Simple Methods for Peak Detection in Time Series Microarray Data.

I. Azzini* R. Dell'Anna
Bioinformatics Group, SRA, ITC-Irst
Via Sommarive 18, I-38050 POVO
(TN) - Italy
+39 0461 405324

F. Ciocchetta F. Demichelis
A. Sboner
Bioinformatics Group, SRA, ITC-Irst
Department of Information and C.T.
Trento University, Italy

E. Blanzieri A. Malossini
Department of Information and C.T.
Trento University
Via Sommarive 14, I-38050 POVO
(TN) - Italy
+39 0461 882097

ABSTRACT
In this work we describe a set of analytical and statistical methods enabling the identification of genes which spike expression within narrow time-windows of *Plasmodium falciparum* Intraerythrocytic Developmental Cycle. Moreover we assign a functional group to each gene by using SVM algorithm and discuss about possible biological role.

General Terms
Algorithms, Measurement.

Keywords
Malaria, DNA microarray, Discrete Mathematics, Support Vector Machine (SVM).

1. INTRODUCTION
To develop new drugs and vaccines that disable the malaria parasite *Plasmodium falciparum (P. falciparum)* [16], researchers need a better understanding of the regulatory mechanisms that drive the malarial life cycle. In [2] the first comprehensive transciptome analysis of the *P. falciparum* asexual cycle, or Intraerythrocytic Developmental Cycle (IDC), which is associated with the clinical symptoms of malaria, is provided. Data in [2] show that: 1. at least 60% of the genome is transcriptionally active during this stage; 2. the *P. falciparum* has evolved an extremely specialized mode of transcriptional regulation that produces a continuous cascade of gene expression, beginning with genes corresponding to general cellular processes, and ending with *Plasmodium*-specific functionalities, most of which are poorly understood. In another recent work on the *P. falciparum* biology [12] (and discussed also in [3]), researchers' attention is mainly placed on the poor knowledge about the *P. falciparum* gene functionality. In fact the malaria genome sequencing consortium estimates that more than 60% of the 5,409 predicted ORFs lacks sequence similarity to genes from any other known organism [7].

The simple program regulating the life of *P. falciparum* may hold the key to its downfall, as any perturbation of the regulatory program may have harmful consequences for the parasite [17]. The simple cascade of gene regulation that directs the asexual development of *P. falciparum* is unprecedented in eukaryotic biology [2]. The transcriptome of the IDC resembles a ''just-in-time'' manufacturing process whereby induction of any given gene occurs once per cycle and only at specific time points when required [2].

Therefore, we argued that anomalous behavior of gene expression around specific time points could be intriguing, for instance those around stage transitions of *P. falciparum*. In particular we posed our attention on genes presenting abrupt variations in the transcriptional profile. Among possible abrupt variations we were interested on "peaks" ("spikes"), as they preserve signal periodicity, which is an IDC transcriptome characteristic, as highlighted by [2].

To achieve this goal, we initially used a simple method that locates possible short and relevant "*peaks*" ("*spikes*") within the gene expression time series [1,5,7]. Usually the approach to time series analysis is to approximate them with a continuous interpolating function. However, in this study we chose to preserve the information about each single point. In fact, our goal was to identify open reading frames (ORFs) that present a relevant peak also with very short duration with respect to the overall length of IDC (48 hours). Our method is built around the discrete derivative operator. Additional methods were used to complete the investigation of candidate genes avoiding direct visual inspection of all available time series microarray data.

We used a Machine Learning technique (specifically SVM) to assign these genes to a functional groups, accordingly to the "The *P. falciparum* Functional Groups" (Table S2 given in [2,4]).

The possibility that we incurred in artifacts (peaks due to experimental setting) was considered. To verify the plausibility of the detected peaks we applied different tests: a statistical test, image visual inspection, etc. Test results and discussion are also here reported.

2. PRELIMINARY ANALYSIS
Given the intrinsic complexity of the experiments involving DNA microarray (see for example [15]), we spent some time to investigate reliability of the contest datasets [2,4]. In particular on a selected sample set of available data: 1. we executed a visual inspection of microarray images (the "Primary Data" in [4]), 2. we used TIGR SpotFinder [10] to analyze these images, and finally, 3. we verified the results of GenePixPro3.0 quality control algorithm. GenePixPro 3.0 [8] is the software used in [2] to

* To whom correspondence should be addressed: azzini@itc.it

acquire and analyze the DNA microarray data. The results and the considerations obtained from this step of our work suggested to use the "QC_Dataset" [2,4] that contains already normalized data. It is the set of oligonucleotides that passed all quality control filters, and it was obtained from the "Complete_Dataset" [2,4]. This choice presents some positive aspects: oligos with many missing data, that may affect the results of our methods, are not present; gene expression values obtained from corrupted images are also not included. The "QC_Dataset" contains 5080 of the 7091 oligonucleotides provided by Bozdech et al. [2].

3. METHODS OF ANALYSIS

3.1 Derivative and Integral Methods

Following [13], we considered the "QC_Dataset" as the matrix depicted in Table 1. We called this matrix **E**, denoting with E(o,t) an element of **E**. The variable o indexes the oligos from Oligo1 to Oligo5080, and for the variable t, $t \in TP$, where TP={TP1-TP22, TP24-TP28, TP30-TP48}. TP23 and TP29 were not provided by [2,4]. Missing value in Table 1 where coded as NA.

Table 1. The data matrix obtained by QC_Dataset

Oligo	TP1	…	TP48
Oligo1	$\log_2$(Cy5/Cy3)	…	$\log_2$(Cy5/Cy3)
…	…	…	…
Oligo5080	Log_2(Cy5/Cy3)	…	$\log_2$(Cy5/Cy3)

In order to find within the **E** matrix gene expression with rapid changes over time (candidate peaks), we used a naïve method (M1) based on the derivative operator. Afterwards, another set of methods (M2-M6) has been used to soundly complete M1 results. This strategy allows finding genes that have a well-shaped peak. M1 method proceeds as described in Figure 1: it simply computes (Step 2) for every oligo (70mer oligonucleotides) in **E**, the difference between the maximum and minimum of the discrete derivative, obtained from Step 1. Then it sorts the resulted array **S** to derive a dataset stratification (for example in increasing order).

Given in input Matrix **E**,

Do $\forall o$, $\forall t \in TP$ such that $E(o,t)$ and $E(o,t+1) \neq NA$:

{ **Step 1.** (Discrete Derivative). Compute:

$$\frac{\Delta E(o,t)}{\Delta t} = \frac{E(o,t+1)-E(o,t)}{(t+1)-t} = \Delta E(o,t)$$

Step 2. (Score). Compute:

$$\left[\max_t(\Delta E(o,t)) - \min_t(\Delta E(o,t))\right] = S_o \}$$

Step 3. (DataSet Stratification). Sort vector $\mathbf{S} = \{ S_o \ \forall o\}$.

Figure 1. The derivative method M1.

All methods (briefly M_i) are reported in Table 2. The methods can be split in three main classes: *Derivative Methods* (M1, M2, M3), *Integral Methods* (M4, M5), and *Other Methods* (M6). Methods M4, reported in Figure 2, computes a score, to be associated to any oligo in **E**, as the fraction of area under a possible peak. Method M6 is a modification of the methods described in [6]; it is a statistical method which is able to detect edges in gene expression time series.

Given in input Matrix **E**, Do $\forall o$, {

Step 1. (Integral). Compute $\sum_{t \in TP} |E(o,t)| = A_1$, $E(o,t) \neq NA$:

Step 2. (Discrete Derivate). If E(o,t) and E(o,t+1)≠NA, compute:

$$\frac{\Delta E(o,t)}{\Delta t} = \frac{E(o,t+1)-E(o,t)}{(t+1)-t} = \Delta E(o,t)$$

Step 3. (Maximum Localization). Find:

$$\tau = \arg\max_{t \in TP}(\Delta E(o,t))$$

Step 4. (Local Integral). Compute

$$\sum_{t=\tau-u}^{t=\tau+u} |E(o,t)| = A_2 \ \ (u = 1,2)$$

Step 5. (Score). Compute $\frac{A_2}{A_1} = S_o$ }

Step 6. (DataSet Stratification). Sort vector $\mathbf{S} = \{ S_o \ \forall o\}$.

Figure 2. The integral method M4.

Table 2. The method set used in this work.

Methods	Description
Derivative	
M1	Figure 1.
M2	As M1, with Step 1 in Figure 1 replaced by: $\frac{\Delta E(o,t)}{\Delta t} = \frac{E(o,t+2)-E(o,t)}{(t+2)-t} = \Delta_2 E(o,t)$
M3	As M1, with Step 2 in Figure 1 replaced by: $\left[\left\|\max_t(\Delta E(o,t))\right\| - \left\|\min_t(\Delta E(o,t))\right\|\right] - \left[\max_t(\Delta E(o,t)) - \min_t(\Delta E(o,t))\right] = S_o$
Integral	
M4	Figure 2.
M5	As M4, "argmax" replaced by "argmin" (Step 3, Figure 2)
Other Methods	
M6	Major Edge Detection, reported in [7].

All methods assign a score S_o to each oligo. The higher the score S_o the higher the probability to find a significant peak with respect to average signal amplitude. For each method M_i, we directly analyzed the oligos with highest score, observing that only oligos with very high score show a well-shaped peak. So we decided to consider only the top ten scored oligos for each methods M_i. In this set of 60 oligos, there are 23 distinct oligos. So we could have a direct visual control on all candidate peaks. In next sections we describe the characteristics of these oligos, and the assigned functional classification.

3.2 Functional Classification using SVM

In order to obtain a functional classification of the 23 oligos detected by the methods M_i, we used a Multiclass Support Vector Machine (M-SVM) [11]. In particular, we adopted the pairwise classification approach, where for each possible pair of functional classes a SVM classifier is trained. For N classes, this results in (N-1)*N/2 binary classifiers. When we tried to classify oligos, we evaluated all the possible binary classifiers, and the class is obtained by a majority voting scheme, i.e. the class with

the highest number of votes gives the label. We used the package e1071 for the R system which provides the M-SVM algorithm [14]. We used the training set provided in Table S2 [2,4] which is a selection of oligos whose functional classification is known. The 530 samples, labeled in 14 different functional classes (see the Table S2 in [2], and Figure 3), are used to train a linear M-SVM. Since oligos with missing features are not included in the SVM training algorithm, we completed the training set adopting the following procedure. When the missing value is not at the beginning or at the end of the time series we applied the following schema (A-B are time point values and "?" denotes a missing value): *1*.A ? B→A (A+B)/2 B; *2*. A ? ? B→A (A+B)/2 (A+B)/2 B, and *3*. A ? ? ? B→A (A+B)/2 (A+B)/2 (A+B)/2 B. When the missing value is at one of the edges of the series we applied: *1*. ? A B→2A-B A B and *2*. A B ?→A B 2B-A. Moreover, we provided a method to rank the oligos based on the voting scheme absolute values.

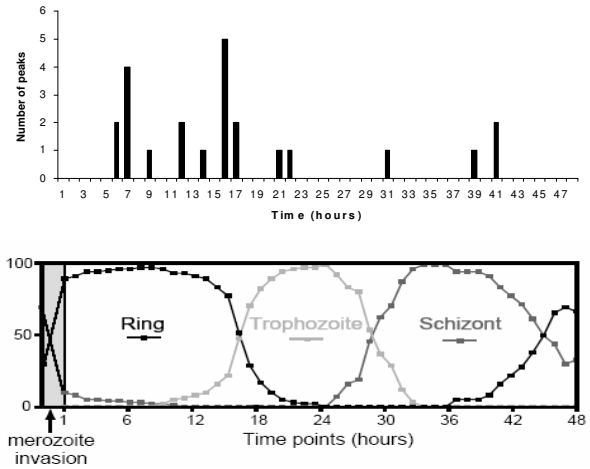

Figure 3. Peak temporal distribution, compared with the IDC temporal evolution [2].

4. RESULTS

In this section we present M_i and SVM results, and discuss about their consistency.

4.1 M_i Results

The 23 oligos found by our methods are reported in Table 3, in Appendix. The temporal distribution of the 23 peak positions is reported in Figure 3 and compared with the major morphological stages throughout the IDC. For this distribution of the peaks we performed the Kolmogorv-Smirnov test in order to reject the hypothesis of uniform distribution. We could reject this hypothesis with confidence level of 0.99. To give an idea about the shape of the detected peaks some representative oligos from Table 3 are reported and discussed hereinafter. In Figure 4 the single oligonucleotides f65819_1 showing a peak at TP7 is depicted. This oligo has a Signal to Noise Ratio (SNR) equal to 3.46 [4]. This peak involves three time points, i.e. is 2 hours long. We observe that TP7 is represented by more than one array hybridization [2].

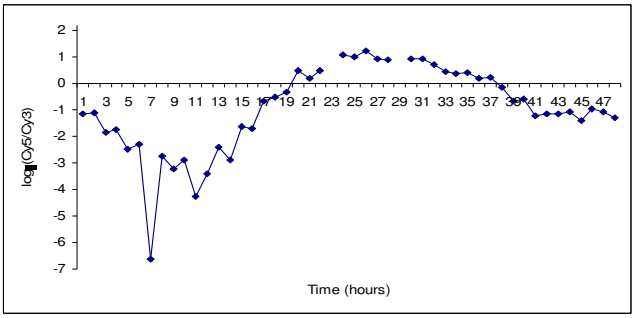

Figure 4. The single oligonucleotide f65819_1.

The peak is placed on a periodical expression profile typical of many ORFs during the IDC [2]. We have no biological information about this ORF, however we may observe that the peak occurs when the Trophozoite phase begins.

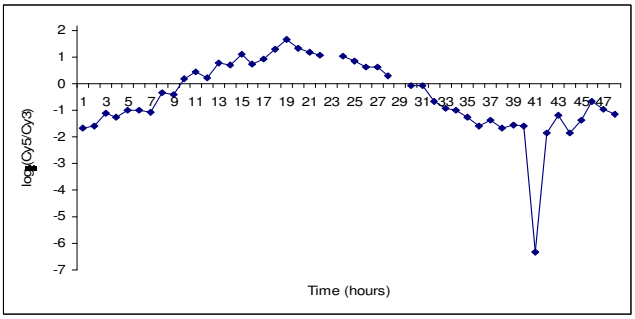

Figure 5. The single oligonucleotide MAL7P1.88-f71224_1.

In Figure 5 a second ORF is displayed, which peak is similar to the previous one (SNR=2.6). The description for this ORF is "*hypothetical protein*". In this case the peak in gene expression profile is placed at TP39, i.e. approximately when the Ring stage starts. In Figure 6 we report another single oligo (SNR=8.41), in this case the peak holds for about 4 hours; furthermore for this ORF we know its functional group: early ring transcripts. In Figure 7. there is a ORF distributed along two oligos (SNR= 14.37 and 8.28). Among the 3 multiple oligonucleotides identified by our methods, these ones have the highest Pearson's correlation value (0.96). The description in this case is: "*PFB0935w ::cytoadherence linked asexual protein 2 ::100% identity to 100% of probable secreted protein PFB0935w-malaria parasite (Plasmodium falciparum)*." We can observe that the peak, even if with different amplitude, is present in the expression of both oligos.

4.2 M-SVM Results

As stated above, we trained the M-SVM on the samples contained in Table S2 [4,5], where we inserted the missing features as explained in Section 3.2. The leave-one-out cross-validation accuracy for the M-SVM is 73%. We applied the classification model of the M-SVM to all the oligos in the "Complete_Dataset" [2,4]. Then, using the absolute values of the voting scheme as a measure of the confidence classification for each oligo, we ranked them in descending order and examined the first 100 oligos.

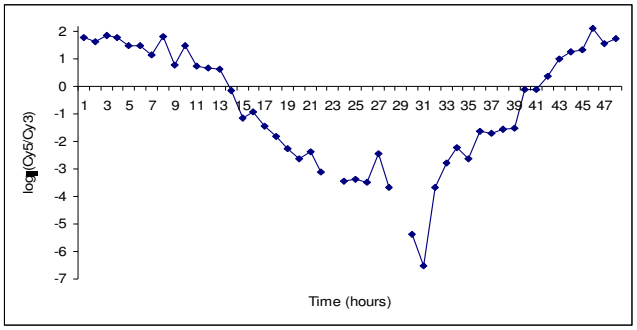

Figure 6. The single oligonucleotide PFL0035c-l1_28.

Similar genes can be found in Table 3.

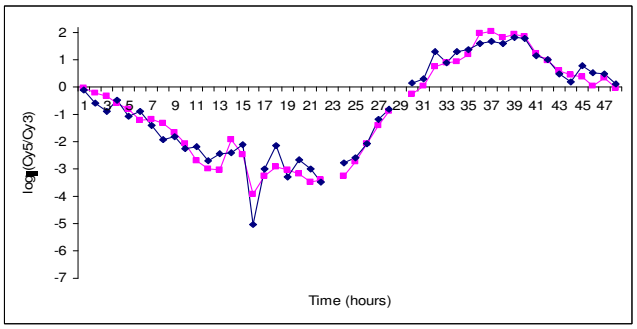

Figure 7. The multiple oligonucleotides PFB0935w (b593, b597).

This table is very large and thus it is not reported here but is available upon request. One of the 23 oligos located by M_i is among the first 100 oligos.

The functional classifications of the 23 oligos predicted by our M-SVM are shown in Table 3. From this table we can draw some observations. The multiple oligos (*red*, *blue* and *green*) with high Pearson correlation fall in the same functional class. The *brown* oligos have relatively low Pearson correlation and then, as we expected, they are in different functional groups. Furthermore, the 3 oligos with known functional group are correctly classified.

4.3 Relations Between the Two Methodologies

In this section we analyze the relations between the functional groups predicted by M-SVM and the peak distribution. From Table 3 we can derive a 'BoxPlot' where the peak positions TP (y axis) are considered as function of the functional group (x axis) predicted by SVM (Figure 8). The functional group order is derived by the phaseogram given in [2]: starting with genes involved in general cellular process and ending with genes devoted to P. Falciparum specific functionality. The two classes M and OM are not considered in [2]. The classes (CM, M, TC, TM) contain only one oligo. The classes (DR, MI, ET, OM) contain oligos with peaks distributed along very short biological time.

5. DISCUSSION

The work described in this paper was mainly dedicated to search in the gene expression time series of the *P. falciparum* asexual cycle anomalous features, associated to temporal steps meaningful from the biological viewpoint.

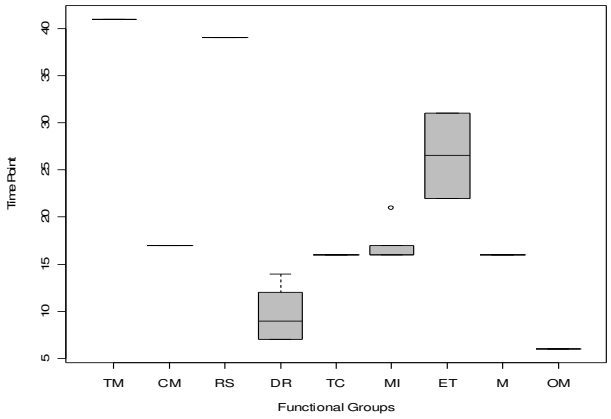

Figure 8. BoxPlot of Functional Groups and Time Points.

Features like peaks constitute one of the simpler anomalies in periodic signals; they do not alter signal periodicity and can be easily characterized by a set of consistent methods. Therefore we targeted our investigation on peaks.

The methods we employed could detect artifacts: to reject this possibility we performed a pool of tests among which a statistical one. We restricted our analysis to the "QC_Dataset", in which gene expression measured on corrupted images are not present. We discarded the oligos with a SNR lower than 2. Moreover we verified by visual inspection that the images corresponding to the temporal positions of the peaks we found do not present any anomalies.

In the temporal distribution of peak positions as reported in Figure 3, the majority of populated channels appeared to be located in time steps related to some stage transitions of *P. falciparum*. The result of Kolmogorv-Smirnov test performed on this distribution allows us to state with a high level of confidence (99%) that this peak position distribution does not come from an uniform one, suggesting that peaks are not due by random experimental errors. In addition, the different concentration of detected peaks at different time points does not seem to be due to the limited number of oligos detected in our study.

The synchronization of in vitro culture of parasites could hardly weaken the reliability of the shortest peaks, i.e. peaks two hours long. However, as already reported in 4.1, some peaks are obtained for more than one array hybridization, other are present in oligos multiple of those detected in our study too, and, in some cases, the peak duration is longer than 2 hours. These elements reduce the possibility of artifacts in the peak set detected.

In general the first part of our study seems confirm the existence of a set of oligos presenting abrupt variations in their transcriptional profiles located mainly in specific time points, relevant for the IDC temporal evolution. These results need to be confirmed by new experiments, possibly characterized by a higher time resolution (less than 1 hour).

The difficulty of relating the existence of these peaks to the cascade of gene regulation that directs the asexual development of *P. falciparum* is also due to the fact that the functionality of most of ORFs found out is unknown. Therefore, in the second part of our study we concentrated on the functional classification of the

23 oligos detected using a SVM. This should help the assessment of the biological meaning of the peaks.

It is worth noting that another proof of the possible existence of a biological meaning for the peaks detected in our study comes from Figure 8, in which, for some of the classes devoted to specific functionality of P. Falciparum, peaks distribute along very short biological ranges which do not overlap each other.

The M-SVM methods extended to the complete dataset showed interesting features. As a matter of fact, the analysis of the first 100 genes showed that several oligos classified as unknown in the original work [2] are well-known to be involved in the Early Ring Transcript phase. This is the case of RESA1, RESA2, STARP, PM-I (emoglobinase) and some dnaJ-like or resa-like genes. Our M-SVM model correctly assign them to the ET phase. It seems interesting an hypothetical gene (ks826_2 and i6740_1) annoted as coronin binding. Similar coronin proteins are involved in the cytoskeleton dynamics, they interacts with actin and play important roles in the generation of FAGOSOMI. It seems interesting that both the oligos are assigned to the Early Ring Transcription phase when the formation of PV (vacuolo parasitoforo) occurs. A large number of oligos classified with high confidence seems to belong to the Early Ring Transcript phase where cellular and molecular processes are specific to the plasmodium. A crucial phase for the design of drugs is the Merozoite Invasion. Tables with the Oligos classified as Merozoite Invasion and Early Ring Transcript with the highest confidence, are very large and then are available upon request.

6. REFERENCES

[1] Bar-Joseph, Z., Analyzing time series gene expression data, *Bioinformatics in press,* 2004.

[2] Bozdech, Z., Llinas, M., Pulliam, B.L., Wong, E.D., Zhu J., and DeRisi, J.L., The Transcriptome of the Intraerythrocytic Developmental Cycle of *Plasmodium falciparum* PLoS Biol. 2003 October; 1 (1): e5 DOI: 10.1371/journal.pbio.0000005.

[3] Broudy, T., *The Modern Age of Malaria Research: Finding New Ways to Combat an Old Disease* Affimetrix Research Community, www.affymetrix.com, September 2003.

[4] CAMDA 2004 Conference, *Contest Datasets*: www.camda.duke.edu/camda04/datasets.

[5] Erdal, S., O. Ozgur., Armbruster, D., Ferhatosmanoglu, H., Ray, W.C., A Time Series Analysis of Microarray Data *Fourth IEEE Symposium on Bioinformatics and Bioengineering (BIBE'04) May 19 - 21*, 2004 Taichung, Taiwan, ROC.

[6] Filkov, V., Skiena, S., Zhi, J., Analysis techniques for microarray time-series data. *J Com Biol* 2002;9(2):317-30.

[7] Gardner, M.J., Hall, N., Fung, E., White, O., Berriman, M., et al. (2002) Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature 419: 498–511.*

[8] GenePix Pro: *The image analysis software for microarrays, tissue arrays and cell* arrays: www.axon.com.

[9] Griffiths, A.J.F., et al. Modern Genetic Analysis. *New York: W. H. Freeman & Co.*: 1999.

[10] Institute for Genomics Research (TIGR): www.tigr.org.

[11] Kreßel, U., Pairwise classification and support vector machine. *In B. Schölkopf, C.J.C. Burges and A.J. Smola, editors, Advances in Kernel methods–SV Learning*, pages 255-268, Cambridge, MA, 1999, MIT Press.

[12] Le Roch, K.G., Zhou, Y., Blair, P.L., Grainger, M., Moch, J.K., Haynes, J.D., De La Vega, P., Holde,r A.A., Batalov, S., Carucci, D.J., Winzeler, E.A., Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science*. 2003 Sep 12;301(5639):1503-8. Epub 2003 Jul 31.

[13] Molla, M., Waddell, M., Page, D., Shavlik, J., (2004). Using Machine Learning to Design and Interpret Gene-Expression Microarrays. *AI Magazine*, 25, pp. 23-44. (*To Appear in the Special Issue on Bioinformatics*).

[14] R project: www.r-project.org ISBN 3-900051-01-1.

[15] Sebastiani, P., Gussoni, E., Kohane I.S., Ramoni, M., (2003), Statistical Challenges in Functional Genomics. (With discussion) *Statistical Science*. 18, 33-70.

[16] Suh, K.N., Kain, Kevin, C., Keystone, J. S., Malaria *CMAJ. 2004 May 25; 170 (11)*: 1693 1702 DOI: 10.1053/cmaj.1030418.

[17] Ward G. (editor), *Monitoring Malaria: Genomic Activity of the Parasite in Human Blood Cells* Public Library of Science, Open-access article, PLoS Biol. 2003 October, University of Vermont.

APPENDIX

Table 3. Functional Classification and Time Point of the 23 oligos. Oligo with the same color are multiple.

oligos	Known Functional Groups	Predicted Functional Class	Time Point
opfblob0072		Cytoplas transl Machin -CM	TP17
n128_25		DNA Replication -DR	TP7
f65819_1		DR	TP7
m364_2		DR	TP12
m12963_1		DR	TP12
n159_34		DR	TP7
ks244_7		DR	TP12
n128_61		DR	TP9
opfm60504		DR	TP14
l1_28	ET	Early ring Transcripts –ET	TP31
ks75_15	ET	ET	TP22
c154		Mitochondrial -M	TP16
b593		Merozoite Invasion -MI	TP16
b597		MI	TP16
n176_5		MI	TP16
opfh0008		MI	TP21
opfblob0105		MI	TP17
b541		MI	TP16
n132_108		OM Organel transl Mach-OM	TP6
m50253_2		OM	TP14
ks1030_4	OM	OM	TP6
n128_33		Ribonucleotide Synthesis -RS	TP7,TP39
f71224_1		RS	TP39
opfh0022		TCA Cycle -TC	TP16
e17542_1		Transcription Machinery -TM	TP41