCAAGaccat TGAACAA tccccatcat 76	1.2	0.0032
YIL037C	641 agcacattat TGAAAAA tgttttaaaa 647	1.1	0.0705
YOR343C	618 aagcgagact ACAATAA tacagcaagg 624	1.1	0.0000
AFR1	419 accactgata TGAAACA agtacgatcc 425	1	0.3031
PHO81	165 caactgcatg GGAAAAA ttatatagat 171	1	0.0041
YOL155C	131 CACGcgcctt AGGAACA taaAAATAAA 137	1	0.0031
YIL083C	529 attgcgtgcc TGAAATC tgcggatttg 535	1	0.0007
CIK1	655 GAAGCAActt TGAAACA caactacgac 661	0.9	0.3031
YPL192C	619 aaaacagaac TGAAACA agtatattgg 625	0.9	0.3031
SCH9	14 gtacaaatca AGAAACA tagtcctgtg 20	0.9	0.0314
CHS1	442 aatacatgca GGAAACA tacattacct 448	0.9	0.0179
YOR129C	459 taagcaagta TGTAACA ttgatagaca 465	0.7	0.0381
	*******		

Fig. 2. Putative binding sites for the transcription activator Ste12 found by the weighted Gibbs sampling method. The column under Ratio lists the relative binding ratios obtained from the genome-wide location analysis, and the column under Prob lists the probabilities of the binding sites given by the motif model represented by the PSSM in Figure 3.



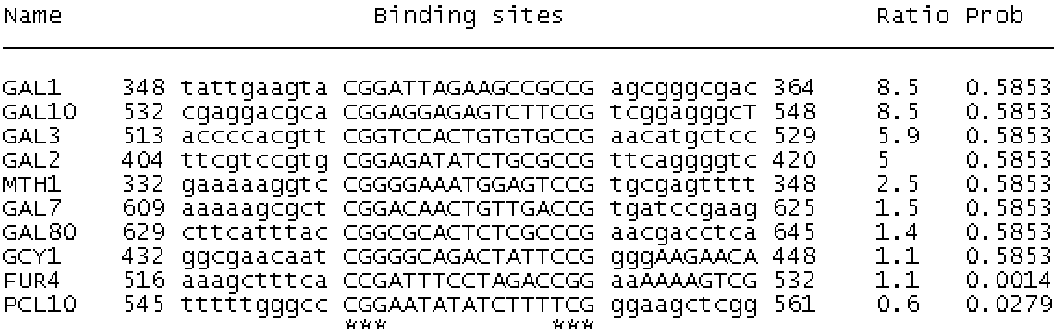
Pos	A	C	G	T
1	0.087732	0.017586	0.050137	0.844545
2	0.028292	0.036817	0.90678	0.028111
3	0.778292	0.045558	0.078109	0.098041
4	0.912907	0.042062	0.01692	0.028111
5	0.83948	0.038565	0.01692	0.105034
6	0.176893	0.760593	0.01692	0.045594
7	0.872697	0.035069	0.01692	0.075314

Fig. 3. The weighted PSSM calculated from the putative binding sites in Figure 2.

The complete results of both tests are available at `http://www.ntu.edu.sg/home/ChenXin/Gibbs`.

4. DISCUSSION AND FUTURE RESEARCH

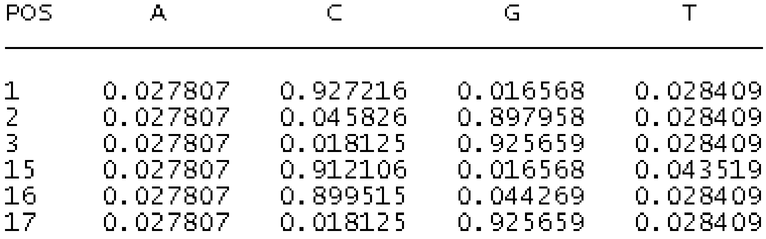
The selection of a suitable threshold value on expression level (or binding ratio) in order to retrieve a set of co-regulated genes and the construction of an accurate PSSM from a set of promoter sequences to



```
Name                    Binding sites                          Ratio Prob
--------------------------------------------------------------------------

GAL1     348 tattgaagta CGGATTAGAAGCCGCCG agcgggcgac 364    8.5   0.5853
GAL10    532 cgaggacgca CGGAGGAGAGTCTTCCG tcggagggcT 548    8.5   0.5853
GAL3     513 accccacgtt CGGTCCACTGTGTGCCG aacatgctcc 529    5.9   0.5853
GAL2     404 ttcgtccgtg CGGAGATATCTGCGCCG ttcaggggtc 420    5     0.5853
MTH1     332 gaaaaaggtc CGGGGAAATGGAGTCCG tgcgagtttt 348    2.5   0.5853
GAL7     609 aaaaagcgct CGGACAACTGTTGACCG tgatccgaag 625    1.5   0.5853
GAL80    629 cttcatttac CGGCGCACTCTCGCCCG aacgacctca 645    1.4   0.5853
GCY1     432 ggcgaacaat CGGGGCAGACTATTCCG gggAAGAACA 448    1.1   0.5853
FUR4     516 aaagctttca CCGATTTCCTAGACCGG aaAAAAGTCG 532    1.1   0.0014
PCL10    545 tttttgggcc CGGAATATATCTTTTCG ggaagctcgg 561    0.6   0.0279
                        ***           ***
```

Fig. 4. Putative binding sites for the transcription activator Gal4 found by the weighted Gibbs sampling. The relative binding ratios and probabilities of the binding sites are displayed as in Figure 3.



```
POS      A          C          G          T
------------------------------------------------

1     0.027807   0.927216   0.016568   0.028409
2     0.027807   0.045826   0.897958   0.028409
3     0.027807   0.018125   0.925659   0.028409
15    0.027807   0.912106   0.016568   0.043519
16    0.027807   0.899515   0.044269   0.028409
17    0.027807   0.018125   0.925659   0.028409
```

Fig. 5. The weighted PSSM calculated from the putative binding sites in Figure 4.

represent the true motif model are two delicate problems usually ignored by a Gibbs sampling strategy in motif discovery. In this paper, we try to tackle these problems by a sequence weighting scheme in order to improve the prediction accuracy of the basic Gibbs sampling algorithm.

Gibbs sampling via sequence weighting can be effectively applied to find motifs when the gene expression data is available. As we have noticed before, several computational methods that take advantage of gene expression variation have been developed [3, 12, 9, 4], but all differ from ours in various aspects. For example, MDscan [12] uses a word-enumeration strategy to exhaustively search for motifs, and is thus a deterministic combinatorial approach. Moreover, it needs a threshold value on expression level in order to extract highly expressed genes, and also treats all putative binding sites equally when representing a motif model, regardless of their expression variations. Our method does not require a preset threshold value.

On the other hand, many computational methods have been proposed to identify motifs in the promoter regions of genes that exhibit similar expression patterns across a variety of experimental conditions [3]. Here, our proposed method focuses on a single experimental condition (relative to a control condition). Previous studies [9] 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.

To summarize, we have proposed in this paper a sequence weighting scheme for enhancing the motif finding accuracy of the basic Gibbs sampling algorithm. It was achieved by estimating a PSSM from the promoter sequences weighted proportionally to the fold changes in the expression of their downstream genes. Our preliminary experiments on simulated and real biological data have clearly shown the advantage of this sequence weighting scheme in a Gibbs sampling. In the future, we would like to test this method on more real data sets with gene expression profiles and extend the method to gene expression data across multiple experimental conditions.

ACKNOWLEDGMENTS

Research supported in part by NSF grant CCR-0309902, NIH grant LM008991-01, NSFC grant 60528001, National Key Project for Basic Research (973) grant 2002CB512801, and a fellowship from the Center for Advanced Study, Tsinghua University.

References

1. 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.
2. H. Bussemaker, H. Li, and E. Siggia. Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis. *Proc. Natl. Acad. Sci. USA* , **97**, 10096-10100, 2001.
3. H. Bussemaker, H. Li, and E.D. Siggia. Regulatory element detection using correlation with expression. *nat. Genet.*, **27**, 167-171, 2001.
4. E. Conlon, X. Liu, J. Lieb, and J. Liu. Integrating regulatory motif discovery and genome-wide expression analysis. *PNAS*, **100**, 3339-3344, 2003.
5. 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.
6. M. Gupta and J. Liu. Discovery of conserved sequence patterns using a stochastic dictionary model. *J. Am. Stat. Assoc.*, **98**, 55-66, 2003.
7. `http://www.fas.harvard.edu/~junliu/Software/gibbs9_95.tar`
8. 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.
9. S. Keles, M. Laan, and M. Eisen. Identification of regulatory elements using a feature selection method. *Bioinformatics*, **18**, 1167-1175, 2002.
10. C. Lawrence, S. Altschul, M. Boguski, J. Liu, A. Neuwald, J. Wootton. Detcting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208-214, 1993.
11. X. Liu, d. Brutlag, and J. Liu. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac Symp Biocomput.*, 127-138, 2001.
12. 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.
13. A. Neuwald, J. Liu, and C. Lawrence. Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci.*, **4**, 1618-1632, 1995.
14. 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.
15. I. Rigoutsos and A. Floratos. Combinatorial pattern discovery in biological sequnce: the TEIRESIAS algorithm. *Bioinformatics*, **14**, 55-67, 1998.
16. B. Ren, *et al.* Genome-wide location and function of DNA binding proteins. *Science*, **290**, 2306-2309, 2000.
17. R. Siddharthan, E. Siggia, E. Nimwegen. PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS Computational Biology*, **1**, e67, 0534-0555, 2005.
18. G. Thijs, M. Lescot, K. Marchal, S. Rombauts, B. Moor, P. Rouze, and Y. Moreau. A higher order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics*, **17**, 1113-1122, 2001.
19. M. Tompa, N. Li, *et al.* Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotechnology*, **23**, 2005.
20. M. Djordjevic, A. Sengupta, and B. Shraiman. A biophysical approach to transcription factor binding site discovery. *Genome Res.*, **13**, 2381-2390, 2003.