A Gibbs Sampling Algorithm for Motif Discovery Using a Linear Mixed Model

Daming Lu
Computer Science Department
University of New Orleans
New Orleans, U.S.A
1-504-451-4037

dlu@cs.uno.edu

ABSTRACT

The identification of motifs in the gene promoters is a critical step in the delineation of the genetic regulatory framework of an organism. In this paper, a new linear mixed model is introduced. This model is a combination of the conventional Position Weight Matrix (PWM) model and a novel Mutual Information (MI) model. PWM can contain individual position frequencies whereas MI can reflect pair wise relation between positions. A training stage is carried out to determine the weight of each model. After that this trained model is embedded into a Gibbs sampling algorithm for motif discovery. After analyzing a set of DNA sequences using this program, putative motifs are gained and compared with experimental verified motifs as well as other popular motif finding software. Results show that this new mixed model can improve motif discovery accuracy to some extent.

Categories and Subject Descriptors

I.2.8 [**PROBLEM SOLVING, CONTROL METHODS AND SEARCH**]

General Terms

Algorithms, Design.

Keywords

Gibbs sampling; Motif discovery; Transcription factors; Mutual information; Mixed motif model.

1.INTRODUCTION

Transcription factors can bind to short DNA segments in the regulatory regions to control their expression. The common pattern of these short DNA segments bound by transcription factor is called motif. Identifying various motifs and the relationships among them is one of the most challenging problems in contemporary bioinformatics area. To date, many computational approaches have been developed to identify motifs via finding overrepresented and conserved DNA segments (putative motif instances) in the regulatory regions of a group of candidate genes [1-6]. These approaches can be classified into two categories: the experimental biology approach and the computational biology approach. The former is of high accuracy, but, as tradeoff, is both time and labor consuming. By contrast, computational approach is of high throughput and efficiency, but it requires large data and is lack of accuracy. Motif discovery via computational methods became feasible as large-scaled public biological database and high performance computers appeared. Consequently, many computational methods were invented and gained satisfactory achievements, including Gibbs sampling [7], EM algorithm [8], etc. These methods applied relative entropy as criterion to evaluate the truthfulness of functional DNA sequences as well as to ascertain their locations. Derived from these methods, many computer software were created. MEME succeeded in finding multiple motifs by applying expectation-maximization algorithm to fit a two-component finite mixture model to the data [8]. AlignAce managed to find motifs that often correspond to the DNA binding preferences of transcription factors using Gibbs sampling [9]. Bioprospector has the feature of modeling gapped and palindrome pattern motifs [10].

In this paper, Gibbs sampling algorithm, which is a popular method for motif finding and has already been implemented by [9] and [10], is chosen for motif identification. Gibbs sampling is essentially a general stochastic strategy for determining the parameters of a statistical model relative to a given data set [11]. This strategy starts with certain initialization of parameter values and iteratively changes the value of one parameter at a time by assuming that the remaining parameters are correct and invoking Bayes Theorem until all parameters converge to stable (if not optimal) values. For changing the value of one parameter, Gibbs sampling employs a position weight matrix (PWM) as the motif model and selects the maximum likelihood value (highest score) for this specific parameter.

However, recent research showed that PWM model is not proper as a representation for motifs because it assumes that all possible positions are independent whereas relationship between positions often plays a significant role [12]. Motivated by this, a novel mutual information (MI) motif model is introduced to reflect the relationship between pair wise positions. Mutual information is a concept in information theory that provides a general measure of dependencies between variables. A larger

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
ISB'10, February 15-17, 2010, Calicut, India.
Copyright 2010 ACM 978-1-60558-722-6/10/02…$10.00.

mutual information indicates a stronger association between two random variables.

We construct a linear mixed model of PWM and MI with the intent to both capture frequencies of individual motif position and reflect relationship between pair wise motif positions as well. Barrios, etc. collected 186 σ^{54}-dependent promoter sequences with conserved positions around -24 and -12, [13]. The weights (parameters) in this linear model are trained by half of the sequences. Then this well-trained model is implanted into Gibbs sampling algorithm as a substitute for the conventional PWM. The other half of the DNA sequences are used for testing. The result is compared with those from both MEME and Bioprospector in the aspects of sensitivity and specificity.

2.MOTIF REPRESENTATION

In this section, we will introduce the conventional PWM model as well as the novel MI model. For convenience, some notations further used in this section are introduced first. Let $B = \{A, C, G, T\}$ be the four nucleotide letters of which DNA sequences are composed and w be the motif width. A sequence $D = d_1, d_2, .. d_w$ is used for matching the two models.

2.1 PWM Model

PWM is a universal way to represent DNA motifs. In PWM, it assumes that all positions in a given motif are completely independent. PWM composes 4 rows to represent 4 different types of nucleotide acids. The length of PWM is equal to motif width w. So a $4 \times w$ matrix is constructed. The score for any specific sequence D is the sum of the matrix values for this sequence [14].

$$ScorePWM\ (D) = \sum_{i=1}^{w} \log P(d_i, i), \qquad d_i \in B$$

where $P(d_i, i)$ is the probability of base d_i at position i.

2.2MI Model

In PWM model, the scores for each position are summed up together to get the total score, which implies that each position contributes independently. The fact, however, is that dependency between positions are significant in short motifs. Zhang and Marr elaborated that there exists weak pair wise correlations within the signals (positions) and they presented a weight array method (WAM) that can help to better discriminate these signals [15]. WAM calculates the conditional probability of nucleotide d_i at position i, given nucleotide d_{i-1} at position $i-1$. It works as follows:

$$ScoreWAM\ (D) = \log(P(d_1, 1)) + \sum_{i=2}^{w} \log P(d_i, i \mid d_{i-1}, i-1),\ d_i \in B$$

WAM suffers from taking only one position ahead into account for calculating a specific position while variable gaps exist among motifs [16].

Mutual information is a quantity that measures the mutual dependence of the two variables and it can reflect pair wise correlation between two positions with gaps [17]. Its mathematical definition is shown as follows:

$$MI(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log \frac{p(x, y)}{p(x)\,p(y)}$$

where $p(x, y)$ is the joint probability of *X* and *Y*.

Applied into our case, the expression is :

$$ScoreMI\ (D) = \sum_{i=1}^{w-1} \sum_{j=i+1}^{w} P(d_i, i\,; d_j, j) \log \frac{P(d_i, i\,; d_j, j)}{P(d_i, i)\ P(d_j, j)} \quad d_i, d_j \in B$$

where $P(d_i, i\,; d_j, j)$ means the joint probability of base d_i at position i and base d_j at position j.

2.3Linear Mixed Model

The above two models are called base models or ground models, meaning each includes a probability estimate over the observed sequences. A linear mixed model is a combination of the two base models and can breed a better performance. In this paper, we combine the two models linearly as follows:

$$Model_{mixed} = \frac{w_{PWM}\,Model_{PWM} + w_{MI}\,Model_{MI}}{w_{PWM} + w_{MI}}$$

In this equation, each model contributes one portion to the mixed model. $Model_{PWM}$ and $Model_{MI}$ have different orders of magnitude so we normalize each first before combine them. Assume we have a sequence $D = d_1, d_2, .. d_l,\ \ l > w$, we calculate the PWM score and the MI score of all possible motif instances starting from $i = 1, 2, ..., l - w + 1$ and store the scores into two arrays, $arrPWM$ and $arrMI$. In the next stage we normalize each array as follows:

$$NarrPWM\,[i] = \frac{arrPWM\,[i] - \min(arrPWM\,)}{\max(arrPWM\,) - \min(arrPWM\,)};\ i = 1, ..., l - w + 1$$

$$NarrMI\,[i] = \frac{arrMI\,[i] - \min(arrMI)}{\max(arrMI) - \min(arrMI)};\ i = 1, ..., l - w + 1$$

Then we select the new motif instance with the highest score from the linear mixed model.

3.MOTIF DISCOVERY

3.1Gibbs Sampling Overview

Gibbs sampling is an algorithm to generate a sequence of samples from the joint probability of multivariable. It requires a random starting point of parameters of interest, $x_1, x_2, ..., x_k$. The marginal distribution is essentially too complicated to get, so is the joint distribution, but the conditional distribution of each variable is comparably easy to gain. Assume that the full conditional distributions $p(x_i \mid x_j; j \neq i),\ i = 1, 2, .. k$ are available. From a random starting point, Gibbs sampling algorithm draws samples $x_1, x_2, ..., x_k$ in the following manner:

$$
\begin{aligned}
x_1^{(t+1)} &\sim p(x_1 \mid x_2 = x_2^{(t)},\ldots,x_k = x_k^{(t)}) \\
x_2^{(t+1)} &\sim p(x_2 \mid x_1 = x_1^{(t+1)}, x_3 = x_3^{(t)},\ldots,x_k = x_k^{(t)}) \\
&\vdots \qquad \vdots \qquad \vdots \\
x_i^{(t+1)} &\sim p(x_i \mid x_1 = x_1^{(t+1)},\ldots,x_{i-1} = x_{i-1}^{(t+1)}, x_{i+1} = x_{i+1}^{(t)},\ldots,x_k = x_k^{(t)}) \\
&\vdots \qquad \vdots \qquad \vdots \\
x_k^{(t+1)} &\sim p(x_k \mid x_1 = x_1^{(t+1)},\ldots,x_{k-1} = x_{k-1}^{(t+1)})
\end{aligned}
$$

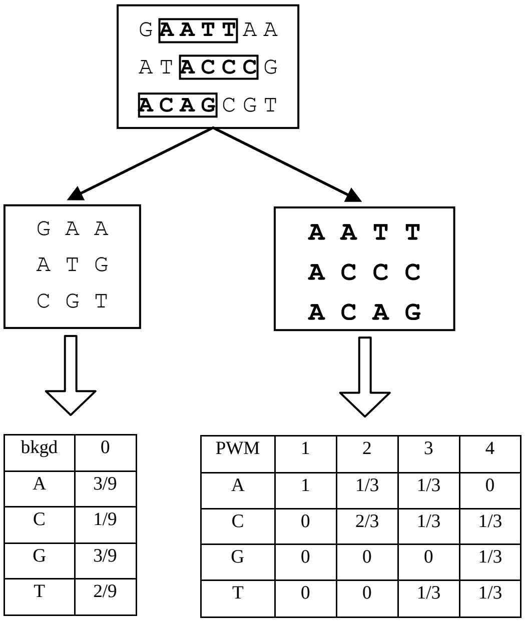
where *t* means the *t*-th iteration.

Geman and Geman [18] showed that when $t \to \infty$, the distribution of $(x_1^{(t)}, x_2^{(t)},\ldots,x_k^{(t)})$ converges to $p(x_1, x_2,\ldots x_k)$.

3.2Update Step

At each update step, 3 new matrices for background, PWM and MI, are constructed, respectively. To make it more clear, we provide an instance as follows:

Suppose we have 3 multi-aligned sequences with predicted motifs (see Figure 1).



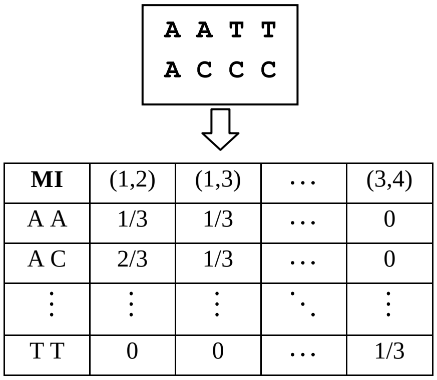
bkgd	0
A	3/9
C	1/9
G	3/9
T	2/9

PWM	1	2	3	4
A	1	1/3	1/3	0
C	0	2/3	1/3	1/3
G	0	0	0	1/3
T	0	0	1/3	1/3

Figure 1. An illustration. Sequence length is 7, motif width is 4. Construct background matrix *bkgd* and PWM matrix *PWM*.

We construct background matrix as well as PWM matrix as in Figure 1. and also construct MI matrix as in Figure 2.

Obviously, background matrix is 4 by 1, PWM matrix is 4 by *motifWidth* and MI matrix is 16 by *motifWidth*(motifWidth-1)/2*. In practice, we substitute pseudo zero for real zero in the matrix as smoothing parameter.



MI	(1,2)	(1,3)	$\cdots$	(3,4)
A A	1/3	1/3	$\cdots$	0
A C	2/3	1/3	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
T T	0	0	$\cdots$	1/3

Figure 2. An illustration to construct MI matrix. It has $4 \times 4 = 16$ rows and $w \times (w-1)/2$ columns.

3.3Training the Weight Parameters

In order to train the weights for each model, w_{PWM} and w_{MI} respectively, we chose a set of DNA sequences with known motifs (half of the 186 sequences). We scanned the range from $w_{PWM} = 1, w_{MI} = 0$, which degenerates to conventional Gibbs sampling algorithm, to $w_{PWM} = 0, w_{MI} = 1$, which is a model of all mutual information without individual base frequency. The step length is set to be 0.02.

As is shown in Figure 3, the performance goes up when the weight of MI model increases. We show the performance via the online software, WebLogo [19].

At the point $w_{PWM} = 0.86, w_{MI} = 0.14$, the performance reaches its maximum. Then it goes down to almost zero as w_{MI} goes up. Our explanation is that , as Zhang and Marr mentioned in [15], pair wise correlations within signals are rather weak. So if we give this weak correlation a high weight, the result would be quite poor. After scanning the range [0,1], we choose $w_{PWM} = 0.86, w_{MI} = 0.14$.

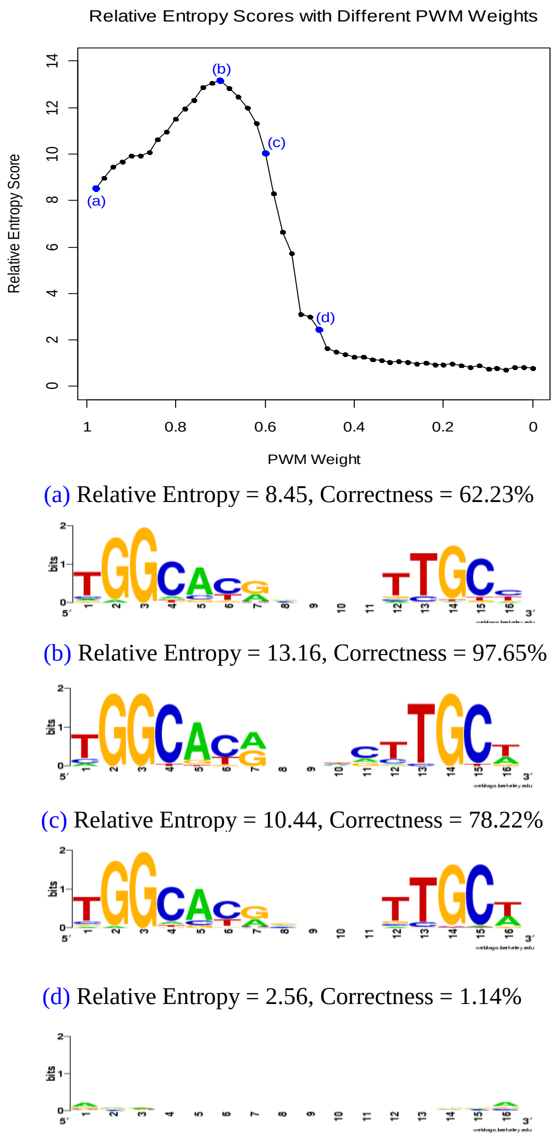
3.4Model Initialization

In initialization, a 4-parameter Dirichlet distribution is sampled to fill the cells in PWM [21]. The Dirichlet distribution is shown as follows:

$$D(\theta \mid \alpha) = Z^{-1}(\alpha)\prod_{i=1}^{k}\theta_i^{\alpha_i - 1}\delta(\sum_{i=1}^{k}\theta_i - 1)$$

where $\alpha = \alpha_1,\ldots,\alpha_k$ are constants that specify this Dirichlet distribution. θ_i satisfies $0 \le \theta_i \le 1$ and $\sum \theta_i = 1$.

Similarly, a 16-parameter Dirichlet distribution is sampled for the MI model.



(a) Relative Entropy = 8.45, Correctness = 62.23%

(b) Relative Entropy = 13.16, Correctness = 97.65%

(c) Relative Entropy = 10.44, Correctness = 78.22%

(d) Relative Entropy = 2.56, Correctness = 1.14%

Figure 3. Relative Entropies and correctnesses with different PWM/MI weights

3.5Convergence Criterion

Convergence criterion of Gibbs sampling algorithm is a profound topic in statistics. Zellner A. and Min C. presented three new operational convergence criteria for Gibbs sampler and related procedures that are useful for determining whether they not only have converged but also have converged to provide reliable results, [20]. But they are complicated and thus hard to implement. As a replacement in bioinformatics, relative entropy has been used to measure information content of a motif. Its performance was proved to be satisfactory [10]. So our convergence criterion is to choose the highest motif score, which is expressed through relative entropy as follows:

$$MotifScore = \exp\{ \sum_{all\ positions\ i} \sum_{all\ bases\ d} [P(d,i) \times \log(\frac{P(d,i)}{P(d)})]\}$$

where $P(d,i)$ means base d at position i in the motif model matrix and $P(d)$ means base d in the background matrix.

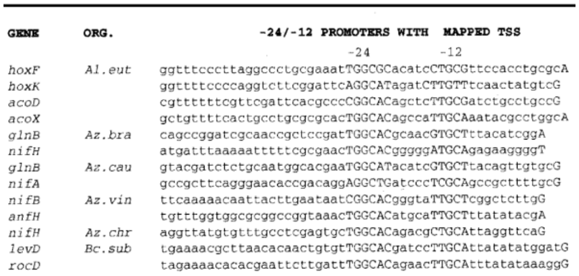
In practice, we judge it as convergence when above 95% predicted motifs stay unchanged between two iterations and choose the predicted motif with highest *MotifScore*. The brief algorithm is shown in Table 1.

Table 1. Gibbs Sampling Algorithm

```
Input: A set S of DNA sequences
Output: A candidate motif for S
begin
Randomly select initial motif instances in the sequences of S.
Initialize model PWM and MI as in 3.4
while not converged do
    Select sequence s from S
    Construct motif models from motif instances in S /{s}
    Score all possible motif instances in s using motif models
    Select a new motif instance x for s based on the scores
    Update new motif instance x for s
    if converged
     then record MotifScore for this run and end while
Output motif instances for S
end
```

4.DATA AND RESULT

Barrios, etc. collected 186 sigma-54 upstream promoters in 1999, with -24/-12 promoter sequences annotated. We use half of these data for training and the other half for testing. The consensus sequences sample shown in Figure 4 are depicted following a previously reported definition on *E.Coli* in [22], where any nucleotide occurring with a frequency of more than 6 standard deviations from the expected random occurrence of each nucleotide (0.25) is denoted as highly conserved (upper case), between three and six standard deviations is denoted weakly conserved (lower case), and below three standard deviations is not significant (N).



GENE	ORG.	-24/-12 PROMOTERS WITH MAPPED TSS
		-24 -12
hoxF	*Al.eut*	ggtttcccttaggccctgcgaaatTGGCGCacatcCTGCGttccacctgcgcA
hoxK		ggttttccccaggtcttcggattcAGGCATagatCTTGTTtcaactatgtcG
acoD		cgtttttcgttcgattcacgcccCGGCACagctcTTGCGatctgcctgccG
acoX		gctgttttcactgcctgcgcgcacTGGCACagccaTTGCAaatacgcctggcA
glnB	*Az.bra*	cagccggatcgcaaccgctccgatTGGCACgcaacGTGCTttacatcggA
nifH		atgatttaaaaattttcgcgaacTGGCACgggggATGCAgagaaggggT
glnB	*Az.cau*	gtacgatctctgcaatggcacgaaTGGCATacatcGTGCTtacagttgtgcG
nifA		gccgcttcagggaacaccgacaggAGGCTGatcccTCGCAgccgcttttgcG
nifB	*Az.vin*	ttcaaaaacaattacttgaataatCGGCACgggtaTTGCTcggctcttgG
anfH		tgtttggtggcgcggccggtaaacTGGCACatgcaTTGCTttatatacgA
nifH	*Az.chr*	aggttatgtgtttgcctcgagtgcTGGCACagacgCTGCAttaggttcaG
levD	*Bc.sub*	tgaaaacgcttaacacaactgtgtTGGCACgatccTTGCAttatatatggatG
rocD		tagaaaacacacgaattcttgattTGGCACagaacTTGCAtttatataaaggG

Figure 4. DNA sequences for training and testing

Around the -24 element, from position -31 to -20, there are eight highly conserved and three weakly conserved nucleotides with the sequence mrNrYTGGCACG. Around the -12 element, from position -15 to -8, there are five highly conserved and one weakly conserved nucleotides with the sequence TTGCWNNw. Therefore, the mrNrYTGGCACG-N4-TTGCWNNw sequence is the updated consensus for the -24/-12 promoters.

Our Gibbs algorithm is run 50 iterations with pre-trained weight parameters. Two other popular Gibbs-based motif discovery programs, Bioprospector and MEME, are also run on the same dataset as competitors. 20 random sequences generated from the background are inserted into the testing dataset as noise. As for Bioprospector, the parameters are set as -W 16 -d 1 -r 2 -a 1 , meaning it is an one-direction search with a motif width 16. Assuming there exists exactly one motif in each sequence, Bioprospector reports 2 top motifs. As for MEME, we choose one per sequence for the occurrences of a single motif that are distributed among the sequences. The other parameters remain as default. The result is shown in Table 2.

Table 2. Comparison of results from three methods

	sensitivity	specificity
this method	0.86	0.93
Bioprospector	0.83	0.91
MEME	0.79	0.88

The definitions of sensitivity and specificity are as follows:

$$sensitivity = \frac{\#of\ True\ Positives}{\#of\ True\ Positives + \#of\ \ False\ Negatives}$$

$$specificity = \frac{\#of\ True\ Negatives}{\#of\ True\ Negatives + \#of\ \ False\ Positives}$$

5.CONCLUSION

This article introduces a linear mixed model to represent conseved motif of functional DNA sequences. The mixed model gets over the defect that the PWM model alone contains only single position information or the MI model alone contains only pair wise information. During the training stage, the weights for both models are chosen for specific biological data for better performance. The result from the testing data shows that our method outperformed Bioprospertor and MEME in some data runs. It indicates that this method is feasible in motif discovery. Thus, the new model can be further applied in the field of motif discovery. The future involves the utilization of mutual information to overcome gaps between two motifs and the implementation of more complicated convergence criterion for Gibbs sampling. A computer program written in PERL is available. Please send email to dlu@cs.uno.edu.

6.ACKNOWLEDGEMENTS

The author would like to express great gratitude to Dr. Zhansheng Duan and Prof. Dongxiao Zhu for their comments on the manuscript.

7.REFERENCES

[1] Liu, Y., Liu, X.S, Wei, L. Altman, R.B., and Batzoglou, S. *Eukaryotic Regulatory Element Conservation Analysis and Identification using Comparative Genomics.* Genome Research. **14**, 451–458, 2004.

[2] Moses, A.M., Chiang, D.Y., and Eisen, M.B. *Phylogenetic Motif Detection by Expectation-maximization.* Pacific Symposium on Biocomputing. **9**, 325–335, 2004

[3] Prakash, A. and Tompa, M. *Discovery of Regulatory Elements in Vertebrates through Comparative Genomics*. Nature Biotechnology. **23**, 1249–1256, 2005.

[4] Prakash, A., Blanchette, M., Sinha, S., and Tompa, M. *Motif Discovery in Heterogeneous Sequence Data*. Pacific Symposium on Biocomputing. **9**, 348-359, 2004.

[5] Sinha, S., Blanchette, M. and Tompa, M. *PhyME: A Probabilistic Aalgorithm for Finding Motifs in Sets of Orthologous Sequences*. BMC Bioinformatics, **5**, 170, 2004.

[6] Wang, T. and Stormo, G.D. *Combining Phylogenetic Data with Co-regulated Genes to Identify Regulatory Motifs.* Bioinformatics, **19**, 2369–2380, 2003.

[7] Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F., and Wootton, J.C. *Detecting Subtle Sequence Signals: a Gibbs Sampling Strategy for Multiple Alignment.* Science **262**, 1993.

[8] Bailey, T.L., and Elkan, C. *Unsupervised Learning of Multiple Motifs in Biopolymers Using Expectation Maximization.* Machine Learning **21**, 51-80, 1995.

[9] Hughes, J.D., Estep, P.W., Tavazoie, S., and Church, G. M. *Computational Identification of Cis-regulatory Elements Associated with Groups of Functionally Related Genes in Saccharomyces Cerevisiae.* Journal of Molecular Biology **296,** 1205-1214, 2000.

[10] Liu, X., Brutlag, D.L., and Liu, J. S. *BioProspector: Discovering Conserved DNA Motifs in Upstream Regulatory Regions of Co-expressed Genes.* Pacific Symposium on Biocomputing., 127-138, 2001.

[11] Liu, J., Neuwald, A., Lawrence C. *Bayesian Models for Multiple Local Sequence Alignment and Gibbs Sampling Strategies*. Journal of the American Statistical Association, **90**, 432, 1156–1170, 1995.

[12] Bulyk, M.L., Johnson, P.L.F., and Church, G.M. *Nucleotides of Transcription Factor Binding Sites Exert Interdependent Effects on the Binding Affinities of Transcription Factors.* Nucleic Acids Research. **30**, 1255-1261, 2002.

[13] Barrios, H., Valderrama, B., and Morett, E. *Compilation and Analysis of* σ^{54} *-dependent Promoter Sequences*. Nucleic Acids Research. **27**, 4305-4313, 1999.

[14] Stormo, G.D. *DNA Binding Sites: Representation and Discovery*. Bioinformatics **16**, 16-23, 2000.

[15] Zhang, M.Q, and Marr, T.G. *A Weight Array Method for Splicing Signal Analysis.* Computer Applications in the Biosciences **5**, 499-509, 1993.

[16] Hu, Y. *Finding Subtle Motifs with Variable Gaps in Unaligned DNA Sequences.* Computer Methods and Programs in Biomedicine **70**, 11-20, 2003.

[17] Fatemeh Z., Hayedeh A., Mehdei S., Abbas N., and Bahram G. *New Scoring Schema for Finding Motifs in DNA Sequences.* BMC Bioinformatics **10**, 93-113, 2009.

[18] Geman S. and Geman D. *Stochastic Relaxation, Gibbs Distributioni, and the Bayes Restoration of Images.* IEEE Transaction on Pattern Analysis and Machine Intelligence, **6**, 721-741, 1984.

[19] Crooks, G.E., Hon, G., Chandonia, J.M, and Brenner, S.E. *WebLogo: A sequence logo generator.* Genome Research, **14**, 1188-1190, 2004.

[20] Zellner A. and Min C. *Gibbs Sampler Convergence Criteria*. Journal of the American Statistical Association, **90**, 921-927.

[21] Sjolander, K., Karplus, K., Brown, M., Hughey, R., Krogh, A., Mian, I.S., and Haussler, D. *Dirichlet Mixtures: A Method for Improving Detection of Weak but Significant Protein Sequence Homology.* Computer Applications in the Biosciences, **12**, 327-345, 1996.

[22] Diane K. H., and William R. M. *Compilation and analysis of Escherichia coli promoter DNA sequences*. Nucleic Acids Research, **11**, 2237-2255, 1983