

# Using Symmetric Causal Independence Models to Predict Gene Expression from Sequence Data

Rasa Jurgelenaite<sup>1</sup>, Tom Heskes<sup>1</sup>, and Tjeerd Dijkstra<sup>2</sup>

<sup>1</sup> Institute for Computing and Information Sciences, Radboud University Nijmegen,  
PO Box 9010, 6500 GL Nijmegen, The Netherlands  
{rasa, tomh}@cs.ru.nl

<sup>2</sup> Department of Parasitology, Leiden University Medical Center,  
PO Box 9600, 2300 RC Leiden, The Netherlands  
t.dijkstra@lumc.nl

**Abstract.** We present an approach for inferring transcriptional regulatory modules from genome sequence and gene expression data. Our method, which is based on symmetric causal independence models, is both able to model the logic behind transcriptional regulation and to incorporate uncertainty about the functionality of putative transcription factor binding sites. Applying our approach to the deadliest species of human malaria parasite, *Plasmodium falciparum*, we obtain several striking results that deserve further (biological) investigation.

**Key words:** Transcriptional regulatory networks, symmetric causal independence models, *Plasmodium falciparum*

## 1 Introduction

One of the major challenges facing biologists is to understand the transcriptional regulation of genes, which is critical for the development, complexity and homeostasis of all living organisms. The introduction of DNA microarray technology [26], which enables researchers to simultaneously measure the concentration of RNA transcripts from a single sample of cells or tissues, has offered the possibility to infer large-scale transcriptional regulatory networks for various organisms. The algorithms developed for this purpose can be grouped into two general strategies: an influence strategy, which seeks to identify regulatory influences between RNA transcripts, and a physical strategy, which seeks to identify the proteins that regulate transcription and the DNA motifs to which the proteins bind [11]. In this paper, we propose a method following the latter strategy, which has the advantage of being able to combine genome sequence data and RNA expression data to enhance the specificity and sensitivity of predicted interactions.

The physical strategy methods that make use of both RNA expression data and genome sequence data rely on the assumption that genes with similar expression profiles share common regulatory mechanisms. Based on the way in which the two sources of data are related, we can distinguish three groups of these methods. The first group includes the methods that first cluster genes on

the basis of their expression patterns and then search for putative motifs in the upstream regions of the genes in each cluster. The early methods following this approach searched for individual transcription factor binding site patterns in upstream regions of the coexpressed genes (see e.g. [5, 8, 28]), while the more recent algorithms search for DNA target sites for cooperatively binding transcription factors [12, 18]. The methods in the second group work in the opposite direction, first identifying a set of candidate motifs and then trying to explain RNA expression using these motifs [7, 15, 22]. Finally, the algorithms in the last group use both sources of data together. These methods use one or more iterations of the following procedure: first, genes are clustered or grouped according to their expression data, then the search for motifs in the upstream regions of the coexpressed genes is performed, and, finally, the motifs identified are used to build models that predict the expression pattern of the gene (see e.g. [2, 27]).

A key feature of transcriptional regulation of gene expression in eukaryotes is that genes are often regulated by more than one transcription factor [30]. A number of approaches have been proposed to address the combinatorial nature of transcriptional regulation. One approach is based on the assumption that the influence of different transcription factors on gene expression is additive. The studies based on this approach use a simple linear regression to relate transcription factor binding sites to gene expression values [7, 17]. A probabilistic model by Segal et al. [27] assumes that genes are partitioned into modules, which determine the gene expression profile. The strength of the association of a gene with a module is the sum of its weighted motifs, where each weight specifies the extent to which the motif plays a regulatory role in the module. These approaches, however, cannot identify synergistic motif combinations that control gene expression patterns. Algorithms have been developed to model the synergy between two transcription factors that bind to sites located anywhere in the upstream region [22] or sites that are spatially close to each other [7, 12]. Beer and Tavazoie [2] present an approach which utilizes AND, OR and NOT logic to capture combinatorial effects of transcription factors in the regulation of gene expression. This method is not only able to infer combinatorial rules that involve more than two transcription factors, but it also includes constraints on motif strength, orientation and relative position. A similar method has been reported by Hvidsten et al. [15]. To link transcription factor binding site combinations to genes with particular expression profiles, the method extracts IF-THEN rules which correspond to AND logic.

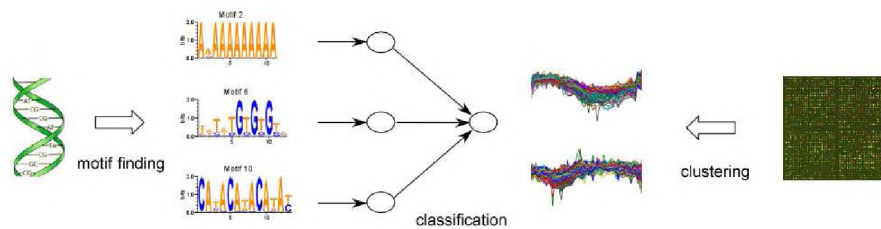
Although the methods that model combinatorial effects of the motifs have appealing properties, their drawback is their inability to cope with uncertainty in the transcription factor binding sites that are identified. The robustness of the method in the face of uncertainty is important, as non-functional transcription factor binding sites can be readily found throughout the genome, including promoters [31]. We present an approach which is both able to model the logic behind transcriptional regulation and to incorporate uncertainty about the functionality of putative transcription factor binding sites. Our probabilistic method, which is based on symmetric causal independence models, extends the earlier methods

that infer combinatorial rules in two important directions. First, we use a broad class of Boolean functions, symmetric Boolean functions, to capture combinatorial effects of transcription factors in the regulation of gene expression. Second, the motifs contribute to the regulation of a gene through hidden variables; thus, the method is able to cope with non-functional transcription factor binding sites.

In this paper, we apply our method to *Plasmodium falciparum*, which is the deadliest species of the parasite that causes malaria in humans. Human malaria infects between 300 and 500 million people and causes up to 2.7 million deaths annually, mostly among young children in Sub-Saharan Africa [6]. In many endemic countries, malaria is also responsible for economic stagnation [23]. A good understanding of transcriptional regulation in this organism is important for devising new ways to disrupt the parasite’s life cycle.

## 2 Methodology

In this section, we present our approach based on symmetric causal independence models for inferring transcriptional regulatory modules from genome sequence and gene expression data. The underlying assumption in this approach is that genes in the different clusters share common regulatory mechanisms. When trying to separate the genes in one cluster from all others, we aim to find motifs and their interactions that are specific to specific regulatory mechanisms. We start our method (Figure 1) with a ‘data pre-processing’ step, where we use a motif-finding algorithm to identify putative transcription factor binding motifs and we cluster genes according to their expression profiles. Then, for each cluster of genes that exhibited significant changes, we learn a symmetric causal independence model, which, given the binding sites of a gene, classifies the gene as belonging to the cluster or not. Finally, we analyze the results of experiments and identify potential transcription factors binding to the motifs that play a regulatory role in gene expression. All these steps are described in detail further in this section.



**Fig. 1.** Overview of the proposed approach.

## 2.1 Finding transcription factor binding motifs

We extracted the DNA sequence 1000 bp upstream from the initiation codon of each of 5404 *Plasmodium falciparum* genes using PlasmoDB release 5.2. In instances where the upstream regulatory region overlapped with another open reading frame, we extracted only the sequence between the open reading frames. To find over-represented motifs, the extracted sequences were analyzed using the AlignACE program [13]. We set the GC background parameter to 0.13 (the fractional GC background for these regions), the number of columns to align to 10 and the number of expected sites to 5.

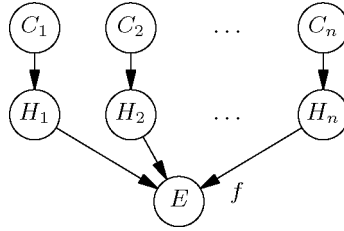
## 2.2 Clustering of the RNA expression data

We used a *Plasmodium falciparum* 3D7 strain RNA expression data set [4]. We downloaded data that were normalized and median-centered and we only used data for those oligonucleotides that have a corresponding open reading frame assigned from PlasmoDB. We discarded the genes for which more than 20% of measurements were missing. A number of open reading frames had more than one oligonucleotide measured; we averaged the measurements of these open reading frames. After the data had been  $\log_2$  transformed, we imputed missing values using the weighted K-nearest neighbours method. We chose to use this data imputation method as it has been shown to provide a more robust and sensitive missing value estimation in microarray data than a singular value decomposition based method or the commonly used row average method [29]. The weighted K-nearest neighbours method uses a weighted average of values from the  $K$  genes closest to the gene of interest as an estimate for the missing value. Based on the results reported in [29], we chose the value of  $K$  to be 15 and Euclidean distance as a metrics for gene similarity.

We used the K-means algorithm [19] with random initializations to cluster the genes according to their RNA expression data. Since the K-means algorithm is known to sometimes get stuck in a local optimum, we ran the algorithm 10 times for each number of clusters. To select the optimal number of clusters we used the so-called C-index [14], which has been shown to outperform 13 other indices for determining the number of clusters in binary data sets when the data are clustered using the K-means algorithm [10].

## 2.3 Learning symmetric causal independence models

The global structure of a *symmetric causal independence model* is shown in Figure 2; it expresses the idea that causes  $C_1, \dots, C_n$  influence a given common effect  $E$  through hidden variables  $H_1, \dots, H_n$  and a symmetric Boolean function  $f$ . All variables in this model are binary; the hidden variable  $H_i$  is considered to be a contribution of the cause variable  $C_i$  to the common effect  $E$ . The function  $f$  represents in which way the hidden effects  $H_i$ , and indirectly also the causes  $C_i$ , interact to yield the final effect  $E$ . To learn more about symmetric causal independence models and learning them, see [16].



**Fig. 2.** Symmetric causal independence model

In this paper, we use symmetric causal independence models as a technique to model combinatorial effects of transcription factor binding motifs in the regulation of gene expression. Transcription factor binding sites are causes in this model, where the positive state of this variable is presence or absence of the motif, depending on the motif’s effect on expression of genes in the cluster. The positive state of the effect variable represents gene belonging to the cluster, and the negative state represents gene belonging to any other cluster.

We used a greedy approach to select the motifs whose absence or presence contributes to the difference between the expression of genes belonging to a given cluster and the expression of the other genes. First, we ranked all motifs based on their mutual information scores, where the mutual information measures the mutual dependence of the variable  $M$  that represents a motif and the class variable  $C$  and is defined as:

$$I(M; C) = \sum_{m \in M} \sum_{c \in C} \Pr(m, c) \log \frac{\Pr(m, c)}{\Pr(m) \Pr(c)}.$$

Then, we built a model from the  $h$  highest ranked motifs. We started from a model containing only a leaky cause, then iteratively added the next highest ranked motif and evaluated the model thus obtained. If the new model did not have a higher score than the previous model, the motif was removed from the model. Since there are  $2^{n+1}$  symmetric Boolean functions for a model with  $n$  variables that represent motifs, evaluating all the resulting models is too expensive computationally. Therefore, we restricted the interaction function space to the Boolean threshold functions. This restriction means that for every added motif we only had to evaluate two models, a gene model with the interaction function  $\tau_k$  and a gene model with the interaction function  $\tau_{k+1}$ , where  $\tau_k$  is the interaction function from the model with the highest score. We evaluated each model using the classification accuracy on the validation set.

To solve the problem of unbalanced data (different class size, see Table 1), we added as many copies of every gene from the smaller class as was needed for this class to amount for at least half of the genes. To learn the parameters of the gene model, we ran 25 iterations of the EM algorithm, computed the classification accuracy on the validation set after each iteration and chose those parameters that provided the highest score.

## 2.4 Evaluation of the results

We used two error estimation methods, cross-validation and bootstrap, to evaluate the models learned. The cross-validation scheme was used to examine the predictive performance of the models, whereas the bootstrap approach was used to evaluate the reliability of the model parameters. For both methods, we performed 100 runs, and the data was split into training, validation and test sets. The validation set was used to choose the number of iterations of the EM algorithm and the threshold function; the results reported were obtained using an independent test set.

We used the results of the bootstrap approach to test for potential synergistic motif pairs. From the results of the bootstrap approach, we estimated  $\hat{\theta} = (\theta_1, \theta_2, \dots)$ , where  $\theta_i$  is the probability that motif  $M_i$  will be chosen as a feature in the model. We introduce a variable  $X_{jk}$  that specifies four possible combinations of occurrence of the motifs  $M_j$  and  $M_k$ . Our null hypothesis was that  $X_{jk}$  follows multinomial distribution, with each trial resulting in one of 4 possible outcomes with probabilities  $p_1 = (1 - \hat{\theta}_j)(1 - \hat{\theta}_k)$ ,  $p_2 = \hat{\theta}_j(1 - \hat{\theta}_k)$ ,  $p_3 = (1 - \hat{\theta}_j)\hat{\theta}_k$ ,  $p_4 = \hat{\theta}_j\hat{\theta}_k$ , and the number of trials  $n$  being equal 100. To measure the discrepancy between the observed and expected counts, we used Pearson's chi-square statistic:

$$X^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

where  $i$  is a possible outcome and the expected count  $E_i = np_i$ .

To compare our classifier to a classifier which assigns all genes to the biggest class, we used a binomial test described in [24]. The test uses the number of cases  $n$  for which the two classifiers produce a different output, and the number of cases  $s$  where the output of the examined classifier was correct, while the output of the reference classifier was wrong. Under the null hypothesis that the two classifiers perform equally well, we compute:

$$p = 2 \sum_{i=s}^n \frac{n!}{i!(n-i)!} 0.5^n.$$

## 2.5 Identifying potential transcription factors binding to the motifs

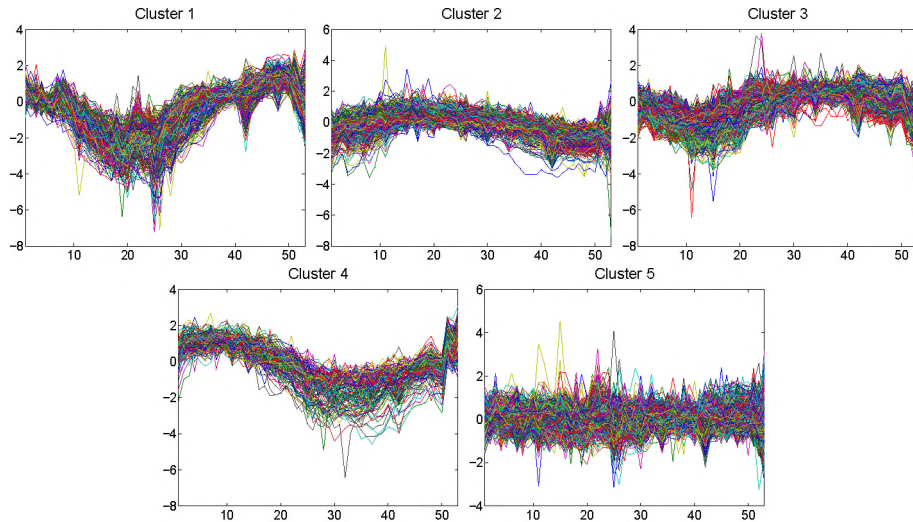
To identify potential transcription factors binding to the motifs, we used comparative genome analysis, which is based on the fact that sequence similarity might reflect functional similarity. Identification, which was done separately for each motif, involves three steps. Firstly, we used STAMP [20], a web tool for exploring DNA-binding motif similarities, to find a number of the closest matches for a given motif in 13 supported databases. Secondly, for each match found, we checked whether the database where the motif is stored reports a transcription factor binding to it. Finally, if the transcription factor is known, we used BLAST [1, 25] to find the most similar protein sequences from the Plasmodium falciparum protein database.

### 3 Experimental Results

#### 3.1 Transcription factor binding motifs found and clusters obtained

AlignACE found 100 transcription factor binding motifs in the given upstream sequences. The motifs that were found to be the most important features for classifying the genes will be discussed later in this section.

We chose the number of clusters to be 5, as the C-index curve had an ‘elbow’ at this value. Figure 3 presents the clusters obtained, which are comparable to the four characteristic stages of intraerythrocytic parasite morphology discussed by Bozdech et al. [4], as the vast majority of genes induced in every one of the stages belong to one of four clusters. Cluster 5 is a cluster of genes whose expression did not show a significant change. The correspondence among the characteristic stages and the clusters and the cluster sizes are given in Table 1.



**Fig. 3.** Clusters of *Plasmodium falciparum* RNA expression data.

#### 3.2 Models learned

We learned the models for the first four clusters, i.e. the clusters of genes whose expression changed throughout the intraerythrocytic stage.

The classification accuracy of the gene models learned using the cross-validation procedure explained in 2.4 is reported in Table 1. The p-values for the null hypothesis that the gene models perform equally well as a classifier which assigns all genes to a bigger class are less than  $10^{-10}$ .

Table 2 lists the motifs that were most often selected as features of the gene model. Due to space limitations, we report only those motifs that were selected as

**Table 1.** A brief description of the clusters, the number of the genes assigned, the corresponding characteristic stage of intraerythrocytic parasite morphology; and classification accuracy obtained using the cross-validation procedure.

Cluster	Number of genes	Corresponding stage	Accuracy obtained (%)	Baseline accuracy (%)
1	329	schizont	60.48	50.79
2	1033	ring/early trophozoite	61.52	52.52
3	985	trophozoite/early schizont	59.16	50.90
4	144	early ring	63.21	51.30
5	1344	-	-	-

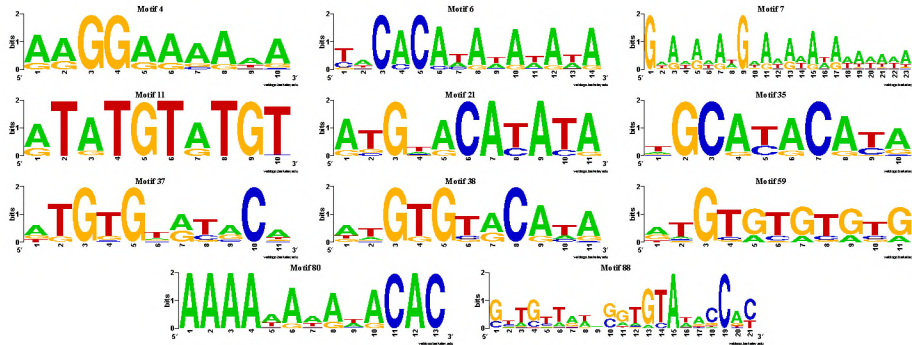
features of the model in more than 50 bootstrap runs. Some of the motifs appear in more than one cluster; however, their weighting is different (not shown) and they can be either ‘present’ or ‘absent’ (the presence or absence is a positive state of the corresponding variable in the model). Sequence logos of the motifs, which were generated using the WebLogo program [9], are shown in Figure 4. A study of the positive states of the variables representing the motifs selected as features of the model in more than 20 bootstrap runs reveals a distinct pattern. The variables in models for cluster 2 and cluster 4 represent the absence of the motifs, while the variables in models for cluster 1 and cluster 3 mainly represent the presence of the motifs. Even though there are 6 motifs that break this pattern in clusters 1 and 3, these motifs are found in a very small number of genes (from 1 to 5 % of genes); the other motifs selected are much more common in genes. The summary of these results is presented in Table 3.

**Table 2.** Motifs that were selected as features of the model in more than half of the bootstrap runs; the number of runs the motif was selected is given in parentheses. ‘Present’ motifs are written in roman, ‘absent’ motifs are written in roman.

Cluster	Motifs selected more than 50 times
1	Motif 38 (100), Motif 37 (95), Motif 6 (65), Motif 59 (65), Motif 11 (63), Motif 21 (55)
2	<b>Motif 6</b> (98), <b>Motif 35</b> (93), <b>Motif 37</b> (89), <b>Motif 38</b> (89), <b>Motif 11</b> (75), <b>Motif 21</b> (68), <b>Motif 59</b> (68), <b>Motif 7</b> (64), <b>Motif 80</b> (59)
3	Motif 6 (100), <b>Motif 88</b> (82), Motif 38 (67), Motif 59 (60), Motif 21 (51)
4	<b>Motif 6</b> (99), <b>Motif 11</b> (93), <b>Motif 4</b> (55)

Interpretation of the probabilities of the hidden variables is somewhat tricky as they highly depend on the number of input variables and the interaction function in the model, which currently vary a lot from one bootstrap run to another. Nevertheless, there is a pattern which suggests that probabilities of





**Fig. 4.** Sequence logos of the motifs that were selected as features of the model in more than half of the bootstrap runs.

**Table 3.** Positive states of variables representing the motifs that were selected as features of the model in at least 20 bootstrap runs.

Cluster		Motifs selected		Positive state: absence		Positive state: presence	
1	14	2	12				
2	14	13	1				
3	10	4	6				
4	15	15	0				

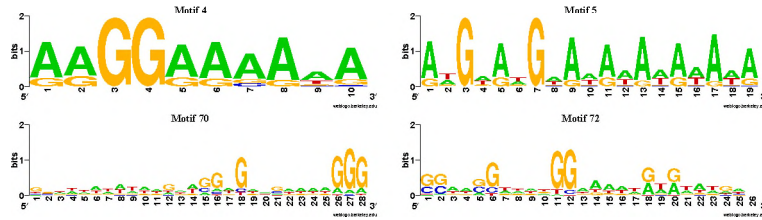
hidden variables contain information about functionality of putative transcription factor binding sites. The pattern emerges when we compare probabilities of hidden variables with the average of probabilities of the other hidden variables in the model. The probabilities in clusters of ‘absent’ motifs were almost the same, while the probabilities in clusters of ‘present’ motifs differed much more and the majority of motifs had the tendency to have corresponding probabilities below or above the average.

To find statistically significant occurrences of motif pairs, we tested all possible pairs of the motifs selected as causes in the model in at least 20 bootstrap runs (see 2.4 for the description of the method). We rejected the null hypothesis at the significance level of 0.05 (corrected for multiple testing) for two motif pairs from cluster 4: for the pair of motifs 70 and 72 (with p-value of 0.0055), and the pair of motifs 4 and 5 (with p-value of 0.0174). These motifs were selected together to be features in the model more often than it would be expected. Sequence logos for potential synergistic motif pairs are shown in Figure 5.

### 3.3 Potential transcription factors binding to the motifs

We present the most significant findings for the motifs reported in Figure 4.

Motifs 6, 11 and 35 have the same closest match - the binding site of fruit fly’s transcription factor Topoisomerase 2, reported in FlyReg database [3]. The most



**Fig. 5.** Sequence logos of potential synergistic motif pairs.

significant alignment in *Plasmodium falciparum* is PF14\_0316, putative DNA topoisomerase 2, whose protein sequence is nearly identical (E value of 0.0).

Another gene of *Plasmodium falciparum* that is a potential transcription factor binding to at least two of the motifs discussed is PF14\_0175, which is annotated as a hypothetical protein in PlasmoDB. One of the closest matches for motif 7 is MCM1+SFF\_M01051 reported in TRANSFAC database [21]. The most significant alignment for MCM1, which is yeast transcription factor involved in cell-type-specific transcription and pheromone response and plays a central role in the formation of both repressor and activator complexes, is PF14\_0175 (E value of  $10^{-5}$ ). Another motif to which this transcription factor could bind is motif 80; this possible connection was found through a different transcription factor in a different organism. Motif FOXP1\_M00987 reported in TRANSFAC is a close match to motif 80. Mouse transcription factor FOXP1 which binds to this motif is thought to repress expression of epithelial genes in the lung and reduce expression from promoters of mouse CC10 gene G002818. The most significant alignment for variants T04812 and T04813 of FOXP1 in *Plasmodium falciparum* is PF14\_0175 (E value of  $10^{-8}$ ).

A gene which is found as potential transcription factor for one third of the motifs analyzed is PFL0465c, zinc finger transcription factor (krox1). For motif 4, the connection was found through motif Helios\_M01004 reported in TRANSFAC and mouse transcription factor IKAROS family zinc finger 2, Helios, whose functions include zinc ion binding, DNA binding and nucleic acid binding (E value of  $7 \cdot 10^{-6}$ ). For motif 21, the connection was found through motif CF2-II\_M00012 reported in TRANSFAC and fruit fly transcription factor CF2-II, a late activator in follicle cells during chorion formation (E value of  $10^{-6}$ ).

## 4 Discussion and Future Work

We have presented an approach which is both able to model the logic behind transcriptional regulation and to incorporate uncertainty about the functionality of putative transcription factor binding sites. Another advantage of our technique is that it does not require other biological knowledge than genome sequence data and RNA expression data to validate the results. Since we do not use expression data while searching for putative regulatory motifs, the accuracy of the models

in predicting gene expression pattern is an unbiased measure of the soundness of the models learned.

Experimental results revealed the lack of consistency in the properties of the models learned. This inconsistency could be caused by the lack of additional constraints on the motifs, such as position relative to the translation start, orientation and functional depth. Therefore, the next step in our research is to implement normal and binomial approximations to Poisson binomial distribution, which will help to reduce computational complexity of the EM algorithm. Reduced computational complexity will enable us to test more interaction functions and to examine the additional constraints on the motifs.

We will also continue our discussions with biologists to find the explanation to the experimental results, especially, the pattern of clusters of ‘present’ and ‘absent’ motifs, and potential transcription factors binding to the motifs.

**Acknowledgments.** We would like to thank Michael A. Beer, Saeed Tavazoie, Zbynek Bozdech and Mahony Shaun for helpful discussions.

## References

1. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25** (1997) 3389–3402
2. Beer, M.A., Tavazoie, S.: Predicting gene expression from sequence. *Cell* **117** (2004) 185–198
3. Bergman, C.M., J.W Carlson, S.E. Celniker: *Drosophila* DNase I footprint database: A systematic genome annotation of transcription factor binding sites in the fruitfly, *D. melanogaster*. *Bioinformatics* **21** (2005) 1747–1749
4. Bozdech, Z., Llinas, M., Pulliam, B., Wong, E.D., Zhu, J., DeRisi, J.: The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*. *PLoS Biology* **1** (2003) 85–100
5. Brazma, A., Jonassen, I., Vilo, J., Ukkonen E.: Predicting gene regulatory elements in silico on a genomic scale. *Genome Research* **8** (1998) 1202–1215
6. Bremen, J.: The ears of the hippopotamus: manifestations, determinants, and estimates of the malaria burden. *American Journal of Tropical Medicine and Hygiene* **64** (2001) 1–11
7. Bussemaker, H.J., Li, H. and Siggia, E.D. Regulatory element detection using correlation with expression. *Nature Genetics* **27** (2001) 167–171
8. Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P.O., Herskowitz, I.: The transcriptional program of sporulation in budding yeast. *Science* **282** (1998) 699–705
9. Crooks G.E., Hon G., Chandonia J.M., Brenner S.E.: WebLogo: A sequence logo generator. *Genome Research* **14** (2004) 1188–1190
10. Dimitriadou, E., Dolničar, S. Weingessel, A.: An examination of indexes for determining the number of clusters in binary data sets. *Psychometrika* **67** (2002) 137–160
11. Gardner, T.S., Faith, J.: Reverse-engineering transcription control networks. *Physics of Life Reviews* **2** (2005) 65–88
12. GuhaThakurta, D., Stormo, G.: Identifying target sites for cooperatively binding factors. *Bioinformatics* **17** (2001) 608–621

13. Hughes, J.D., Estep, P.W., Tavazoie S., Church, G.M.: Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *Journal of Molecular Biology* **296** (2000) 1205–1214
14. Hubert, L.J., Levin, J.R.: A general statistical framework for accessing categorical clustering in free recall. *Psychological Bulletin* **83** (1976) 1072–1082
15. Hvidsten, T.R., Wilczyński, B., Kryshchak, A., Tiuryn, J., Komorowski, J., Fidelis K.: Discovering regulatory binding-site modules using rule-based learning. *Genome Research* **15** (2005) 856–866
16. Jurgelenaite, R., Heskes, T.: EM algorithm for symmetric causal independence models. *Proceedings of the Seventeenth European Conference on Machine Learning* (2006) 234–245
17. Keleş, S., van der Laan, M., Eisen, M.B.: Identification of regulatory elements using a feature selection method. *Bioinformatics* **18** (2002) 1167–1175
18. Liu, X., Brutlag, D., Liu, J.S.: BioProspector: discovering conserved DNA motifs in upstream regulatory regions of coexpressed genes. *Pacific Symposium on Biocomputing* (2001) 127–138
19. MacQueen, J.B.: Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* **1** (1967) 281–297
20. Mahony, S., Benos, P.V.: STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Research* (2007) in press
21. Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., Voss, N., Stegmaier, P., Lewicki-Potapov, B., Saxel, H., Kel, A.E., Wingender, E.: TRANSFAC and its module TRANSCmpel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Research* **34** (2006) D108–110
22. Pilpel, Y., Sudarsanam, P., Church, G.M.: Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature Genetics* **29** (2001) 153–159
23. Sachs, J.D., Malaney, P.: The economic and social burden of malaria. *Nature* **415** (2002) 680–685
24. S. Salzberg: On comparing classifiers: pitfalls to avoid and a recommended approach, *Data Mining and Knowledge Discovery* **1** (1997) 317–327
25. Schäffer, A.A., Aravind, L., Madden, T.L., Shavirin, S., Spouge, J.L., Wolf, Y.I., Koonin, E.V., Altschul, S.F.: Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Research* **29** (2001) 2994–3005
26. Schena, M., Shalon, D., Davis, R.W., Brown, P.O.: Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270** (1995) 467–470
27. Segal, E., Yelensky, R., Koller, D.: Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. *Bioinformatics* **19** (2003) 1273–1282
28. Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J., Church, G.M.: Systematic determination of genetic network architecture. *Nature Genetics* **22** (1999) 281–285
29. Troyanskaya O., Cantor M., Sherlock G., Brown P., Hastie T., Tibshirani R., Botstein D., Altman R.B.: Missing value estimation methods for DNA microarrays. *Bioinformatics* **17** (2001) 520–525
30. Wagner, A.: Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. *Bioinformatics* **15** (1999) 776–784
31. Werner, T.: Models for prediction and recognition of eukaryotic promoters. *Mammalian Genome* **10** (1999) 168–175