Mining Regulatory Elements in the *Plasmodium falciparum* Genome Using Gene Expression Data

Chengyong Yang &
Erliang Zeng
School of Computer Science
Florida International University
Miami FL 33199, USA
(+1) 305.348.1749
cyang01,ezeng001@cs.fiu.edu

Kalai Mathee
Department of Biological Sciences
Florida International University
Miami FL 33199, USA
(+1) 305.348.1261
matheek@fiu.edu

Giri Narasimhan
School of Computer Science
Florida International University
Miami FL 33199, USA
(+1) 305.348.3748
giri@cs.fiu.edu

ABSTRACT

There is very little information available with regard to gene regulatory relationships 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 of all the conditions were fed into the Iterative Signature Algorithm (ISA), which outputs modules composed of sets of genes associated with co-regulating conditions. Potential TFBMs were discovered by applying AlignACE on the resulting gene sets. In the second approach, MotifRegressor was used to generate 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. Currently, a total of 637 and 840 motifs have been discovered 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.

1. INTRODUCTION

The challenge of CAMDA'04 is to analyze the gene expression data generated by DeRisi's laboratory using transcripts from the organism *Plasmodium falciparum*, harvested at 46 different time points during its intraerythorcytic developmental life cycle [2]. *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.

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 and MotifRegressor. Using the limited information on known motifs, we were able to validate some of our results. We further analyzed the data using the *Iterative Signature Algorithm* (ISA) to obtain clusters of genes that are potentially co-regulated at specific sets of time points [8]. In particular, we started with collections of genes that were related either by their functional annotation or by the presence of a significant motif, and applied the "refinement" procedure of ISA.

Finally, we input all the information into a relational database (implemented using MySQL) with the goal of facilitating complicated querying of the available data (Figure 1). We show, with examples, how a biologist can generate simple "conjectures" using this database, which could then be used to perform directed experimentation in a laboratory.

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 [10]. However, they did not use gene expression data to validate their results. The current work is much more comprehensive and provides a launch-pad for more biologically meaningful analysis. While our results are extensive and available at our website, because of space-limitations, we will confine our discussions to a few choice examples.

In Section 2, we introduce some of 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, *modules*) 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 collections of genes that were known to be (or conjectured to be) co-regulated. 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 [1, 8]. A first 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]. A second set was generated by starting from genes that shared the same motif (as detected by the

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Conference'04, Month 1–2, 2004, City, State, Country.

Copyright 2004 ACM 1-58113-000-0/00/0004…$5.00.

methods described below). A third set was generated by starting from random initial sets. In total, 217 modules were obtained, with gene sets ranging in size from 10 to 500.

AlignACE: AlignACE (**Align**s Nucleic **A**cid **C**onserved **E**lements) is a Gibbs sampling algorithm for detecting motifs that are over-represented in a set of DNA sequences [7, 11]. 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.

MotifRegressor: MotifRegressor is another motif-detection tool. It 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 [3, 9].

Data: Quality expression data (QC data) was downloaded form the CAMDA website. The gene expression data was available for 46 hours post-infection (hpi). Standard R package routines (based on the k nearest neighbor method) were used to impute missing values [15]. Regulatory Sequence Analysis Tools were used to extract upstream sequences for the ORFs [16].

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 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 that were identical, similar, or overlapping 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 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.

Database Queries: Queries to the database were written in perl DBI. The implemented database facilitated the design of non-trivial queries and made it possible to mine for information related to gene regulation.

Figure 1 shows the scheme used for the analyses of the data.

3. EXPERIMENTAL RESULTS

There are very few regulatory elements in *P. falciparum* that have been reported [6]. We sought to validate our results using the known motifs. We discuss some of the interesting motif groups found.

G Box Motif Group: Recently, a novel G-rich regulatory element named G-box was identified upstream of several *P. falciparum hsp* genes [10]. Since the genome of *P. falciparum* is AT-rich (only 15% GC content), the G-box is considered a unique regulatory element. We investigate 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 2). Furthermore, the G-box motif was found to be significant in all 46 time points, and was not confined to just the *hsp* gene, suggesting that the G-box is a common regulatory element, and is not stage-specific.

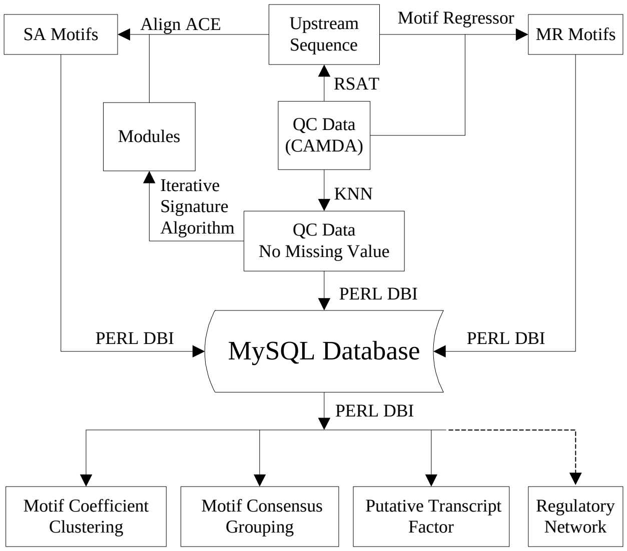


Figure 1: Flowchart for mining TFBMs for *P. falciparum*.

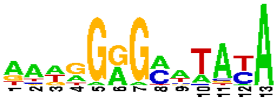
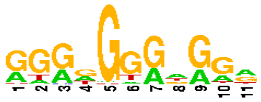
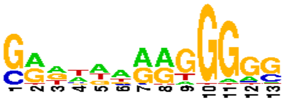
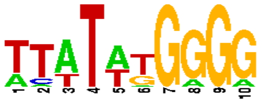
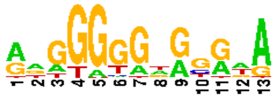
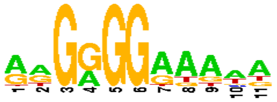
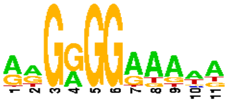
Locus	hpi	Logo[4]
PFI0875w (HSP)	1-3, 6, 22, 27-37, 41-48	
MAL8P1.143 (hypothetical)	1-48	
PF08_0032 (hypothetical)	1-3, 6, 27-37, 41-48	
PF11_0175 (HSP 101)	11-18, 26-33	
PF11_0188 (HSP 90)	1-48	
PF11_0351 (HSP 70)	1-48	
PFL0740c (hypothetical)	1-48	

Figure 2. G-box motifs appeared in the upstream sequences of the Hsp genes given in column 1. The motifs shown using the Logo format were obtained by using AlignACE on modules that included the hpi mentioned in the second column.

Next, we compared the motif sequences found by our analyses with the published sequence, (A/G)NGGGG(C/A) [10]. However, the AlignACE method found several longer motifs containing the published sequence for G-box. The variations of the motifs found are shown in Figure 2. One of the motifs contained a string of Ts preceding the G box; we will refer to this as the TG-box. The analysis of the TG-box shows that it has significant expression only during two time periods - high expression at hpi 11 and a low expression at hpi 33. An inspection of the genes that shared this module (module m7_7_2000Motifs consisting of 10 genes) showed that they exhibited similar expression patterns, and sparse in 7 chromosomes (chromosomes 4, 7-8, 11-14) with diverse functions. The TG-box motif is found during late ring stage with all the genes being strongly induced, and during the schizont stage, where all the genes were strongly repressed; it was not found to be significant in the other two intraerythrocytic stages of the parasite. It may be noted that MotifRegressor was unable to find the TG-box motif.

Motifs in *var* genes: It is known that there are nearly 50 diverse *var* genes distributed throughout the parasite genome coding for variants of PfEMP1, *P. falciparum* erythrocyte membrane protein 1; they are responsible for both antigenic variation and cytoadherence of infected erythrocytes in malaria [17, 18]. The ability of the parasite to switch the expression of PfEMP1 allows it to escape specific immune responses directed against it, and changes in antigenic phenotype correlate with altered properties of PfEMP1 [17, 18]. 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 (Fig. 3): 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* (CACGGACACATGCAGTAACCGAGAATTATTATATATAA ATAT), *SPE2* (TGTGCATAGTGGTGCG) and *CPE* (ATGTTGTACAT) [18]. These were found by transfection experiments, and not by the use of sequence analysis or motif prediction software [18].

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 3). 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.

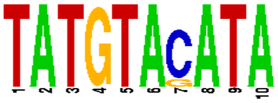
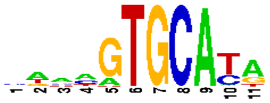
Locus	Stage	Motif effect	Logo
PFL0935c PF14_048 PFI1830c PF10_0406 PFL1955w PFA0765c PFD0615c PFB0010w PF08_0103	Ring	Induce	
PFD0230c PF08_010 PFL0935c PF10_040 PFB0010w PFI1830c PFA0765c	Schizont	Repress	

Figure 3. Some significant motifs from the *var* genes.

Discovery of Multiple Motifs: A total of 637 significant motifs were predicted by the MOTIFREGRESSOR software across the 46 time points. These were then clustered by motif coefficients, as suggested by Conlon *et al*. [3]. 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 within the *P. falciparum* IDC cycle, indicating that they are regulated periodically during the IDC cycle. We also note that many motifs were found at the time points at which they were known to have the strongest effect.

Motifs of EBA140: We also analyzed the gene for erythrocyte-binding antigen 140 (EBA140). This is a particularly interesting gene, since the corresponding protein shares structural features and homology with EBA175 that is implicated in merozoite invasion using a sialic acid-dependent receptor on human erythrocytes [2, 14]. Eight significant motifs were identified in the upstream region of *eba140.* These eight motifs were also found in one other gene MAL13P1.61 encoding a hypothetical protein. Incidentally, both genes are adjacent on chromosome 13. This analysis suggests that both these genes are tightly co-regulated.

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, the putative transcription factor and the EBA140 gene were co-expressed; they also shared common motifs which were at upstream locations -752 and -1330 in EBA140 (Fig 4). These two elements have very similar core sequence ("ACACA"). These two motifs were also shared by 77 others 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.

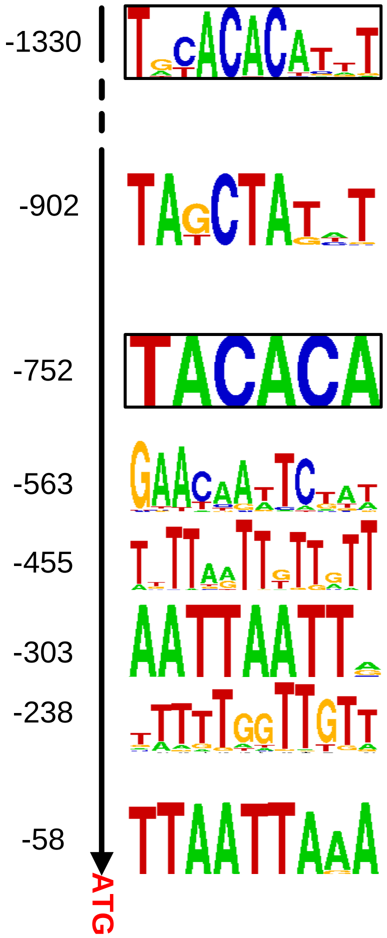


Figure 4. Motifs found in upstream of gene EBA140.

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 a domain-specific expert we can easily generate more biologically meaningful conjectures using such a database.

4. DISCUSSION AND CONCLUSIONS

Using the ISA approach, modules were generated, which 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 [5, 13], the ISA algorithm does not require the genes in a cluster be correlated under all the conditions. It also allows genes to be part of multiple modules, since some genes may be 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. Although the MotifRegreesor approach could not effectively discover some motifs such as the TG-box, it may be ascribed to the fact that the genome of *P. falciparum* is AT-rich. Our results on the TG-box and the G-box motifs support the conclusion that this organism may have unique regulatory mechanisms different from other known eukaryotic organisms [10].

The novel database of information relating to *P. falciparum* regulatory elements in IDC cycle can be a useful tool that can 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.

We will provide a website [http://biorg.cs.fiu.edu/CAMDA2004], which will contain all the motifs and modules discovered by our analyses. It will also contain supplementary data, with a list of possible conjectures resulting from our analyses.

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. ELZ is supported by a Florida International University Presidential Graduate Fellowship. Research of GN was supported in part by NIH Grant P01 DA15027-01.

REFERENCES

[1]. Bergmann, S., *et al.* Iterative signature algorithm for the analysis of large-scale gene expression data. *Physical Review E*, *67* (3). -.

[2]. Bozdech, Z., *et al.* The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum. *Plos Biology*, *1* (1). 85-100.

[3]. Conlon, E.M., *et al.* Integrating regulatory motif discovery and genome-wide expression analysis. *Proceedings of the National Academy of Sciences of the United States of America*, *100* (6). 3339-3344.

[4]. Crooks, G.E., *et al.* WebLogo: A sequence logo generator. *Genome Research*, *14* (6). 1188-1190.

[5]. Eisen, M.B., *et al.* Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, *95* (25). 14863-14868.

[6]. Horrocks, P., *et al.* Control of gene expression in Plasmodium falciparum. *Mol Biochem Parasitol*, *95* (2). 171-181.

[7]. Hughes, J.D., *et al.* Computational identification of cis-regulatory elements associated with groups of functionally

related genes in Saccharomyces cerevisiae. *Journal of Molecular Biology*, *296* (5). 1205-1214.

[8]. Ihmels, J., *et al.* Revealing modular organization in the yeast transcriptional network. *Nature Genetics*, *31* (4). 370-377.

[9]. Liu, X.S., *et al.* An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat Biotechnol*, *20* (8). 835-839.

[10]. Militello, K.T., *et al.* Identification of regulatory elements in the Plasmodium falciparum genome. *Mol Biochem Parasitol*, *134* (1). 75-88.

[11]. Roth, F.P., *et al.* Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature Biotechnology*, *16* (10). 939-945.

[12]. Sturn, A., *et al.* Genesis: cluster analysis of microarray data. *Bioinformatics*, *18* (1). 207-208.

[13]. Tamayo, P., *et al.* Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci U S A*, *96* (6). 2907-2912.

[14]. Thompson, J.K., *et al.* 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.

[15]. Troyanskaya, O., *et al.* Missing value estimation methods for DNA microarrays. *Bioinformatics*, *17* (6). 520-525.

[16]. van Helden, J. Regulatory Sequence Analysis Tools. *Nucl. Acids. Res.*, *31* (13). 3593-3596.

[17]. Voss, T.S., *et al.* Genomic distribution and functional characterisation of two distinct and conserved Plasmodium falciparum var gene 5' flanking sequences. *Mol Biochem Parasitol*, *107* (1). 103-115.

[18]. Voss, T.S., *et al.* Identification of nuclear proteins that interact differentially with Plasmodium falciparum var gene promoters. *Mol Microbiol*, *48* (6). 1593-1607.

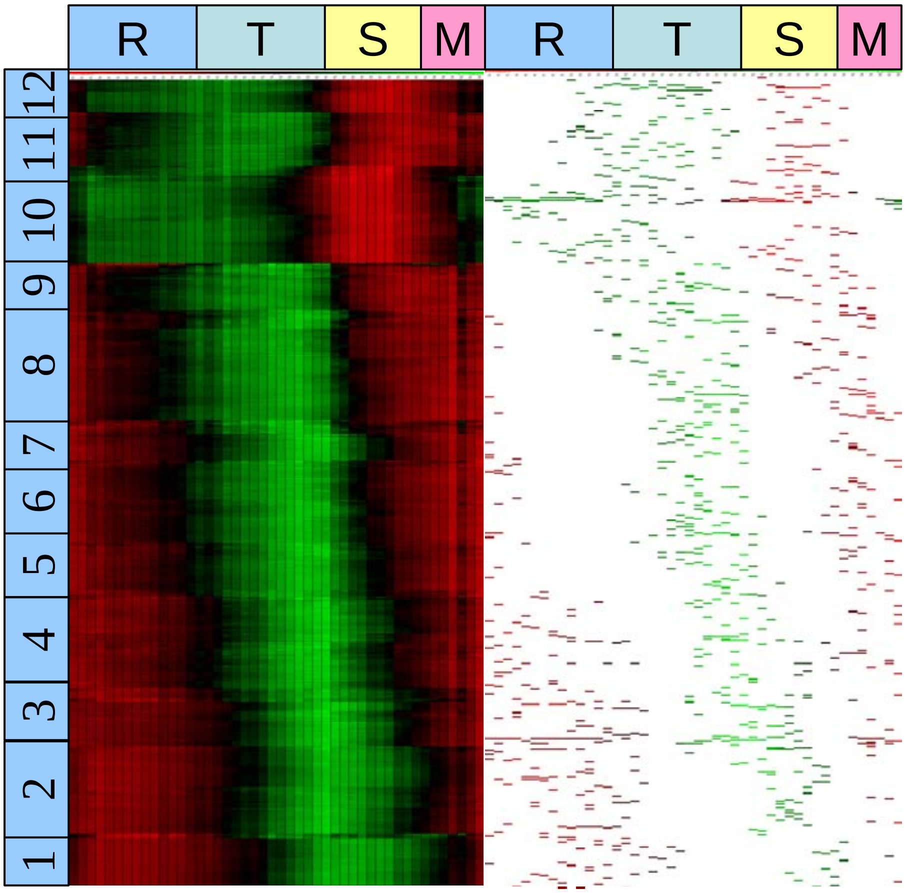


Figure 5. Motif clusters from cell cycle expression time series experiments. The 637 significant motifs reported by Motif Regressor over one cell cycle are clustered by motif coefficients over 46 time points. This figure was produced using Genesis software package by applying hierarchical clustering with Euclidean distance metric on the motif coefficient data [12]**. This figure is best seen in color. 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.**