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], Stéphane Rombauts[2], Bart De Moor[1], Pierre Rouzé[3], and Yves Moreau[1]

[1] ESAT-SISTA/COSIC, KULeuven
Kardinaal Mercierlaan 94, 3001 Leuven, Belgium
Tel: +32 16 321884, Fax: +32 16 321970
[2] Department of Plant Genetics, VIB, University Gent
Ledeganckstraat 35, 9000 Gent, Belgium
[3] INRA associated laboratory, VIB, University Gent
Ledeganckstraat 35, 9000 Gent, Belgium

gert.thijs@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 co-expressed genes. Here we present a modification to the original Gibbs Sampling algorithm [16]. We introduce a new variable $\gamma_{k,c}$, which is the probability of finding c copies of the motif in sequence S_k. 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 co-expressed 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.

Introduction

Microarray technology allows biologist to monitor the mRNA expression levels of several thousands of genes in one experiment [6, 17, 27, 28, 30]. 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 [29, 31, 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 [2, 7, 11, 21, 29, 30, 32]. 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 co-expression indicates co-regulation.

Co-regulated 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 co-expressed genes [3, 7, 26, 39].

Many researchers have been working on computational methods to detect regulatory elements. The algorithms to find regulatory elements can be divide into two classes: word analysis methods [14, 33, 34, 35] and methods based on probabilistic sequence models [1, 12, 16, 18, 22, 26, 37]. The word analysis methods are based on the frequency analysis of oligonu-

cleotides 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 we present a modification of the original Gibbs Sampling algorithm by Lawrence et al. [16]. We present a probabilistic framework to estimate the expected number of copies of a motif in each sequence in the dataset. We have also introduced 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. We have succesfully tested our implementation on different data sets of intergenic sequences.

Algorithm and Implementation

Finding multiple copies

Applying clustering to the gene expression profiles of a microarray experiments gives several groups of co-expressed genes. Following the basic model, we can assume that co-expression indicates co-regulation, but it is expected that only a subset of the co-expressed 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:

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

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

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 [12] 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 [15] and GeneMark.hmm [19], 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:

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

The full description and details of this background model are described elsewhere [8].
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 only sequences in the training set that are in the intergenic region and that do not overlap with coding sequences. At the moment we have only constructed an independent background model for *Arabidopsis thaliana*. But 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.

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 $\tilde{S} = \{S_i | i \neq z\}$ 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 $W_x = P_x/P_x^0$

$$P_x = P(S_z|x, \theta_W, B_m), \qquad P_x^0 = P(S_z|B_m)$$

(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.

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*. 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 found. 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, too much noise will be introduced in the motif model.
- Overlap (O): This parameter defines the allowed overlap between the different motifs.

Results

G-box sequences

We examined several data sets of upstream DNA sequences to test the performance of our implementation of the motif sampler. To validate the motif sampler we construct a data set with a known regulatory element involved in light regulation in plants, G-box. This data set consists of 33 sequences containing 500bp upstream of the translation start. The sequences were selected from PlantCARE [25]. 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 postions of the motif in the sequences.

We experimented with several combinations of the parameters. Here we show the results when we searched for a motif of length 8 that can have 4 copies and we used the *Arabidopsis* background model of order 3. The first motif found in this experiment has the consensus: `kCCACGTG`, that is the G-box consensus. The other motifs found `kCACmCAn`, `mAAGnAAA`, `TTGrAmAA`, `CAAnwTCm` and `CnmnnmTC` are of minor importance, since the likelihood of these motifs was much smaller than the likelihood of the G-box motif.

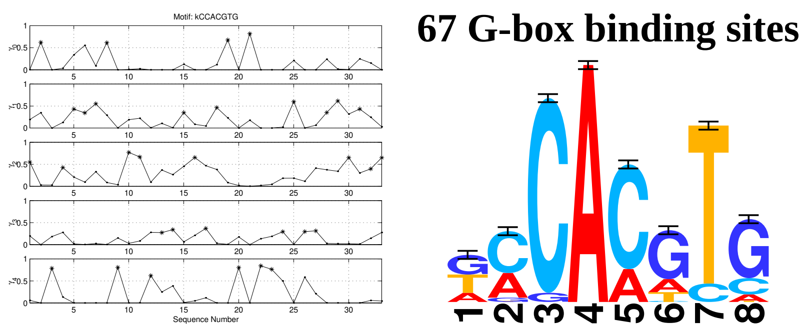

Figure 1: Weighting factors $\gamma_{k,c}$, with $c = 0, \ldots, 4$ and $k = 1, \ldots, 33$, associated with the found motif, `kCCACGTG`, in the 33 G-box sequences. The stars indicate the expected number copies of the motif for each sequence. The logo of the G-box is given on the right.

Figure 1 shows the weighting factors $\gamma_{k,c}$ associated with the motif found in the 33 G-box sequences and the logo of the motif on the right. The weighting factors $\gamma_{k,c}$ are ordered from top to bottom in ascending order of c, from 0 to 4. The X-axis corresponds to the sequence numbers. The expected number of copies in a sequence is indicated by a star in the corresponding row of the plots. The first subplot corresponds to $c = 0$ and shows the probability of not finding the motif in the sequences. There are 4 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. The next sub plots show the probability of finding

1 to 4 copies of the motif. We can see that there are several sequences indicated to have 3 or 4 copies. Since in most of the sequences there is only one known position of the G-box, the other indications can only be said to be putative G-boxes and they have to be checked experimentally. Since 4 copies are allowed per sequence more noise will introduced in the motif model and the logo will be more degenerate.

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 [20]. The genes were selected from several bacterial species of the alpha-proteobacteria *Azospirillum brasilense*, *Paracoccus denitrificans*, *Rhodobacter sphaeroides*, *Rhodobacter capsulatus*, *Sinorhizobium meliloti* and *Escherichia coli*(γ-proteobacteria. The data set contains 10 sequences with varying length.

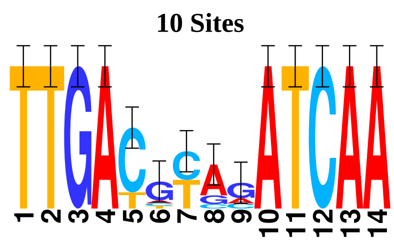

Figure 2: FNR binding site logo

The background model was in the case compiled from the sequence data of order 1. The order is here 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 2. 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 motif but only two unknown motifs are found. They both occur only in 3 sequences out of 10: `sGGyCGAATGGTCG` and TTCATGACAGTCCT.

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.

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 co-expressed genes. As a test case we use the data from Reymond *et al.*, where the gene expression in response to mechanical wounding was measured [24]. 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 co-expressed genes we use a clustering algorithm developed in our group [4]. As input parameters to the clustering, we use a small radius: 1.2 and the minimal number of genes in a cluster is set to be at least 3. This results in 8 small clusters of co-expressed genes. We search for motifs in all 8 clusters. We select the sequence 500bp upstream of translation start for each gene. We use the *Arabidopsis* background model of order 3 as background model in the experiments. 2 different motif lengths, 8 and 12, are tried and each experiment was repeated 10 times, to distinguish between reliable motifs and motifs that are found just by chance. There are only motifs found in the 4 clusters containing more than 3 genes. Table 2 gives an overview of the most important results.

	Consensus	PLantCARE	Function
1	TAArTAAGTCAC	TGAGTCA	tissue specific GCN4-motif
		CGTCA	MeJA-responsive element
	ATTCAAATTTmA	ATACAAAT	element associated to GCN4-motif
	CTTCTTCGATCT	TTCGACC	elicitor responsive element
		(T)TGAC(C)	Box-W1, elicitor responsive element
2	CCCGCGTTTCAA	CCCCCG	enhancer like element
	ATTTTGACyCGy	TGACG	MeJa responsive element
		(T)TGAC(C)	Box-W1, elicitor responsive element
	mACGTCAC	CGTCA	MeJA responsive element
		ACGT	ABA response element (ABRE)
3	wATATATATmTT	TATATA	TATA-box like element
4	yTGACCGTCCsA	CCGTCC	meristem specific activation of H4 gene
		CCGTCC	A-box, light or elicitor responsive element
		TGACG	MeJA responsive element
		CGTCA	MeJA responsive element
	CACGTGGmAAsA	CACGTG	G-box light responsive element
		ACGT	ABA response element

Table 2: Results motif search in 4 clusters. In the second column the consensus of the motif is given, in the next column is the corresponding motif in PlantCARE and finally there is a short explanation of the motif.

To assign a functional interpretation to the motifs, the consensus of the motifs are compared with the entries in PlantCARE. Several interesting motifs are found: methyl jasmonate(MeJa) responsive elements, elicitor-responsive elements and the abscisic acid reponse 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 [9, 23, 13, 10]. 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 these regulators.

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 contribution is the inclusion of an higher order background model instead of using single nucleotide frequencies as the background model. The second 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. In the examples we show that the algorithm is able to find over-represented motifs in sequences upstream of translation start.

The algorithm is accessible through a simple web interface. We have developed an algorithm that only has a limited number of parameters. We also try to make the interface easy to use and to understand. The parameter definitions are straightforward and easy to interpret.

To further improve the usability and performance of the motif sampler we will work on several add-ons. First we would like to implement a method to automatically detect the best 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 expected number of copies is one of our priorities. This is of course closely related to the scores associated with the motifs. The ultimate goal is to develop a robust algorithm with as few as possible user parameters. When the parameters are well handled, 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.

Acknowledgments

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

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] A. Brazma, J. Vilo, E. Ukkonen, and K. Valtonen. Data mining for regulatory elements in yeast genome. In *ISMB 97*, pages 65–74, 1997.

[3] Philip 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. *Nucl. 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] Thijs G., Lescot M., Marchal K, Rombauts S., De Moor B., Rouz P., and Moreau Y. Use of an higher order background model to detect over-represented motifs in dna sequences. *submitted*, 2000.

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

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

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

[12] J. D. Hughes, Preston 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.

[13] Rouster J., Leah R., Mundy J., and Cameron-Mills V. 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.

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

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

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

[21] 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.

[22] 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.

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

[24] 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.

[25] 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.

[26] 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.

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

[28] 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.

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

[30] 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.

[31] Z. Szallasi. Genetic network analysis in light of massively parallel biological data acquisition. In *Proceedins PSB99*, volume 4, pages 5–16, 1999.

[32] 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.

[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 *Pacific Symposium on Biocomputing 2000*, Honolulu, Hawai, 2000. in press.

[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 *Pacific Symposium on Biocomputing 2000*, volume 5, pages 467–486, 2000.