PlasmoTFBM: An Intelligent Queriable Database for Predicted Transcription Factor Binding Motifs in *Plasmodium falciparum*

Chengyong Yang[1], Erliang Zeng[1], Kalai Mathee[2], Giri Narasimhan[1, 3]

[1] *Bioinformatics Research Group (BioRG), School of Computer Science, Florida International University, Miami, FL 33199.*

[2] *Department of Biological Sciences, Florida International University, Miami, FL 33199.*

[3] *Corresponding Author*

ABSTRACT: There is very little information available with regard to gene regulatory circuitries in *Plasmodium falciparum*. In an attempt to discover transcription factor binding motifs (TFBMs) in *P. falciparum*, we considered two approaches. In the first approach, gene expression data from asexual intraerythrocytic developmental cycle generated every hour for 48 hour post-infection were fed into the ISA (Iterative Signature Algorithm), which outputs modules composed of sets of genes associated with co-regulating conditions. Putative TFBMs were discovered by applying the AlignACE program on the resulting gene sets. In the second approach, the MotifRegressor program was used to predict potential motifs associated with induced and repressed genes for each time point and then clustered based on the strength of their correlation to the gene expression (i.e., motif coefficients) across different time points. A total of 637 and 840 putative motifs were predicted by the MotifRegressor and ISA-AlignACE programs, respectively. All this information was uploaded into a database, thus making it easy to devise complex queries. Using published information on known motifs, we were able to validate some of our results. In addition, modules consisting of putative transcription factors and related genes were also investigated. This work provides a bioinformatics methodology to analyze transcription regulation and TFBMs across the whole genome. By constructing a comprehensive relational database and an intelligent, user-friendly query system, biologically meaningful conclusions can be drawn easily even by an investigator with no prior knowledge of databases.

Keywords: Transcription factor, Regulatory elements, Motifs, G-box, *SPE* elements, *CPE* elements, *var* genes, Heat shock protein, SERA, EBA140, *Plasmodium falciparum*

1. INTRODUCTION

The challenge of CAMDA'04 was to analyze the gene expression data, which was generated by DeRisi's laboratory using transcripts from the organism *Plasmodium falciparum*, harvested at 46 different time points during its intraerythorcytic developmental life cycle [Bozdech, et al. 2003]. *P. falciparum* is one of four species of the parasitic protozoan genus Plasmodium, and is responsible for the vast majority of malaria episodes, affecting 200-300 million individuals and causing 0.7-2.7 million deaths per year worldwide [http://www.who.int/malaria].

In this paper, we focused on mining for information related to gene regulation and transcription factor binding motifs (TFBM), which is important considering the fact that direct experimental identification of TFBMs is slow and laborious. We used two recently developed algorithms to predict potential TFBMs: AlignACE [Hughes, et al. 2000; Roth, et al. 1998] and MotifRegressor [Conlon, et al. 2003; Liu, et al. 2002]. Using the limited information on known motifs, we were able to validate some of our results.

The AlignACE (**Align**s Nucleic **A**cid **C**onserved **E**lements) program is best applied on sets of co-regulated genes. Standard clustering tools such as hierarchical, K-means clustering, and self-organizing maps assign genes to unique clusters by relying on the similarity of the expression profiles of the co-regulated genes across all conditions for their identification [Han, et al. 2001]. However, many genes play multiple roles under various conditions in complex, interrelated biological processes. We, therefore, obtained clusters of potentially co-regulated genes by using the Iterative Signature Algorithm (ISA), which (a) allows for clustering of genes that exhibit similarity of the expression profiles only at specific sets of time points, and (b) allows for genes to be part of multiple clusters [Ihmels, et al. 2002]. This permits the ISA approach to explore complex interrelationships among genes. It outputs a set of transcription modules, each of which is a self-consistent unit consisting of potentially co-regulated genes and the regulating conditions [Ihmels, et al. 2004].

One of the difficulties with the motif discovery programs is that they produce a large number of predicted TFBMs along with associated scores representing the statistical significance of the predictions. However, drawing biologically useful inferences or conjectures remains a difficult problem. In this paper, we present a new approach that will facilitate the process of drawing meaningful conclusions that are likely to be useful to a biologist. This is achieved by constructing a comprehensive relational database for ***Plasmo**dium falciparum* with the predicted **T**ranscription **F**actor **B**inding **M**otifs called **PlasmoTFBM** (Figure 1), and an intelligent, user-friendly query system.

The PlasmoTFBM database contains the following information:

1. All the discovered TFBMs, along with their significance scores, the software using which they were found, and the genes whose upstream sequences contained them along with their location in those upstream sequences.
2. Clusters of co-regulated genes (referred to as transcription modules, or simply modules), and the time points at which they were found to be co-regulated.
3. All genes and ORFs in the genome, their chromosomal location, their functional annotation, and their expression information at all the time points during the development of the parasite.

We show, with examples, how an investigator can generate "conjectures" using this database, which could then be used to perform directed laboratory experimentation.

The only other related work on studying genome-wide TFBMs in *P. falciparum* is by Militello *et al.*, where they applied the AlignACE software to the upstream sequences of heat shock proteins [Militello, et al. 2004]. The current work provides a more comprehensive analysis by using gene expression data to support the results. While our extensive results are available at our website [http://biorg.cs.fiu.edu/CAMDA2004], because of space-limitations, in this paper we will confine our discussions to a few select examples.

In Section 2, we introduce some of the methods used in this paper. In Section 3, we briefly describe the experiments that were performed and present a small cross-section of the results. In Section 4, we conclude with some discussions.

2. METHODS

Transcription Modules: For this paper, we define a transcription *module* (or simply, *module*) as a set of co-regulated genes along with a set of conditions (time points) during which they appear to be co-regulated. We started with three collections of genes that were known to be (or conjectured to be) co-regulated (described in detail in the following paragraph). These collections were then refined using the ISA. The modules output by this algorithm satisfy a *self-consistency* property, which implies that the set of genes and the set of conditions show a strong correlation with each other.

Transcription modules were generated in several different ways, each time by applying the ISA algorithm [Bergmann, et al. 2003; Ihmels, et al. 2002]. A first set was generated by starting from a specific interesting gene. For this paper, 13 putative transcription factors were chosen. They are MAL13P1.213, MAL7P1.86, MAL8P1.131, PF07_0057, PF10_0143, PF13_0043, PF14_0469, PFA0525w, PFB0290c, PFB0730w, PFE0305w, PFE0415w, and PFI1260c. A second set of modules was generated by starting from collections of genes known to be involved in the same function (e.g., heat shock proteins); such sets were obtained from the PlasmoDB website (http://www.plasmodb.org) [Bahl, et al. 2003]. A third set was generated by starting from random initial sets. User-defined thresholds for the ISA method were chosen as follows: gene thresholds were selected from 1.0 to 2.5 with a step of 0.1, and condition threshold was fixed at 2 (it was held constant because its choice had a negligible effect on the output over a comparable range, as was also observed in [Ihmels, et al. 2004]). In total 217 transcription modules were obtained with gene sets ranging in size from 10 to 500. All the 217 modules can be found on the supplemental website at [http://biorg.cs.fiu.edu/CAMDA2004/].

AlignACE: AlignACE is a Gibbs sampling algorithm for detecting motifs that are over-represented in a set of DNA sequences [Hughes, et al. 2000; Roth, et al. 1998]. A C++ implementation was downloaded from their website [http://atlas.med.harvard.edu]. The upstream sequences of co-regulated genes obtained from the transcription modules (described above) were downloaded, and AlignACE was used to search for motifs in them. For our experiments with AlignACE, the GC content was set at 19.36%, the GC-content of the *P. falciparum* genome [Gardner, et al. 2002].

MotifRegressor: MotifRegressor is a second motif-detection tool used in this work. It first uses MDscan as a feature extraction tool to construct candidate motif matrices and then applies regression analysis to select motifs that are strongly correlated with changes in gene expression [Conlon, et al. 2003; Liu, et al. 2002]. For our experiments, the upstream sequences were cleaned up so that single repeats of at least 10 bases and double repeats of at least 16 bases were removed. As mentioned below, MotifRegressor was applied separately on gene expression data from the 46 time points. MotifRegressor has the advantage of using a more sophisticated background model (third-order Markov model), and selects for motifs that explain the data and correlate with the expression behavior of interest. It also provides significance scores for the discovered motifs.

For most part, we used the default settings for MotifRegressor. In this procedure, upstream sequences are first ordered by their relative gene expression values, then the top 50 sequences are chosen as a seed to obtain matrices for w-mer motifs (here we used $5 \leq w \leq 15$). Using a semi-Bayesian scoring function, the 50 highest-scoring motifs are obtained and then refined by using the 250 sequences with the highest relative gene expression values. Sequence *Motif-Matching Score* is generated in this step to determine how well the upstream sequence of a gene g matches a motif m. For motifs reported by MDscan, gene expression values were regressed on sequence motif-matching score using a stepwise linear regression procedure. The candidate motifs with a significant p value ($P \leq 0.01$) are retained.

Data: The gene expression data that passed all quality control filters (QC data) were downloaded form the CAMDA website. The gene expression data was available for every hour up to 46 hours post-infection (hpi). Standard R package routines (based on the K nearest neighbor method) were used to impute missing values [Troyanskaya, et al. 2001]. Regulatory Sequence Analysis Tools were used to extract upstream sequences for the ORFs [van Helden 2003]. For the analysis, the length of the upstream sequences used was 2000 bp.

Generating potential TFBMs: The QC data and the corresponding upstream sequences were analyzed. The ISA algorithm was applied on available collections of related genes. The resulting transcription modules were used as initial sets to run AlignACE resulting in one set of motifs. Then, the MotifRegressor software was ran on the gene expression data for each of the 46 time points separately, to obtain 46 sets of significant motifs. Motifs with identical consensus sequences were merged using Perl scripts (the cleaning step). There were 1077 motifs generated from MotifRegressor and 936 from AlignACE. After the cleaning step, 637 MotifRegressor and 840 AlignACE motifs remained.

Database: A relational database called PlasmoTFBM was designed and implemented using MySQL to store all the available information. This includes the gene expression data, generated significant motifs and modules, gene annotation information including the functional information and the chromosomal location. Figure 1 shows the scheme used for the analyses of the data.

Web Query and Visualization: Web query interface was implemented using PHP (**P**HP: **H**ypertext **P**reprocessor). Although it is possible to design complex queries for the PlasmoTFBM database using Perl DBI, it requires non-trivial expertise to be able to use it effectively. The motivation for the query system was as follows. Most biologists perform research on a small set of genes, usually a set of genes that are involved in a specific function or a specific pathway. Such a biologist would be interested in knowing whether this database has results that are relevant to their genes of interest, i.e., what other genes are co-regulated with the ones in questions, what motifs might they share, what developmental stage or functional pathway might they be involved in, what transcriptional factors may be regulating the genes of interest, and finally, what biologically meaningful conjectures can result from the analyses and that may be relevant to the genes of interest. Answers to such questions may be the starting point for further investigations for the biologist. A handy web-based query system could automate some of the analyses.

Consider the following example. Assume that the genes of interest are *MAL13P1.60* and *MAL7P1.86*. The *MAL13P1.60* encodes the protein erythrocyte-binding antigen 140 (EBA140), which is implicated in merozoite invasion using a sialic acid-dependent receptor on human erythrocytes [Baum, et al. 2003; Bozdech, et al. 2003; Thompson, et al. 2001]. The open reading frame, *MAL7P1.86*, codes for a putative alpha subunit of transcription initiation factor IIE (TF IIE). In addition, both genes are expressed highly during the merozoite stage of the parasite's development. A biologist may be interested in studying their relationship: Are they co-regulated (i.e., is there a transcription module that contains both of them, is there a set of time points or conditions under which their expression profiles are correlated)? Do they share any motifs in their upstream regions? Can any other relationships be conjectured?

In the first step, all transcription modules that include some subset of the genes of interest are computed, sorted by the number of genes of interest that they contain. Next, the user may choose a subset of the generated modules for further exploration. Suppose that the user decides to explore the module MAL7P1.86_g2_c6, which contains both the genes of interest. The query system also outputs the conditions (hpi 1, 27, and 41-45 in this example) and gene sets associated with the selected module. The user could then ask for the list of all the motifs found in the module MAL7P1.86_g2_c6, and under the selected conditions (say, hpi 27, 42 and 45 in this example).

Visualization tools are provided to visualize the final results in a more meaningful way. All motifs of interest are displayed using the WebLogo notation [Crooks, et al. 2004]. The user may select a specific set of genes, and the motifs for each gene of interest is then displayed in a graphical manner by showing their location as a function of their distance from the translation start site (ATG) in the upstream sequence of the gene. See Figure 2 for an example, where we looked at group of genes that code for serine-repeat antigens (SERA) [Rosenthal 2004]. The average gene expression profile along with the standard deviation is also displayed for the selected genes. All images are generated dynamically using PHP and the GD graphics library. Thus the web interface makes it possible to mine information and visualize some interesting results simply through a series of mouse clicks, and without the user having to learn a complicated database or a query language.

3. EXPERIMENTAL RESULTS

There are very few regulatory elements in *P. falciparum* that have been reported [Horrocks, et al. 1998]. We sought to validate our results using the known motifs. We discuss some of the interesting motif groups found.

G Box Motifs: Recently, a novel G-rich regulatory element named G-box was identified upstream of several *P. falciparum hsp* genes [Militello, et al. 2004]. Since the genome of *P. falciparum* is AT-rich (only 15% GC content), the G-box is considered a unique regulatory element. We investigated motifs in seven genes corresponding to heat-shock proteins (Hsp) or putative Hsps. The G-box was also found by our analyses in all these seven *hsp* genes (Figure 3). Furthermore, our analysis showed that the G-box motif was found to be significant at all 46 time points, and was not confined to just the *hsp* gene family, suggesting that the G-box is a common regulatory element, and is not stage-specific.

Next, we compared the motif sequences found by our analyses with the published sequence, (A/G)NGGGG(C/A) [Militello, et al. 2004]. However, the AlignACE method found several longer motifs containing the published sequence for G-box. The variants of these motifs found are shown in Figure 3.

Motifs in *var* genes: It is known that there are nearly 50 diverse *var* genes distributed throughout the parasite genome coding for variants of *P. falciparum* erythrocyte membrane protein 1 (PfEMP1); they are responsible for both antigenic variation and cytoadherence of infected erythrocytes in malaria [Voss, et al. 2000, 2003]. The ability of the parasite to switch the expression of PfEMP1 allows it to escape specific immune responses, and changes in its antigenic phenotype correlate with the altered properties of PfEMP1 [Voss, et al. 2000, 2003]. Thus understanding the regulatory mechanisms of PfEMP1 variants and other genes is very critical.

It was observed previously that most of the *var* genes were expressed in the early ring stage, but only one *var* gene variant is induced in the trophozoite stage, while the others are silent. We queried our database to find the motifs contained in the *var* genes. Our analysis showed the presence of two significant motifs (Figure 4): one was observed in a cluster of *var* genes at hpi 11 associated with inducing effect, while another motif at hpi 38 associated with repressing effect.

Previous studies of *var* genes have shown that nuclear proteins bind to conserved sequence motifs called *SPE1*

(CACGGACACATGCAGTAACCGAGAATTATTATATATAAATAT) and *SPE2* (T**GTGCATA**GTGGTGCG) and *CPE* (ATGT**TGTACAT**) [Voss, et al. 2003]. These were found by transfection experiments, and not by the use of sequence analysis or motif prediction software [Voss, et al. 2003].

We used the motif sequence information and queried our database. We found motifs in our database that were subsequences of the *SPE2* and *CPE* elements reported previously (Figure 4). The portions of SPE2 and CPE that overlapped with our motifs are underlined above. In addition, our analysis showed that similar motifs were significant in a group of *var* genes that were induced at the ring stage. In contrast, the extended *SPE2* element was found in a group of *var* genes that were repressed at the schizont stage. However, these motifs were not unique to the group of *var* genes, but were also present in other genes at the ring and schizont stages. The analysis of the SPE1 sequence did not generate any potentially useful interpretations.

Discovery of Multiple Motifs: The MotifRegressor program predicted a total of 637 significant motifs across the 46 time points. The motifs were then clustered by motif coefficients, as suggested by Conlon *et al*. [Conlon, et al. 2003]. In brief, for each motif, at each time point, the gene expression values were regressed against the upstream sequence motif-matching scores (reported by the MDscan component of MotifRegressor). Consequently, each motif can be represented by a vector of 46 simple regression coefficients. The 637 motifs were then hierarchically clustered into 12 groups based on the Euclidean distances between their coefficient vectors. The motif coefficients can be interpreted as the putative influence of a particular motif on the expression of downstream genes. Figure 5 shows the clusters of motifs with the plot on the left showing the motif coefficients across all time points. The plot on the right side shows the time points when the corresponding motifs were discovered as being significant. As can be seen in the figure, a majority of the motifs showed a periodic behavior, indicating that they are regulated periodically during the *P. falciparum* IDC. The above analysis showed that many motifs were found at the time points at which they were known to have the strongest effect (See supplemental material "Time point distribution of motif clusters" at the website [http://biorg.cs.fiu.edu/CAMDA2004/]).

Motifs of EBA140: Next we analyzed the motifs in the gene for erythrocyte-binding antigen 140 (*eba140* or *MAL13PI.60*) that lies in *P. falicparum* chromosome 13 [Gardener et al, 2002]. As described before, this is a particularly interesting gene, since the corresponding protein shares structural features and homology with EBA175 which, in turn, is implicated in merozoite invasion using a sialic acid-dependent receptor on human erythrocytes [Baum, et al. 2003; Bozdech, et al. 2003; Thompson, et al. 2001]. Eight significant motifs were identified in the upstream region of *eba140*. The adjacent gene on chromosome 13 is *MAL13PI.61* encoding a hypothetical protein that is divergently transcribed, and therefore share the upstream promoter region with *eba140*. Analysis suggests that both these genes are tightly co-regulated, and it is not clear which of the genes (or both) is regulated by the putative motifs reported in their common upstream regions.

Querying the database helped us to locate a module that contained *eba140* and a putative transcription factor, MAL7P1.86, which has a peak expression at hpi 42 (early merozoite stage). AlignACE, when applied to this module had discovered a motif shared by the upstream sequences of the genes *eba140* and *MAL7P1.86*. At the spanned time period, this *MAL7P1.86* and the *eba140* genes were co-expressed; they also shared common motifs, which were at upstream locations -752 and -1330 in *eba140* (Figure 6). These two elements have very similar core sequence ("ACACA"). These two motifs were also shared by 77 other genes that are highly expressed at 41 hpi. One possible conjecture is that these genes are regulated by MAL7P1.86 by interacting with these two TFBMs. This would then suggest that MAL7P1.86 is auto-regulated. Alternatively, one could also conjecture that these genes are activated by an unknown transcription factor that interacts at these motifs.

It is worth pointing out that the above analysis on *eba140* and *MAL7P1.86* was easily performed as a sequence of straightforward queries of our database. Our belief is that with the help of domain-specific experts we can easily generate more biologically meaningful conjectures using a database such as PlasmoTFBM.

4. DISCUSSION AND CONCLUSIONS

Using the ISA approach, transcription modules were generated. Each module consists of a set of potentially co-regulated genes along with a set of time points at which the regulation is potentially occurring. Correlation and dependencies between the conditions can be used to elucidate system-level transcriptional relationships. Compared to other existing clustering approaches [Eisen, et al. 1998; Tamayo, et al. 1999], the ISA algorithm does not require the genes in a cluster to be correlated under all the conditions. It also allows genes to be part of multiple modules, which is a likely event since many genes are involved in different pathways at different time points.

We applied two existing motif detection tools on the CAMDA data sets. Both methods found a large number of potential transcription factor binding motifs. Our results on the G-box motifs support the conclusion that this organism may have unique regulatory mechanisms different from other known eukaryotic organisms [Militello, et al. 2004]. By design, the two approaches will find sequence motifs that are enriched in the input sequences (AlignACE) or best match the expression pattern (MotifRegressor). However, false positives are inevitable. AlignACE, in particular, is prone to give high scores to over-represented sequences in low-complexity regions, even though more stringent clusters from the ISA approach were used. Mechanisms to remove spurious results are extremely critical, but difficult and are themselves error-prone. In this current work, we rely on the significance scores provided by AlignACE (MAP Scores) and MotifRegressor (Sequence Motif-Matching Score) to provide the necessary guidance to decrease the number of false positives. More sophisticated mechanisms to improve the quality of the results are planned for the future.

We have implemented a novel database called PlasmoTFBM containing information relating to *P. falciparum* regulatory elements in IDC, which can be a useful tool to facilitate further biological research on the organism. Some sample questions that can be answered with relative ease with the use of our database include: (a) Find the set of genes X on chromosome A between loci L_1 and L_2. (b) Find motifs that are significant for set X during the schizont stage. (c) Locate a transcription factor Y co-regulated with X during the early merozoite stage or late schizont stage. (d) Does transcription factor Y share any motifs that are significant during hpi 18-21? Thus, it is possible to “bootstrap” any information available from the biological experiments to generate new and useful (and plausible) conjectures that can then drive future directed laboratory experiments.

Considering that very few regulatory elements were previously known for *P. falciparum*, the PlasmoTFBM database provides a useful pool of potential targets for investigators. It is well known that genes can be regulated both at the transcriptional and the translational stages. Recent research has suggested that the post-transcriptional gene regulation may be a predominant mechanism used by *P. falciparum* [Coulson, et al. 2004; Hall, et al. 2005]. However, this does not diminish the importance of transcriptional regulation. Thus our database could still play an important role in revealing the putative regulatory elements involved in the transcriptional stage.

We provide a website [http://biorg.cs.fiu.edu/CAMDA2004], which will contain all the motifs and modules discovered by our analyses. We also provide a website [http://biorg.cs.fiu.edu/TFBM/] for web query and data visualization.

ACKNOWLEDGEMENTS

Authors CY and EZ contributed equally to this paper. We thank Dr. X. Shirley Liu for help with MotifRegressor. We thank Gaolin Zheng and Haifeng Wang for helpful discussions. EZ is supported by a Florida International University Presidential Graduate Fellowship. GN was supported in part by NIH Grant P01 DA15027-01.

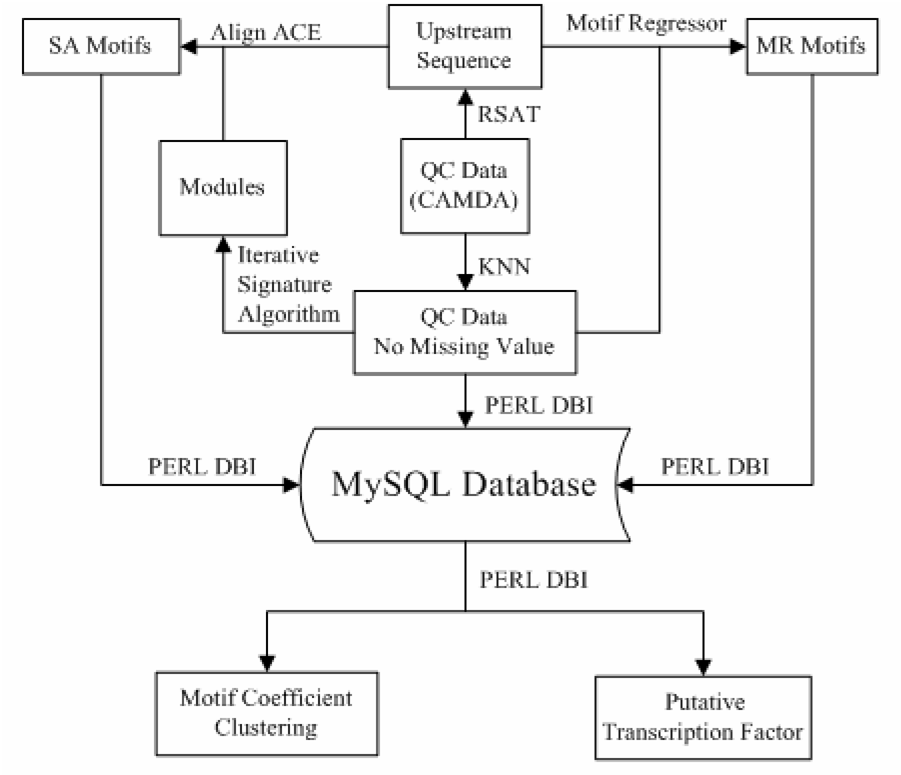


Figure 1: Flowchart for mining TFBMs for *P. falciparum*.

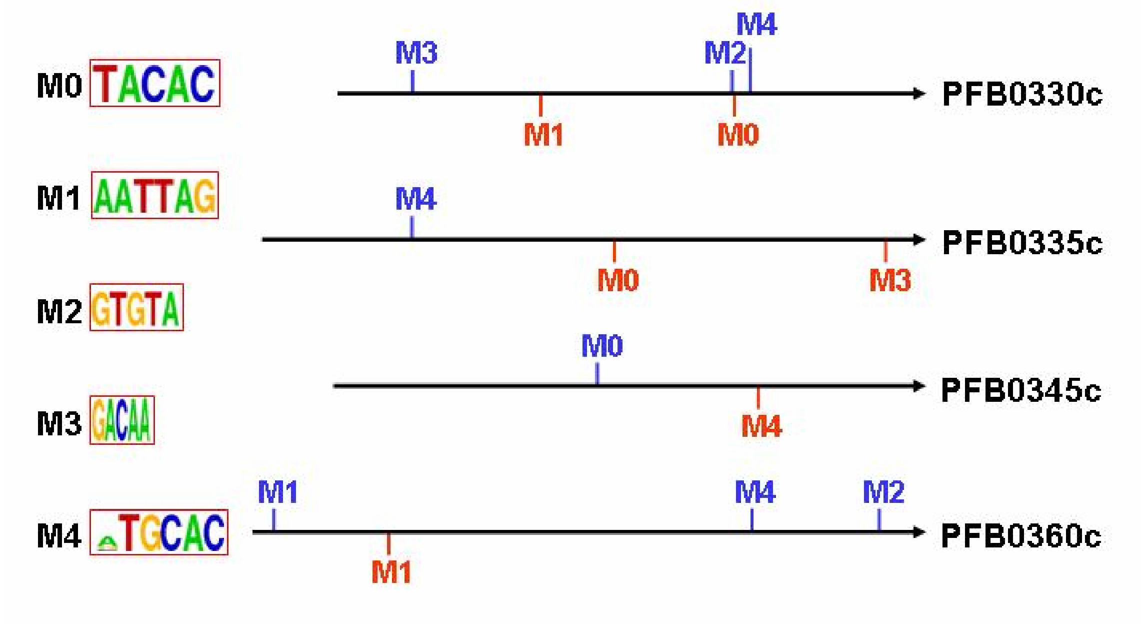


Figure 2: Motifs visualized in the upstream sequence of all the SERA genes of interest. The line indicates the upstream sequence with the translation start site at the right end. Motifs labeled as M0, M1, M2, M3, and M4 correspond to motifs Motif.P29.5.3BG, Motif.N29.6.15BG, Motif.P31.5.22BG, Motif.P33.5.10BG, and Motif.P35.6.3BG, respectively. Red color represents motifs in the forward direction while blue color represents those in the reverse direction.

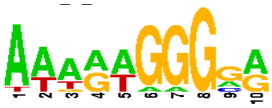
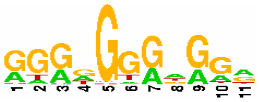
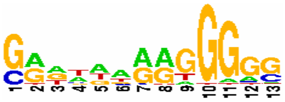
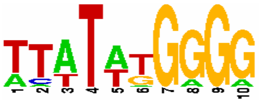
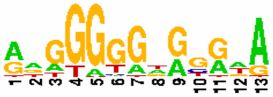
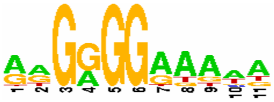
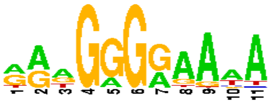
Locus	hpi	WebLogo	Motif Score
PFI0875w (HSP)	26-34, 39-45		15.58
MAL8P1.143 (hypothetical)	1-48		113.62
PF08_0032 (hypothetical)	1-3, 6, 27-37, 41-48		38.32
PF11_0175 (HSP 101)	11-18, 26-33		15.28
PF11_0188 (HSP 90)	1-48		50.08
PF11_0351 (HSP 70)	1-48		222.77
PFL0740c (hypothetical)	1-48		132.64

Figure 3. G-box motifs appeared in the upstream sequences of the *hsp* genes given in column 1. The motifs shown using the WebLogo format [Crooks, et al. 2004] were obtained by using AlignACE on modules that included the hpi mentioned in the second column. The AlignACE method provided the motif scores mentioned in the last column [Hughes, et al. 2000].

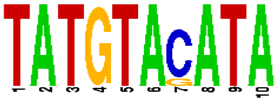
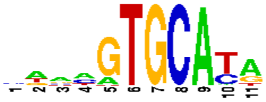
Locus	Stage	Motif effect	WebLogo	Motif Score
PFL0935c PF14_048 PFI1830c PF10_0406 PFL1955w PFA0765c PFD0615c PFB0010w PF08_0103	Ring	Induce		3720.20
PFD0230c PF08_010 PFL0935c PF10_040 PFB0010w PFI1830c PFA0765c	Schizont	Repress		4249.90

Figure 4. Some significant motifs from the *var* genes. The first one contains part of the CPE motif, while the second one contains a part of the SPE2 motif [Voss, et al. 2003]. WebLogo was used to display the motif. The motif scores are the result of using MotifRegressor program [Conlon, et al. 2003].

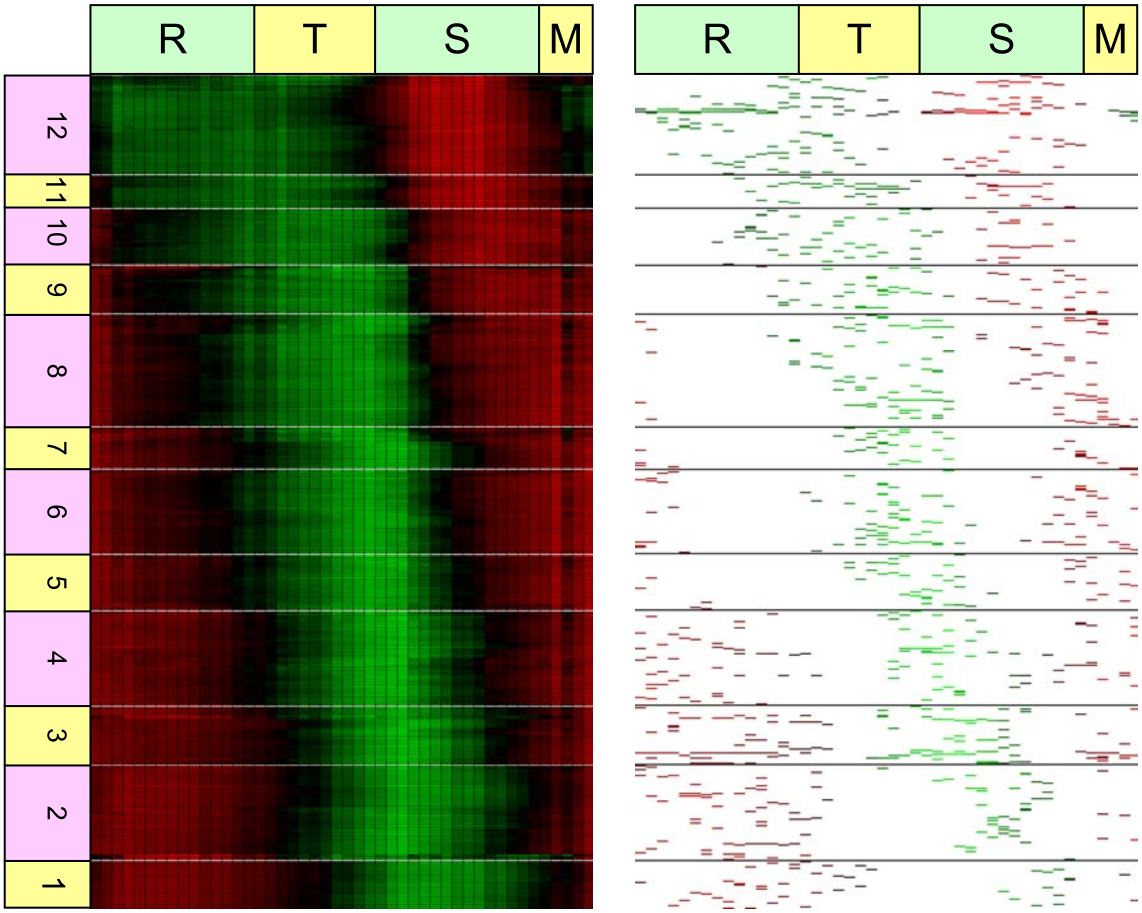


Figure 5. Motif clusters from cell cycle expression time series experiments. The 637 significant motifs reported by MotifRegressor over one cell cycle are clustered by motif coefficients over 46 time points. This figure was produced using Genesis software package by applying hierarchical clustering with Euclidean distance metric on the motif coefficient data [Sturn, et al. 2002]. Red shades correspond to positive motif coefficients (and, therefore positive correlations with the expression of the downstream genes), while green shades correspond to negative coefficients. The figures indicate the stages of the parasite (R-Ring, T-Trophozoite, S-Schizont, M-Merozoite) and the 12 clusters of motifs obtained.

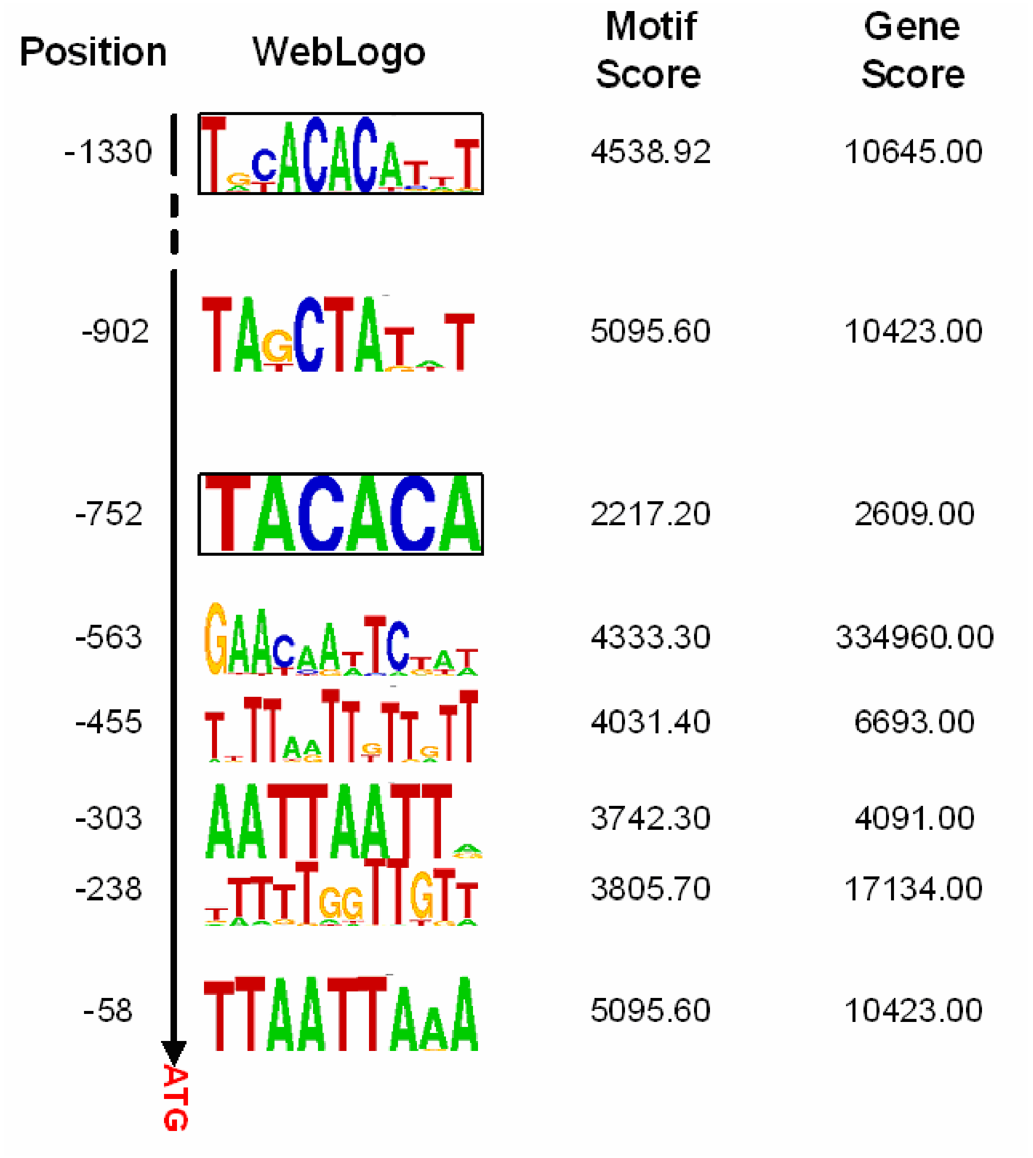


Figure 6. Motifs found in upstream of gene *eba140*. Boxed motifs are motifs shared by genes *eba140* and the divergently transcribed *MAL7P1.86* encoding a putative transcription factor (as well as other 77 other genes). Motif scores were as reported by the MotifRegressor program. The gene score shown on the last column indicates how well the upstream sequence of a gene matches a motif in terms of both degree of matching and number of sites [Conlon, et al. 2003].

5. REFERENCES

Bahl A., Brunk B., Crabtree J., Fraunholz M.J., Gajria B., Grant G.R., Ginsburg H., Gupta D., Kissinger J.C., Labo P., Li L., Mailman M.D., Milgram A.J., Pearson D.S., Roos D.S., Schug J., Stoeckert C.J. Jr, and Whetzel P. (2003). "PlasmoDB: the Plasmodium genome resource. A database integrating experimental and computational data." *Nucleic Acids Res*. **31**(1):212-5.

Baum, J., Thomas, A.W. and Conway, D.J. (2003). "Evidence for diversifying selection on erythrocyte-binding antigens of *Plasmodium falciparum* and *P. vivax*." *Genetics*. **163**(4): 1327-36.

Bergmann, S., Ihmels, J. and Barkai, N. (2003). "Iterative signature algorithm for the analysis of large-scale gene expression data." *Phys Rev E Stat Nonlin Soft Matter Phys*. **67**(3): 031902-1-18.

Bozdech, Z., Llinas, M., Pulliam, B.L., Wong, E.D., Zhu, J.C. and DeRisi, J.L. (2003). "The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*." *PLoS Biol*. **1**(1): 85-100.

Conlon, E.M., Liu, X.S., Lieb, J.D. and Liu, J.S. (2003). "Integrating regulatory motif discovery and genome-wide expression analysis." *Proc Natl Acad Sci U S A*. **100**(6): 3339-44.

Coulson, R.M., Hall, N. and Ouzounis, C.A. (2004). "Comparative genomics of transcriptional control in the human malaria parasite *Plasmodium falciparum*." *Genome Res*. **14**(8): 1548-54.

Crooks, G.E., Hon, G., Chandonia, J.M. and Brenner, S.E. (2004). "WebLogo: A sequence logo generator." *Genome Res*. **14**(6): 1188-90.

Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998). "Cluster analysis and display of genome-wide expression patterns." *Proc Natl Acad Sci U S A*. **95**(25): 14863-8.

Gardner, M.J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R.W., Carlton, J.M., Pain, A., Nelson, K.E., Bowman, S., Paulsen, I.T., James, K., Eisen, J.A., Rutherford, K., Salzberg, S.L., Craig, A., Kyes, S., Chan, M.S., Nene, V., Shallom, S.J., Suh, B., Peterson, J., Angiuoli, S., Pertea, M., Allen, J., Selengut, J., Haft, D., Mather, M.W., Vaidya, A.B., Martin, D.M., Fairlamb, A.H., Fraunholz, M.J., Roos, D.S., Ralph, S.A., McFadden, G.I., Cummings, L.M., Subramanian, G.M., Mungall, C., Venter, J.C., Carucci, D.J., Hoffman, S.L., Newbold, C., Davis, R.W., Fraser, C.M., and Barrell, B. (2002). "The genome sequence of the human malaria parasite *Plasmodium falciparum*." *Nature*. **419**(6906): 498-511.

Hall, N., Karras, M., Raine, J.D., Carlton, J.M., Kooij, T.W., Berriman, M., Florens, L., Janssen, C.S., Pain, A., Christophides, G.K., James, K., Rutherford, K., Harris, B., Harris, D., Churcher, C., Quail, M.A., Ormond, D., Doggett, J., Trueman, H.E., Mendoza, J., Bidwell, S.L., Rajandream, M.A., Carucci, D.J., Yates, J.R., 3rd, Kafatos, F.C., Janse, C.J., Barrell, B., Turner, C.M., Waters, A.P. and Sinden, R.E. (2005). "A comprehensive survey of the Plasmodium life cycle by genomic, transcriptomic, and proteomic analyses." *Science*. **307**(5706): 82-6.

Han, J. and Kamber, M. (2001). "Data Mining: Concepts and Techniques." *Morgan Kaufmann Publishers*.

Horrocks, P., Dechering, K. and Lanzer, M. (1998). "Control of gene expression in *Plasmodium falciparum*." *Mol Biochem Parasitol*. **95**(2): 171-81.

Hughes, J.D., Estep, P.W., Tavazoie, S. and Church, G.M. (2000). "Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*." *J Mol Biol*. **296**(5): 1205-1214.

Ihmels, J., Friedlander, G., Bergmann, S., Sarig, O., Ziv, Y. and Barkai, N. (2002). "Revealing modular organization in the yeast transcriptional network." *Nat Genet*. **31**(4): 370-7.

Ihmels, J., Bergmann, S., and Barkai, N. (2004). "Defining transcription modules using large-scale gene expression data." *Bioinformatics*. **20**(13): 1993-2003.

Liu, X.S., Brutlag, D.L. and Liu, J.S. (2002). "An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments." *Nat Biotechnol*. **20**(8): 835-9.

Militello, K.T., Dodge, M., Bethke, L. and Wirth, D.F. (2004). "Identification of regulatory elements in the *Plasmodium falciparum* genome." *Mol Biochem Parasitol*. **134**(1): 75-88.

Rosenthal, P.J. (2004). "Cysteine proteases of malaria parasites." *Int J Parasitol*. **34**(13-14): 1489-99.

Roth, F.P., Hughes, J.D., Estep, P.W. and Church, G.M. (1998). "Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation." *Nat Biotechnol*. **16**(10):939-45.

Sturn, A., Quackenbush, J. and Trajanoski, Z. (2002). "Genesis: cluster analysis of microarray data." *Bioinformatics*. **18**(1): 207-8.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S. and Golub, T.R. (1999). "Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation." *Proc Natl Acad Sci U S A*. **96**(6): 2907-12.

Thompson, J.K., Triglia, T., Reed, M.B. and Cowman, A.F. (2001). "A novel ligand from *Plasmodium falciparum* that binds to a sialic acid-containing receptor on the surface of human erythrocytes." *Mol Microbiol*. **41**(1): 47-58.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. and Altman, R.B. (2001). "Missing value estimation methods for DNA microarrays." *Bioinformatics*. **17**(6): 520-5.

van Helden, J. (2003). "Regulatory sequence analysis tools." *Nucleic Acids Res*. **31**(13): 3593-96.

Voss, T.S., Kaestli, M., Vogel, D., Bopp, S. and Beck, H.P. (2003). "Identification of nuclear proteins that interact differentially with *Plasmodium falciparum var* gene promoters." *Mol Microbiol*. **48**(6): 1593-607.

Voss, T.S., Thompson, J.K., Waterkeyn, J., Felger, I., Weiss, N., Cowman, A.F. and Beck, H.P. (2000). "Genomic distribution and functional characterisation of two distinct and conserved *Plasmodium falciparum var* gene 5' flanking sequences." *Mol Biochem Parasitol*. **107**(1): 103-15.