A Combined Model and a Varied Gibbs Sampling Algorithm Used for Motif Discovery

Xiaoming Wu, Bo Wang, Changxin Song, Jingzhi Cheng

School of Life Science and Technology
Xi'an Jiaotong University,
Xi'an, 710049 China

xjtuwxm@sina.com

Abstract

The conserved sequences in gene regulatory regions dominate gene regulation. Discovering these sequences and their functions is important in post genome era. A novel model is constructed to represent conserved motifs of DNA sequences. This model is a combination of PWM and WAM models. The advantage is the new model not only can comprise individual base frequencies in the motifs, but also can embody relationship of neighbourhood bases. In addition, a varied Gibbs sampling algorithm is applied with consideration of the different motif occurrences in each sequence. This variation is more accordant with the true situation of gene transcription controlling mechanism. By combining the model and the discovery algorithm, a program is constructed. After analysed a set of DNA sequences of upstream regions of genes using this program, putative motifs are discovered and are compared to experimental verified regulatory sequences. Results showed that this combination is ideal for motif discovery and the practice is meaningful for gene regulation research.

Keywords: gene regulatory elements; motif discovery; gene expression analysis; mixture motif model.

1 Introduction

Uncovering the hidden mechanism of gene transcription control is a huge work in post genome era. Various methods have been invented to decipher the information encoded in DNA sequences. The approaches come from two ways: the biological experimental way or computational biology way. Biology experiment is accurate to locate the functional DNA subsequence in the genome sequences, but it is time and labour consuming. Conversely, computational way is high throughput and time saving, but it needs large amount of DNA sequences as prerequisite and is not very accurate. Motif discovery by computer programs became feasible as the publicly available of biosequences databases and high performance computers appearance. Consequently, many fundamentally computational methods to discover functional biosequences have occurred. Those methods include greedy algorithm developed by Hertz and Stormo(Hertz and Stormo, 1999a; Stormo and Hartzell, 1989), Gibbs sampling method introduced by Lawrence (Lawrence et al., 1993), and EM method used by Elkan (Bailey and Elkan, 1995). These methods use relative entropy as criteria to evaluate the truthfulness of functional DNA sequences and to locate their locations. Recently, other novel or more complex methods for motif discovery have occurred (Bajic and Seah, 2003; Bajic et al., 2003; Brazma et al., 1998; Buhler and Tompa, 2001; Buhler and Tompa, 2002; Jonassen et al., 1995). They either use many sequence features to discover important transcription elements (Bajic and Seah, 2003), or use suffix tree to discover high frequent patterns (Vilo, 2002). Buhler use random projection to find conserved biosequences and also obtained a good result(Buhler and Tompa, 2002). Li presented an efficient approximation and give out a polynomial time approximation scheme to this problem, but most result would be the local optimal, though very similar but not equal to the optima results (Li et al., 1999). Although these methods achieve certain success, and many computer programs have been developed based on them, the problem of motif discovery from DNA sequences still remains difficult because of its complex nature.

Up to now, many algorithms and their variations use Position Weight Matrix (PWM) models to represent motifs, but new research shows that this model is not very accurate in some cases, since it do not consider the independence of neighbourhood bases (Bulyk et al., 2002). However, a Weight Array Model (WAM) model has the characteristic to embody the relation between consecutive bases (Zhang and Marr, 1993), but it requires prior information about which positions are non-independent (Stormo, 2000). In this article, we tried to combine the two wildly used models forming a mixture model to represent motifs and to search them from a DNA sequences set.

In another aspect, the search strategy differs largely also. Some basic algorithms like consensus(Hertz and Stormo, 1999b), EM (Lawrence and Reilly, 1990) and Gibbs sampler (Lawrence et al., 1993) brought solutions to this problem, but the result was not satisfactory enough. The enhanced computer programs based on them such as MEME (Bailey and Elkan, 1995), AlignAce (Hughes et al., 2000), and Bioprospector (Liu et al., 2001) are more powerful in dealing with true data, since these programs are enhanced by using more complex models and considering more parameters. After considering the above algorithms, we found a varied Gibbs sampling method

similar to Bioprospector used has some advantages, so we used it serving as the discovery algorithm.

2 DNA Conserved Sequences and Models to Represent Them

2.1 The Existence of Motifs

The genome of organism can be looked as a long DNA sequence and each part of the sequence has its own functions and characters. The sequence is made up of coding areas called genes that encode proteins, as well as no-coding areas that hold important regulatory function. Each gene consists exon, intron and UTRs in both ends. In eukaryotic, regulatory sequences are special DNA sequences usually located in upstream of a gene, controlling the transcription of the gene. If the functions of a set of genes were similar to each other, the regulatory sequences would resemble each other also, since these genes are subject to same regulation. In organisms, the regulatory sequences are binding sites of special proteins that serve as transcription factors or promoters, or other enhancers. The common length of binding site is 6~10bp (Sinha and Tompa, 2002). Many methods have been used to represent the common sequences, including IUPAC code, regular expression, consensus sequences, HMM model, neural network and so on. In this study, we brought about a mixture model comprising two basic models: PWM and WAM.

2.2 PWM Model

PWM is a universal way to represent DNA motifs. In a PWM, there is a matrix element for all possible bases at every position in the motif, the score for any particular sequence is the sum of matrix values for the sequence(Stormo, 2000). PWM composes 4 rows to represent 4 types of nucleotides acids of DNA sequences, the length of the PWM equals the length of the motif. The score of a given sequence $S = s_1 s_2 \ldots s_l$ matching the model can be calculated by:

$P_{PWM}(S) = \prod_{i=1}^{l} p_{s_i}^{(i)}$, Here, $p_{s_i}^{(i)}$ is the probability of base s_i in the i th position of the motif. The score also means the probability of the sequence generated by the PWM model representing the motif. However, if the values in the matrix cells are bases frequencies of genome data, or user provided data, it can be used as background model parameters.

A simplified way to calculate the score is to use the logarithm of the value, and the above formula can be written as $R^{\log}_{PWM}(S) = \sum_{i=1,l} \log(p_{s_i})$.

2.3 WAM Model

In PWM, the scores for each position are added together to get the total score, which implies that each position contributes independently. (Stormo, 2000). Indeed, there are often strong local dependencies within short DNA motif and this dependence among positions could be important(Bulyk et al., 2002). WAM (Zhang and Marr, 1993) incorporates dependencies between adjacent positions and can grasp features of some motifs.

The probability of a particular sequence $S = s_1 s_2 \ldots s_l$ being generated by WAM model is:

$$P_{WAM}(S) = p_{s_1}^{(1)} \prod_{i=2}^{l} p_{s_{i-1}, s_i}^{(i-1,i)}$$

Here, $p_{j,k}^{i-1,i}$ is the conditional probability of generating nucleotide s_k at position i, given nucleotide s_j at position $i-1$(Burge and Karlin, 1997). Also, if the data in WAM is counted from a complete genome sequence, or is uniformed, it can be served as a background model. Similarly, the logarithm of this probability can be calculated by $R^{\log}_{WAM}(S) = \log(p_1) + \sum_{i=2,i} \log(p_{s_{i-1}, s_i}^{(i-1,i)})$

2.4 The Mixture Model

The above two models are called base or ground models: each includes a probability estimate over the observed sequences. A mixture model is a combination of the two base models and can get better results (Eskin et al., 2001). In this study, we have combined PWM and WAM to form the mixture model. We considered the two parts not equally important and gave the PWM part more weight than WAM part, so we define the combined model as: $M_{mix} = \frac{w_a M_{pwm} + w_b M_{wam}}{w_a + w_b}$ in this equation, both model contribute one portion to the mix model, w_a, w_b are weights of the two parts. It is also a normalized model of the two and can embodiment individual position as well as adjacent base frequencies of a motif.

2.5 Evaluate a Sequence by Mixture Model

The score that a sequence matches the mix model can be calculated from the scores of the same sequence matching the basic models. It is an addition of two log-odd ratios of the probability of the sequence generated by the models versus that generated by background models. The score can be calculated by:

$$Score_{mix} = \frac{w_a}{w_a + w_b} \log \frac{P_{PWM}(S)}{P_{b,PWM}(S)} + \frac{w_b}{w_a + w_b} \log \frac{P_{WAM}(S)}{P_{b,WAM}(S)}$$

In this equation, $P_{b,PWM}(S)$ is the probability of sequence S generated by a background PWM model, which is a 4 dimensional vector representing the occurrence likelihood of 4 bases. It also can be seen as a zero order Markov model(Thijs et al., 2001). Similarly, $P_{b,WAM}(S)$ is the probability of sequence S generated by a background WAM model, which can be initialised by random data.

3 Motif Extracting Using Varied Gibbs Sampling

3.1 Coexpressed Genes and Repetitive Motif Model

In the progress of evolution, many related genes are derived from a common ancestor, also the regulatory sequences of the genes. The instances of a motif may vary because of different numbers of biology processes such as duplication or translocation. Some genes may have more copies of regulatory sequences, consequently have a strong response to the signal, while others would have little copies and as a result have a weak response to the same stimulate. Through biological experiments such as microarray or other methods like inter-species sequence comparison, those genes comprised motif instance can be identified and their upstream regions can be separated too.

A more clearly description is: in the upstream of each gene, there would or would not exist a motif in each sequence, the location or the motif appearance is not known (Liu et al., 2001). In Figure 1, there are n DNA sequences, each containing $k, (k \geq 0)$ copies of a motif instance. Sequence S_2 has two copies of motif and S_{n-1} has no copy of motif. Motif discovery is to find out all the motif instances and their locations in each sequence.

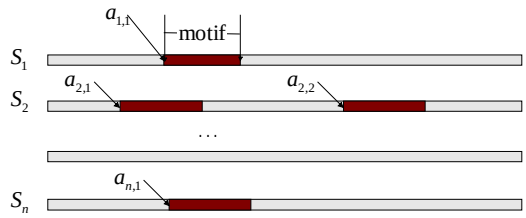

Figure 1. Repetitive motif model

3.2 Algorithm Brief

3.2.1 Model Initialise

A mixture model with two matrices was constructed and initialised in the first place. The cells in PWM were filled by sampling a 4 parameter Dirichlet distribution (Sjolander et al., 1996) described by:

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

Here $\alpha = \alpha_1, \ldots \alpha_k$, with $\alpha_i > 0$, are constants specifying the Dirichlet distribution, and θ_i satisfies $0 \leq \theta_i \leq 1$ and sum to 1. $Z(\alpha) = \frac{\prod_i \Gamma(\alpha_i)}{\Gamma(\sum_i \alpha_i)}$ is normalizing factor of the distribution. As to PWM model, α is a 4 dimensional vector representing the different base occurrence probability of each position. The means of the distribution is equal to the normalized parameter, and the value of α is inversely proportional to the distribution variance (Durbin et al., 1998). This step provided the prior information of PWM model.

WAM part is composed of 16 rows and $l-1$ columns. Similarly, sampling a Dirichlet distribution of 16 parameters provided the prior information of the WAM model. The final model is a mixture of the two preceding basic models.

3.2.2 Motif Discovery Principle

After the initial model was built, it can be used for the sequences discovery. Above all, a set of DNA sequences believed to contain common binding site pattern were provided, in these sequences, most similar and most frequent sub-sequences would be the instances of the regulatory motif, since only regulatory element has high number copies of instances in given sequences, and similar to each other because of same biological functions. As a result, from the given data set, there is a high chance to produce a motif model to represent the instances. The more similarity the finding sub-sequences are, the more conserved the model to describe the motif is. The probability of a model generated from observed sequence can be calculated by Bayesian inference:

$$P(M \mid S) = \frac{P(S \mid M) P(M)}{P(S)} \propto P(S \mid M)$$

Here, S is the sequence set, M is the mixture model. This formula means the probability of a model generated from a given data set, proportion to the probability of the model generating the sequence. If a model is conserved, it would have a high chance to generate sub-sequences that are instances of a motif, therefore, have a high chance to generate the given sequences set, so has a high score of $P(S \mid M)$. A varied Gibbs sampling method in this study can be used to discover the model which can maximized the probability.

3.3 Specific Steps

The steps are similar to the original Gibbs sampling methods, but differ in some specific manipulations. This new method also includes a sampling step and a model update step. The steps are given below.

1) According to the mixture model with prior information, search each given DNA sequence, and find out corresponding motif instances in each sequence. Altogether there are n piece of subsequence to form a multi alignment matrix.

2) Randomly select an integer i , $i \in [1, n]$ remove the i th motif instance from the alignment, and obtain a new alignment. Use the new alignment to update the mixture model. Since the mixture model comprises two parts, the update step includes two parts also. As to the PWM we update each element according to a relation $p_{b,j} = 0.8 \frac{c_{b,j}}{K} + 0.2 p_{b,j}$, here, $p_{b,j}$ is the data in b th row and j th column, $c_{b,j}$ is the total number of base $b, b \in \{A, C, G, T\}$ in position j of the alignment. K is the total subsequence number of the alignment. This operate can be regarded as adding

a maximum likelihood (ML) estimate of observed data into the original value.

As to the step of WAM update, the new value of each matrix element was calculated by: $q_{i,j} = 0.2\frac{e_{i,j}}{K} + 0.8q_{i,j}$, here $e_{i,j}$ is the occurrence of i th dinucleotide bases $bp(i) \in \{NN \mid N \in \{A,C,G,T\}\}$ in j th position of the alignment, K is the total number of dinucleotide bases in this position. This is also a revising of origin value by adding a maximum likelihood estimation of consecutive bases. Since the ML estimation would bias from the true value because of lack of data, only 20% was updated and 80% was retained of the value.

3) Searching new motif instance from the i th DNA sequence according to the updated model, and constructing new alignment matrix. This step would discover multiple or no motif instance from a DNA sequence. Particularly, each length l subsequence is extracted and scored according to the mixture model; the ratio of the score to the average is then calculated. If all the ratios are less than a threshold, we then regard that this sequence has no copy of the motif. Otherwise, if some ratios are above the threshold, the corresponding sub-sequences are regarded as motif instance and are selected to construct the alignment. The threshold is adjusted in each cycle by criteria: if above 20% sequences are found no copies of the motif, the threshold is decreased in the next cycle.

4) Repeat step 2, until the model converge.

Finally, a mixture model and motif instances in each sequences would be obtained.

4 Data and Result

According to the above steps, a motif discovery program was developed. The test data used was a set of DNA sequences comprising CRP binding site. CRP is a protein of *E.coli*; it takes an important role in metabolism by combining to special DNA sequences and forming DNA-protein complex which regulates some gene transcription. Stormo has collected 18 pieces of DNA sequence; all of them have the ability to combine to CRP. The location of the binding site in each DNA sequences was validated by experiments (Stormo and Hartzell, 1989). The consensus sequence is TGTGAnnnnnnTCACA; the length is 16.

In order to simulate the true situation that some sequences have no motif instance, we have added two computer generated sequences according to a background base distribution. Altogether there are 20 sequences to form the data set, and each sequence is at the length of 105bp. Then we used these data serving as input data to perform the discovery. After running the program, results were obtained and were listed in table 1.

Table1: Putative binding sites obtained from 20 DNA sequences

Gene names	Motif copies	Motif locations by different method				Discovered motifs	Scores
		MEME	Bioprospector	This method	Experiment verified		
CE1CG	1	64	64,71	64	20,64	TTTGATCGTTTTCACA	64.72
ECOARABOP	1	58	58	58	20,58	TTTGCACGGCGTCACA	63.08
ECOBGLR1	1	79	79	79	79	TGTGAGCATGGTCATA	62.5
ECOCRP	1	66	66	66	66	TGCAAAGGACGTCACA	63.53
ECOCYA	1	53	38	53	53	TGTTAAATTGATCACG	62.64
ECODEOP2	1	10,63	63	10	10,63	TTTGAACCAGATCGCA	64.26
ECOGALE	2	27,45	27,45	45,54	45	TGTCACACTTTTCGCA	61.44
ECOILVBPR	2	42	42	25,42	42	TCTGCAATTCAGTACA	59.75
ECOLAC	2	12	12,15	12,83	12,83	TGTGAGTTAGCTCACT	64.73
ECOMALE	1	17	17	17	17	TGTAACAGAGATCACA	66.15
ECOMALK	2	51	59	64,32	32,64	CGTGATGTTGCTTGCA	60.96
ECOMALT	1	44	47	44	44	TGTGACACAGTGCAAA	64.47
ECOOMPA	1	51	51	51	51	CCTGACGGAGTTCACA	64.14
ECOTNAA	1	74	74	74	74	TGTGATTCGATTCACA	64.76
ECOUXU1	1	20	71	20	20	TGTGATGTGGTTAACC	62.49
PBR322	1	56	35	56	56	TGTGAAATACCGCACA	64.44

TRN9CAT	2	90	4	3,87	3,87	TGAGACGTTGATCGGC	56.04
TDC	1	81	81	81	81	TGTGAGTGGTCGCACA	64.35
RNDSEQ1	0	24	34				
RNDSEQ2	0	49	56				

The table listed the locations and the found motifs in each sequence, altogether there are 18 sequences identified motifs. The program did not found any motif instances from the two artificial sequences. Actually, there are 24 motifs in this data set, and the program found out 23 copies where of which 21 copies are true motif. There are also 2 false positives and 3 true negatives. The Sensitivity $Se = \frac{TP}{TP + FN}$ is 0.87, Specificity $Sp = \frac{TP}{TP + FP}$ is 0.91.

To make comparison, we used other programs to discover motifs from the same data set. The first program used is Bioprospector (Liu et al., 2001), the service is at http://bioprospector.stanford.edu. This program discovered 23 motifs, of which 12 motifs are exactly matches and 12 are missed. The sensitivity and specificity of this program are 0.5 and 0.52.

Another program we used is MEME(Bailey and Elkan, 1994), the service is available at the web site http://meme.sdsc.edu/. The discovered motif instance locations are listed in the third column of table 1. This program discovered 22 motifs in which 17 of them exactly match the experiment verified data. 8 sites were not found. The sensitivity and specificity are 0.71 and 0.77. The above comparison of this method to other two methods showed some superiority.

Another test is plant promoter sequence discovery. PlantProm (Shahmuradov et al., 2003) is a DNA sequence database including 130 tata-less promoter and 175 tata promoter sequences of monocot and dicot plant genes. Some important regulatory sequences have annotated and the features of them have obtained. We used this method and other two methods to make discovery. The result consensus sequences of the three methods listed in table 2. Table 3 and table 4 are nucleotide frequencies matrices of two identified regulatory elements.

Table 2. Discovered motif by 3 methods

Method	**Discovered motifs of two data sets**	
	Tata promoters	**Tata-less promoter**
This method	TATAAATA	CACAATA
Bioprospector	CTATAAAT	CCAAAACC
MEME	TATCTTCCG	CCAAACC

Table 3. Nucleotide Frequencies Matrix for TATA box

	1	2	3	4	5	6	7	8
A	0.03	**0.95**	0.00	**1.00**	**0.62**	**0.97**	**0.38**	**0.73**
C	0.01	0.00	0.04	0.00	0.00	0.00	0.01	0.08
G	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.10
T	**0.96**	0.05	**0.96**	0.00	**0.38**	0.01	**0.61**	0.09
	T	**A**	**T**	**A**	**A/T**	**A**	**T/A**	**A**

Table 4. Nucleotide Frequencies Matrix for CCAAT box

	1	2	3	4	5
A	**0.40**	0.02	**1.00**	**1.00**	0.00
C	0.18	**0.98**	0.00	0.00	0.00
G	0.15	0.00	0.00	0.00	0.00
T	**0.27**	0.00	0.00	0.00	**1.00**
	a/t	**C**	**A**	**A**	**T**

In this experiment, the discovered motifs match the promoter elements matrix very well, so they can be considered as the true motifs. The other two programs also obtained motifs, but their matches to the matrices were less nice comparing to our method.

5 Conclusion

This article brought out a combined model to represent conserved motif of functional DNA sequences. The combined model is a mixture of two basic models: PWM and WAM. It gets over the defect that the basic model only contained either single position information or just neighbourhood base information. In addition, a varied Gibbs sampling algorithm was employed to the discover algorithm. This algorithm suits the situation of DNA sequence comprised no copy or multiple copies of motif. Through the analysis of a set of CRP binding gene sequences, the algorithm found out most motif instances of the binding site. The results excel that obtained by MEME or Bioprospector algorithm using default parameters. We also used the program to discover two sets of plant promoter sequences, and the motifs found accorded with the reported matrices. Results of the two study cases indicate that this method is feasible in motif discovery. Therefore the new model with the varied Gibbs sampling algorithm can be further applied in the field such as motif discovery or co-expressed gene analysis.

6 Reference

Bailey, T. L., and Elkan, C. (1994): Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, pp. 28-36.

Bailey, T. L., and Elkan, C. (1995): Unsupervised learning of multiple motifs in biopolymers using

expectation maximization. *Machine Learning* **21**, 51-80.

Bajic, V., and Seah, S. (2003): Dragon Gene Start Finder identifies approximate locations of the 5' ends of genes. *Nucleic Acids Research* **31**, 3560-3563.

Bajic, V., Seah, S., Krishnan, A. C. S., Koh, J., and Brusic, V. (2003): Computer model for recognition of functional transcription start sites in polymerase II promoters of vertebrates. *Journal of Molecular Graphics & Modeling* **21**, 323-332.

Brazma, A., Jonassen, I., Eidhammer, I., and Gilbert, D. (1998): Approaches to the automatic discovery of patterns in biosequences. *Journal of Computational Biology* **5**, 279-305.

Buhler, J., and Tompa, M. (2001): Finding motifs using random projections. In Proceedings of the Fifth Annual International Conference on Computational Molecular Biology (RECOMB01), pp. 69-76.

Buhler, J., and Tompa, M. (2002): Finding Motifs Using Random Projections. *Journal of Computational Biology* **9**, 225-242.

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. *Nucl. Acids. Res.* **30**, 1255-1261.

Burge, C., and Karlin, S. (1997): Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology* **268**, 78-94.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998): *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press. Cambridge.

Eskin, E., Grundy, W. N., and Singer, Y. (2001): Using mixtures of common ancestors for estimating the probabilities of discrete events in biological sequences. *Bioinformatics* **17**, 65S-73.

Hertz, G., and Stormo, G. (1999a): Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* **15**, 563-577.

Hertz, G. Z., and Stormo, G. D. (1999b): Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* **15**, 563-577.

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

Jonassen, I., Collins, J. F., and Higgins, D. (1995): Finding flexible patterns in unaligned protein sequences. *Protein Science* **4**, 1587-1595.

Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F., and Wootton, J. C. (1993): Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **262**, 5131.

Lawrence, C. E., and Reilly, A. A. (1990): An expectation maximization (EM) algorithm for the identifcation and characterization of common sites in unaligned biopolymer sequences. *PROTEINS: Structure Function and Genetics* **7**, 41-51.

Li, M., Ma, B., and Wang, L. (1999): finding similar regions in many sequences. Proc. 31st ACM Symp. Theory of Computing.

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., pp. 127-138.

Shahmuradov, I. A., Gammerman, A. J., Hancock, J. M., Bramley, P. M., and Solovyev, V. V. (2003): PlantProm: a database of plant promoter sequences. *Nucl. Acids. Res.* **31**, 114-117.

Sinha, S., and Tompa, M. (2002): Discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucl. Acids. Res.* **30**, 5549-5560.

Sjolander, K., Karplus, K., Brown, M., Hughey, R., Krogh, A., Mian, I. S., and Haussler, D. (1996): Dirichlet mixtures: A method for improving detection of weak but significant protein sequence homology. *Computer Applications in the Biosciences.* **12**, 327-345.

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

Stormo, G. D., and Hartzell, G. W. (1989): Identifying protein-binding sites from unaligned DNA fragments. *Proc. Natl. Acad. Sci. USA* **86**, 1183-1187.

Thijs, G., Lescot, M., Marchal, K., Rombauts, S., De Moor, B., Rouze, P., and Moreau, Y. (2001): A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics* **17**, 1113-1122.

Vilo, J. (2002): Pattern Discovery from Biosequences, pp. 149: *Department of Computer Science, University of Helsinki*, University of Helsinki, Finland.

Zhang, M. Q., and Marr, T. G. (1993): A weight array method for splicing signal analysis. *Comput. Appl. Biosci.* **9**, 499-509.