

Linking Gene Expression Patterns and Transcriptional Regulation in *Plasmodium falciparum*

Aidan J. Peterson
Fox Chase Cancer Center
333 Cottman Avenue
Philadelphia, PA 19111
(215) 728-4067
Aidan.Peterson@fccc.edu

Andrew V. Kossenkov
Fox Chase Cancer Center
333 Cottman Avenue
Philadelphia, PA 19111
(215) 728-1492
Andrew.Kossenkov@fccc.edu

Michael F. Ochs
Fox Chase Cancer Center
333 Cottman Avenue
Philadelphia, PA 19111
(215) 728-3112
Michael.Ochs@fccc.edu

ABSTRACT

Elucidation of the genome sequence of *P. falciparum*, the primary causative agent of human malaria, has opened new avenues for exploring the biology of this microorganism. The CAMDA 2004 dataset offers a detailed view of gene expression during the intra-erythrocyte stage of the parasite life cycle. We used Bayesian Decomposition to model expression patterns in the time series data. We examined the results over a range of potential solutions with the goal of choosing a number of patterns that modeled the experimental data faithfully, and where the genes associated with the patterns provided biological insight. When the data was modeled with seven or eight patterns, each pattern represented a smooth temporal expression pattern whose contributing genes were enriched for Gene Ontology (GO) terms. We discuss the features of the different patterns in terms of GO term enrichment to explore the correlation between genes with functionally related products and co-regulation. As control of gene expression has not been elucidated in *P. falciparum*, we must work backwards from microarray expression profiles that represent the output of the transcriptional program. We use the genomic sequences of genes linked by Bayesian Decomposition to uncover sequence elements related to stage-specific transcriptional control.

1. INTRODUCTION

Malaria is caused by *Plasmodium* parasites that infect and destroy several human cell types during their life cycle. The global effort to reduce the impact of malaria is intimately tied to the study of the complex biology of the protozoan parasite. The genome sequence for *P. falciparum*, the species responsible for the majority of malaria cases, was released in 2002 [7], permitting genome-scale efforts to catalog the proteome and transcriptome of *Plasmodium* [4,6,9]. One fundamental area of *Plasmodium* biology that has not been characterized is the control of gene expression. The CAMDA 2004 dataset provides a high-quality representation of the gene expression behavior for most of the known and predicted genes of *P. falciparum* during the intra-erythrocyte development cycle (IDC). We wish to discover patterns in the expression data that are likely to reflect biological co-regulation. Such groups of co-regulated genes will then be used to explore common predicted regulatory features of the clustered genes as an approach to explore the transcriptional control logic of *Plasmodium*. A better understanding of the biological processes that *Plasmodium* uses during the IDC should lead to improved therapeutics and eventual amelioration of this devastating disease.

The microarray data of Bozdech et al [4] and le Roch et al [9] reveal robust variation in transcript levels through the IDC. Microarray results can be considered a direct readout of the transcriptional program *output*, yet we know very little about the *inputs* directing transcription in this organism. At the genomic level, the transcriptional profiles of the vast majority of *Plasmodium* genes do not show a discernable relationship to chromosome position, which suggests that regulatory mechanisms act on individual genes. Several promoters have been studied to determine which sequence regions control expression of a reporter transcript, and several studies have identified DNA-binding activities from *Plasmodium* nuclear extracts. The basal transcription machinery proteins are present in the genome, as are chromatin components and proteins containing motifs commonly associated with chromatin regulation. Sequence analysis of the *P. falciparum* genome, however, has exposed a surprising paucity of recognizable transcription factors [1,5].

The most conservative model drawn from the available data is that *P. falciparum* uses a set of DNA-binding proteins to control gene expression, but that these factors have not yet been identified experimentally or observed in the genome sequence. It has been suggested that *Plasmodium* relies heavily on post-transcriptional mechanisms to control protein expression. The most direct support for this notion is the discrepancy between the stage-specific qualities of the proteome [6] and the expression of the majority of genes in a single life cycle [4]. The absence of recognized regulatory transcription factor coding regions, an exaggerated number of potential RNA-binding proteins in the genome, and rRNA expression switching during the life cycle have also been used to suggest post-transcriptional control [1,5]. These areas deserve active exploration, but such mechanisms would have to operate *in addition* to the robust transcriptional control revealed by microarray studies.

2. METHODS

2.1 Bayesian Decomposition

Analysis of the microarray data was done with Bayesian Decomposition (BD) [10] in order to identify key regulatory time points within the data. BD allows the identification of overlapping patterns within the data (here, overlapping times of expression) linked to specific genes. This permits the algorithm to both identify groups of genes that initiate expression while other genes continue ongoing expression and to identify genes

that are regulated at multiple points during the parasitic life cycle. This is critically important for promoter analysis, since the limits of the genetic alphabet (G, A, T, C) make identification of DNA binding motifs difficult. The inclusion of promoter regions not truly involved in transcription factor binding quickly leads to loss of signal for identification of promoter elements.

BD was applied to the Overview data set comprising measurements of mRNA levels for 3719 oligos at 46 separate time points varying from 1 to 48 hours post infection. Expression levels were provided as ratios between the mRNA level at the time point and a pooled reference sample comprising a mixture of mRNA from all time points. Time points at 23 and 29 hours were removed due to quality control problems. BD was run positing 3 to 12 patterns, with duplication in all runs with different random seeds. We estimated the noise at 20% of signal (i.e., a multiplicative noise) and missing values were assigned a ratio of 1 with an uncertainty of 100, allowing the algorithm to essentially ignore their contribution during modeling.

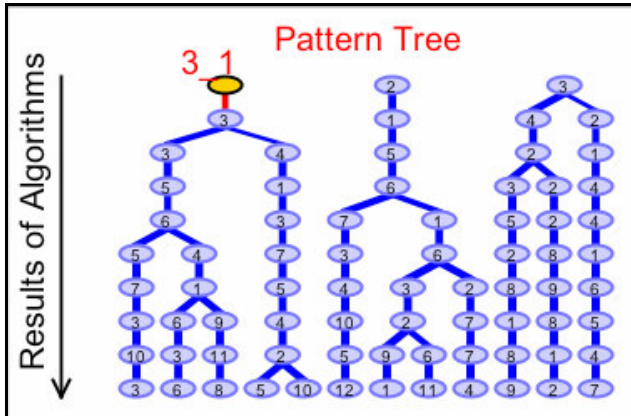


Figure 1. Pattern tree relating expression patterns found by BD for the IDC gene expression cycle. Shown here for 3 to 12 patterns, the ClutrFree tree links the most related patterns, providing a way to visualize the stability and splitting of the patterns as the number of fitted patterns is increased.

2.2 Pattern Visualization

Output files from BD analyses were imported into the ClutrFree [3] program to visualize the expression pattern elements and to analyze the relationships between the patterns. ClutrFree was also used to export the membership matrices describing the weights of each pattern assigned to each oligo element to reconstitute its overall expression profile.

The published GO term annotations for the *P. falciparum* genome [2,7] were combined with the member lists for each pattern determined by BD. For each set of patterns, the enrichment of GO terms in each pattern, relative to the number of times the term appears in the annotated data set, was recorded as a p-value. The p-values were generated from a hypergeometric distribution that takes into account the enrichment ratio and the number of occurrences of each term. A simple metric was devised to track the GO term information content for each pattern. The number of Process and Function terms enriched better than p-values of 0.05 and 0.01 were summed for each pattern. The average score per pattern was used to compare enrichment strength for a range of fitted patterns.

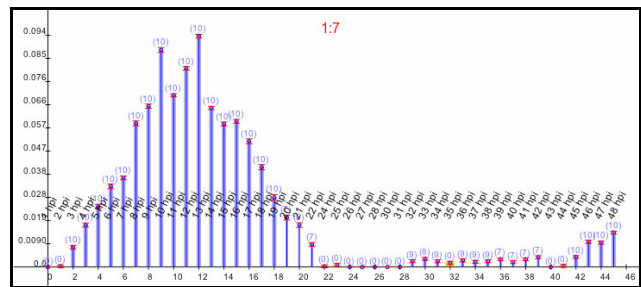
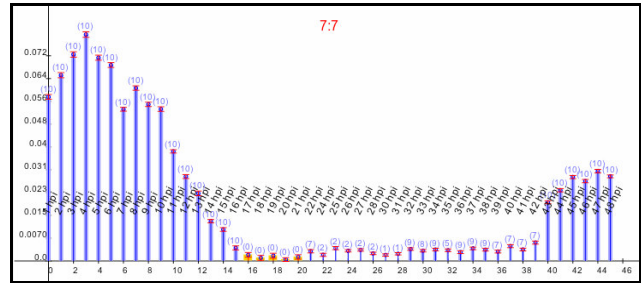
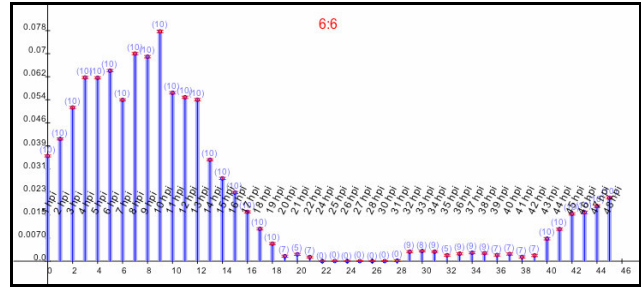


Figure 2. Representative patterns showing relationship between "parent" and "children" temporal patterns. In this example, an expression pattern from the 6 pattern set (top graph) and two patterns from the 7 pattern set (middle and bottom graphs) are shown. The pattern numbers are arbitrary in the sense that the patterns are found de novo during each BD run, but they correspond to patterns 6 of 6, 7 of 7 and 1 of 7 on the pattern tree in Figure 1, where they occupy a branch point in the tree. The peak centered near 8 hours post infection (8 hpi) splits into two peaks near 4 and 12 hpi when an additional pattern is fit by BD.

2.3 Identification of Regulatory Sequences

To discover potential regulatory DNA sequences, we use the pattern information from BD to place genes in co-regulation groups. A list of oligos with strong membership in each of the stable patterns was converted to a gene list, and the genomic sequence near the annotated gene sequence was extracted from PlasmoDB data files. AlignACE, a standard program to detect elements enriched in a subset of sequences [8], will be used to identify sequence patterns that correspond to each expression pattern. We are exploring various parameters including the length of putative promoter sequence to search, and inclusion of alternate potential transcriptional start sites for many *P. falciparum* genes.

3. RESULTS

3.1 Temporal Expression Peaks

BD was performed to find the prominent component patterns in the expression profiles of the oligo elements in the overview dataset of Bozdech et al [4]. The analysis was carried out over the range of 3 to 12 pattern vectors, with several independent runs for each number of patterns. Figure 1 shows a pattern tree representation from ClutrFree that charts the correlations of patterns. The patterns appear as unimodal smooth curves distributed along the time course, despite the fact that no smoothing function was used by the algorithm. The patterns with peaks closest to the first and last time points have a single peak if the pattern is allowed to wrap to the next cycle of the IDC. Visual analysis of the patterns using ClutrFree revealed that as the number of basis vectors fit by the algorithm increased, "parent" patterns split into two patterns, each with temporal peaks offset to either side of the peak of the parent pattern. An example is shown in Figure 2, comparing one of six patterns to the two most related of seven patterns. As the number of patterns increased beyond eight, they began to exhibit features that are unlikely to reflect true temporal gene expression patterns. For example, one of the nine patterns has a broad peak composed of amplitudes that are erratic from hour to hour (Fig 3). Biological patterns are expected to be smooth because the population synchrony is not exact, so the erratic peaks almost certainly reflect sample-to-sample variation in the data. Our conclusion is that fewer than nine patterns should be used to fit the data since additional patterns are less likely to represent true biology.

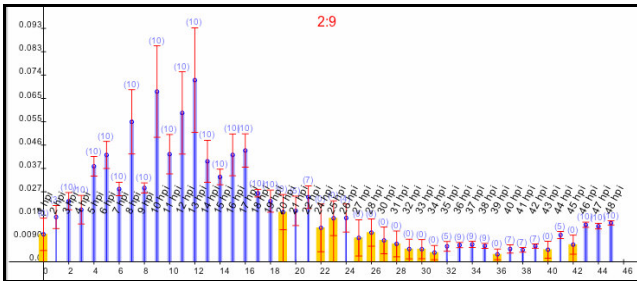


Figure 3. Example of a noisy pattern. This pattern has noisy amplitudes that are erratic from time point to time point over a broad peak. This pattern likely is modeling noise in the dataset rather a meaningful temporal expression pattern.

For three through seven patterns, different simulation runs (with different starting points in solution space) yield virtually identical results, indicating that the solutions are robust. In other words, the same stable solution is reached starting from different random starting points in the model space. The stable positions and shapes of the temporal patterns indicates that the individual expression patterns are not uniformly distributed across the IDC life cycle. This reproducibility of pattern recovery was generally true out to 12 pattern vectors, except for 8 or 10, where several patterns are different for independently determined solutions. In the instance of the eight basis vectors, one pattern element resembles the "noisy" pattern that is consistently present in the 9 pattern set, and another solution splits a temporal peak into daughter peaks, as is observed for pattern number increases in the lower range. This indicates that the sampling algorithm is able to identify two mathematically acceptable solutions, as seen in some other contexts [11]. In summary, the fit patterns are stable over different runs out to seven patterns, and above this number, alternate solutions become possible and thus are found in different fitting runs. This observation suggests that 7 or 8 patterns is the

appropriate range to take advantage of the robust patterns in the data.

3.2 Gene Ontology enrichment

In addition to the visual inspection of pattern features described above, we considered how well the various patterns clustered genes with related gene functions. The number of GO terms enriched above significance thresholds was determined for each pattern in each set of pattern solutions, and an average pattern information score was determined for each pattern number from 3 to 12. A plot of this GO enrichment score versus number of basis vectors (Figure 3) shows that the enrichment per pattern peaks at 7 to 8 patterns, and becomes erratic with higher pattern numbers. To the extent that GO term clustering represents meaningful co-regulation, this analysis suggests that 7 or 8 patterns present biologically motivated solutions. This consideration is especially important if the groups defined by clustering will be used to search for novel shared features such as regulatory DNA sequence elements in co-regulated genes.

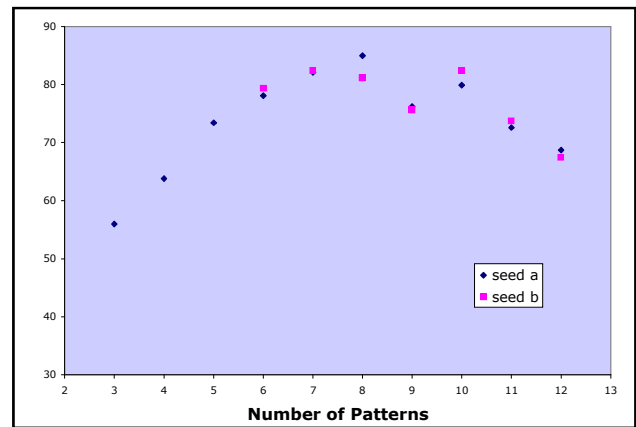


Figure 4. GO term enrichment as a function of the number of basis vectors in Bayesian Decomposition. The average GO enrichment per pattern increases as additional patterns are allowed from 3 to 8 patterns, indicating that the newly formed patterns better fit the biological grouping of the genes. At 9 patterns and above, the average enrichment declines, indicating that the additional finer patterns have less correspondence to biological groups. Seeds a and b indicate different random starting points for fitting the same data.

GO term enrichment and clustering has been noted for gene expression patterns in the published CAMDA data set [4] and in gene expression profiles detected with a short-oligo array set [9]. We examined the GO term enrichment lists for each of the seven stable patterns modeled by BD to determine if they represent meaningful biological themes. Table 1 presents the GO terms enriched in each of the seven patterns, with the pattern peaks arranged chronologically. Some of the enriched terms make sense in terms of the invasion, replication, and maturation cycle that occurs in the erythrocytes. Metabolism, energy generation, and protein synthesis dominate immediately after invasion, followed by DNA replication prior to cell division, and the schizonts express transporters, kinases, and surface molecules in preparation for the next round of invasion. Enrichment of other features, such as RNA binding and mRNA processing terms, are not easily

explained, but suggest broad areas of study that are needed to understand *Plasmodium* replication during the IDC.

We will present at the meeting our analysis of the promoter elements for each of the seven patterns identified in the *Plasmodium* life cycle. These patterns are strongly linked to between 10 and 470 genes (greater than 60% gene behavior explained by one pattern), with several hundred additional genes showing regulation linked to two patterns (greater than 80% gene behavior explained by two patterns). Successful identification of common DNA binding motifs within these elements will support the hypothesis that transcriptional regulation in *Plasmodium* includes use of transcription factors, as in other eukaryotes.

4. CONCLUSION

We used BD to discover pattern elements in the *P. falciparum* IDC transcriptome. Reflecting the nature of the individual input elements, the most robust expression pattern elements have a prominent peak during the IDC. As with any pattern discovery method, choosing an appropriate number of component patterns (or clusters) is not straightforward. We found that fitting seven or eight patterns produced smooth temporal patterns sufficient to reconstruct the data, and maximize the GO term enrichment of the patterns. As one test of the predictive value of the patterns, we are testing whether membership in an IDC expression pattern predicts expression behavior in other stages, such as the gametocyte stages studied by le Roch et al [9].

The common underlying principle in studying co-regulation of genes is that common biological events often underpin coincident expression. By coupling pattern membership information of genes with genomic sequence, we will apply this principle to explore transcriptional control in *Plasmodium*. We will test the prediction that regulatory DNA elements are enriched in genes grouped according to pattern elements discovered using BD.

Table 1. GO terms enriched in the seven stable expression patterns determined by BD. Enriched GO terms from the Process and Function categories were sorted by p-value, and the top twenty for each pattern are listed. Each of the terms listed has a p-value of 0.01 or less.

Peak expression Parasite stage	GO terms enriched
4 hpi early Ring Stage	protein biosynthesis macromolecule biosynthesis RNA metabolism RNA processing RNA binding structural constituent of ribosome structural molecule biosynthesis RNA splicing mRNA splicing mRNA processing RNA helicase pre-mRNA splicing factor transcription nucleic acid binding ATP dependent RNA helicase RNA dependent adenosinetriphosphatase

	transcription, DNA-dependent mRNA binding
12 hpi Ring Stage	structural constituent of ribosome RNA binding macromolecule biosynthesis protein biosynthesis hexose metabolism monosaccharide metabolism RNA splicing threonine endopeptidase proteasome endopeptidase catabolism structural molecule glucose metabolism mRNA splicing RNA processing hexose catabolism alcohol catabolism glucose catabolism monosaccharide catabolism carbohydrate catabolism RNA helicase
21 hpi Trophozoite	protein biosynthesis macromolecule biosynthesis RNA binding nucleic acid binding biosynthesis helicase ligase RNA metabolism structural constituent of ribosome ligase, forming phosphoric ester bonds RNA ligase ligase, forming carbon-oxygen bonds ligase, forming aminoacyl-tRNA and related compounds tRNA ligase amino acid activation ATPase ATP dependent helicase RNA processing RNA splicing structural molecule
31 hpi early Schizont	cell cycle structural molecule DNA metabolism structural constituent of ribosome mitotic cell cycle protein biosynthesis DNA replication and chromosome cycle oxidoreductase S phase of mitotic cell cycle DNA replication macromolecule biosynthesis ligase, forming phosphoric ester bonds biosynthesis RNA ligase ligase, forming carbon-oxygen bonds ligase, forming aminoacyl-tRNA and related compounds tRNA ligase catabolic carbohydrate metabolism

	DNA repair nucleic acid binding
38 hpi late Schizont	phosphorus metabolism phosphate metabolism phosphorylation protein modification protein amino acid phosphorylation cell invasion kinase phosphotransferase, alcohol group as acceptor protein kinase vesicle-mediated transport protein serine/threonine kinase protein serine/threonine phosphatase signal transduction protein amino acid dephosphorylation dephosphorylation protein phosphatase cell communication cytoskeleton organization and biogenesis transferase, transferring phosphorus-containing groups phosphoric monoester hydrolase
42 hpi late Schizont/ early Ring	phosphorus metabolism phosphate metabolism phosphorylation protein modification protein amino acid phosphorylation cell invasion cell communication response to drug response to chemical substance response to abiotic stimulus protein binding cytoskeletal protein binding actin binding multidrug efflux pump multidrug transporter drug transporter protein kinase cytoskeleton organization and biogenesis phosphotransferase, alcohol group as acceptor protein serine/threonine kinase
47 hpi Ring	cysteine-type endopeptidase response to drug response to chemical substance multidrug efflux pump multidrug transporter response to abiotic stimulus drug transporter cell communication transcription transcription, DNA-dependent cell-cell adhesion cell adhesion cysteine-type peptidase response to external stimulus ATPase transcription from Pol I promoter ATP dependent helicase

--	--

5. ACKNOWLEDGEMENTS

We thank Ghislain Bidaut for providing an updated version of the ClutrFree software, and Tom Moloshok and Sinoula Apostolou for assistance and discussions.

6. REFERENCES

- [1] Aravind, L., L.M. Iyer, et al. *Plasmodium* Biology: Genomic Gleanings. *Cell* 115 (2003): 771-785.
- [2] Ashburner, M., C.A. Ball, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25 (2000): 25-9.
- [3] Bidaut, G. and M.F. Ochs. ClutrFree: cluster tree visualization and interpretation. *Bioinformatics* (2004) doi:10.1093/bioinformatics/bth307.
- [4] Bozdech, Z., M. Llinás, et al. The Transcriptome of the Intraerythrocytic Developmental Cycle of *Plasmodium falciparum*. *PLoS Biology* 1 (2003): 1-16.
- [5] Coulson, R.M.R., N. Hall and C.A. Ouzounis. Comparative Genomics of Transcriptional Control in the Human Malaria Parasite *Plasmodium falciparum*. *Genome Res.* (2004) doi/10.1101/gr.2218604.
- [6] Florens, L., M.P. Wahburn, et al. A proteomic view of the *Plasmodium falciparum* life cycle. *Nature* 419 (2002): 520-6.
- [7] Gardner M.J., N. Hall, et al. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* 419 (2002): 498-511.
- [8] Hughes, J.D., P.W. Estep, et al. Computational Identification of *Cis*-regulatory Elements Associated with Groups of Functionally Related Genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.* 296 (2000): 1205-14.
- [9] Le Roch, K.G., Y. Zhou, et al. Discovery of Gene Function by Expression Profiling of the Malaria Parasite Life Cycle. *Science* 301 (2003): 1503-8.
- [10] Moloshok, T.D., R.R. Klevecz, et al. Application of Bayesian Decomposition for analyzing microarray data. *Bioinformatics* 18 (2002): 566-75.
- [11] Ochs, M.F., R.S. Stoyanova, et al. A new method for spectral decomposition using a bilinear Bayesian approach. *J. Magn. Reson.* 137 (1999): 161-76.
- [12] Voss, T.S., M. Vogel, et al. Identification of nuclear proteins that interact differentially with *Plasmodium falciparum* var gene promoters. *Mol. Microbiol.* 48 (2003): 1593-607.

