



Modeling within-motif dependence for transcription factor binding site predictions

Qing Zhou and Jun S. Liu*

Department of Statistics, Harvard University, 1 Oxford ST, Cambridge, MA 02138, USA

Received on August 28, 2003; revised on October 31, 2003; accepted on November 3, 2003

Advance Access publication January 29, 2004

ABSTRACT

Motivation: The position-specific weight matrix (PWM) model, which assumes that each position in the DNA site contributes independently to the overall protein–DNA interaction, has been the primary means to describe transcription factor binding site motifs. Recent biological experiments, however, suggest that there exists interdependence among positions in the binding sites. In order to exploit this interdependence to aid motif discovery, we extend the PWM model to include pairs of correlated positions and design a Markov chain Monte Carlo algorithm to sample in the model space. We then combine the model sampling step with the Gibbs sampling framework for *de novo* motif discoveries.

Results: Testing on experimentally validated binding sites, we find that about 25% of the transcription factor binding motifs show significant within-site position correlations, and 80% of these motif models can be improved by considering the correlated positions. Using both simulated data and real promoter sequences, we show that the new *de novo* motif-finding algorithm can infer the true correlated position pairs accurately and is more precise in finding putative transcription factor binding sites than the standard Gibbs sampling algorithms.

Availability: The program is available at <http://www.people.fas.harvard.edu/~junliu/>

Contact: jliu@stat.harvard.edu

INTRODUCTION

A transcription factor (TF) functions by binding to the recognition site in the promoter region of a gene that it regulates. The common pattern of the recognition sites of a TF is called a binding motif. Since experimental procedures to determine the exact binding sites are too expensive and time-consuming, computational methods have been developed in the past two decades for discovering novel motif patterns and TF binding sites in a set of promoter sequences. Some early methods based on the site consensus used enumeration (Galas *et al.*, 1985) or an heuristic progressive alignment procedure (Stormo and Hartzell, 1989) to find motifs. A formal statistical model for the position-specific weight matrix (PWM)-based method was described in Lawrence and Reilly (1990) and a complete

Bayesian method was given in Liu (1994) and Liu *et al.* (1995). Based on a missing data formulation, the EM algorithm (Lawrence and Reilly, 1990; Bailey and Elkan, 1994; Grundy *et al.*, 1996) and the Gibbs sampler (Lawrence *et al.*, 1993; Liu, 1994; Liu *et al.*, 1995; Neuwald *et al.*, 1995) were employed for motif discovery. By iteratively masking out aligned sites, AlignACE is more effective in finding multiple distinct motifs (Roth *et al.*, 1998). BioProspector (Liu *et al.*, 2001) uses the third order Markov chain to model the background sequences so as to improve the motif specificity and can search for motifs with two components. Dynamic programming was utilized by Gupta and Liu (2003) to sample motif sites with gaps according to a stochastic dictionary model.

In the traditional PWM-based methods, it is assumed that the positions within a motif site are mutually independent. Although recent biological experiments have shown that nucleotides of TF binding sites exert interdependent effects on the binding affinities of TFs (Bulyk *et al.*, 2002; Benos *et al.*, 2002b), it is conceivable that the PWM model is still a reasonably good approximation of the ‘true’ protein–DNA interaction model (Benos *et al.*, 2002a) and there is no need for a more complex model for discovering novel TF binding sites. On one hand, a more comprehensive model may allow for a better fit to the data. On the other hand, the more complex model may over-fit the data and result in an inferior predictive power. Our study here demonstrates that employing models that allow for correlated position pairs can indeed improve the sensitivity and specificity of both *de novo* motif discoveries and site predictions based on known TF motifs.

Barash *et al.* (2003) proposed a Bayesian network approach to ‘mine’ the optimal TF binding model from all possible ones that allow for various interactions, as well as an EM algorithm for simultaneous model mining and *de novo* motif finding. Although the Bayesian network can describe very complex correlations, it induces a parameter space that is too large and may not be well supported by the available sequence data. Since the model space is exceedingly large, learning a complex Bayesian network is very time-consuming. Furthermore, prescribing a sensible prior distribution on the model space is rather non-trivial. Here, we propose to search for an appropriate motif model among the set of models that allow only for non-overlapping correlated position pairs. Pair correlations

*To whom correspondence should be addressed.

are not necessarily between two neighboring positions, but two different pairs are not allowed to have common positions. We give a smaller *a priori* probability to the model with more parameters than that with fewer ones.

Since different combinations of correlated positions imply motif models with different numbers of parameters, we employ the Bayes factor (BF) as a criterion for model selection. We describe a Markov chain Monte Carlo (MCMC)-based algorithm to sample jointly the motif structure and the binding sites and, by using simulations based on the CRP data (Stormo and Hartzell, 1989), show that the pair-correlation model suffers little, if not at all, from the problem of overfitting. Applying the algorithm to known TF binding sites from TRANSFAC (Wingender *et al.*, 2000), we find that 25% of the motifs have correlated positions with statistical significance. Simulation studies show that not only can this method not only find true correlated positions within a TF site, but also decrease false negative (FN) and false positive (FP) error rates for *de novo* motif predictions. We also show that this method is more precise in predicting binding sites in the promoter sequences of a set of E2F controlled genes.

METHODS

Let S denote the set of N sequences (e.g. upstream regions of a set of potentially co-regulated genes) of lengths L_1, \dots, L_N , respectively, from which one wishes to find the binding sites of a TF. Let $A = \{A_{ij}, i = 1, 2, \dots, N, j = 1, 2, \dots, L_i\}$ be the site location indicator array, where $A_{ij} = 1$ implies that the j th base on the i th sequence is the starting position for a motif site. We use a first-order Markov chain to model non-site positions of the sequences and the parameter of this background model θ_0 (i.e. the transition matrix) is estimated from S prior to the motif search. Thus, we effectively assume that θ_0 is known *a priori*. We further denote the aligned motif sites by $S(A)$ and the non-site background sequences by $S(A^c)$. The motif width w , if treated as an unknown, can be inferred by an approach similar to that used in Gupta and Liu (2003). Since this paper focuses on the discovery of within-site correlated positions, we treat w as known in the following sections.

The generalized weight matrix model

To describe the generalized weight matrix (GWM) model that allows for correlated within-site positions, we let the motif positions be numbered from 1 to w and let (i, j) represent that positions i and j are correlated. The occurrences of the nucleotides in these two positions are described by a 4×4 probability matrix corresponding to the 16 dinucleotide pairs. We set the constraint that no two correlated pairs of positions (i, j) and (k, l) can have overlapping positions. For example, for a motif of width 6, one possible model is (2, 6)(4, 5), implying that positions 2 and 6 are correlated, positions 4 and 5 are correlated and all other positions are independent. Its GWM is $(\theta_1, \theta_{(2,6)}, \theta_3, \theta_{(4,5)})$, where θ_i is a probability vector of length 4 for position i , and $\theta_{(j,k)}$

a probability matrix for the correlated positions j and k . The probability that 'ACATTG' occurs under this model is then $\theta_1(A) \times \theta_{(2,6)}(\text{CG}) \times \theta_3(A) \times \theta_{(4,5)}(\text{TT})$. Note that (1, 3)(3, 4) is not a legitimate model since '3' is a common position for the two pairs.

Let \mathcal{H} be the set of all legitimate models and let $\mathcal{H}_m \subset \mathcal{H}$ be the set of models that have m pairs of correlated positions. To put a prior on these models, we consider the following two criteria: (1) if $H_1, H_2 \in \mathcal{H}_m$, then $P(H_1) = P(H_2)$; (2) if $H_1 \in \mathcal{H}_m$ and $H_2 \in \mathcal{H}_{m+1}$, then $P(H_1)/P(H_2) = \binom{w-2m}{2}$. In words, we penalize the model with an additional pair of correlated positions by the inverse of the number of ways of inducing the pair. Thus, for any $H \in \mathcal{H}_m$, its prior probability is

$$P(H) \propto \left[\prod_{j=1}^m \binom{w-2(j-1)}{2} \right]^{-1} = \frac{2^m \cdot (w-2m)!}{w!}, \quad (1)$$

and the parameters for the motif pattern (i.e. frequencies of mono- or di-nucleotides) are represented as Θ_H .

The joint posterior distribution

Under the above framework, we can write the joint distribution of the observed data and all the unknown variables involved:

$$P(S, A, \Theta_H, H, p_0) = P(S | A, \Theta_H, H) P(A | p_0) \times P(\Theta_H | H) P(p_0) P(H). \quad (2)$$

Here p_0 is called the 'site abundance' parameter in that we assume that a randomly selected segment of width w has *a priori* probability p_0 of being a motif site. It is also possible and sometimes even desirable to let different sequence positions have different 'site abundance' to reflect the scientist's prior knowledge (e.g. comparative genomics results or other information) on motif site locations. However, here we focus on the simpler case with constant motif abundance.

All the early methods such as the Gibbs motif sampler (Henceforth, GMS; Liu *et al.*, 1995; Neuwald *et al.*, 1995), BioProspector (Liu *et al.*, 2001), MEME (Grundy *et al.*, 1996) and AlignACE (Roth *et al.*, 1998) use a fixed site abundance p_0 typically ranging from 1/200 to 1/5000. It is suggested by recent studies (Liu *et al.*, 2002) that a bad choice of p_0 will have a significantly negative effect on the accuracy of the motif search. To overcome this shortcoming, we treat p_0 as an unknown parameter and give it a prior distribution $\text{Beta}_{a,b}(p_0) \propto p_0^{a-1} (1-p_0)^{b-1}$, where a and b are the 'pseudo-counts'.

ALGORITHM

We developed a Gibbs motif sampling algorithm from the joint posterior distribution (2) of the site locations and the motif model, henceforth GMS-MP. For comparison purposes, we use GMS-P to denote the Gibbs motif sampler that updates

p_0 but assumes independence for all the site positions. With a random start of site locations \mathbf{A} , our algorithm (GMS-MP) cycles through the following steps:

- (i) Updating site locations. For $i = 1, \dots, N$ and $j = 1, \dots, L_i$, draw A_{ij} according to the current site locations in other sequences and conditional on all other parameters.
- (ii) Updating site abundance p_0 . We sample from the conditional distribution

$$[p_0 | \mathbf{A}] = \text{Beta}(|\mathbf{A}| + a, L - |\mathbf{A}| + b),$$

where $|\mathbf{A}|$ is the total number of sites and L is the total length of the sequences.

- (iii) Updating model. We first propose to modify the current model H by either adding a correlated pair of positions or deleting an existing correlated pair. This proposal is accepted or rejected according to the Metropolis–Hastings rule (Liu, 2001).

Step (i) can be implemented similarly to the GMS (Liu *et al.*, 1995). After cycling through all the sequence positions in step (i), we update p_0 according to the posterior distribution $\text{Beta}(|\mathbf{A}| + a, L - |\mathbf{A}| + b)$. A phase shifting step via the Metropolis–Hastings rule (Liu, 1994) is conducted once every 20 iterations of site-updates in order to escape from suboptimal modes. Multiple Metropolis–Hastings moves can be implemented in each iteration to speed up sampling of the model space. After each iteration, the joint posterior probability (2) is calculated and the one that maximizes this quantity is recorded as the output of the algorithm.

Since different models may imply different dimensional Θ_H , we need to calculate the ‘collapsed’ probability in order to conduct step (iii):

$$P(\mathbf{S}, \mathbf{A} | H) = \int P(\mathbf{S}, \mathbf{A} | \Theta_H, H) P(\Theta_H | H) d\Theta_H, \quad (3)$$

which, after combining with the prior $P(H)$, gives rise to the conditional posterior probability of a specific model:

$$P(H | \mathbf{S}, \mathbf{A}) \propto P(\mathbf{S}, \mathbf{A} | H) P(H). \quad (4)$$

In the likelihood (3), $P(\Theta_H | H)$ is the prior for Θ_H . If we put a four-dimensional Dirichlet prior on each independent position and a 16-dimensional Dirichlet prior on each correlated pair, Equation (3) can be computed exactly as products and ratios of gamma functions. Suppose that the current model H^t has m ($0 \leq m \leq [w/2]$) correlated position pairs. Two operators are defined to propose the model modification.

- (A) Add one pair: randomly choose two positions from the $(w - 2m)$ independent ones and form a correlated pair;
- (B) Delete one pair: choose one of the m existing correlated pairs at random and release the two positions to be independent.

The above two operations are proposed with equal probabilities for $1 \leq m \leq [w/2] - 1$ possible pairs. If $m = 0$ (or $[w/2]$), we only propose to add (delete) one pair. The proposed model, denoted by H^* , will be accepted with probability $\min(1, r)$ according to the Metropolis–Hastings rule,

$$\begin{aligned} r &= \frac{P(H^* | \mathbf{S}, \mathbf{A}^t) \cdot T(H^t | H^*)}{P(H^t | \mathbf{S}, \mathbf{A}^t) \cdot T(H^* | H^t)} \\ &= \text{BF}(H^*; H^t) \cdot \frac{P(H^*)}{P(H^t)} \cdot \frac{T(H^t | H^*)}{T(H^* | H^t)}, \end{aligned} \quad (5)$$

where $\text{BF}(H^*; H^t) = P(\mathbf{S}, \mathbf{A}^t | H^*) / P(\mathbf{S}, \mathbf{A}^t | H^t)$ is the BF, $P(H^t)$ is the prior probability of model H^t (1), and $T(H^* | H^t)$ is the probability of proposing model H^* from H^t .

RESULTS

GWM for known TF binding sites

To verify that the GWM model can indeed improve the specificity of motif site predictions, we carried out an experiment similar to that in Barash *et al.* (2003). From the TRANSFAC database (Wingender *et al.*, 2000), Barash *et al.* (2003) had selected 95 TFs for which there were 20 or more known binding sites (data available from the supplementary material of Barash *et al.*, 2003). Then, they designed a cross-validation test to compare the relative performances of their Bayesian network approach to the standard PWM method. We modified their strategy slightly as follows: for a set with N known binding sites of a TF, we learn a GWM model on $N - 1$ sites and compute the predictive log-probability score for the remaining site based on the learned model. This step was repeated N times with each site serving as the test site once. Simultaneously, we randomly selected 2000 human genes based on the Ensembl (<http://www.ensembl.org>) annotation and extracted their upstream 1 kb. For the model we learned in the leave-one-out experiment, the log-probability score for all the w -mer’s on these random upstream sequences was calculated. Those ‘random’ w -mer’s that had a score greater than the 10th percentile of the test site scores (i.e. at 90% sensitivity) were recorded as FPs. The same procedure was also repeated for the PWM model.

Among the 95 sets of TF binding sites, there are 22 sets for which the posterior probability of the most likely correlated model (GWM) is at least six times that of the independent model (PWM). At the 90% sensitivity level, the GWM results in a smaller FP value than the PWM model for 17 of the 22 TFs, and a slightly larger FP value for the remaining five TFs. The results are summarized in Table 1. There are 10 sets for which the posterior probability of the most likely model is between two and six times that of the PWM model and the GWM performed better than the PWM in 7 of these 10 sets.

The ROC curves of the two worst cases (V\$IRF7_01 and V\$FAC1_01) are plotted in Figure 1a and b. It can be seen that

Table 1. The GWMs discovered from the known binding sites of 22 TFs.

TF	<i>N</i>	RFP	ΔLD	Structure
V\$IRF7_01	28	0.414	0.78	(5, 4)(7, 6)
V\$FAC1_01	24	0.497	2.69	(14, 11)(2, 1)(12, 13)
V\$PAX8_01	35	0.555	0.77	(14, 15)(3, 1)(7, 6)
V\$VMBYB_01	24	0.741	0.02	(10, 9)
V\$TBP_01	47	0.867	1.11	(7, 8)(2, 1)
V\$PAX6_01	47	1.011	0.53	(19, 16)(3, 4)
V\$FOXJ2_01	41	1.065	1.94	(16, 18)(1, 3)
V\$BRACH_01	40	1.200	2.58	(1, 2)(19, 21)(4, 3)
V\$SPZ1_01	30	1.345	1.00	(10, 1)(12, 6)
V\$ARNT_01	20	1.355	1.85	(12, 8)(9, 6)(1, 2)(11, 7)(4, 5)
P\$ABF_Q2	49	1.493	1.04	(5, 6)
V\$S8_01	59	1.505	2.61	(12, 10)(7, 5)(11, 9)(3, 1)(8, 6)(2, 4)
V\$ELK1_02	31	1.565	0.70	(3, 4)
V\$STAT5A_01	33	1.589	1.12	(1, 15)(7, 14)
P\$P_01	36	1.798	0.61	(8, 9)
V\$HAND1E47_01	29	2.067	1.82	(14, 13)(1, 2)
V\$SOX9_B1	73	2.117	0.86	(2, 14)(6, 5)(1, 3)
V\$ATF_01	25	2.134	1.89	(2, 1)(10, 11)
V\$AHRARNT_01	24	2.517	3.90	(12, 10)(3, 5)(7, 6)(4, 1)(11, 13)(9, 8)
V\$PBX1_02	40	3.941	1.17	(9, 6)(4, 5)
P\$ABF1_01	20	6.319	4.37	(11, 10)(9, 13)(15, 12)
V\$PPARG_01	72	7.654	2.64	(8, 11)(10, 7)(5, 14)(9, 12)(6, 13)

Notation *N* refers to the number of known sites; RFP stands for the ratio of FPs from the PWM to FPs from the GWM; ΔLD is the difference of *LD* between the GWM and the PWM; and 'structure' gives the significant pairs of correlated positions.

although the GWM showed a larger FP at the 90% sensitivity, the overall difference between the GWM and PWM is not significant. The GWMs in these two cases performed even better at some other sensitivity levels. A similar situation happened in another case, V\$TBP_01. V\$VMBYB_01 is a short (its width is 10) and weak motif. The correlated position pair (9, 10) is at the end of the motif and shows a preference for 'AA', which may have weakened the specificity of the motif due to the abundance of simple repeats in these sequences. As positive examples, the GWM for V\$PPARG_01 contains five correlated position pairs and its posterior probability is 3.5×10^{15} times that of the PWM. The only nucleotide combinations of the position pair (6, 13) are 'AA', 'CC', 'GG', 'TT' and 'GT'. P\$ABF1_01 has a significantly correlated position pair (10, 11) in which 'A' and 'C' tend to co-occur. Figure 1c and d show the ROC curves for the two positive examples, where the FP rates with the GWM are much smaller than those with the PWM across different sensitivity levels.

We have also computed the ΔLD measure as suggested by Barash *et al.* (2003), which showed that all the 22 GWMs improved in terms of *LD*. A positive ΔLD implies that the 'centers' of the two likelihood-ratio score distributions, one for the true sites and another for the random sites, are further apart based on the GWM model than based on the PWM model. But this measure does not account for the change of the variance of the score distributions under the new model.

Thus, using FP rates and the ROC curves shown here seems to be a more objective way of comparing the models.

De novo motif discovery for two real datasets

CRP binding sites This dataset contains 18 sequences of 105 bp each that are known to be bound by cyclic-AMP receptor protein (CRP) in *Escherichia coli* (Stormo and Hartzell, 1989). In order to make the problem more difficult, some random sequences were generated and mixed with the original 18 sequences. The three datasets we used in this study contain 0, 20 and 40 random sequences, respectively. The GMS, GMS-P and GMS-MP were applied to these three datasets. In GMS, the site abundance parameter was set at $p_0 = 1/100$ and $1/400$, respectively (i.e. one binding site per $1/p_0$ bp is expected). BioProspector (Liu *et al.*, 2001) was also applied to the datasets.

It is seen from Table 2 that GMS-P and GMS-MP found more true sites than GMS and BioProspector in all settings. For the original CRP data (dataset 1), there is approximately one site per 100 bases. Consistent with this fact, the GMS with $p_0 = 1/100$ achieved a comparable result with GMS-P and GMS-MP, whereas GMS with $p_0 = 1/400$ and BioProspector had higher FN rates. For dataset 3, which contains 40 additional random sequences, the GMS with $p_0 = 1/400$ and BioProspector have comparable FPs with GMS-P and GMS-MP, whereas GMS with $p_0 = 1/100$ had a much higher FP rate. Although, the maximum *a posteriori* model found by GMS-MP has (6, 10) as correlated positions, it is not significant and, thus, no definitive conclusion can be drawn. We have also conducted simulations to test the effects of updating p_0 and observed a similar result: adjustable p_0 enables the sampler to result in low, stable and ballanced overall error rates, whereas a fixed p_0 at wrong value can be detrimental.

To test whether the pair correlation model suffers the overfitting problem, we created test datasets as follows. Each dataset consists of 20 sequences of length 105 bp, among which *K* sequences were randomly selected among the 18 sequences in the original CRP dataset and the remaining $20 - K$ sequences were generated from our background model. Fifty datasets were generated for $K = 16, 14, 12, 10$ and 9 , respectively. Both GMS-P and GMS-MP were applied to each dataset 10 times and the maximum *a posteriori* results from the two algorithms were compared. We observed that GMS-MP performed similar to GMS-P in all cases, indicating that overfitting is not a problem for GMS-MP. For example, when $K = 12$, GMS-P and GMS-MP found the true motif in 44 and 43, respectively, of the 50 simulated datasets; when $K = 10$, these numbers decreased to 11 and 12, respectively; when $K = 9$, neither algorithm could find the true motif in any of the datasets.

E2F binding sites We extracted 123 E2F-induced human genes from Ishida *et al.* (2001), Muller *et al.* (2001) and

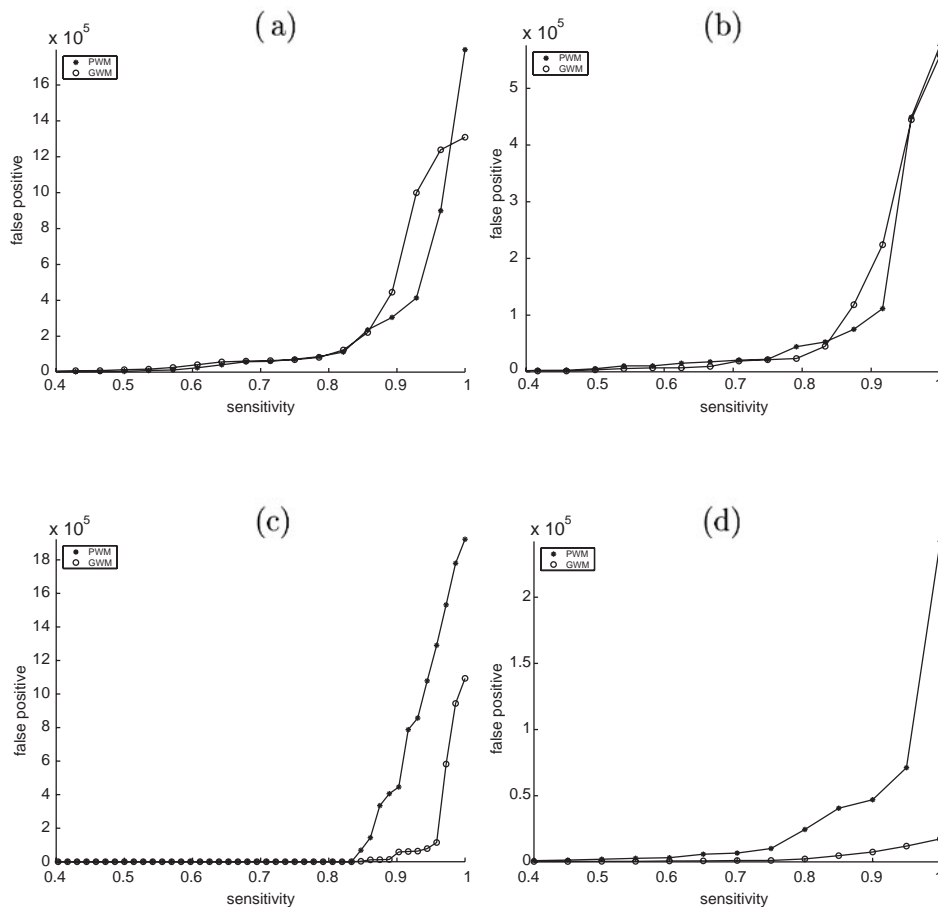


Fig. 1. ROC curve comparisons between the GWMs and the PWMs for (a) V\$IRF7_01, (b) V\$FAC1_01, (c) V\$PPARG_01 and (d) P\$ABF1_01.

Table 2. Results for the CRP data

Dataset	GMS1	GMS2	BioProspector	GMS-P	GMS-MP
CRP	16 (2)	9 (1)	11 (0)	16 (1)	16 (1)
CRP + 20bg	8 (8)	8 (8)	4 (9)	8 (8)	9 (9)
CRP + 40bg	3 (28)	2 (11)	2 (9)	3 (11)	5 (13)

GMS1 has a fixed $p_0 = 1/100$ and GMS2 has a fixed $p_0 = 1/400$. The number of experimentally determined sites that are also predicted by the algorithms (true positives) are listed in the table. The number of sites that are predicted by the algorithm but not validated experimentally (FPs) are in parentheses. Since the CRP binding motif has a palindromic form, only forward strands were searched.

Ren *et al.* (2002), and obtained their upstream 300 bp according to the Ensembl annotation. Since these are only E2F-induced genes, not all the upstream sequences contain E2F binding sites. The known E2F weight matrix (denoted by W_1) from TRANSFAC was used to scan these sequences. Each candidate segment was scored by the ratio of its likelihood under the known E2F weight matrix to that under the background model. We found 15 upstream sequences that had at

least one high score (>1000) site, among which there are 26 putative sites with a score greater than 300. The weight matrix constructed by these sites (denoted by W_2) is very similar to the known E2F weight matrix with a consensus of SGCGCSAAA. Under this threshold (300), we estimated by simulation that the expected numbers of FP and FN sites are about 2.7 and 3, respectively.

The GMS with fixed p_0 at 1/100 and 1/1000, respectively, GMS-P and the GMS-MP were applied to the 15 sequences. For each algorithm, the weight matrix was built by its predicted sites. We compared these weight matrices to the known E2F weight matrix (W_1) and the weight matrix constructed by the 26 putative sites via scanning (W_2) based on the relative entropies between them. We also tried BioProspector, but it did not find this motif. From the results in Table 3, we can see that the GMS-MP reported weight matrices with a smaller entropy distance compared with those reported by the GMS. The number of sites found by the GMS-MP is closer to the number of putative sites obtained by scanning (26). GMS1 outputs 55 sites, suggesting that $p_0 = 1/100$ was perhaps too large and many FP sites were produced. In contrast, GMS2

Table 3. Results for the E2F data

Algorithm	GMS1	GMS2	GMS-P	GMS-MP
$I(\bullet W_1)$	2.61	3.37	1.81	1.83
$I(\bullet W_2)$	2.01	4.34	1.99	1.37
n	55	16	21	32

GMS1 has a fixed $p_0 = 1/100$ and GMS2 has a fixed $p_0 = 1/1000$. Notation ‘ n ’ is the number of output sites by the algorithms and ‘ I ’ is the relative entropy.

seems to have a high FN rate, and the weight matrix built by its predicted sites has the largest entropy distances to W_1 and W_2 . The most probable model found by GMS-MP has positions 1 and 2 significantly correlated (the corresponding GWM is shown in Fig. 2). The model’s posterior probability is 20.3 times that of the PWM.

Simulation studies for *de novo* motif discovery

In the following studies, we first generated the background sequences according to a first-order Markov chain with parameters estimated by more than 2000 upstream 1 kb human sequences based on the Ensembl database, and then randomly inserted motif sites generated by a motif model into these background sequences according to a site abundance parameter p_0 . The first group of sequence datasets were generated using three motif models (M1–M3) of width 10. These models all have two fixed pairs of correlated positions, (3, 8) and (5, 7), which have joint distributions [CG, GC] = [0.5, 0.5] and [AA, TT] = [0.5, 0.5], respectively, but 2, 4 and 6, weakly conserved independent positions (i.e. the most conserved letter has a frequency less than 0.7), respectively. Twenty sequence datasets were generated for each model, and each dataset contained 40 sequences of 500 bp each with one motif site per sequence on average ($p_0 = 1/500$).

The second group of sequence datasets were simulated using four correlated-position models (M4–M7) with a fixed marginal PWM as shown in Figure 3. Model M4 has positions (9, 10) correlated with the joint distribution [AA, CC, GG, TT] = [0.2, 0.3, 0.2, 0.3]; M5 has an additional correlated pair (5, 6) with the joint distribution [AT, CG, GC] = [0.3, 0.3, 0.4]; M6 adds the third correlated pair (4, 11) with distribution [AT, GC] = [0.6, 0.4]; and M7 includes the fourth pair (3, 12) with distribution [AT, TA] = [0.5, 0.5]. For each motif model, 20 sequence datasets were generated, each containing 40 sequences of length 200 bp with one motif site per sequence on average.

For each sequence dataset generated, we ran GMS-P and GMS-MP for 2000 iterations and picked the maximum *a posteriori* motif configurations to compare. It is seen from Table 4 that, compared with the GMS-P, the GMS-MP had a much lower FN and FP rate for the binding site prediction in all the cases. Furthermore, GMS-MP was precise in



Fig. 2. (a) Sequence logo plot for the E2F sites predicted by the GMS-MP. The traditional consensus for the E2F motif is the one from positions 2 to 10. (b) The joint distribution of the position pair (1, 2), which has been found to be significantly correlated by the GMS-MP.



Fig. 3. Sequence logo plot (Schneider and Stephens, 1990) for the motif model used in the second simulation, where all the positions are marginally weak.

Table 4. Comparison between GMS-P and GMS-MP for *de novo* motif finding

Algorithms	Models	Fixed correlations			Fixed PWM			
		M1	M2	M3	M4	M5	M6	M7
GMS-P	Fail%	0	40	45	80	70	40	50
	FN%	17.5	30.3	49.3	71.1	71.3	62.7	71.8
	FP	16.3	23.7	29.3	10.5	12.5	10.6	10.7
GMS-MP	Fail%	0	5	30	80	60	20	10
	FN%	6.7	11.5	13.4	33.5	7.7	1.1	0.3
	FP	9.3	17.7	21.7	15.0	11.3	8.0	8.4

M1–M3 are motifs with fixed correlated position pairs and different numbers of weak positions, and M4–M7 are motifs with a fixed weak PWM (Fig. 3) but varying numbers of correlated position pairs. For $i = 1, 2, 3$, model M_i has $2i$ weak positions; and for $i = 4, \dots, 7$, M_i has $i - 3$ pairs of correlated positions. Fail% is the percentage of times that the algorithm fails to find the motif (defined as missing more than 80% of true sites); FN% is the percentage of FN sites and FP is the number of FP sites.

finding the true correlated positions, with an average error rate less than 0.05. It is especially striking to see that GMS-MP drastically improved the success rates of *de novo* motif finding in models with a fixed PWM but increased numbers of correlated position pairs. Since all four models (M4–M7) have the same PWMs, the performance difference between GMS-P and GMS-MP must be due to the ability of GMS-MP to capture additional information in correlated positions.

This also demonstrated that with the new GWM model, the Gibbs sampling strategy is effective in finding weakly conserved motif sites if these sites show additional within-site positional correlations.

DISCUSSION

As is true for all *de novo* motif discovery algorithms, there always exists the problem of being trapped in local optima. There are at least two strategies to alleviate this problem. If we have some additional information such as Chromatin Immunoprecipitation microarray data or other case-control gene expression data to rank the genes according to the likelihood of their upstream regions containing the motif sites, we can scan through all the *w*-mers on the few (20, say) top sequences to find the overrepresented ones as the initial weight matrices to start our search. This is similar to the strategy used in MDScan (Liu *et al.*, 2002). A more time-consuming strategy, which has been implemented in MEME (Grundy *et al.*, 1996), is to use every *w*-mer in the whole sequence set to initiate the search. Our algorithm implemented a more flexible strategy that allows the user to input his/her favorite sequence pattern (consensus). The initial weight matrix is then built by assigning 0.55 probability to the consensus letter on each position and 0.15 for the remaining letters. We have tested on a number of datasets that if the input pattern is close to the true one, our algorithm will find the true sites very rapidly. This flexible implementation also allows users to perform a similar strategy to that of MDScan and MEME by writing a short Perl script shell.

In theory, a proper account of correlated positions within TF binding sites can help us better understand protein–DNA interactions and improve the sensitivity of statistical models used for TF binding site predictions. To achieve this end, we have introduced a pair-correlation model for TF binding motifs and designed a MCMC algorithm GMS-MP to simultaneously infer the motif pattern, the correlated positions and the site abundance parameter. Using both real and simulated datasets, we have shown that, where there are traces of within-site correlations, the GMS-MP performed significantly better than the standard PWM-based Gibbs sampling methods. Compared with the Bayesian network approach (Barash *et al.*, 2003), our model is simpler, prescribing priors is easier, and the computation is much faster—the step of sampling pair correlations takes up only about 3% of the total computing time, which is much faster than learning a Bayesian network. Our method also does not suffer any problems with overfitting, which is likely to occur, due to our employment of a rather conservative prior distribution on the model space.

ACKNOWLEDGEMENTS

This work is supported in part by NSF Grants DMS-0204674 and DMS-0244638. We thank Shirley Liu and Shane Jensen for insightful comments.

REFERENCES

- Bailey, T.L. and Elkan, C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Proceedings of the 2nd International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, Menlo Park, California, 28–36.
- Barash, Y., Elidan, G., Friedman, N. and Kaplan, T. (2003) Modeling dependence in protein–DNA binding sites. *RECOMB'03*. Berlin, Germany, April 2003.
- Benos, P.V., Bulyk, M.L. and Stormo, G.D. (2002a) Additivity in protein–DNA interactions: how good an approximation is it? *Nucleic Acids Res.*, **30**, 4442–4451.
- Benos, P.V., Lapedes, A.S. and Stormo, G.D. (2002b) Probabilistic code for DNA recognition by proteins of the EGR family. *J. Mol. Biol.*, **323**, 701–727.
- Bulyk, M.L., Johnson, P.L.F. and Church, G.M. (2002) Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res.*, **30**, 1255–1261.
- Galas, D.J., Eggert, M. and Waterman, M.S. (1985) Rigorous pattern-recognition methods for DNA sequence: analysis of promoter sequences from *Escherichia coli*. *J. Mol. Biol.*, **186**, 117–128.
- Grundy, W.N., Bailey, T.L. and Elkan, C.P. (1996) ParaMEME: a parallel implementation and a web interface for a DNA and protein motif discovery tool. *Comput. Appl. Biosci.*, **12**, 303–310.
- Gupta, M. and Liu, J.S. (2003) Discovery of sequence patterns using a stochastic dictionary model. *J. Amer. Stat. Assoc.*, **98**, 55–66.
- Ishida, S., Huang, E., Zuzan, H., Spang, R., Leone, G., West, M. and Nevins, J.R. (2001) Role for E2F in control of both DNA replication and mitotic functions as revealed from DNA microarray analysis. *Mol. Cell. Biol.*, **21**, 4684–4699.
- Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.N. and Wootton, J. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.
- Lawrence, C.E. and Reilly, A.A. (1990) An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins*, **7**, 41–51.
- Liu, J.S. (1994) The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *JASA*, **89**, 958–966.
- Liu, J.S. (2001) *Monte Carlo Strategies in Scientific Computing*. Springer-Verlag, New York.
- Liu, X., Brutlag, D.L. and Liu, J.S. (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac. Symp. Biocomput.*, **6**, 127–138.
- Liu, X., Brutlag, D.L. and Liu, J.S. (2002) An algorithm for finding protein–DNA interaction sites with applications to chromatin immunoprecipitation microarray experiments. *Nat. Biotechnol.*, **20**, 835–839.
- Liu, J.S., Neuwald, A.N. and Lawrence, C.E. (1995) Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *JASA*, **90**, 1156–1170.
- Muller, H., Bracken, A.P., Vernell, R., Moroni, M.C., Christians, F., Grassilli, E., Prosperini, E., Vigo, E., Oliner, J.D. and Helin, K. (2001) E2Fs regulate the expression of genes involved in differentiation, development, proliferation, and apoptosis. *Genes Dev.*, **15**, 267–285.

- Neuwald,A.F., Liu,J.S. and Lawrence,C.E. (1995) Gibbs motif sampling—detection of bacterial outer-membrane protein repeats. *Protein Sci.*, **4**, 1618–1632.
- Ren,B., Cam,H., Takahashi,Y., Volkert,T., Terragni,J., Young,R.A. and Dynlacht,B.D. (2002) E2F integrates cell cycle progression with DNA repair, replication, and G(2)/M checkpoints. *Genes Dev.*, **16**, 245–256.
- Roth,F.R., Hughes,J.D., Estep,P.E. and Church,G.M. (1998) Finding DNA regulatory motifs within unaligned non-coding sequences clustered by whole genome mRNA quantization. *Nat. Biotechnol.*, **16**, 939–945.
- Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.
- Stormo,G.D. and Hartzell,G.W. (1989) Identifying protein-binding sites from unaligned DNA fragments. *Proc. Natl Acad. Sci., USA*, **86**, 1183–1187.
- Wingender,E., Chen,X., Hehl,R., Karas,H., Liebich,I., Matys,V., Meinhardt,T., Pruss,M., Reuter,I. and Schacherer,F. (2000) TRANSFAC: an integral system for gene expression regulation. *Nucleic Acids Res.*, **28**, 316–319.