BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 12 2009, pages 1484–1491
doi:10.1093/bioinformatics/btp179

Genome analysis

Gene regulation in the intraerythrocytic cycle of *Plasmodium falciparum*

Rasa Jurgelenaite[1,*,†], Tjeerd M. H. Dijkstra[1], Clemens H. M. Kocken[2] and Tom Heskes[1]

[1]Institute for Computing and Information Sciences, Radboud University Nijmegen, Nijmegen and [2]Department of Parasitology, Biomedical Primate Research Center, Rijswijk, The Netherlands

Received on December 2, 2008; revised and accepted on March 26, 2009

Advance Access publication March 31, 2009

Associate Editor: Martin Bishop

ABSTRACT

Motivation: To date, there is little knowledge about one of the processes fundamental to the biology of *Plasmodium falciparum*, gene regulation including transcriptional control. We use noisy threshold models to identify regulatory sequence elements explaining membership to a gene expression cluster where each cluster consists of genes active during the part of the developmental cycle inside a red blood cell. Our approach is both able to capture the combinatorial nature of gene regulation and to incorporate uncertainty about the functionality of putative regulatory sequence elements.

Results: We find a characteristic pattern where the most common motifs tend to be absent upstream of genes active in the first half of the cycle and present upstream of genes active in the second half. We find no evidence that motif's score, orientation, location and multiplicity improves prediction of gene expression. Through comparative genome analysis, we find a list of potential transcription factors and their associated motifs.

Contact: r.jurgelenaite@cmbi.ru.nl

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Malaria is caused by protozoan parasites of which *Plasmodium falciparum* causes up to 2 million deaths, mainly children under the age of 5, annually in sub-Saharan Africa (Bremen, 2001). Malaria is a poverty-related disease and vaccines are not yet available (Tetteh and Polley, 2007). In addition, resistance to the most commonly available antimalarials (chloroquine and antifolate drugs) has spread worldwide. A thorough understanding of the complex biology of the parasite, with developmental stages in humans and Anopheles mosquitoes, will help in designing more effective control strategies to combat malaria. Basic knowledge of one of the processes fundamental to *Plasmodium* biology, gene regulation including transcriptional control, is still lacking. The published *P.falciparum* genome sequence has not provided much grip on transcription control, since not many transcription factors could be identified and intergenic regions appeared to mainly consist of A+T sequences (Gardner *et al*., 2002).

In this article, we present a bioinformatics approach which combines gene expression data with genome sequence data to identify regulatory sequence elements and to learn more about gene regulatory mechanisms. A key feature of transcriptional regulation of gene expression in eukaryotes is that genes are often regulated by more than one transcription factor (Wagner, 1999). It has been suggested that combinatorial gene regulation might be the general mode of transcriptional regulation in *P.falciparum* (Essien *et al*., 2008; Van Noort and Huynen, 2006). A number of approaches have been proposed to address the combinatorial nature of transcriptional regulation. One group of approaches is based on the assumption that the influence of different transcription factors on gene expression is additive. The studies based on this assumption use linear regression to relate regulatory sequence elements to gene expression values (Bussemaker *et al*., 2001; Keleş *et al*., 2002). These approaches, however, cannot identify synergistic regulatory element 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 (Pilpel *et al*., 2001) or sites that are spatially close to each other (GuhaThakurta and Stormo, 2001). Beer and Tavazoie (2004) presented an approach which utilizes AND, OR and NOT logic to capture combinatorial gene regulation. 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 (Werner, 1999). We present an approach which is both able to capture the combinatorial nature of gene regulation and to incorporate uncertainty about the functionality of putative regulatory sequence elements. Our probabilistic method, which is based on noisy threshold models, a type of Bayesian network, extends the earlier methods that infer combinatorial rules in two directions. First, we consider a larger class of Boolean functions, Boolean threshold functions, to capture combinatorial effects. Second, the regulatory sequence elements contribute to the regulation of a gene through hidden variables that capture the probability that a sequence element is functional; thus, the method is able to cope with non-functional regulator binding sites.

*To whom correspondence should be addressed.

†Present address: Center for Molecular and Biomolecular Informatics, Radboud University Nijmegen Medical Center, Nijmegen, The Netherlands

Differently from other recent methods that use both expression and genome sequence data to explain gene regulation, we do not use expression data while searching for putative regulatory motifs. Therefore, the accuracy of the models in predicting the gene expression pattern is an unbiased measure of the soundness of the models learned, which allows us to test various hypotheses and to validate the results when knowledge about the gene regulation processes is not available. This property of our method is critical given how little is known about gene regulation in *P.falciparum*.

2 METHODS

Our approach to infer regulatory modules from genome sequence and RNA expression data is shown in Figure 1. The underlying assumption in the approach is that genes within a cluster share common regulatory mechanisms. We start with a preprocessing step, where we use a motif-finding algorithm to identify putative regulatory sequence elements and we cluster genes according to their expression profiles. Then, for each cluster of genes that exhibited significant changes in their expression, we learn noisy threshold models, which model combinatorial effects of gene regulators, and, given the putative regulatory sequence elements for a gene, classify the gene as belonging to the cluster or not.

The global structure of a noisy threshold model can be seen in Figure 1; it expresses the idea that causes (regulatory sequence elements) influence a given common effect (gene membership to a gene expression cluster) through hidden variables and a Boolean threshold function. All variables in this model are binary; the positive state of a cause variable corresponds to either the absence or presence of the motif, and the positive state of the effect variable represents that the gene belongs to the cluster. Hidden variables are considered to be a contribution of their parent variables, regulatory sequence elements, to the gene expression pattern; the absent causes do not contribute to the effect. The Boolean threshold function τ_k returns true when there are at least k trues among the hidden variables. The commonly used OR and AND functions are the extremes of a spectrum of threshold functions: the OR function is a threshold function τ_k with $k=1$ and the AND function is a threshold function τ_k where k equals the number of causes in the model.

2.1 Finding regulatory motifs

We extracted the DNA sequence 1000 bp upstream from the initiation codon of all *P.falciparum* protein coding genes using PlasmoDB release 5.2 (Bahl *et al*., 2003). 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 (Hughes *et al*., 2000). Default AlignACE parameters were used, except that the fractional GC content was set to 0.13 (the GC background of *P.falciparum* upstream regions) and the expected number of sites was set to five. Sequence logos of the motifs were generated using the WebLogo program (Crooks *et al*., 2004).

2.2 Clustering gene expression data

We used a *P.falciparum* 3D7 strain RNA expression dataset (Llinas *et al*, 2006). 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 >20% of the 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 (Troyanskaya *et al*., 2001). 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 Troyanskaya *et al*. (2001), we chose the value of K to be 15 and the Euclidean distance as a metric for gene similarity.

We used the K-means algorithm 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 c, where $c=2,\ldots,50$. To select the optimal number of clusters, we used the so-called C-index (Hubert and Levin, 1976), which has been shown to outperform 13 other indices for determining the number of clusters in binary datasets when the data are clustered using the K-means algorithm (Dimitriadou *et al*., 2002).

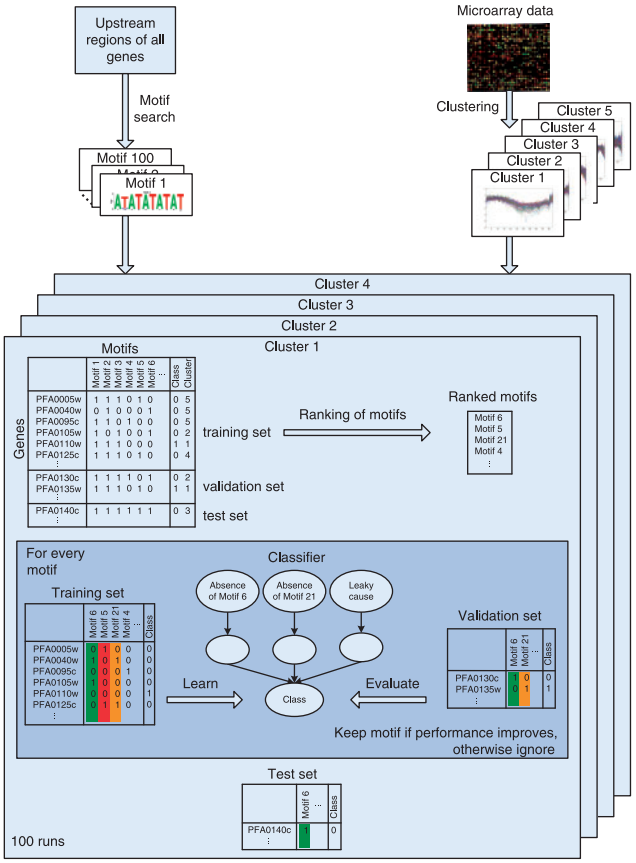

Fig. 1. Overview of the proposed approach. After data preprocessing is completed, we perform 100 runs of two error estimation methods, cross-validation and bootstrap, for every cluster. Every run starts with motif ranking, where motifs are ranked based on mutual information scores computed between the motifs and the class. An iterative greedy procedure used to learn a noisy threshold model adds the next highest ranked motif (orange color) to the current model. A motif that improves classification performance (green color) is kept in the model, a motif that does not improve the classification performance (red color) is removed.

2.3 Learning noisy threshold models

We split the data into training, validation and test sets. The training set was used to rank the motifs and learn a noisy threshold model, the validation set was used to choose the model parameters, and the test set was used to evaluate the classification performance of the model.

We used an iterative greedy approach to learn a noisy threshold model that separates the genes in cluster i from all other genes based on motif absence/presence. 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:

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

Variables M and C are binary, their true values denote the presence of at least one binding site for a motif in the upstream region of a gene and the belonging of the gene to the cluster for which a model is learned, respectively. Then, to learn a noisy threshold model, we started from a model containing no causes except a default so-called leaky cause (Henrion, 1989) and iteratively added the next highest ranked motif. If the new model did not have a higher classification accuracy on the validation set than the previous model, the motif was removed from the model. For each newly added motif, we evaluated two models, a model with the interaction function τ_k and a model with the interaction function τ_{k+1}, where τ_k is the interaction function from the model with the best classification accuracy in the previous iteration. To learn the probabilities of hidden variables in a noisy threshold model, we ran 10 iterations of the expectation-maximization (EM) algorithm described in Jurgelenaite and Heskes (2008), computed the classification accuracy on the validation set after each iteration and chose the number of iterations that provided the best classification accuracy.

To solve the problem of unbalanced data (different class size), 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 in both classes.

To test whether additional information about the regulatory sequence elements is useful for predicting gene expression, we performed two more sets of experiments. In the first set of experiments, we learned models where constraints on motif's orientation, location with respect to ATG and functional depth were added if they improved classification performance on the validation set. Likewise, in the second set of experiments, we learned models where information about additional copies of a motif was included into the model if it improved the classification performance on a validation set. For more information, see Supplementary Material.

2.4 Evaluation of the models learned

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, the threshold function values and the motifs that were selected as model features. We performed 100 runs of both error estimation methods. In the cross-validation scheme, one-hundredth of the genes was used as test data and one-fourth of the genes was used as validation data. In the bootstrap approach, we created a training set by sampling n genes with replacement from the original data, where n is the number of genes, and we formed a validation set from the genes that were excluded.

To compare two models learned under different assumptions as well as to compare our model to a classifier that assigns all genes to the bigger class, we used the altered form of the exact version of the McNemar's test (Salzberg, 1997). See Supplementary Material for a description of the significance test.

To identify motifs significant for the classification of the genes, we looked for motifs that were selected as causes in the model in a significant number of bootstrap runs. For more information, see Supplementary Material.

2.5 Identifying potential transcription factors

Few transcription factors in *P.falciparum* have been identified by sequence similarity searches, which are based on the hypothesis that sequence similarity across species reflects functional similarity. To identify potential transcription factors, we used an extended comparative genome analysis approach. The approach includes transcription factor binding motifs significant for the classification of the genes and includes three steps. First, we used STAMP (Mahony and Benos, 2007), a web tool for exploring DNA-binding motif similarities, to find a number of the closest matches for every significant motif. Using STAMP, we searched 10 databases of both eukaryotic and prokaryotic transcription factor binding motifs: TRANSFAC 10.4 (Matys *et al*., 2006), JASPAR 2.0 (Sandelin *et al*., 2004), FlyReg (Bergman *et al*., 2005), AthaMap (Steffens *et al*., 2004), PLACE (Higo *et al*., 1999), DPInteract (Robison *et al*., 1998), RegTransBase (Kazakov *et al*., 2007), two databases of yeast transcription factor binding motifs (Harbison *et al*., 2004; MacIsaac *et al*., 2006) and a database of human transcription factor binding motifs (Xie *et al*., 2005). Secondly, for each match found, we checked whether the database where the motif is stored reports a transcription factor that binds to it. Thirdly and finally, if the transcription factor is known, we used NCBI BLAST (Altschul *et al*., 1997) with default parameters to find the most similar protein sequences in the *P.falciparum* protein database.

3 RESULTS

3.1 Inferred significant motifs

Since the average length of upstream sequences that contribute to regulation is not known, we tested two sets of upstream sequences with different lengths: 1000 and 1500 bp. The models learned from 1000 bp upstream sequences showed slightly better classification performance; therefore, we report and analyze the results obtained using the set of 1000 bp upstream sequences.

We chose the number of clusters to be five, as the C-index (Hubert and Levin, 1976) curve had an 'elbow' at this value. The clusters are comparable with the four characteristic stages of intraerythrocytic parasite morphology discussed by Bozdech *et al*. (2003), as the vast majority of genes induced in every one of the stages belongs to one of four clusters. Cluster 5 is a cluster of genes whose expression is more or less constant throughout the intraerythrocytic cycle. The correspondence between the characteristic stages and the clusters is given in Table 1. From now on, we refer to the clusters by the names of the corresponding characteristic stages.

AlignACE found 100 motifs and about 160 000 sites for all of them. We used these motifs to learn the models for the first four clusters, i.e. the clusters of genes whose expression changed throughout the intraerythrocytic cycle. The classification accuracy of the noisy threshold models learned using the cross-validation procedure is reported in Table 1; we attain an accuracy of around 60% for each of the stages with little difference in accuracy between the stages. The predictability of the gene expression from upstream sequence elements is around the same as the predictability for *Saccharomyces cerevisiae* reported in Yuan *et al*. (2007). All 39 significant motifs (motifs selected as causes in a significant number of bootstrap runs) are shown in Figure 2.

The percentage of significant motifs that are present in upstream regions gradually increases in genes expressed in later stages throughout the intraerythrocytic cycle. Not surprisingly, the increase is strongest in the motifs that were significant for predicting the expression of genes in three or four clusters. It is interesting to note that the percentage of the motifs that predict gene expression positively correlates with the erythrocytic malaria parasites' total parasite protein content (Krugliak *et al*., 2002) as shown in Supplementary Figure 1 and mRNA half-lives (Shock *et al*., 2007).

Figure 3 summarizes the parameters of the noisy threshold models learned using the bootstrap procedure. The median of the threshold function values for the clusters of TES and S genes is approximately

Table 1. Summary of the clusters and classification accuracy (%) of the noisy threshold models learned to explain them

Cluster	Number of genes	Characteristic stage	Obtained accuracy	Baseline accuracy	p-value
1	144	Early ring (ER)	58.5	50.4	3×10^{-3}
2	1033	Ring/early Trophozoite (RET)	60.8	52.5	10^{-14}
3	985	Trophozoite/early Schizont (TES)	58.6	50.9	4×10^{-11}
4	329	Schizont (S)	60.5	50.8	2×10^{-6}
5	1344	–	–	–	–

The p-values are computed for the null hypothesis that the noisy threshold models are just as good as a baseline classifier which assigns all genes to the bigger class.

Fig. 2. Motifs significant for predicting the expression of genes in different clusters. The motifs are ordered from top to bottom in terms of how often they appear as a feature explaining membership to an expression cluster. Font size indicates relative importance of a motif: large, medium and small font sizes indicate a motif selected as a feature in at least 75 bootstrap runs, less than in 75 but at least in 50 bootstrap runs and less than in 50 but in a significant number of bootstrap runs, respectively.

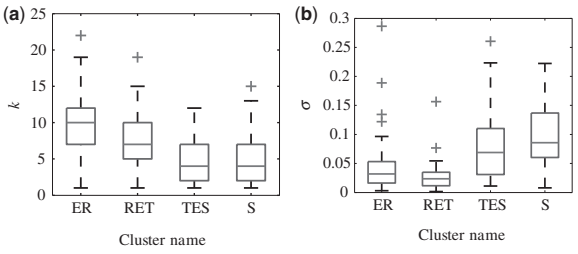

Fig. 3. Summary of the parameters of noisy threshold models learned in 100 bootstrap runs: box and whisker plots of the threshold functions τ_k (**a**), and of the standard deviation σ of the probabilities of hidden variables in a model (**b**). Median (horizontal line), quartiles (box), 5% and 95% confidence intervals (whiskers) and outliers are indicated.

four, meaning that presence of four or more functional regulatory elements predicts expression of genes in these clusters. These results match the findings of Van Noort and Huynen (2006), who reported that most *Plasmodium* genes have between three to seven different regulatory elements in their upstream regions. The median of the standard deviation (SD) of the probabilities of hidden variables provides evidence that noisy threshold models capture the probability that a sequence element is functional. There is little variation in the probabilities of the hidden variables in the models where cluster membership is explained by motif absence, which allows us to conclude that motifs have very similar explanatory power. However, there is much more variation in the probabilities of the hidden variables in the models where cluster membership is mostly explained by motif presence; this can be explained by different rates of functional binding sites for different motifs.

To verify that the models represent gene regulatory mechanisms and could not be learned from sequence elements found just anywhere in the genome, we ran two additional sets of experiments. We learned models that predict gene expression using over-represented motifs found either in coding regions or in 1000 bp downstream sequences. In both cases, the models for all four clusters were not significantly better than a baseline classifier which assigns all genes to the bigger class (data not shown).

3.2 Pattern of present/absent motifs

Figure 2 reveals a distinct pattern in that all motifs for the earlier ER and RET stages, with the exception of Motif 1, explain membership by their absence, while many of the motifs for the later TES and S stages explain membership by their presence. Interestingly, the motifs that break this pattern, with the exception of Motif 1, are found in a small number of genes (from 1% to 5% of the genes), in sharp contrast to the other significant motifs, which are more common. We put this observation to the test by learning noisy threshold models in which only either the presence or the absence of the motifs could be selected as the causes. As Table 2 shows, the models that follow the pattern were about as good as the original models, whereas models that break the pattern did not perform better than the baseline classifier.

3.3 Additional information about motifs

The classification accuracy of the models with constraints on motifs and p-values for the null hypothesis that the models perform equally well as the original models are shown in Table 3. Even though

Table 2. Classification accuracy (%) of models in which only either the presence or the absence of the motifs were selected as causes

Cluster	Models following the pattern	*p*-value	Models breaking the pattern	*p*-value
ER	59.5	10^{-3}	50.9	0.44
RET	61.3	10^{-15}	52.4	0.56
TES	58.5	5×10^{-11}	51.1	0.39
S	59.9	6×10^{-6}	49.3	0.85

The *p*-values are computed for the null hypothesis that the noisy threshold models are just as good as a baseline classifier which assigns all genes to the bigger class.

Table 3. Classification accuracy (%) of models with additional constraints and models with additional binding sites

Cluster	Original models	Models with constraints	*p*-value	Models with additional binding sites	*p*-value
ER	58.5	56.4	0.90	59.1	0.25
RET	60.8	61.4	0.22	61.6	0.12
TES	58.6	60.2	0.01	58.9	0.31
S	60.5	61.4	0.21	60.3	0.60

there is a slight tendency for a few motifs in the RET cluster to have a positional preference—e.g. in a number of models, Motif 6 is constrained to 250–500 bp upstream, and Motif 1 is constrained to 500–1000 bp—this seems to be circumstantial. These constraints are not preserved in the models for the other clusters and the *p*-values corrected for multiple testing (corrected for eight models learned in this section) are not significant. Yuan *et al.* (2007) reach the same conclusion in their paper, where the authors showed that the orientation and position information for predicted transcription factor binding sites in *S.cerevisiae* do not help in predicting coexpression of genes.

The right side of Table 3 lists the classification accuracy of the models with additional binding sites and *p*-values for the null hypothesis that the models perform equally well as the original models without information about the number of motif binding sites per gene. The results suggest that, even if some of the genes have multiple functional copies of motifs in their upstream regions, additional copies of motifs do not help to predict gene expression.

3.4 Correspondence to functionally tested motifs

In Figure 2, we presented a list of 39 motifs that are deemed significant by our approach. We obtained support for the hypothesis that many of these motifs are functional by using STAMP (Mahony and Benos, 2007) to compare functionally tested *Plasmodium* sequence motifs (Calderwood *et al.*, 2003; Chow and Wirth, 2003; De Silva *et al.*, 2008; Dechering *et al.*, 1999; Gunasekera *et al.*, 2007; Horrocks and Lanzer, 1999; Lanzer *et al.*, 1992; Militello *et al.*, 2004; Ruvalcaba-Salazar *et al.*, 2005) with the putative regulatory motifs we used for learning the models. The results of this comparison, presented in Figure 4, are encouraging as the vast majority of the best matches were the putative regulatory motifs found to be significant for predicting gene expression patterns.

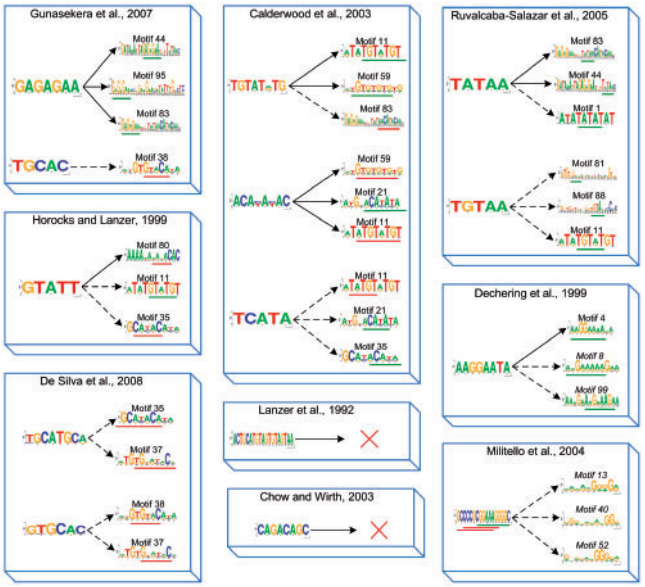

Fig. 4. We report up to three best matches for a functionally tested sequence motif if the *E*-value of the match is $<10^{-2}$. Dashed arrows indicate weak matches whose *E*-values are $>10^{-3}$ for functionally tested motifs that are up to 5 bp long and $>10^{-5}$ for functionally tested motifs that are >5 bp; solid arrows indicate the better matches. The locations of the alignments are underlined by green (forward alignment) and red (reverse alignment) lines in longer sequences of the matches. The names of the significant motifs are written in roman, the names of the motifs that we did not find to be significant are written in italic.

A few interesting observations that emerge from this comparison are discussed further.

The TATA box sequence (TATAA) which was shown to be bound by the TATA box binding protein (Ruvalcaba-Salazar *et al.*, 2005) is a part of two motifs, i.e. Motif 83 and Motif 44. It is interesting to note that in both motifs the TATAA sequence is preceded by the GAGAGAA sequence, which has been reported as a putative enhancer element in *P.falciparum* (Gunasekera *et al.*, 2007), and in both cases the distance between these two elements is 3 nt.

Even though the TCATA sequence was not found in any of the motifs, the three closest matches, Motifs 11, 21 and 35, had the same TCATA sequence-like sequence, the ACATA sequence.

We did not find close matches to the dual palindromic G-boxes GCCCCGCGGAAAGGGGC, which were shown to activate gene expression in *Plasmodium* species (Militello *et al.*, 2004). Different from the matches to the other functionally tested sequences, all three closest matches to the dual palindromic G-boxes, Motifs 13, 40 and 52, are motifs that were found to be not significant for predicting gene expression pattern.

3.5 Potential transcription factors

Comparative genome analysis (Supplementary Fig. 2) resulted in the disclosure of 10 potential transcription factors in *P.falciparum*. Our criterion for labelling a gene as a potential transcription factor was the gene being found as a match via more than one transcription factor binding motif from the other organisms. The rationale for this criterion was based on the fact that our approach ended up with <60 genes in *P.falciparum* whose *E*-value was ≤ 1

therefore, the probability of a gene being found by chance via different matches is very small. Nine genes were found repeatedly using the comparative genome analysis approach: *PFL0465c*, *PF14_0175*, *PF13_0198*, *PF13_0072*, *PF11_0294*, *MAL3P7.34*, *PFB0540w*, *PF10_0143* and *MAL13P1.176*. The analysis showed one more potential transcription factor, *PF14_0316*, putative DNA topoisomerase-II, which did not meet the criterion above, but was an identical match to the gene in the fruit fly; furthermore it was found via the closest matches for two very important motifs for predicting gene expression patterns, Motif 6 and Motif 11. All 10 potential transcription factors were found via matches in eukaryotic organisms. STAMP found a number of significant matches to the motifs in two prokaryotic transcription factor binding site databases that were searched; however, transcription factors binding to these motifs did not have homologues in *P.falciparum*. The alignments which produced the potential transcription factors and *E*-values of the motif and protein sequence matches are shown, respectively, in Supplementary Figure 2 and Supplementary Tables 1–10.

The natural question to ask is how many of these genes appeared as significant matches only because they are paralogues of a true transcription factor. To answer this question, we used NCBI BLAST (Altschul *et al*., 1997) to examine how similar the protein sequences of these potential transcription factors are. Three of the genes, *PF14_0316*, *PFL0465c* and *PF11_0294*, are not similar to any other gene from the list. Two of the genes, *PF13_0198* and *MAL13P1.176*, are paralogues of each other, and both of them are similar to the hypothetical protein *PF14_0175* with *E*-values of 5×10^{-6} and 2×10^{-8}, respectively. Even though the protein sequences of the other genes have some similarities between themselves and with the two paralogues, none of the sequences are closely related as there is only one match whose *E*-value is $<10^{-2}$: *MAL3P7.34* is similar to *PFB0540w* with an *E*-value of 6×10^{-5}.

4 DISCUSSION

Previous bioinformatics approaches have yielded some information on transcription regulatory elements and transcription factors in *P.falciparum*. Calebaut *et al*. (2005) predicted general transcription factors associated with RNA polymerase II. Several approaches searched for regulatory elements in the upstream sequences of a gene family (Militello *et al*., 2004) or clusters of co-expressed genes (Elemento *et al*., 2007; Van Noort and Huynen, 2006; Young *et al*., 2008). Also, comparative genomics was used to discover motifs in *Plasmodium* (Imamura *et al*., 2007; Wu *et al*., 2008).

Differently from other bioinformatics approaches applied to *P.falciparum*, our method is able to model the logic behind gene regulation and to incorporate uncertainty about the functionality of putative regulatory sequence elements. The classification accuracy of the noisy threshold models, which is an unbiased measure of the soundness of the models, allowed us to test a number of different models and, consequently, different hypotheses about gene regulation. Our main findings are as follows. First, we report a prioritized list of 39 regulatory motifs relevant for gene regulation and we show that prevalence of these motifs increases with progression through the developmental cycle. Second, we show that several factors (other than DNA sequence of the motif) are unlikely to contribute to gene regulation. Third, we provide a list of 10 potential transcription factors with their associated binding motifs.

Of the 39 significant motifs 11 are implicated in the regulation of genes expressed in the second phase of the asexual development in the blood cell (motifs that have a 'presence' label in column 3 or 4 in Fig. 2). The second phase consists of the final 18 h of the intraerythrocytic cycle, in which nucleic division is followed by merozoite formation and release (Florens *et al*., 2002). Given that hundreds of genes are involved in nucleic division, it is likely that these motifs contribute to the regulation of the expression of the genes involved in mitosis. Supporting evidence for this hypothesis could be our identification of *PF14_0316* as a transcription factor. *PF14_0316* is annotated as a putative topoisomerase-II in PlasmoDB 5.4 and it has been shown by Kelly *et al*. (2006) that treatment of asexual stage *P.falciparum* parasites with the topoisomerase-II inhibitor etoposide results in chromosomal cleavage.

There are two peculiar findings of our approach. First, cluster membership of genes expressed in the first two stages is predicted solely by the absence of motifs (with exception of Motif 1, see Fig. 2). Second, we found that the relative abundance of motifs upstream of cluster five genes, the cluster of genes that do not significantly change expression during the intraerythrocytic cycle, was halfway between the abundance of motifs upstream of the early and late stages genes. One possible explanation is that we found no evidence of regulation of the genes expressed in the initial phase of the intraerythrocytic cycle, in which principal modifications of the host cell occur that allow the parasite to transport molecules in and out of the cell, to prepare the surface of the red blood cell to mediate cytoadherence and to digest the cytoplasmic contents in its food vacuole. For this hypothesis, the number of motifs not being the same or less than in genes of the first stage could be explained by these motifs being relevant in different combinations in other life-cycle stages. A second possible explanation for these findings could be that the lack of regulatory motifs is a kind of gene regulation. Motif 4 seems to support this explanation as it is found in 42% of the genes of the ER cluster as compared with 63% of the genes in the other four clusters; furthermore, the percentage of genes that have Motif 4 almost does not vary throughout these four clusters. For this hypothesis, the number of motifs in cluster 5 is a baseline number of motifs found in the genes. We did not find any evidence for a third possible explanation (tested separately but results not significant)—that a fraction of the genes in cluster 5 should belong to one of the other four clusters.

Our model allowed us to test two sets of upstream sequences with different lengths. We found that models learned from the set of 1000 bp upstream sequences showed better classification performance than models learned from the set of 1500 bp upstream sequences. This result might look surprising, given that a *cis*-acting sequence element has been found as far away as 1600 bp upstream (Osta *et al*., 2002). However, genes that have regulatory elements far from the translation start site are probably an exception since 43% of the genes have another open reading frame starting <1500 bp upstream, and, additionally, 21% of the genes share a region where the distance between the translation start sites of the two genes is <3000 bps.

Our second main finding is the irrelevance of ancillary information for predicting gene expression. Our model allowed us to test whether extra information besides the DNA sequence of the motif, namely, score, orientation, location and multiplicity of the motif, makes a significant contribution to explaining cluster

membership. In our experiments, this additional information about the motifs did not improve the prediction of gene expression. This does not imply, however, that these motif properties are biologically irrelevant.

Our third main findings are 10 potential transcription factors and their associated DNA binding motifs. Although it is too optimistic to expect that all 10 are involved in transcriptional regulation, we believe that some of them can be true transcription factors, given that one of them, *PFL0465c*, is annotated as a zinc finger transcription factor in PlasmoDB and another, *PF10_0143*, is annotated as a putative transcriptional activator. There are three reasons why we think it is very unlikely that genes *PF13_0198* and *MAL13P1.176* are transcription factors: (i) they are transmembrane proteins localizing to the rhoptry organelles; (ii) both of them are similar to the hypothetical protein *PF14_0175*, which is similar to a number of transcription factors in other organisms; (iii) their expression pattern is different from that of the other potential transcription factors. The fact that 8 out of the 10 potential transcription factors show no significant change in their expression during the intraerythrocytic cycle suggests that transcription factors specific to the intraerythrocytic cycle are likely to be regulated post-transcriptionally.

The analysis to identify potential transcription factors also revealed that many of the significant motifs were similar to the same motifs in other organisms, and, thus, similar among themselves. This allows us to hypothesize that at least some of the motifs are bound by transcription factors that bind a family of similar but distinct motifs.

Funding: Netherlands Organization for Scientific Research (NWO) under project number FN4556; Vici grant (639.023.604 to T.H.).

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Bahl,A. *et al*. (2003) PlasmoDB: the Plasmodium genome resource. A database integrating experimental and computational data. *Nucleic Acids Res.*, **31**, 212–215.

Beer,M.A. and Tavazoie,S. (2004) Predicting gene expression from sequence. *Cell*, **117**, 185–198.

Bergman,C.M. *et al*. (2005) Drosophila DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, Drosophila melanogaster. *Bioinformatics*, **21**, 1747–1749.

Bozdech,Z. *et al*. (2003) The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum. *PLoS Biol.*, **1**, 85–100.

Bremen,J. (2001) The ears of the hippopotamus: manifestations, determinants, and estimates of the malaria burden. *Am. J. Trop. Med. Hyg.*, **64**, 1–11.

Bussemaker,H.J. *et al*. (2001) Regulatory element detection using correlation with expression. *Nat. Genet.*, **27**, 167–171.

Calderwood,M. *et al*. (2003) Plasmodium falciparum var genes are regulated by two regions with separate promoters, on upstream of the coding region and a second within the intron. *J. Biol. Chem.*, **278**, 34125–34132.

Callebaut,I. *et al* . (2005) Prediction of the general transcription factors associated with RNA polymerase II in Plasmodium falciparum: conserved features and differences relative to other eukaryotes. *BMC Genomics*, **6**, 100.

Chow,C. and Wirth,D. (2003) Linker scanning mutagenesis of the Plasmodium gallinaceum sexual stage specific gene pgs28 reveals a novel downstream cis-control element. *Mol. Biochem. Parasitol.*, **129**, 199–208.

Crooks,G.E. *et al*. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.

De Silva,E.K. *et al*. (2008) Specific DNA-binding by apicomplexan AP2 transcription factors. *Proc. Natl. Acad. Sci. USA*, **105**, 8393–8398.

Dechering,K. *et al*. (1999) Isolation and functional characterization of two distinct sexual-stage- specific promoters of the human malaria parasite Plasmodium falciparum. *Mol. Cell. Biol.*, **19**, 967–978.

Dimitriadou,E. *et al*. (2002) An examination of indexes for determining the number of clusters in binary data sets. *Psychometrika*, **67**, 137–160.

Elemento,O. *et al.* (2007) A universal framework for regulatory element discovery across all genomes and data types. *Mol. Cell*, **28**, 337–350.

Essien,K. *et al*. (2008) Computational analysis of constraints on noncoding regions, coding regions and gene expression in relation to Plasmodium phenotypic diversity. *PLoS ONE*, **3**, e3122.

Florens,L. *et al*. (2002) A proteomic view of the Plasmodium falciparum life cycle. *Nature*, **419**, 520–526.

Gardner,M.J. *et al*. (2002) Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature*, **419**, 498–511.

GuhaThakurta,D. and Stormo,G. (2001) Identifying target sites for cooperatively binding factors. *Bioinformatics*, **17**, 608–621.

Gunasekera,A. *et al*. (2007) Regulatory motifs uncovered among gene expression clusters in Plasmodium falciparum. *Mol. Biochem. Parasitol.*, **153**, 19–30.

Harbison,C. *et al*. (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.

Henrion,M. (1989) Some practical issues in constructing belief networks. In Kanal,L.N. *et al*. (eds), *Uncertainty in Artificial Intelligence 3*, Elsevier, p. 161–174.

Higo,K. *et al*. (1999) Plant cis-acting regulatory DNA elements (PLACE) database. *Nucleic Acids Res.*, **27**, 297–300.

Horrocks,P. and Lanzer,M. (1999) Mutational analysis identifies a five base pair cis-acting sequence essential for GBP130 promoter activity in Plasmodium falciparum. *Mol. Biochem. Parasitol.*, **99**, 77–87.

Hubert,L. and Levin,J. (1976) A general statistical framework for accessing categorical clustering in free recall. *Psychol. Bull.*, **83**, 1072–1082.

Hughes,J. *et al*. (2000) Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J. Mol. Biol.*, **296**, 1205–1214.

Imamura,H. *et al*. (2007) Sequences conserved by selection across mouse and human malaria species. *BMC Genomics*, **8**, 372.

Jurgelenaite,R. and Heskes,T. (2008) Learning symmetric causal independence models. *Mach. Learn.*, **71**, 133–153.

Kazakov,A.E. *et al*. (2007) RegTransBase – a database of regulatory sequences and interactions in a wide range of prokaryotic genomes. *Nucleic Acids Res.*, **35**, 407–412.

Keleş,S. *et al*. (2002) Identification of regulatory elements using a feature selection method. *Bioinformatics*, **18**, 1167–1175.

Kelly,J. *et al*. (2006) Evidence on the chromosomal location of centromeric DNA in Plasmodium falciparum from etoposide-mediated topoisomerase-II cleavage. *Proc. Natl. Acad. Sci. USA*, **103**, 6706–6711.

Krugliak,M. *et al*. (2002) Intraerythrocytic Plasmodium falciparum utilizes only a fraction of the amino acids derived from the digestion of host cell cytosol for the biosynthesis of its proteins. *Mol. Biochem. Parasitol.*, **119**, 249–256.

Lanzer,M. *et al*. (1992) A sequence element associated with the Plasmodium falciparum KAHRP gene is the site of developmentally regulated protein-DNA interactions. *Nucleic Acids Res.*, **20**, 3051–3056.

Llinas,M. *et al*. (2006) Comparative whole genome transcriptome analysis of three Plasmodium falciparum strains. *Nucleic Acids Res.*, **34**, 1166–1173.

MacIsaac,K. *et al*. (2006) An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC Bioinformatics*, **7**, 113.

Mahony,S. and Benos,P. (2007) STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res.*, **35**, 253–258.

Matys,V. *et al*. (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34**, 108–110.

Militello,K. *et al*. (2004) Identification of regulatory elements in the Plasmodium falciparum genome. *Mol. Biochem. Parasitol.*, **134**, 75–88.

Osta,M. *et al*. (2002) A 24 bp cis-acting element essential for the transcriptional activity of Plasmodium falciparum CDP-diacylglycerol synthase gene promoter. *Mol. Biochem. Parasitol.*, **121**, 87–98.

Pilpel,Y. *et al*. (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.*, **29**, 153–159.

Robison,K. *et al*. (1998) A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete Escherichia coli K-12 genome. *J. Mol. Biol.*, **284**, 241–254.

Ruvalcaba-Salazar,O.K. *et al*. (2005) Recombinant and native Plasmodium falciparum TATA-binding-protein binds to a specific TATA box element in promoter regions. *Mol. Biochem. Parasitol.*, **140**, 183–196.

Salzberg,S. (1997) On comparing classifiers: pitfalls to avoid and a recommended approach. *Data Min. Knowl. Discov.*, **1**, 317–327.

Sandelin,A. *et al*. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, **32**, 91–94.

Shock,J. *et al*. (2007) Whole-genome analysis of mRNA decay in Plasmodium falciparum reveals a global lengthening of mRNA half-life during the intra-erythrocytic development cycle. *Genome Biol.*, **8**, R134.

Steffens,N. *et al*. (2004) AthaMap: an online resource for in silico transcription factor binding sites in the Arabidopsis thaliana genome. *Nucleic Acids Res.*, **32**, 368–372.

Tetteh,K. and Polley,S. (2007) Progress and challenges towards the development of malaria vaccines. *BioDrugs*, **21**, 357–373.

Troyanskaya,O. *et al*. (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 520–525.

van Noort,V. and Huynen,M. (2006) Combinatorial gene regulation in Plasmodium falciparum. *Trends Genet.*, **22**, 73–78.

Wagner,A. (1999) Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. *Bioinformatics*, **15**, 776–784.

Werner,T. (1999) Models for prediction and recognition of eukaryotic promoters. *Mamm. Genome*, **10**, 168–175.

Wu,J. *et al*. (2008) Discovering regulatory motifs in the Plasmodium genome using comparative genomics. *Bioinformatics*, **24**, 1843–1849.

Xie,X. *et al*. (2005) Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, **434**, 338–345.

Young,J. *et al*. (2008) In silico discovery of transcription regulatory elements in Plasmodium falciparum. *BMC Genomics*, **9**, 70.

Yuan,Y. *et al*. (2007) Predicting gene expression from sequence: a reexamination. *PLoS Comp. Biol.*, **3**, e243.