Finding DNA Regulatory Motifs with Position-dependent Models

HuihaiWu
University of Surrey, Guildford, United Kingdom
Email: h.wu@surrey.ac.uk

Prudence W.H. Wong, Mark X. Caddick and Chris Sibthorp
University of Liverpool, Liverpool, United Kingdom
Email: {pwong, caddick, c.j.sibthorp}@liverpool.ac.uk

***Abstract*—We consider the problem of de novo DNA motif discovery. The position weight matrix (PWM) model has been extensively used, yet this model makes the assumption that nucleotides at different positions are independent of each other. Recent results have shown that nucleotides bound by transcription factors often exhibit adjacent or nonadjacent dependencies. We address this problem by devising positional dependency models capable of capturing adjacent dependencies and non-adjacent dependencies (SPWDM). Our algorithms are based on Gibbs sampling to update the model parameter and dependencies structure. We compare two scoring functions: χ^2-score and a conditional probability based score. We also improve several Gibbs sampling stages. Experiments are carried out on simulated and real data, showing that the SPWDM model makes improvement over pure PWM. The modifications to the Gibbs sampling algorithm are also shown to be effective.**

***Index Terms*—DNA motif discovery, Position weight matrix, Position-dependent Model, Gibbs sampling**

I. Introduction

Two key areas in bioinformatics and genome analysis are gene identification and prediction of their functions. These relate to the promoters and other regulatory elements. Interpreting the “transcriptional regulatory code” represents a major current challenge. Transcription factors (TFs) activate or repress gene expression through binding regulatory regions adjacent to the target gene. One TF may bind to the regulatory regions of several genes and the corresponding binding sites would be similar in terms of length and pattern, which is called a *motif*. These DNA motifs are fairly short, 5 to 20 base-pairs (bp), they recur throughout the genome associated with co-regulated genes, often recur several times within a regulatory domain and are generally not orientation specific. *Motif finding* is to find such patterns within a sequence set associated with a group of regulated genes. Finding DNA motifs has been addressed for decades but still remains a major challenge in computational biology.

The main difficulty is that the degree of conservation of binding sites can vary significantly. The alignment of a set of binding sites can be seen as the representation of a motif. A motif is usually represented by a string [1] or a matrix. We focus on the matrix representation. In a position weight matrix (PWM), the frequency of each nucleotide at each specific position is recorded in the form of scores which are usually based on probabilities or log ratios of frequencies.

Probabilistic approaches have been used extensively in motif finding. The model parameters are estimated using maximum-likelihood principle or Bayesian inference, such as Expectation Maximization (EM) algorithm [2] and Gibbs sampling [3]. Although employing local search, they are efficient and effective with genome background information.

Position dependencies. Traditional PWM assumes nucleotide at each position is independent of each other. Recent researches [4], [5] show that nucleotides at different positions often exhibit adjacent or non-adjacent dependencies, and non-adjacent dependence could be important since three-dimensional folding of proteins, and binding between DNA and protein frequently involve interactions between nucleotides at non-adjacent positions within the binding sites. Such interdependencies within DNA motifs can also be demonstrated by TF binding domain information [6].

Positional dependencies were firstly exploited in the signal identification problem [7], [8], [9]. There are a few *de novo* motif finding approaches taking nucleotide interdependencies into account. In [10], the authors employed Bayesian networks (BNs) to model the motifs, where positional interdependencies are represented by a directed acyclic graph (DAG) and the motif strength is scored by maximum-likelihood. BN suffers the problem of huge number of parameters which increase exponentially in the number of edges in the DAG. To tackle this problem, the authors applied more succinct structures and structural EM algorithm to implement parameter and structure learning. Note that PWM and Markov model are special cases of BNs.

This work supported by the project: “Tools for Motif Recognition in Fungi” at the Department of Computer Science and School of Biological Sciences in the University of Liverpool, funded by BBSRC.

Taking advantage of PWM, [11] proposed the generalized weight matrix (GWM) to allow pairing motif positions such that one position can only be paired once. A Markov chain Monte Carlo (MCMC)-based algorithm was applied to jointly sample the motif structure and the location of binding sites.

Gibbs sampling. Gibbs sampling was first applied to motif finding by Lawrence and Liu et al. [3], [12], [13]. Since then, numerous enhancements have been made and software packages have been developed such as AlignACE [14], Bio-Prospector [15] and GMS [16]. In [15], threshold sampler is proposed to implement sampling multi-copies for each gene sequence. In [17], the number of binding site copies for each sequence was estimated individually and a maximum number per sequence was set. In [18], a neural network was used to define the binding energy of a TF. To score the predicted motifs, in [14], a measure called group specificity score was devised to gauge how well a motif targets the upstream regions of the genes used to find it relative to the upstream regions of all of the genes in the genome.

Aspergillus. The *Aspergilli* are an important group of organisms, which offer a wealth of genome resources, including ten genome sequences. The three recently published *Aspergillus* genome sequences used in this study are those of *A. nidulans*, an important model system, *A. fumigatus*, an important opportunistic pathogen and allogen, and *A. oryzae* which is of biotechnological importance, playing a major role in Japanese food industry. We test our motif models and algorithms on *Aspergillus* and other species.

Our contributions. We address the *de novo* motif discovery problem taking positional dependencies into account. Using similar ideas as GWM, we develop three models (and algorithms) to capture adjacent and non-adjacent dependencies (with different scoring functions). We employ Gibbs sampling approach. To improve overall performance, we have introduced several modifications to the Gibbs sampling algorithm. Experiments are carried out on simulated and real data including *Aspergillus*' and Tompa's data set [19].

The rest of the paper is organized as follows. In Section II, we formulate the *de novo* DNA motif finding problem, and elaborate the motif models and Gibbs sampling algorithms in details. In Section III, we discuss the experiment settings and results. Finally we conclude in Section IV.

II. Methodology

We first formulate the *de novo* motif finding problem. Consider a set of n co-regulated gene sequences $S = \{S_1, ..., S_n\}$. Let ℓ_i be the length of S_i and $S_i = \langle b_{i,1}, b_{i,2}, ..., b_{i,\ell_i} \rangle$. A binding site of width ω started at the j-th nucleotide of S_i is denoted by $s_{i,j} = \langle b_{i,j}, b_{i,j+1}, ..., b_{i,j+\omega-1}\rangle$, where $1 \leq j \leq \ell_i - \omega + 1$.We denote by $A_i = \{s_{i,j_1}, s_{i,j_2}, ..., s_{i,j_{k_i}}\}$ the set of k_i binding sites in S_i. An alignment of a motif can be represented as $A = \bigcup_{i=1}^{n} A_i$, which contains all the binding sites in all sequences. For simplicity, an alignment is usually represented as a sequence of ω nucleotide variables $\langle X_1, X_2, ..., X_\omega\rangle$, where the nucleotide variable X_j is a random variable that takes one of the values in $\{A, C, G, T\}$.

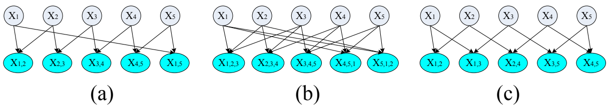

Figure 1. (a) diPWDM. (b) triPWDM. (c) Example of diSPWDM.

Given a set of co-regulated sequences, the problem is to find a set of motifs and the corresponding alignment. In Section II-A, we describe several motif models that capture position-dependencies in the motif. In Section II-B, we describe the Gibbs sampling algorithm.

A. Motif Models

Consider a ω-mer motif. A DNA motif model is $M = \langle\theta, G\rangle$, where *model parameter* θ captures the probabilities of a nucleotide appearing in a certain position, and *model structure* G is a graph representing positional interdependencies. G contains ω nodes, one for each motif position. An edge between two nodes means interdependency of those two motif positions. Fig.1 illustrates the different models.

PWM model. G contains ω nodes and no edges between any two nodes. The model parameter is a $4\times\omega$ probability matrix P. Suppose we have a ω-mer $s = \langle x_1, x_2, ..., x_\omega\rangle$, the probability of it being a binding site under the motif model M is $P(s|M) = \prod_{j=1}^{\omega} P(X_j)$ where $P(X_j)$ is a shorthand for $P(X_j = x_j)$, the probability of nucleotide x_j appearing in position j. Notice that PWM assumes that position variables X_j and X_k are independent for any $j \neq k$.

diPWDM and triPWDM model. A possibility to capture positional dependencies is to combine nucleotides into a super-nucleotides. If we combine two adjacent nucleotides as a di-nucleotide, we obtain the following di-nucleotides: $\{(X_1,X_2), (X_2,X_3), ..., (X_{\omega-1},X_\omega), (X_\omega,X_1)\}$. We call this model *diPWDM*, position weight model with dependencies for di-nucleotides. The graph structure of diPWDM contains ω nodes and each node has an edge to the next node. The probability matrix of diPWDM is a $16\times\omega$ matrix, for 16 possible di-nucleotides. The probability of a ω-mer $s = \langle x_1, x_2, ..., x_\omega\rangle$ being a binding site under the diPWDM model M is (with $X_{w+1} = X_1$):

$$P(s|M) = \prod_{j=1}^{\omega} P(X_j, X_{j+1})\ . \tag{1}$$

Similarly, we define the tri-PWM model for every three neighbouring nucleotides. The graph structure of triPWDM contains ω nodes and each node X_j has two edges, one to node X_{j+1} and one to X_{j+2}. The probability matrix of triPWDM is a $64\times\omega$ matrix. The probability of a ω-mer $s = \langle x_1, x_2, ..., x_\omega\rangle$ being a binding site is

$$P(s|M) = \prod_{j=1}^{\omega} P(X_j, X_{j+1}, X_{j+2})\ . \tag{2}$$

SPWDM model. The graph structures of diPWDM and triPWDM models are fixed and cannot capture dependencies between nucleotides located far apart. To tackle this problem, we adopt a similar idea as Zhou and Liu's [11] that a nucleotide can be paired with another nucleotide in any position of the site (not necessarily a

Algorithm 1 Building structure of diSPWDM model

```
1) For each nucleotide position, set the corresponding node to be 0-connected.
2) Find a pair of positions $j$ and $k$ which has the highest dependency in terms of the scoring matrix (either $score_{\chi}$ or $score_{max}$). Connect the two corresponding nodes and set them to be 1-connected.
3) Find a pair of 1-connected and 0-connected node with the highest dependency score. Set them to be 2-connected and 1-connected, respectively.
4) Repeat Step 3 until there is no 0-connected node. Then we are left with two 1-connected nodes and we connect them to become 2-connected.
```

neighbor). We keep the graph structure as a ring like diPWDM, i.e., each nucleotide position is connected to two other positions. The dimension of the probability matrix of the model parameters is the same as diPWDM, i.e., $16\times\omega$. We call this model *structural PWDM*, abbreviated as diSPWDM.

Consider the following graph structure: $j_1, j_2, ..., j_\omega$ is a permutation of $[1, ..\omega]$ and the node X_{j_1} is connected to X_{j_2} , X_{j_2} connected to X_{j_3}, ..., X_{j_ω} to X_{j_1}. Similar to Eq. (1), the probability of a ω-mer $s = \langle x_1, x_2, ..., x_\omega\rangle$ being a binding site under the diSPWDM model M is computed as follows (with $j_{w+1} = j_1$):

$$P(s|M) = \prod_{k=1}^{\omega} P(X_{j_k}, X_{j_{k+1}})\ . \quad (3)$$

$\langle\theta, G\rangle$ of SPWDM model. Given a set of potential binding sites, we determine the SPWDM model parameter and structure by measuring the dependencies of di-nucleotides. We use two criteria to score the dependencies: χ^2 score and maximal degree of dependency. Using these two scores, we further distinguish two models. We also give an algorithm to determine the model structure based on these two scores.

χ^2 score. Consider a ω-mer motif modelled by ω nucleotide variables $\langle X_1, X_2, ..., X_\omega\rangle$. For any two nucleotide variables X_j and X_k, a contingency table is a 2D table whose entries are the frequencies of the nucleotides in $X = \{A, C, G, T\}$ observed at position j and position k in the motif. We denote this frequencies as $O_{j,k}(\cdot)$. $E_{j,k}(\cdot)$ denotes the expected frequencies of occurrence at positions j and k. For any $x, y \in X$, $E_{j,k}(x,y) = \frac{f_j(x) f_k(y)}{t}$ where $f_j(x)$ is the frequency of nucleotide x appearing in position j and t is the total number of sites. Then $score_{\chi}(j,k) = \sum_{x\in X}\sum_{y\in X}\frac{(E_{j,k}(x,y)-O_{j,k}(x,y))^2}{E_{j,k}(x,y)}$. A high χ^2 score implies a strong dependency. We name this model *diSPWDM*$_\chi$.

Maximal degree of dependency. The second score uses conditional probabilities: $score_{max}(j,\ k) = max_{x,y\in X}\ P(X_j = x|X_k = y)P(X_k = y|X_j = x) = max_{x,y\in X}\frac{(O_{j,k}(x,y))^2}{f_j(x)f_k(y)}$. $P(X_j\,|X_k)$ indicates the degree of X_j depending on X_k. Similar to the χ^2 scoring, a high $score_{max}$ implies a high dependency between the two positions. We name this model *diSPWDM*$_{\text{max}}$.

Algorithm 2 Gibbs sampler to update model parameter and structure $\langle\theta, G\rangle$

```
We denote by $A = \bigcup_{i=1}^{n} A_i$ the alignment of motif, where $A_i$ is a set of candidate binding sites in $S_i$.
1) Background model $B$ is pre-computed from the whole genome. For each potential binding sites, precompute the background information under $B$.
2) Initialize motif model parameter $\theta$ and structure $G$ by randomly selecting a set of binding sites (and the corresponding motif alignment $A$).
3) Sampling step. Randomly select a sequence $S_i$ to perform sampling. For each potential binding site $s_{i,j}$ on $S_i$, compute the conditional probability $P(s_{i,j}|A_{[-i]})$, where $A_{[-i]} = \cup_{j\neq i}A_j$. Then a set of sites is chosen and the model parameter $\theta$ is updated accordingly.
4) Model structure $G$ is updated based on the current alignment using Algorithm 1.
5) Perform phase shifting to avoid getting locked into a suboptimal solution.
6) Repeat Steps 3-5 for a preset number of rounds and report the motif as output.
7) Repeat Steps 2-6 to obtain a preset number of motifs.
```

Building the model structure. After obtaining the dependence scoring matrix, we build the model structure by a greedy procedure (Algorithm 1), which tries obtaining as high total dependencies as possible.

B. Gibbs Sampler

We extend the Gibbs sampler [3], [12], [13] to update iteratively between motif alignment A and motif model $\langle\theta, G\rangle$ until converging to a local optimal motif. In each iteration, we pick a sequence S_i and update the model parameter θ based on the current alignment. Then the model structure G is updated. See Algorithm 2. Due to space limit, we only elaborate the steps with proposed improvement.

Step 3. The core computation is to compute the predictive distribution in this step. A sequence S_i is randomly selected and a set of candidate binding sites in S_i is selected. The predictive distribution $P(s_{i,j}\,|A_{[-i]})$ for site $s_{i,j}$ of S_i is evaluated by a weight

$$W_{i,j} = \frac{P(s_{i,j}|M)}{P(s_{i,j}|B)}\ , \quad (4)$$

where the site probability $P(s_{i,j}|M)$ is computed by Eq. (1), (2), or (3) depending on the models, and $P(s_{i,j}\,|B)$ is computed in Step1 and is the same for all models for a given motif width.

We then sample multiple copies of candidate sites. Here we propose a new method to control the variability (or conserveness) of the predicted motif. To this end, we need a score to measure the degree of variability of a potential site against a motif, and a threshold to determine whether a site is selected.

The threshold. Let us consider the di-SPWDM with model parameter $[p_{x,j}]$, for $1 \leq x \leq 16$ and $1 \leq j \leq w$. The threshold makes use of the maximum value of the j-th column $max_x(p_{x,j})$, and a parameter v, called *variability index*, to control the variability of the predicted motif.

$$threshold = \left(\sum_{j=1}^{\omega} \frac{(\sum_{x=1}^{16} p_{x,j}) - max_x(p_{x,j})}{15 \times \omega \times max_x(p_{x,j})}\right)^{v} . \qquad (5)$$

Scoring. We then evaluate the variability of a site. This involves a user defined parameter m which indicates the number of sites to sample. First, we rank all sites of S_i in decreasing order of their weights $W_{i,j}$ (Eq. (4)). By this ranking we choose a subset with the highest weight, denoted as $\{s_{i,1}, \ldots, s_{i,m}\}$. Second, we score these sites from $s_{i,1}$. Suppose when we score $s_{i,k}$, the value of the current motif probability matrix is $[p_1^{(k)}, \ldots, p_\omega^{(k)}]$.

$$score(s_i, k) = \prod_{j \in [1,\omega]} \frac{min_{r \in [1,k]}(p_j^{(r)})}{max_x(p_{x,j})} . \qquad (6)$$

The term $min_{r \in [1,k]}(p_j^{(r)})$ means that we choose the minimum p_j from the sites $s_{i,1}, \ldots, s_{i,k}$ with weight higher than $s_{i,k}$ (including $s_{i,k}$). This is to reduce the variability among the selected sites in a position specific way. A site is selected if its score (Eq. (6)) is above the threshold (Eq. (5)), otherwise the sampling on this sequence is stopped.

Step 6. Steps 3-5 are repeated for a preset number of rounds. We measure the motif significance to allow ranking in Step 7. The measurement should reflect $W_{i,j}$ and the number of binding sites (larger the better). We adopt the *sum of the weights*, denoted by SW-score, of all candidate sites. SW-score takes into account different model structure G (Eq. (4)). It is able to locate more true positive sites than traditional Information Content (IC) criterion [17].

III. Experiments

We use both simulated and real data for evaluation. The real data include the species *Aspergillus*, human and yeast.

Algorithms tested. We test four algorithms for different underlying motif models. We denote the algorithms as A1 for PWM model, A3 for triPWDM, A2$_\chi$ for diSPWDM$_\chi$, and A2$_{\max}$ for diSPWDM$_{\max}$. The algorithms are coded and tested in MATLAB and the experiments are ran on a PC with a 2.5GHz Xeon CPU and 8GB memory.

Statistical measurement. The performance of the algorithms is measured by the statistical value called performance coefficient (CC), which depends on true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN). The formula used is described in [19]. The higher the value the better the algorithm is.

A. Test on Simulated Data

Data generation. We generate 50 gene sets each consisting of 20 sequences of length 500bp. These artificial sequences are simulated by a background Markov model; a 3rd-order Markov model trained from the upstream region of all genes of *Aspergillus nidulans*.

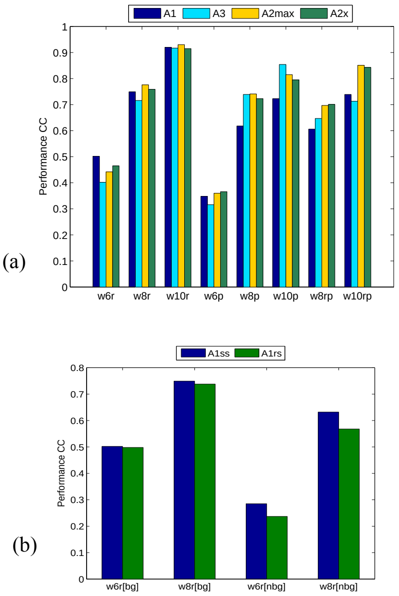

Figure 2. (a) Performance (CC) of A1, A3, A2max and A2χ on simulated data sets. Data sets include w6r, w8r, w10r, w6p, w8p, w10p, w8rp and w10rp. (b) Comparing algorithms with different initializations; A1ss (our scheme): A1 initialized by using seed sites; A1rs (traditional scheme): A1 initialized by using random motif alignment. The data sets tested include w6r[bg]([nbg]) and w8r[bg]([nbg]), where [bg] and [nbg] indicates predicting with and without background information, respectively.

The motif width we use is 6, 8, and 10bp. We limit the number of nucleotide positions having nucleotide dependencies to be at most half of the motif width, e.g., 2-4 for 10bp motifs. We generate three types of motifs based on PWM, triPWDM, and diSPWDM, denoted as type *r*, *p*, and *rp*, respectively. E.g., a *w6r* motif is a width-6 motif under the PWM model. The motifs generated are randomly implanted into the artificial sequences such that each gene set contains one motif. Due to space limit, we skip the details of the parameters used by the algorithm.

Results. Fig. 2(a) shows the overall performance CC of each algorithm on different data sets. Perhaps not suprisingly, different algorithms perform well on different types of data sets, especially for the data set in which the implanted motif is generated under the same motif model. Algorithms A1 and A3 only perform well in the corresponding data type, i.e., *r* and *p*, respectively, and performance deteriorates greatly for other data types. On the other hand, A2$_\chi$ and A2$_{\max}$ perform the best for data type *rp*, while having reasonably good performance on the other data types.

Initialization in Step 2 of Gibbs sampler. A1ss is A1 using our seed site selection and A1rs for traditional random selection. We also carry out two tests, with and without background information for Step3 in calculating the weight. Fig. 2(b) shows that with no background information, A1ss is better than A1rs, and their performance is comparable with background information. The performance with background information is much better than without background.

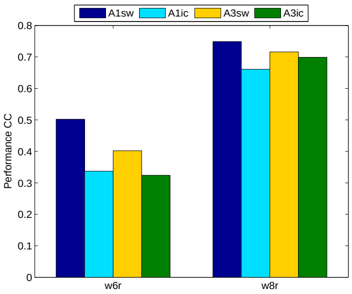

Figure 3. Comparing different motif scores; A1sw and A3sw: results of A1 and A3 using SW score; A1ic and A3ic: results of A1 and A3 using IC score. Data sets tested are w6r and w8r.

Ranking of motifs using SW-score. As discussed in Step 6, we rank motifs using sum of weight SW-score instead of Information Content (IC). We test this using both A1 and A3, with A1ic and A1sw means A1 with IC and SW, respectively (similarly for A3). Fig. 3 supports that ranking by SW-score is better than IC-score.

B. *Test on Real Data*

To examine the applicability of the algorithms in practice, we conducted tests on real data.

Aspergillus. The *Aspergillus* data (A. nidulans, A. fumigatus, A. oryzae) include 11 co-expressed gene sets. Two sets are provided by our project team, while the others are extracted from relevant biological literatures [20], [21], [22], [23], [24]. The number of sequences in a gene set ranges from 18 to 52, and all sequences are upstream region of gene with length 1000 bp. The motif width ranges from 6 to 9 bp. The binding sites of motifs are localized by scanning the sequences according to the reported motif consensus.

To indicate how overrepresented these motifs are, we introduce a quantity called *motif strength*, which equals *SW(motif)/*N_{kbp}, where N_{kbp} is the total number of nucleotides divided by 1000. We notice that it would be difficult to predict motifs for Asp1 and Asp7 while it would be easier for Asp4 and Asp5. On the other hand, it is indicated in the literature that for Asp11, there are strong mutual dependencies between nucleotides at positions 2, 6 and 8 within the motif.

Results. The results are shown in Table I. The best CC value in each gene set is highlighted. We notice that there is no clear winner for all the cases, yet, for cases with reasonable CC value (i.e., except Asp1 and Asp7), A2$_\chi$ achieves the highest CC for 4 out of 9 cases, while the other algorithm achieves the highest for 1 or 2 cases. In particular, for Asp11 that we mentioned above there are strong interdependencies among the motif, A2$_\chi$ outperforms other algorithms (A2$_{max}$ performs better than A1 and A3 as well but for a smaller margin), supporting the claim that A2 is able to capture interdependencies.

Tompa's data set. To give a comparison with other motif finders, we conducted the test on the data used by Tompa et al. [19], which was extracted from TRANSFAC database, the data can be obtained at website: http://bio.cs.washington.edu/assessment/. The data consists of a total of 56 sequence datasets for 4 different species including fly, human, mouse and yeast.

TABLE I. PERFORMANCE CC OF ALGORITHMS A1, A3, A2MAX AND A2χ FOR 11 ASPERGILLUS DATA SETS.

Data	A1	A3	A2max	A2χ	Average
Asp1	**0.091**	0.028	0.046	0.041	0.052
Asp2	0.396	0.341	**0.462**	0.394	0.398
Asp3	0.534	0.455	0.399	**0.54**	0.482
Asp4	0.56	**0.682**	0.484	0.511	0.559
Asp5	0.532	0.5	0.451	**0.785**	0.567
Asp6	**0.309**	0.193	0.192	0.182	0.219
Asp7	0.019	0.062	**0.073**	0.035	0.047
Asp8	0.409	0.274	0.343	**0.415**	0.36
Asp9	**0.682**	0.399	0.5	0.542	0.531
Asp10	0.224	**0.306**	0.215	0.271	0.254
Asp11	0.246	0.215	0.263	**0.304**	0.257

TABLE II. PERFORMANCE CC OF A2χ, A2MAX, MEME, ALIGNACE, IMPROBIZER AND MOTIFSAMPLER ON TOMPA'S DATA SET.

A2χ	A2max	MEME	AlignA-CE	Improb-izer	MotifSa-mpler
0.076	0.068	0.071	0.066	0.055	0.067

We tested the algorithmA2$_\chi$ and A2$_{max}$. The test setting is as follows. As we do not known the motif width in advance, each dataset has been tested with motif width of 6, 8, 10, 12, 14, 16, 18, 20 bp. The number of motifs to search ranges from 200 to 500 depending on the number of nucleotides of each dataset.

The motif finders we compared to here include MEME, AlignACE, Improbizer and MotifSampler. MEME is based on EM algorithm, and AlignACE, Improbizer and Motif-Sampler are based on Gibbs sampler. The detailed test results and parameter settings for these motif finders can be found at website mentioned above. The comparison results is shown in Table II. We note that A2$_\chi$ has slightly better performance than other motif finders.

IV. DISCUSSION AND CONCLUSION

In this work, we explore DNAmotif finding with adjacent and nona-djacent positional interdependencies. For adjacent interdependencies, we devise the triPWDM model which is capable of capturing interdependencies within 3 neighboring nucleotides. For non-adjacent interdependencies, we devise the diSPWDM model (diSPWDM$_\chi$ and diSPWDMmax) which is capable of dynamically capturing pairing dependencies at any two positions in the motif. The strength of the dependencies in diSPWDM$_\chi$ and diSPWDM$_{max}$ is respectively measured by χ^2 score and a score computed by conditional probabilities between pairing nucleotides.

We adopt Gibbs sampling approach to sample the target motifs based on the defined models. For the diSPWDM model, a model structure updating step is incorporated to the Gibbs sampler so as to implement the model structure converging.

To make Gibbs sampler more efficient, we also provide some improvements for it, for instance, the seed site initialization method, to handle the seed site sampling space, and multi-copies sampling technique for sampling sites at one sequence, and a new motif scoring criterion for ranking the predicted motifs.

Our experiments demonstrate the capability and applicability of four algorithms with different models on both simulated data and real data. The results show that the algorithms have their own strengths for predicting motifs under the same model as the one they are based on, but may perform poorly for motifs under different models. Among these, $A2_{\chi}$ and $A2_{max}$ are able to achieve good performance over different type of data sets under different motif models, indicating this approach can be fruitful. We also illustrate the effectiveness of the improvements on Gibbs sampler.

Comparison with other model. As introduced in Section I, the idea of SPWDM is similar to GWM in [11], where both employing dynamic model structure which can be updated during the sampling process. However, compared to the GWM model, SPWDM is able to capture more interdependencies between nucleotides than GWM, because GWM model is constructed only by pairing non-overlapped nucleotides constraining the dependences between paired positions. In SPWDM any nucleotide is allowed to have two connections with other nucleotides, extending the dependencies between positions. Therefore, SPWDM should suffer less overfitting problem. Consequently, algorithms under SPWDM model are able to find position-dependent motifs and at the same time are less likely to miss random motifs. Our model structure updating method is also different. GWM makes use of MCMC Metropolis-Hastings sampling method to sample a group of models and select one with the largest posterior, while we use Gibbs sampling for model converging (both parameter and structure) which is expected to be more time-efficient than the Metropolis-Hastings sampling method.

In practical applications, given the motif width, perhaps the most difficult parameter to be decided is the variability index, which has a big influence on the predicted motifs. One option is to try out a range of values and see how variable the predicted motifs are, then select the most appropriate value within that range.

Our tests on real data show how hard it is to accurately predict motifs on gene sets with a weak target motif and high nucleotide noise. In addition to the background information, we may need more biological prior knowledge to strengthen weak motif predicting. For example, ChIPchip data localizes TF binding sites to much shorter DNA sequences [25], [26], which in turn reduce the noise of data. Information about the phylogenetic relationship between species is increasingly used in *de novo* motif finding and makes the predicted motifs more biologically relevant [27], [28].We can see that one of the future challenges posed for this area is how to make full use of different kinds of prior knowledge to support motif discovering.

References

[1] A. Brazma, I. Jonassen, I. Eidhammer, and D. Gilbert, “Approaches to the automatic discovery of patterns in biosequences,” J. Comput. Biol., vol. 5, no. 2, pp. 279–305, 1998.

[2] T. Bailey and C. Elkan, “Fitting a mixture model by expectation maximization to discover motifs in biopolymers,” in Proc. Int. Conf. Intell. Sys. for Mol. Bio., 1994, pp. 28–36.

[3] C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton, “Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment,” Science, vol. 262, pp. 208–214, 1993.

[4] T. Man and G. Stormo, “Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay,” Nuc. Acids Res., vol. 29, no. 12, pp. 2471–2478, 2001.

[5] M. L. Bulyk, P. L. Johnson, and G. M. Church, “Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors,” Nuc. Acids Res., vol. 30, no. 5, 2002.

[6] E. Xing and R. Karp, “MotifPrototyper: a Bayesian profile model for motif families,” Proc. Nat. Acad. of Sc., vol. 101, no. 29, pp. 10 523–10 528, 2004.

[7] M. Zhang and T. Marr, “A weight array method for splicing signal analysis,” Comp. App. in Biosc., vol. 9, no. 5, pp. 499–509, 1993.

[8] P.Agarwal andV. Bafna, “Detecting non-adjoining correlations with signals in DNA,” RECOMB, 1998.

[9] C. Burge and S. Karlin, “Prediction of complete gene structures in human genomic DNA,” J. Mol. Bio., vol. 268, no. 1, pp. 78–94, 1997.

[10] Y. Barash, G. Elidan, N. Friedman, and T. Kaplan, “Modeling dependencies in protein-DNA binding sites,” RECOMB, 2003.

[11] Q. Zhou and J. S. Liu, “Modeling within-motif dependence for transcriptionfactor binding site predictions,” Bioinformatics, vol. 20, no. 6, pp. 909–916, 2004.

[12] J. S. Liu, A.F. Neuwald, and C. E. Lawrence, “Bayesian models for multiple local sequence alignment and Gibbs sampling strategies,” J. American Stat. Ass., vol. 90, no. 432, pp. 1156–1170, 1995.

[13] A. F. Neuwald, J. S. Liu, and C. E. Lawrence, “Gibbs motif sampling: detection of bacterial outer membrane protein repeats,” Protein Science, vol. 4, no. 8, pp. 1618–1632, 1995.

[14] J.D. Hughes,P.W. Estep,S.Tavazoie, andG.M. Church, “Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae,” J. Mol. Bio., vol. 296, no. 5, pp. 1205–1214, 2000.

[15] X. Liu, D. L. Brutlag, and J. S. Liu, “BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes,” PSB, pp. 127–138, 2001.

[16] W. Thompson, E. C. Rouchka, and C. E. Lawrence, “Gibbs recursive sampler: finding transcription factor binding sites,” Nuc. Acids Res., vol. 31, no. 13, pp. 3580–3585, Jul1 2003.

[17] G. Thijs,K. Marchal,M. Lescot,S. Rombauts,B.D. Moor,P. Rouze, andY. Moreau, “AGibbs sampling method to detectoverrepresented motifs in the upstream regions of coexpressed genes,” J. Comp. Bio., vol. 9, no. 2, pp. 447–464, 2002.

[18] C. T. Workman and G. D. Stormo, “ANN-Spec: a method for discovering transcriptionfactor binding sites with improved specificity,” PSB, pp. 467–478, 2000.

[19] M.Tompa, N. Li, andT. L. B. et al., “Assessing computational tools for the discovery of transcriptionfactor binding sites,” Nat. Biotech., vol. 23, no. 1, pp. 137–144, 2005.

[20] M. Schrettl, H. Kim, M. Eisendle, C. Kragl, W. Nierman, T. Heinekamp, E.Werner, I. Jacobsen,P. Illmer, H.Yi, A. Brakhage, and H. Haas, “SreA-mediated iron regulation in Aspergillus fumigatus,” Mol. Microbio., vol. 70, no. 1, pp. 27–43, 2008.

[21] J. Mogensen, H. B. Nielsen, G. Hofmann, and J. Nielsen, “Transcription analysis using high-density micro-arrays of Aspergillus nidulans wild-type and creA mutant during growth on glucose or ethanol,” Fungal Genetics and Biology, vol. 43, no. 8, pp. 593–603, 2006.

[22] M. R. Andersen, W. Vongsangnak, G. Panagiotou, M. P. Salazar, L. Lehmann, and J. Nielsen, “A trispecies aspergillus microarray: Comparative transcriptomics of three Aspergillus species,” Proc. Nat. Acad. of Sc., vol. 105, no. 11, pp. 4387–4392, 2008.

[23] K. Sakamoto, T. H. Arima, K. Iwashita, O. Yamada, K. Gomi, and O. Akita, “Aspergillus oryzae atfB encodes a transcription factor required for stress tolerance in conidia,” Fungal Genetics and Biology, vol. 45, no. 6, pp. 922–932, 2008.

[24] D. Hagiwara,A.Kondo,T. Fujioka, andK. Abe, “Functional analysis of C2H2 zinc finger transcription factor CrzA involved in calcium signaling in Aspergillus nidulans,” Cur. Gene., vol. 54, no. 6, pp. 325–338, 2008.

[25] X. Liu, D. Brutlag, and J. Liu, “An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments,” Nat. Biotech., vol. 20, no. 8, pp. 835–839, 2002.

[26] X. Chen, L. Guo, Z.Fan, andT. Jiang, “W-AlignACE: an improved Gibbs sampling algorithm based on more accurate position weight matrices learned from sequence and gene expression/ChIP-chip data,” Bioinfo., vol. 24, no. 9, pp. 1121–1128, 2008.

[27] L. A. Newberg,W. A. Thompson, S. Conlan,T. M. Smith, L. A. Mc-Cue, and C. E. Lawrence, “Aphylogenetic Gibbs sampler that yields centroid solutions for cis-regulatory site prediction,” Bioinformatics, vol. 23, no. 14, pp. 1718–1727, 2007.

[28] R. Siddharthan, E. D. Siggia, and E. van Nimwegen, “PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny,” PLoS Comp. Bio., vol. 1, no. 7, p. e67, 2005.

Huihai Wu was born in Jiangsu China, 1975. He received Bachelor in Communication & Information Engineering from the Department of Radio Engineering, the Southeast University, Nanjing China, in 1997, and PhD degree in Bioinformatics from the School of Information Systems, Computing and Mathematics, the Brunel University, United Kingdom, in 2009. He was a Postdoctoral Fellow at the Department of Computer Science, the University of Liverpool, United Kingdom. He is currently serving as Bioinformatics Experimental Officer at the Faculty of Health and Medical Sciences, the University of Surrey, United Kingdom. His research interests include Bioinformatics, systems biology, computational biology and statistical data analysis for biological data.

Prudence W.H. Wong was born in Hong Kong. She received her PhD degree from the University of Hong Kong in 2003. She is currently serving as a Senior Lecturer at the Department of Computer Science, the University of Liverpool, United Kingdom. Before joining the University of Liverpool in 2004, she was a Postdoctoral Fellow at the University of Hong Kong. Her research interests include on-line algorithms and computational biology.

Mark X. Caddick was born in United Kingdom, 1958. He received his BSc(Hons) degree from the University of Liverpool, and PhD degree from the University of Newcastle upon Tyne. He is currently serving as Professor and Head of the Department of Functional and Comparative Genomics, the University of Liverpool, United Kingdom. His research interests include the molecular and genetic analysis of gene expression and associated regulatory mechanisms.