A Gibbs Sampling method to detect over-represented motifs in the upstream regions of co-expressed genes

Gert Thijs[1] Kathleen Marchal[1] Magali Lescot[1] Stephane Rombauts[2] Bart De Moor[1]
Pierre Rouzé[3] Yves Moreau[1]
[1] ESAT-SISTA/COSIC, KULeuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium
[2] Plant Genetics, VIB, University Gent, Ledeganckstraat 35, 9000 Gent, Belgium
[3] INRA associated laboratory, VIB, University Gent, Ledeganckstraat 35, 9000 Gent, Belgium
{Gert.Thijs,Yves.Moreau}@esat.kuleuven.ac.be

ABSTRACT

Microarray experiments can reveal useful information on the transcriptional regulation. We try to find regulatory elements in the region upstream of translation start of coexpressed genes. Here we present a modification to the original Gibbs Sampling algorithm [12]. We introduce a probability distribution to estimate the number of copies of the motif in a sequence. The second modification is the incorporation of a higher-order background model. We have successfully tested our algorithm on several data sets. First we show results on two selected data set: sequences from plants containing the G-box motif and the upstream sequences from bacterial genes regulated by O_2-responsive protein FNR. In both cases the motif sampler is able to find the expected motifs. Finally, the sampler is tested on 4 clusters of coexpressed genes from a wounding experiment in *Arabidopsis thaliana*. We find several putative motifs that are related to the pathways involved in the plant defense mechanism.

1. INTRODUCTION

Microarray technology allows biologists to monitor the mRNA expression levels of several thousands of genes in one experiment [6, 13, 26, 27, 29]. By measuring the mRNA at consecutive time points during a biological experiment, it is possible to construct an expression profile for each gene on the array. This type of experiments has opened new research directions in biology and genetics [28, 30, 36, 38]. It is very interesting to find genes that have a similar behavior under the same experimental conditions. Several clustering algorithms are available to group genes that have a similar expression profile [7, 8, 17, 28, 29, 31]. Given the clusters of genes with a highly similar expression profile, we can search for the mechanism that is responsible for their typical behavior. The basic assumption in the model states that coexpression indicates coregulation.

Coregulated genes are known to share some common motifs, that are binding sites for transcription regulators. A sensible approach to detect these regulatory elements is to search for over-represented motifs in the region upstream of translation start in a set of coexpressed genes [3, 7, 24, 39].

Many researchers have been working on computational methods to detect regulatory elements. The algorithms to find regulatory elements can be divided into two classes: word analysis methods [10, 33, 34, 35] and methods based on probabilistic sequence models [1, 9, 12, 14, 18, 24, 37]. The word analysis methods are based on the frequency analysis of oligonucleotides in the upstream region and on intelligent word counting strategies. A common motif is then compiled by grouping over-represented similar words. When using a probabilistic sequence model the motif is represented by a position probability matrix. The basic model assumes that the motif is hidden in a noisy background sequence. To find the parameters of this model, maximum likelihood estimation is used. Most used methods are *Expectation Maximization* (EM) and *Gibbs Sampling*. EM is a deterministic algorithm and Gibbs Sampling is a stochastic equivalent of EM.

In this paper a modification of the original Gibbs Sampling algorithm by Lawrence et al. [12] is presented. A probabilistic framework is used to estimate the expected number of copies of a motif in a sequence. We introduce also the use of a higher order background model based on an Markov chain. We describe the incorporation of these modifications in the Gibbs Sampling algorithm to find the parameters and have successfully tested our implementation on different data sets of intergenic sequences.

2. ALGORITHM AND IMPLEMENTATION

2.1 Finding multiple copies

Applying clustering to the gene expression profiles of a microarray experiment gives several groups of coexpressed genes. Following the basic assumption, we can assume that coexpression indicates coregulation, but it is expected that only a subset of the coexpressed genes are actually coregulated. When searching for possible regulatory elements in such a set of sequences, this idea has to be taken into account. It is important to have an algorithm that can distinguish between sequences in which there is motif and the ones in which there is not.

In higher organisms regulatory elements can have several copies to increase the influence of the element in the process of transcriptional regulation. We reformulate the probabilistic sequence model in such a way that we can estimate the number of copies of the motif in the sequence. The number of copies of a motif in each sequence is represented by creating a new missing value Q_k, the number of copies of the motif in sequence S_k. Q_k varies between 0 en C_{max}. C_{max} is a user defined parameter to set the maximal number of copies of motifs in a sequence.

First, we define the motif model. The motif is represented by a position probability matrix θ_W:

$$\text{Motif } \theta_W = \begin{pmatrix} q_1^{\mathtt{A}} & q_2^{\mathtt{A}} & \cdots & q_W^{\mathtt{A}} \\ q_1^{\mathtt{C}} & q_2^{\mathtt{C}} & \cdots & q_W^{\mathtt{C}} \\ q_1^{\mathtt{G}} & q_2^{\mathtt{G}} & \cdots & q_W^{\mathtt{G}} \\ q_1^{\mathtt{T}} & q_2^{\mathtt{T}} & \cdots & q_W^{\mathtt{T}} \end{pmatrix},$$

with W the fixed length of the motif. The background model is represented by B_m, with m the order of the model (see next section). Using Q_k and Bayes' theorem we can write an equation to calculate the probability $\gamma_{k,c}$ of finding c copies of the motif in sequence S_k given the motif and background model:

$$\begin{aligned} \gamma_{k,c} &= P(Q_k = c|S_k, \theta_W, B_m) \\ &= \frac{P(S_k|Q_k = c, \theta_W, B_m)P(Q_k = c|\theta_W, B_m)}{P(S_k|\theta_W, B_m)} \\ &= \frac{P(S_k|Q_k = c, \theta_W, B_m)P(Q_k = c|\theta_W, B_m)}{\sum_{c=0}^{C_{max}} P(S_k|Q_k = c, \theta_W, B_m)P(Q_k = c|\theta_W, B_m)} \end{aligned}$$

$\gamma_{k,c}$ describes a discrete probability distribution. The parameters are estimated in each iteration of the algorithm. Finally, the expected number of copies of the motif in sequence S_k is calculated as $E(Q_k) = \sum_{c=1}^{C_{max}} c\gamma_{k,c}$.

2.2 Background model

The second modification is the use of a higher order background model. The most popular motif detection methods accessible on the net, AlignACE [9] and MEME [1], use a simple background model. The background model is described only by the single nucleotide frequency distribution. But if we look more closely at most state-of-the-art gene detection software: Glimmer [5], HMMgene [11] and GeneMark.hmm [15], they all use higher order Markov processes to model coding and non-coding sequences. Starting from the ideas behind these gene prediction algorithms, we developed a background model based on a Markov Process of order m. This means that the probability of the nucleotide b_l at position l in the sequence depends on the m previous bases in the sequence. Such a model is described with a transition matrix. Given a background model of order m, B_m, the probability of sequence S being generated by the background model can be written as:

$$P(S|B_m) = P(b_1, \ldots, b_l) \prod_{l=1}^{L} P(b_l|b_{l-1}, \ldots, b_{l-m})$$

An elaborate evaluation and discussion of the influence of the use of a higher-order background model on motif detection by Gibbs Sampling has been described elsewhere [32].

Important to know is that the background model can be either constructed from the original sequence data or from an independent data set. The latter approach is the more sensible one if the independent data set is carefully created. Carefully created means in this case that there are only sequences in the training set that are in the intergenic region and that do not overlap with coding sequences. At the moment only an independent background model for *Arabidopsis thaliana* is constructed based on the sequences in Araset [19]. Nevertheless the algorithm can also be used for other organisms. In the results section we show some results with upstream sequences selected from plants and also on intergenic sequences from bacteria.

2.3 Algorithm

Both modifications have been included in the iterative procedure of the Gibbs sampling algorithm. First the number of copies is sampled according to the distribution Γ. In the next step the motif model is updated based on the current alignment vector and the probability distribution of the motif positions is reestimated. An alignment vector is then selected by sampling according to this updated distribution Γ. Given the new alignment vector we can reestimate the distribution Γ. Here follows a high level description of the algorithm.

1. Select or compute the background model B_m.
2. Compute the probability P_x^0 for all segments x of length W in every sequence. Since the background model is fixed, it is not necessary to recalculate these values in each iteration.
3. Initialization of the alignment vector $A = \{A_k|k = 1 \ldots N\}$ and the weighting factors $\Gamma = \{\Gamma_k|k = 1 \ldots N\}$:
   $$A_k = \{a_{k,1}, \ldots, a_{k,C}\} \quad \text{and} \quad \Gamma_k = \{\gamma_{k,1}, \ldots, \gamma_{k,C}\}$$
4. Sample each Q_k from the corresponding distribution Γ_k.
5. For each sequence S_z, $z = 1, \ldots, N$
   (a) Create subsets $\tilde{S} = \{S_i|i \neq z\}$ and $\tilde{A} = \{\tilde{A}_i|i \neq z\}$, with $\tilde{A}_i = \{a_{i,1}, \ldots, a_{i,Q_i}\}$
   (b) Calculate θ_W and θ_0 based on $\tilde{S}$ and $\tilde{A}$.
   (c) Assign to each segment x from S_z the weight $W_x = P_x/P_x^0$
   $$P_x = P(x|\theta_W), \qquad P_x^0 = P(x|B_m)$$
   (d) Sample new position a_z from probability distribution W_x.
   (e) Update the distribution Γ_z.
6. Repeat from 4 until convergence is reached.

2.4 Implementation

The core algorithm of the motif sampler is implemented using the mathematical programming environment *Matlab*. To manipulate DNA sequence we use *Perl* and the modules from *BioPerl*, [http://bio.perl.org/]. The algorithm is accessible through a web interface:
http://www.esat.kuleuven.ac.be/˜thijs/Work/MotifSampler.html.

There are 5 parameters to be set:

- Background model (B_m): As the background model either one of the pre-compiled models from *Arabidopsis thaliana* can be selected or the background model can be computed from the sequence data themselves.

- Length (W): The length of the motif is fixed. Reasonable values range from 5 to 15.

- Motifs (N): The number of different motifs to be searched for. The motifs will be searched for in consecutive runs while the positions of the previously found motifs are masked.

- Copies (C): This number sets the maximum number of copies of a motif in every sequence. If this number is set too high, noise will be introduced in the motif model.

- Overlap (O): This parameter defines the allowed overlap between the different motifs.

The final result of the motif sampler consists of three parts: the position probability matrix θ_W, the alignment vector A and the weighting factors Γ. Based upon these values different scores with their own chararcteristics can be calculated: consensus score, information content and log-likelihood.

The consensus score is a measure for the conservation of the motif. A perfectly conserved motif will have a score equal to 2 while a motif with a uniform distribution will have a score equal to 0.

$$\text{Consensus Score} = 2 - \frac{1}{W}\sum_{l=1}^{W}\sum_{b=A}^{T} q_l^b \log(q_l^b)$$

The information content or Kullback-Leiber distance between motif and the single nucleotide frequency tells how much the motif differs from the background. This score will be maximal if the motif is well-conserved and differs considerably from the background distribution.

$$\text{Information Content} = \frac{1}{W}\sum_{l=1}^{W}\sum_{b=A}^{T} q_l^b \log\left(\frac{q_l^b}{q_0^b}\right)$$

As a final score we consider the log-likelihood. The motif and corresponding positions are the results of maximum likelihood estimation. Therefore the log-likelihood is a good measure for the quality of the motif. In this case we are especially interested in the positive contribution of the motif to the global log-likelihood. If we write the probability of the sequence being generated by the background model, $P(S|B_m)$, as P_0, the log-likelihood can be calculated:

$$\log\left(\pi(S,A,Q|\theta_W,B_m)\right) =$$
$$\log\left(\sum_{c=0}^{C_{max}} \gamma_c P(S|A_c,\theta_W,B_m)P(A_c|\theta_W,B_m)\right) =$$
$$\log(P_0) + \log(C) + \underbrace{log\Big(\sum_{c=0}^{C_{max}} \gamma_c \frac{P(S|A_c,\theta_W,B_m)}{P_0}\Big)}_{\text{motif contributions}}$$

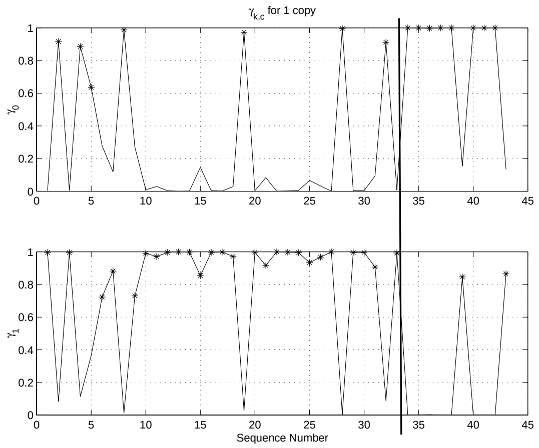

Figure 1: Weighting factors $\gamma_{k,c}$ when searching for 1 copy, associated with kmCACGTG. Each subplot corresponds to a number of copies c, with $c = 0, \ldots, C_{max}$ and indicates the probability of finding a motif c times in each sequence. The stars indicate the expected number copies of the motif.

3. RESULTS

3.1 G-box sequences

We exhaustively tested the performance of our implementation of the motif sampler. To validate the motif sampler we constructed two data sets: one with a known regulatory element involved in light regulation in plants, G-box and one of so-called random sequences in which no G-box is reported. The G-box data set consists of 33 sequences selected from PlantCARE [23] containing 500bp upstream of the translation start. This data set is well suited to give a proof of concept and to test the performance of the motif sampler, since we exactly know the consensus of the motif and also the positions of the motif in the sequences. The random set consists of 87 sequences of 500bp. This set was used to introduce noise to the test sets. Here we show the results of two different tests. The first test shows the influence of the number of copies and the second test illustrates the improvement due to the use of a higher-order background model when noise is added to the data set.

First we experimented with 33 G-box sequences together with 10 sequences from the random set. This set was used to test the influence of the number of copies. When the number of copies is set to 1, a more conserved motif will be found, but a number of occurrences will be missed. Increasing the number of copies will allow to better locate the true number of copies of a motif but more noise is introduced to the initial model and the final model will be more degenerate. This trade-off has to be taken into account when fine-tuning the algorithm. The results shown are based on parameter settings to search for a motif of length 8bp that can have either 1 or 4 copies with the *Arabidopsis* background model of order 3. In both cases a motif was found with a consensus resembling the G-box consensus CACGTG. These motifs are also the motifs with the highest scores.

Figure 1 and 2 show the weighting factors $\gamma_{k,c}$ associated

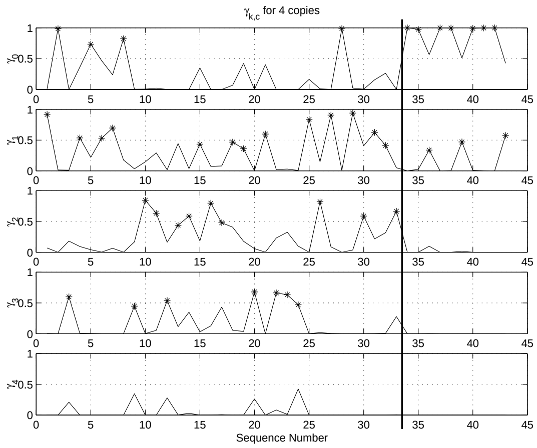

Figure 2: Weighting factors $\gamma_{k,c}$ when searching for 4 copies, associated with kCCACGTG. Each subplot corresponds to a number of copies c, with $c = 0, \ldots, C_{max}$ and indicates the probability of finding a motif c times in each sequence. The stars indicate the expected number copies of the motif.

with 1 copy and 4 copies respectively. The weighting factors $\gamma_{k,c}$ are ordered from top to bottom in ascending order of c. The X-axis corresponds to the sequence numbers: the first 33 sequences are the G-box sequences and the last 10 are the random sequences. The expected number of copies in a sequence is indicated by a star in the corresponding row of the plots. The first subplot (top) corresponds to $c = 0$ and shows the probability of not finding the motif in the sequences. The probability distribution of each individual sequence shows that when searching for 1 copy this distribution tends to be rather extreme (Fig. 1). When allowing 4 copies on the other hand the distribution will be smoother (Fig. 2). Based on the reported motifs in the G-box data set, we know that there are several sequences in which multiple copies of the G-box occur. So when limiting the number of copies to 1 some of the information to construct the motif model will be discarded.

All random sequences are expected to have 0 copies of the motif, however in a few sequences a motif is found. In the case of 1 copy only 2 sequences out of 10 have a single copy of the motif, but with a high probability. If we allow 4 copies, 3 random sequences out of 10 are indicated as having 1 copy of the motif but the probability of finding this copy has decreased.

Let us now consider the G-box sequences. When searching for 1 copy, 7 sequences out of 33 are indicated as not having the motif. In case of 4 copies there are still 4 out of 33 G-box sequences that are indicated as not having a G-box. If we check those 4 sequences in PlantCARE, we find that the G-boxes are not found experimentally but only by homology search. It might be that these motifs are false positives, but this can not be concluded decisively without biological evidence.

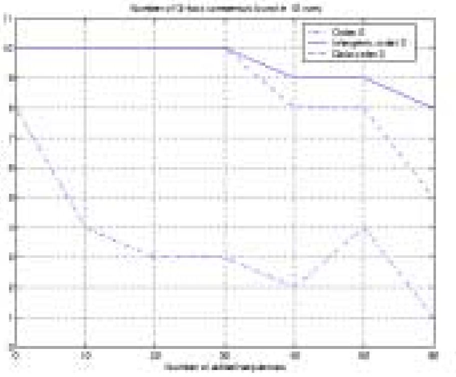

Figure 3: Total number of times the G-box consensus is found in 10 repeated runs of the tests for three different background models. The data set consists of the 33 G-box sequences and a fixed number of added noisy sequences.

Another important issue is the influence of noise on the performance of the motif sampler. Noise is due to the presence of upstream sequences that do not contain the motif. To introduce noise in the data set we added in several consecutive tests each time 10 extra sequences, in which no G-box is reported, to the G-box data set. We exhaustively tested several configuration to see how the noise influences the performance of the motif sampler. To test the significance of the results each test was repeated 10 times. Figure 3 shows the total number of times the G-box consensus in 10 runs for three different background models and an increasing number of added sequences. As can be expected, the number of times the G-box is detected decreases when more noise is added to the original set of 33 G-box sequences. This influence is more dramatic for the single nucleotide background model then for the third order background model.

3.2 Bacterial Sequences

As another test a data set with intergenic sequences from bacteria was created. The selected data set contains a subset of bacterial genes all shown to be regulated by the bacterial O_2-responsive protein FNR [16]. The genes were selected from several bacterial species: *Azospirillum brasilense*, *Paracoccus denitrificans*, *Rhodobacter sphaeroides*, *Rhodobacter capsulatus*, *Sinorhizobium meliloti* and *Escherichia coli*. The data set contains 10 intergenic sequences with varying length.

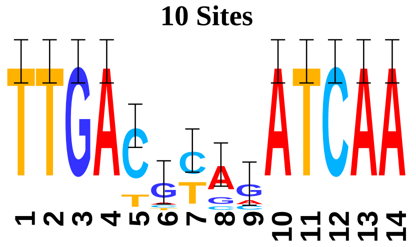

Figure 4: FNR binding site logo

In this case there was no precompiled background model available, therefore a background model of order 1 was compiled from the sequence data. The order of the background model is limited by the number of nucleotides in the data set. The motif sampler could retrieve the consensus sequence of the FNR-consensus sequence, described in literature as an interrupted palindrome of 14 bp. The motif is shown in Figure 4. Table 1 gives a more descriptive overview of the results. The first two columns identify the gene by their accession number and gene name. The given position is the position in the input sequence. The next column is the site as found in the sequences and finally there is a the probability of finding this motif in the sequence. We can see that the motif is found in all 10 sequence with a high probability score. This means that the motif sampler is very confident on finding the motif in the sequences. We searched for more motifs but only two, unknown, motifs were found: `sGGyCGAATGGTCG` and `TTCATGACAGTCCT`. They both occur only in 3 sequences out of 10.

Accn	Gene	Position	Site	Prob.
af016223	ccoN	60	TTGACGCGGATCAA	1.0000
af054871	cytN	255	TTGACGTAGATCAA	1.0000
pdu34353	ccoN	131	TTGACGCAGATCAA	1.0000
pdu34353	ccOt	210	TTGACGCAGATCAA	1.0000
af195122	bchE	82	TTGACATGCATCAA	0.9998
af016236	dorS	8	TTGACGTCAATCAA	1.0000
ae000220	narK	267	TTGATTTACATCAA	0.9986
rlfixnc	fixNc	104	TTGATGTAGATCAA	1.0000
rlfixnd	fixNd	240	TTGACGCAGATCAA	1.0000
pdu34353	fnr	36	TTGACCCAAATCAA	0.9999

Table 1: Bacterial O_2-responsive protein FNR. The first two column describe the gene. The given position is the position of the motif in the input sequences. The fourth column gives the motif in the sequence and the last column gives the probability score of the motif.

3.3 Microarray experiment

The previous examples can be seen as proof of concept data sets to test the performance of the motif sampler. We also used the motif sampler to find motifs in clusters of coexpressed genes. As a test case we use the data from Reymond *et al.*, where the gene expression in response to mechanical wounding was measured [22]. mRNA was extracted from leaves at 8 time points up to 24 hours after the wounding and an expression profile was constructed. To find the groups of coexpressed genes we use a clustering algorithm developed in our group [4]. As input parameters to the clustering, we use a small radius of 1 and the minimal number of genes in a cluster is set to be at least 3. This results in 8 small clusters of coexpressed genes.

To analyze the cluster we select the sequence 500bp upstream of translation start for every gene present in one of the clusters. We looked for 10 different motifs of length 8 and 12bp. To distinguish between stable motifs and motifs that are found just by chance, we repeated each experiment 10 times. The results with the *Arabidopsis* background model of order 3 are shown since they gave the most promising results. Four of the clusters contained only 3 genes and we did not found any interesting motif in these small clusters. The most interesting motifs were only detected in the clusters containing more than 3 genes. Table 2 gives an overview of the most important results. The consensus sequences are a compilation of the consensus sequences of length 8 and 12 bp. Only the relevant part of the consensus is displayed. Together with the consensus the number of times the consensus was found in 10 runs is indicated. The most frequent motifs are shown here.

To assign a functional interpretation to the motifs, the consensus of the motifs was compared with the entries described in PlantCARE. Several interesting motifs are found: methyl jasmonate(MeJa) responsive elements, elicitor-responsive elements and the abcissic acid response element (ABRE). It is not surprising to find these elements in gene promoters induced by wounding, because there is a clear cross-talk between the different signal pathways leading to inducible defense gene expression [2, 21, 25, 20]. Depending on the nature of a particular aggressor (wounding/insects, fungi, bacteria, virus) the plant is able to fine-tune the induction of defense genes either by employing a single signal molecule or by a combination of the 3 regulators jasmonic acid (JA), ethylene and salicylic acid (SA). In the third and fourth cluster there are also some strong motifs found that do not have a corresponding motif in PlantCARE. These motifs look promising but they need some further investigation.

4. DISCUSSION

We have introduced a modified version of the original Gibbs Sampler algorithm to detect regulatory elements in the upstream region of DNA sequences. The first change is the use of a probability distribution to model the number of copies of a motif in each sequence. We propose an iterative sampling scheme to find the most likely motif alignment. The second contribution is the inclusion of an higher order background model instead of using single nucleotide frequencies as the background model. The use of a carefully selected independent data set to construct a background model improves the performance and robustness of the motif sampler.

In this study we showed that our implementation of the Gibbs Sampler is able to find over-represented motifs in a well described test set of upstream sequences. Finally the motif sampler was also used to find motifs in a sets of coexpressed genes.

The algorithm is accessible through a web interface, where a limited number of parameters is user defined. The parameter definitions are kept simple and easy to interpret. There is no need for the users to go through the details of the implementation to understand the reasonable parameter settings.

However the implementation is far from final and needs some further extensions to improve the usability and performance. In this perspective, we will work on several add-ons. First we would like to implement a method to automatically detect the optimal length of the motif. At the moment the length of the motif is still a user-defined and fixed parameter. Also the optimization of the procedure to find the number of copies is important. This is of course closely related to the improvement of the motif scores. The ultimate

Cluster	Consensus	Runs	PLantCARE	Function
1	TAArTAAGTCAC	7/10	TGAGTCA	tissue specific GCN4-motif
(11 seq.)			CGTCA	MeJA-responsive element
	ATTCAAATTT	8/10	ATACAAAT	element associated to GCN4-motif
	CTTCTTCGATCT	5/10	TTCGACC	elicitor responsive element
2	TTGACyCGy	5/10	TGACG	MeJa responsive element
(6 seq.)			(T)TGAC(C)	Box-W1, elicitor responsive element
	mACGTCACCT	7/10	CGTCA	MeJA responsive element
			ACGT	Abcissic acid response element
3	wATATATATmTT	5/10	TATATA	TATA-box like element
(5 seq.)	TCTwCnTC	9/10	TCTCCCT	TCCC-motif, light response element
	ATAAATAkGCnT	7/10	-	-
4	yTGACCGTCCsA	9/10	CCGTCC	meristem specific activation of H4 gene
(5 seq.)			CCGTCC	A-box, light or elicitor responsive element
			TGACG	MeJA responsive element
			CGTCA	MeJA responsive element
	CACGTGG	5/10	CACGTG	G-box light responsive element
			ACGT	Abcissic acid response element
	GCCTymTT	8/10	-	-
	AGAATCAAT	6/10	-	-

Table 2: Results of the motif search in 4 clusters for the third order background model. In the second column the consensus of the found motif is given together with the number of times this motif was found in the 10 runs. Finally the corresponding motif in PlantCARE and a short explanation of the described motif is given.

goal is to develop a robust algorithm with as few as possible user parameters. When the parameters are handled well, we will focus on the development of more specific motif models, like short palindromic motifs that are separated by a small variable gap or the combined occurrence of motifs. Moreover the probabilistic framework in which the motif sampler is implemented is well suited to incorporate prior biological knowledge in the sequence model.

4.1 Acknowledgments

Gert Thijs is research assistant with the IWT; Yves Moreau is a post-doctoral researcher of the FWO; Prof. Bart De Moor is full time professor at the KULeuven. Pierre Rouzé is Research Director of INRA (Institut National de la Recherche Agronomique, France). This work is partially supported by: 1. IWT projects: STWW-980396; 2. Research Council KULeuven: GOA Mefisto-666; 3. FWO projects: G.0240.99 and G.0256.97; 4. IUAP P4-02 (1997-2001); 5. Industrial Contract Research: Data4s. The scientific responsibility is assumed by its authors.

5. REFERENCES

[1] T. L. Bailey and C. Elkan. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning*, 21:51–80, 1995.

[2] G.F. Birkenmeier and C.A. Ryan. Wound signaling in tomato plants. evidence that aba is not a primary signal for defense gene activation. *Plant Physiol*, 117(2):687–693, 1998.

[3] P. Bucher. Regulatory elements and expression profiles. *Current Opinion in Structural Biology*, 9:400–407, 1999.

[4] F. De Smet, G. Thijs, K. Marchal, B. De Moor, and Y. Moreau. Quality-based clustering of gene expression profiles. *submitted*, 2000.

[5] A.L. Delcher, D. Harman, S. Kasif, O. White, and S.L. Salzberg. Improved micorbial gene identification with glimmer. *Nucleic Acid Research*, 27(23):4636–4641, 1999.

[6] J.L. DeRisi, V.R. Iyer, and P. O. Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278:680–, 1997.

[7] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, 95:14863–14868, 1998.

[8] L.J. Heyer, S. Kruglyak, and S. Yooseph. Exploring expression data: Identification and analysis of coexpressed genes. *Genome Research*, 9:1106–1115, 1999.

[9] J.D. Hughes, P.W. Estep, S. Tavazoie, and G.M. Church. 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] L.J. Jensen and S. Knudsen. Automatic discovery of regulatory patterns in promoter regions based on whole cell expression data and functional annotation. *Bioinformatics*, 16(4):326–333, 2000.

[11] A. Krogh. Two methods for improving performance of an hmm and their application for gene finding. In *Proceedings ISMB'97*, pages 179–186, 1997.

[12] C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton. Detecting subbtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, 262:208–214, 1993.

[13] R.J. Lipschutz, S.P.A. Fodor, T.R. Gingeras, and D.J. Lockheart. High density synthetic oligonucleotide arrays. *Nature Genetics Supplement*, 21:20–24, january 1999.

[14] J.S. Liu, A.F. Neuwald, and C.E. Lawrence. Bayesian models for multiple local sequence alignment and gibbs sampling strategies. *Journal of the American Statistical Association*, 90(432):1156–1170, 1995.

[15] A.V. Lukashin and M. Borodowsky. Genemark.hmm: new solutions for gene finding. *Nucleic Acid Research*, 26:1107–1115, 1998.

[16] Kathleen Marchal. *The O_2 paradox of Azospirillum brasilense under diazotrophic conditions.* PhD thesis, FLTBW, KULeuven, 1999.

[17] E. Mjolsness, T. Mann, R. Castaño, and B. Wold. From coexpression to coregulation: An approach to inferring transcriptional regulation among gene classes from large-scale expression data. In *Proceedings NIPS 2000*, volume 12, pages 928–934, 2000.

[18] A.F. Neuwald, J.S. Liu, and C.E. Lawrence. Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Science*, 4:1618–1632, 1995.

[19] N. Pavy, S. Rombauts, P. Déhais, C. Mathé, D.V.V. Ramana, P. Leroy, and P. Rouzé. Evaluation of gene prediction software using a genomic data set: Aplication to *Arabidopsis thaliana* sequences. *Bioinformatics*, 15:887–899, 1999.

[20] H. Pena-Cortes, J.J. Sanchez-Serrano, R. Mertens, L. Willmitzer, and S. Prat. Abscisic acid is involved in the wound-induced expression of the proteinase inhibitor ii gene potato and tomato. *Proc. Natl. Acad. Sci USA*, 86:9851–9855, 1989.

[21] P. Reymond and E.E. Farmer. Jasmonate and salicylate as global signals for defense gene expression. *Curr Opin Plant Biol*, 1(5):404–411, 1998.

[22] P. Reymond, H. Weber, M. Damond, and E. E. Farmer. Differential gene expression in response to mechanical wounding and insect feeding in Arabidopsis. *Plant Cell*, 12:707–719, 2000.

[23] S. Rombauts, P. Déhais, M. Van Montagu, and P. Rouzé. PlantCARE, a plant cis-acting regulatory element database. *Nucleic Acids Research*, 27:295–296, 1999.

[24] F.P. Roth, J.D. Hughes, P.W. Estep, and G.M. Church. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole genome mRNA quantitation. *Nature Biotechnology*, 16:939–945, 1998.

[25] J. Rouster, R. Leah, J. Mundy, and V. Cameron-Mills. Identification of a methyl jasmonate-responsive region in the promoter of a lipoxygenase 1 gene expressed in barley grain. *Plant Journal*, 11(3):513–523, 1997.

[26] M. Schena. Genome analysis with gene expression microarrays. *BioEssays*, 18(5):427–431, 1996.

[27] M. Schena, D. Shalon, R.W. Davis, and P.O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270:467–470, 1995.

[28] G. Sherlock. Analysis of large-scale gene expression data. *Curr. Opin. Immunol.*, 12:201–205, 2000.

[29] P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast S. Cerevisiae by microarray hybridization. *Molecular Biology of the Cell*, 9:3273–3297, 1998.

[30] Z. Szallasi. Genetic network analysis in light of massively parallel biological data acquisition. In *Proceedings PSB'99*, volume 4, pages 5–16, 1999.

[31] S. Tavazoie, J.D. Hughes, M.J. Campbell, R.J. Cho, and G.M. Church. Systematic determination of genetic network architecture. *Nature Genetics*, 22(7):281–285, 1999.

[32] G. Thijs, M. Lescot, K. Marchal, S. Rombauts, B. De Moor, P. Rouzé, and Y. Moreau. A higher order background model improves the detection by Gibbs sampling of potential promoter regulatory elements in DNA sequences. Technical Report 00-128, ESAT-SISTA/COSIC, KULeuven, 2000. submitted, Genome Research.

[33] J. van Helden, B. André, and L. Collado-Vides. Extracting regulatory sites from upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *Journal of Molecular Biology*, 281:827–842, 1998.

[34] J. van Helden, A.F. Rios, and J. Collado-Vides. Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Research*, 28(8):1808–1818, 2000.

[35] A. Vanet, L. Marsan, A. Labigne, and M.F. Sagot. Inferring regulatory elements from a whole genome. an analysis of helicobacter pylori sigma(80) family of promoter signals. *Journal of Molecular Biology*, 297(2):335–353, 2000.

[36] X. Wen, S. Fuhrman, G. S. Michaels, D. B. Carr, S. Smith, J. L. Barker, and R. Somogyi. Large-scale temporal gene expression mapping of central nervous system development. *Proc. Natl. Acad. Sci. USA*, 95:334–339, 1998.

[37] C.T. Workman and G.D. Stormo. Ann-spec: a method for discovering transcription binding sites with improved specificity. In *Proceedings PSB'2000*, volume 5, Honolulu, Hawai, 2000.

[38] M.Q. Zhang. Large-scale gene expression data analysis: A new challenge to computational biologists. *Genome Research*, 9:681–688, 1999.

[39] J. Zhu and M.Q. Zhang. Cluster, function and promoter: analysis of yeast expression array. In *Proceedings PSB'2000*, volume 5, pages 467–486, 2000.