Chapter 3

Simple Methods for Peak and Valley Detection in Time Series Microarray Data

A. Sboner[a,b], A. Romanel[b], A. Malossini[b], F. Ciocchetta[a,b], F. Demichelis[a,b], I. Azzini[a], E. Blanzieri[b] and R. Dell'Anna[a]

[a]*Bioinformatics Group, SRA Division, ITC-irst, Via Sommarive 18, I-38050 Povo (TN), Italy*
[b]*Department of Information and Communication Technology, University of Trento, Via Sommarive 14, I-38050 Povo (TN), Italy*

Abstract Given a set of gene expression time series obtained by a microarray experiment, this work proposes a novel quality control procedure that exploits six analytical methods, each of which allows for the identification in an automated way of genes that have expression spikes within narrow time-windows and over a chosen amplitude threshold. The output of these methods, suitably combined in an automated way, provides an exhaustive list of genes and time points in which abrupt variations have been detected. The quality control on these genes is then performed by a biologist, who classifies the spikes either as biologically relevant or as artifacts. In the latter case, spikes must be eliminated by a smoothing procedure. In this chapter, we first describe the six methods and their iterative and automated implementation. As a case study, we discuss the application of the panel of these six methods to the transcriptome of *Plasmodium falciparum* intraerythrocytic developmental cycle. Assuming that spikes detected in this set have been labeled as artifacts by a biologist, in the second part of the chapter we discuss the effect of our smoothing procedure for different types of data analysis.

Keywords: malaria, DNA microarray, discrete mathematics, support vector machine (SVM), quality control

1. INTRODUCTION

To develop new drugs and vaccines that disable the malaria parasite *Plasmodium falciparum* (*P. falciparum*) [19], researchers need a better understanding of the regulatory mechanisms that drive the malarial life cycle. In [2], the first comprehensive transcriptome 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, and (2) *P. falciparum* has developed an extremely specialized mode of transcriptional regulation. A continuous cascade of gene expression is produced, beginning with genes corresponding to general cellular processes, and ending with *Plasmodium*-specific functionalities, most of which are poorly understood. In other recent works on the biology of *P. falciparum* [3,15], attention is mainly devoted to the poor knowledge of the *P. falciparum* gene functionalities. In fact, the malaria genome sequencing consortium estimates that more than 60% of the 5,409 predicted open reading frames (ORFs) lack sequence similarity with genes from any other known organism [8].

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 [20]. 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].

Quality control in microarray data analysis aims at discarding flawed data at an early stage of the analysis. The typical quality control procedure is performed after measurements on the raw digital image, in order to ensure that the measurements are not affected by image artifacts and thus increasing signal-to-noise ratio. However, given the experimental structure of present datasets, namely the time series component, it is possible to use such temporal information in order to further detect expression points that could be still affected by noise. Abrupt variations in the transcriptional profile can indicate anomalous behavior that needs to be assessed by a biologist as (a) being artifacts, or (b) carrying relevant biological information. Among abrupt variations, we were particularly interested in peaks and valleys, as they preserve signal periodicity, which (as shown in [2]) is an IDC transcriptome characteristic. Usually, time-series analysis [1,5,6] first approximates temporal signals by a continuous interpolating function. However, in this study we chose to preserve the actual information contained in each time point. In fact, our goal is to identify ORFs that show a relevant variation with very short duration with respect to the overall length of IDC (48 hours). To achieve this goal, we set up five different simple methods based on the discrete derivative and integral operators. An additional sixth method directly matches abrupt variations on transcriptional profiles. These six methods separately perform gene expression investigation in an iterative and automated way, thus avoiding time-expensive, direct visual inspection of all available time series microarray data. The output lists of detected genes and time points from the six methods generally overlap but are not coincident because the six methods are concerned with different behaviors

in the temporal signal. Therefore, by merging in an automated way the six different output lists (see Section 3.1), a more comprehensive list of genes and time points at which relevant peaks and valleys are present can be obtained. As mentioned, the detected spikes can be classified by the biologist either as biologically relevant or as artifacts. In the former case, further analysis for biological interpretation of the results is required. In the latter case, peaks and valleys are artifacts that were not detected by conventional quality control procedures. They are therefore removed and substituted by a smoothing procedure that preserves the periodic nature of the overall signal.

The first part of the chapter is devoted to the description of our procedure for peak and valley detection. We discuss the application of the panel of the six methods to the transcriptome of *P. falciparum* IDC. Assuming that any biological relevance of the abrupt variations in this set is ruled out by a biologist, in the second part of the chapter we check whether the smoothing procedure influences further analysis. We find that our smoothing procedure changes the data analysis results. Our quality control procedure can therefore be effective either in further improving the signal to noise ratio of time series microarrays data or in highlighting possible biologically important time points in the same data.

2. PRELIMINARY ANALYSIS

Given the intrinsic complexity of the experiments involving DNA microarray (see for example [10,17]), we investigated thoroughly the reliability of the contest datasets [2,4]. In particular, on a selected sample set of available data, we: (1) performed a visual inspection of microarray images (the "Primary Data" in [4]), (2) used TIGR SpotFinder [12] to analyze these images, and finally, (3) checked the results of GenePixPro3.0 quality control algorithm. GenePixPro3.0 [9] is the software used in [2] to acquire and analyze the DNA microarray data. The results and considerations obtained from this step of our work suggested us to use the "QC_dataset" [2,4]. This is the set of oligonucleotides that passed all quality control filters and was obtained from the "Complete_Dataset" [2,4]. This choice presents some positive aspects: oligos with many missing data, which may affect the results of our methods, are not present; gene expression values obtained from corrupted images are also not included. Moreover, this choice allows us to prove that our quality control procedure is able to further increase the signal to noise ratio. The "QC_dataset" contains 5080 of the 7091 oligonucleotides provided by Bozdech et al. [2].

3. METHODS OF ANALYSIS

3.1. Detection Methods

Following [13], we considered the "QC_dataset" as the matrix depicted in Table 1. We label this matrix **E**, denoting with $E(o,t)$ an element of **E**. The variable o indexes the oligos from Oligo_1 to Oligo_{5080}, and for the variable t, $t \in \text{TP}$, where $\text{TP} = \{\text{TP}_1, \ldots, \text{TP}_{22}, \text{TP}_{24}, \ldots, \text{TP}_{28}, \text{TP}_{30}, \ldots, \text{TP}_{48}\}$. TP_{23} and TP_{29} were not provided by [2,4]. Missing values in Table 1 were imputed with the "loess()" local regression function, provided by the "stats" package of R (version 2.01) [11,2]. The local weighting parameter was reduced to 12%.

In order to find within the **E** matrix gene expressions with rapid changes in time (in particular, candidate peaks and valleys), we exploited six different methods (labeled M_i, $i = 1, \ldots, 6$), concisely reported in Table 2. They can be split into three main classes: derivative methods (M_1, M_2, M_3), integral methods (M_4, M_5), and other methods (M_6).

3.1.1. Method Description. Each method can be described at abstract level as follows. For each transcriptional profile, method M_i detects a time point τ_i in which the expression variation occurs. This is accomplished by means of the score S_o. For methods M_1, M_2, M_4, M_5 and M_6, the higher the score S_o, the higher the probability to find a significant peak (or valley) with respect to the average signal amplitude. Contrary to the other five methods, for method M_3 the closer to zero is S_o, the higher is the probability to find a significant peak (or valley).

All methods M_i differ in the way they calculate S_o. Method M_1 proceeds for each oligo o as described in Figure 1:

Step 1: M_1 computes at each time point t the discrete derivative, calculated as the ratio of finite differences of width one;

Step 2: the maximum absolute value of the discrete derivative of Step 1 is calculated. This value is S_o.

The same procedure characterizes method M_2 with the discrete derivative calculated as the ratio of finite differences of width two.

Method M_3 proceeds for each oligo o as follows:

Table 1. The data matrix **E** obtained by QC_dataset

Oligo	TP_1	...	TP_{48}
Oligo_1	$\log_2$(Cy5/Cy3)	...	$\log_2$(Cy5/Cy3)
...	...	...	...
Oligo_{5080}	$\log_2$(Cy5/Cy3)	...	$\log_2$(Cy5/Cy3)

Given in input matrix **E**,
Do ∀ oligo o,
{

Step 1. (Discrete derivative). $\forall t \in$ TP 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:

$$\max_{t \in TP} |\Delta E(o,t)| = S_o$$

}

Figure 1. The derivative method M_1.

Step 1: the discrete derivative $\Delta E(o,t)$ is calculated at each time point t, as in M_1;

Step 2: the maximum and minimum values of $\Delta E(o,t)$ are calculated;

Step 3: S_o is calculated as reported in Table 2. As already pointed out, the smaller is S_o, the higher is the probability to find a significant peak (or valley). This formula allows us to discriminate between spikes and unit-step like behavior of the signal.

Method M_4 is reported in Figure 2. For each oligo o it proceeds as follows:

Step 1: a normalization is performed, by subtracting from each element $E(o,t)$ the arithmetic mean computed on the whole temporal signal;

Step 2: the discrete integral A_1 of the absolute value of the normalized signal is calculated;

Step 3: the discrete derivative (width one) $\Delta E(o,t)$ of the original signal is calculated;

Step 4: the maximum value of $\Delta E(o,t)$ is calculated and the time point τ at which it occurs is stored;

Step 5: the positive integral A_2 of the normalized signal of width two around τ is calculated;

Step 6: the fraction of area $S_o = A_2/A_1$ is calculated.

The same procedure as in M_4 characterizes method M_5, except for Step 4, in which the minimum value of $\Delta E(o,t)$ is calculated. In other words, M_4 detects peaks while M_5 detects valleys.

Methods M_6, reported in Figure 3, proceeds for each oligo o as follows:

Step 1: time points t, $t+1$, $t+2$ are considered (for each $t \in$ TP) and the values α and β are calculated (see Figure 3). In case of a perfect spike, $\alpha = \beta$. For a first type discontinuity (a unit-step function) α or β is zero. There-

Given in input Matrix **E**,

Do ∀ oligo o,

{

Step 1. (Normalization). Compute**:** $\forall t \in \mathrm{TP}$

$$\overline{E}(o,t) = E(o,t) - \underset{t\in \mathrm{TP}}{\mathrm{mean}}(E(o,t))$$

where *mean* is the arithmetic mean over time

Step 2. (Integral). Compute:

$$\sum_{t\in \mathrm{TP}} \left|\overline{E}(o,t)\right| = A_1$$

Step 3. (Discrete derivate), compute:

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

Step 4. (Maximum localization). Find:

$$\tau = \underset{t\in \mathrm{TP}}{\arg\max}\left(\Delta E(o,t)\right)$$

Step 5. (Local integral). Compute:

$$\sum_{t=\tau-1}^{\tau+1} \left|\bar{E}(o,t)\right| = A_2$$

Step 6. (Score). Compute $\dfrac{A_2}{A_1} = S_o$

}

Figure 2. The integral method M_4.

fore $SV(E(o,t))$ yields one in the first case, and zero in the second. The term $(\alpha+\beta)/2$ weights the asymmetry of non perfect spikes;

Step 2: the properly normalized maximum value of $SV(E(o,t))$ gives score S_o.

Method M_6, therefore, looks for three-point structures in each gene profile, weighting their possible asymmetry and selecting that structure for which the area is maximal.

3.1.2. Spike Detection. In order to single out peaks and valleys in temporal signal, an amplitude threshold, called *pv*, must be given. The spike detection procedure identifies in an automated way those expression variations which are greater than *pv*. In this procedure, each method M_i is separately and

Given in input Matrix **E**,

Do ∀ oligo o,

{

Step 1. (Spike value detection). $\forall t \in \text{TP}$ compute:

$$\alpha = \left|E(o,t+1) - E(o,t)\right|$$
$$\beta = \left|E(o,t+2) - E(o,t+1)\right|$$
$$SV(E(o,t)) = \frac{\min\{\alpha,\beta\}}{\max\{\alpha,\beta\}} \cdot \frac{\alpha+\beta}{2}$$

Step 2. (Score). Compute:

$$\frac{\max_{t\in\text{TP}} SV(E(o,t))}{\sum_{t\in\text{TP}} SV(E(o,t))} = S_o$$

}

Figure 3. The method M_6.

Table 2. The methods M_i for spike detection

Methods	Description
Derivative	
M_1	Figure 1
M_2	As M_1, with Step 1 in Figure 1 replaced by: $\frac{\Delta E(o,t)}{\Delta t} = \frac{E(o,t+2)-E(o,t)}{(t+2)-t} = \Delta_2 E(o,t)$
M_3	As M_1, with Step 2 in Figure 1 replaced by: $[\lvert \max_{t\in\text{TP}}(\Delta E(o,t))\rvert - \lvert \min_{t\in\text{TP}}(\Delta E(o,t))\rvert] - [\max_{t\in\text{TP}}(\Delta E(o,t)) - \min_{t\in\text{TP}}(\Delta E(o,t))] = S_o$
Integral	
M_4	Figure 2
M_5	As M_4, with "argmax" replaced by "argmin" (Step 3, Figure 2)
Other methods	
M_6	Figure 3

iteratively applied to each expressionary time series of the considered dataset. Each method M_i provides its final list of genes and time points at which the expression values are greater than pv. In other words, for each method M_i the iterative procedure can be schematized as follows: (i) For each expressionary time series $E(o,t)$, the time point τ_i is found for which the maximum value

of S_o occurs (or minimum value S_o for method M_3); (ii) if the expression value $E(o, \tau_i)$ is greater than pv, $E(o, \tau_i)$ is substituted by applying in τ_i the "loess()" function with local weighting parameter reduced to 15%, and the value τ_i is stored; (iii) steps (i) and (ii) are repeated until no new τ_i is found in which the expression value is greater than pv; (iv) a set of oligos for which at least one spike is found, and the list of the corresponding time points τ_i ($i = 1, \ldots, k$ with $k =$ number of detected spikes) is provided as output. This iterative and automated procedure for each method M_i is implemented in R [11].

The lists of oligos and time points provided by the six methods are not necessarily the same, as S_o is differently calculated for each method. Therefore, a more comprehensive list is obtained by merging the six contributes and discarding redundancies in the merged list. This combination is performed in an automatic way.

The final number of affected oligos and of detected time points depends on the threshold value (pv) chosen. Therefore, the spike detection is carried out for different pv values. A small value of pv (with respect to the average signal amplitude) does not allow us to discriminate between simple amplitude fluctuations and abrupt variations, while too large a pv value may miss spikes which could be relevant for the quality control procedure (see as an example Section 4). Therefore, by analyzing the obtained numbers and performing a visual inspection of the correspondingly identified expression profiles, the most appropriate pv value can be chosen (see as an example Table 3, Figures 4.A1 and 4.B1).

3.1.3. Smoothing Procedure. At this point, the list of oligos and time points related to the chosen pv value is passed on to a biologist. If he/she does not assign any biological importance to the detected peaks and valleys, the expressionary time series of the original dataset carrying those spikes are substituted by the corresponding smoothed profiles. These smoothed profiles are obtained by applying the "loess()" function, with local weighting parameter reduced to 30%, to each previously detected time point. Figure 4 presents two examples of expression time series identified by the iterative procedure, performed with pv equal to 2. Expression data before and after the described smoothing are reported therein.

3.2. Evaluation of the Detection Methods

In the case of artifact detection, it is necessary to provide evaluation methods in order to assess the impact that their smoothing-out can have on further analysis. In other words, it is necessary to check if the smoothing procedure has some effect on the results of data analysis. We considered a functional classification with support vector machine and the power spectrum analysis.

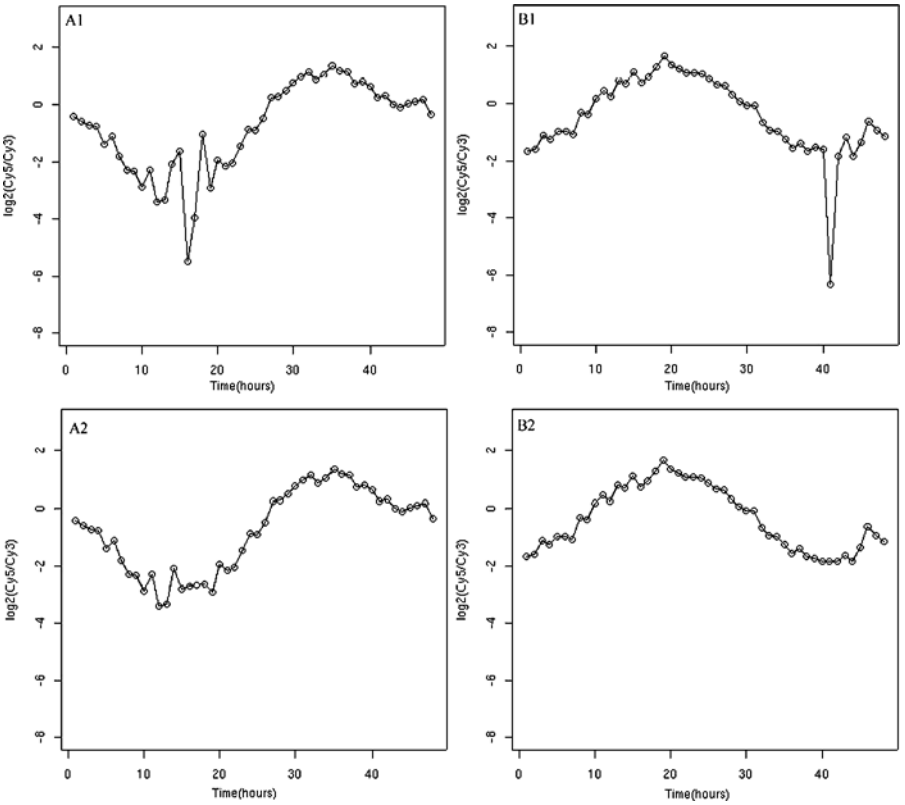

Figure 4. Example of expression time series detected by the iterative procedure, performed with $pv = 2$. Profiles before (A1 and B1 panels) and after (A2 and B2) the described smoothing are reported. Genes reported in A and B are respectively b541 (detected time points: TP_{15}, TP_{16}, TP_{17}, TP_{18}) and f71224_1 (detected time points: TP_{39}, TP_{40}, TP_{41}, TP_{43}).

In the following, we discuss the application of our quality control procedure to the QC_dataset described in [2,4]. The application of the spike detection method to the QC_dataset, and the subsequent evaluation performed by biology experts, led to the smoothing of the original expressionary time series in the time points detected by the methods. Therefore a new dataset, QC_dataset_smooth, was built.

3.2.1. Effect of Spike Smoothing on a MSVM Functional Classification. Support vector machine (SVM) is a state-of-the-art classifier which has been widely used in the analysis of microarray data [7,14,18]. We studied the effect of spike smoothing on a multi-class SVM (MSVM) classifier [13] provided by the package "e1071" of R [11] by considering its influence both on model selection and on functional class prediction. In particular, we adopted

the pair-wise classification approach, where for each possible pair of functional classes an SVM classifier is trained. For N classes, this results in $(N-1)\cdot N/2$ binary classifiers, and the resulting class is chosen by majority voting, i.e. the class with the highest number of votes gives the label. We chose a linear kernel for the MSVM algorithm.

In SVM, model selection the choice of the cost parameter C is required, which sets the trade-off between model complexity and generalization error. Usually, the best cost parameter C is estimated through a cross validation procedure, as in our case a leave-one-out (LOO) cross validation.

For the model selection analysis, as a training set we first used the dataset provided in TableS2 [2,4], hereafter called "raw_dataset". TableS2 describes the known functional classification of 530 genes belonging to the QC_dataset. Afterwards, the second dataset, hereafter called smooth_dataset, in which the same genes are extracted from QC_dataset_smooth, was considered.

We evaluated the effect of the smoothing procedure on model selection, namely the selection of the cost parameter C, by fixing different values of C and computing the LOO accuracy both on the raw_dataset and the smooth_dataset. This analysis aims at verifying if the smoothing procedure changes the LOO accuracy on each C parameter, therefore affecting model selection and thus classification task. Our aim is not at evaluating the predictive accuracy of the model built after smoothing the spikes, for which the small degree of bias (many cases and few parameters) should be instead considered as in [21]. Assuming that the choice of the cost parameter C is performed by selecting the best LOO accuracy, this analysis shows that the smoothing procedure will lead the experimenter to choose a different value of C, and thus different MSVM models for the subsequent analysis.

Afterwards, to assess the impact of the smoothing procedure on functional classification, we trained the MSVM on the raw_dataset and on the smooth_dataset, and used the obtained models to predict the genes without functional annotation in the QC_dataset, and in the QC_dataset_smooth, respectively. Given the raw_dataset and the smooth_dataset, in both cases the parameter C maximizing LOO accuracy was chosen to build the corresponding model. Any two C parameters showing similar LOO accuracy could be used instead. In fact, we do not aim at selecting the two models with highest predictive accuracy; but we want to point out the differences in the functional classification of the two MSVM models due to our smoothing procedure. The idea is to isolate the effects of the smoothing procedure on the functional classification results. However, our choice for the C parameters is that corresponding to the maximum LOO accuracy, thus resembling the standard choice for model selection.

The result shows that even two MSVM models showing similar LOO accuracy classify differently the unknown oligos. We do not know which is the

"right" classification, but we point out that the smoothing procedure plays a great role for the subsequent biological investigations.

3.2.2. Effects on Power Spectrum. We assessed how our smoothing procedure affects the power spectrum used in [2] to select the genes that have a definitely periodic time course. We thus repeated the computational steps therein described to obtain the power spectrum, using the QC_dataset as well as the QC_dataset_smooth, and compared the differences.

4. RESULTS

Table 3 reports, for different values of *pv*, the number of oligos having at least one spike. For each value of *pv*, the list of oligos and time points were obtained as described in Section 3.1 by merging the results obtained by the different six methods M_i and considering only once the time points which are identified highlighted by more than one method.

Table 3 illustrates that the value *pv* equal to 2 sensibly discriminates between irrelevant time variations ($pv < 2$) and too-stringent spike-detection conditions ($pv > 2$). This choice was confirmed by visually inspecting a number of selected expressionary profiles, as those reported in Figures 4.A1 and 4.B1. As reported in Table 3, for $pv = 2$ the automated procedure identified 334 oligos, each presenting abrupt expression variation in at least one time point. Accordingly, a new dataset, "QC_dataset_smooth", was obtained by substituting in the original QC_dataset the 334 transcriptional profiles obtained by our procedure with their smoothed version. For the sake of simplicity, Table 4 only reports those 56 genes with the functional annotation. The complete list is available upon request.

Table 3. Number of oligos with at least one spike detected by the iterative procedure for different values of *pv*

pv	# oligos
1	3305
2	334
3	28
4	8
5	2
6	1
7	0

Table 4. Genes with functional annotation which present at least one detected spike in their expression (see Supplemental table for class acronym definition)

oligo_ID	Class	oligo_ID	Class	Oligo_ID	Class	oligo_ID	Class
a10325_30	ER	f739_1	MI	l1_28	ER	opfblob0060	AM
a10325_32	ER	i10472_1	MI	m14235_3	CT	opfblob0092	MI
a12696_3	MI	i1225_2	MI	m33088_2	AM	opfk12894	ER
a1718_1	DR	i14975_1	MI	m36656_1	MI	opfl0013	AM
b218	MI	i8675_1	AM	m54626_4	CT	opfl0022	AM
b230	MI	j116_7	MI	m60464_2	MI	opfl0029	M
b391	OT	j170_10	MI	n131_10	OT	opfl0141	AM
b444	MI	kn9335_1	DR	n132_124	MI	opfm60467	MI
d49942_9	MI	kn973_2	DR	n132_125	MI	ptrgln	PG
e15509_11	AM	ks1030_4	OT	n134_78	DR	ptrgly	PG
e18550_1	MI	ks26_17	AM	n137_2	CT	z_4_50	MI
e24991_1	MI	ks48_18	ER	n138_34	M	z_4_50	MI
f12313_1	MI	ks510_10	MI	n141_14	MI		
f27464_2	OT	ks510_8	MI	opfb0671	MI		
f49857_1	MI	ks75_15	ER	opfblob0020	ER		

We first assessed the distribution of those time points by computing the histogram reported in Figure 5. We can note that the methods identified more than 100 spikes at time point 18.

In this section we first discuss how the smoothing procedure affects MSVM model selection and functional classification. Consistency considerations are also reported.

Concerning the model selection, Table 5 reports the LOO accuracy values for different values of the cost parameter C.

From Table 5 it is evident that selecting the parameter with maximal LOO accuracy, using the raw_dataset the best parameter should be $C = 0.1$, while using the smooth_dataset the chosen parameter should be $C = 1.0$. Hence, despite of the very few modifications induced by the smoothing procedure (56 out of 530 genes of the training set), two different models should be selected.

We then predicted the functional class of genes without annotation in the QC_dataset as well as in the QC_dataset_smooth, as described in Section 3.2.1. In the confusion matrix we obtained 970 off-diagonal elements (out of 4550), i.e. 970 elements were classified differently by the two classifier obtained by raw_dataset and smooth_dataset. The confusion matrix regarding the prediction of functional expression of unknown genes between the two MSVM models, selected for $C = 0.1$ and $C = 1.0$ respectively, is provided as supplemental material.

Concerning the power spectrum analysis, the smoothing procedure, by eliminating abrupt changes in the signal, removes high frequency components in the Fourier space. Therefore, as expected, the power spectrum shifts towards

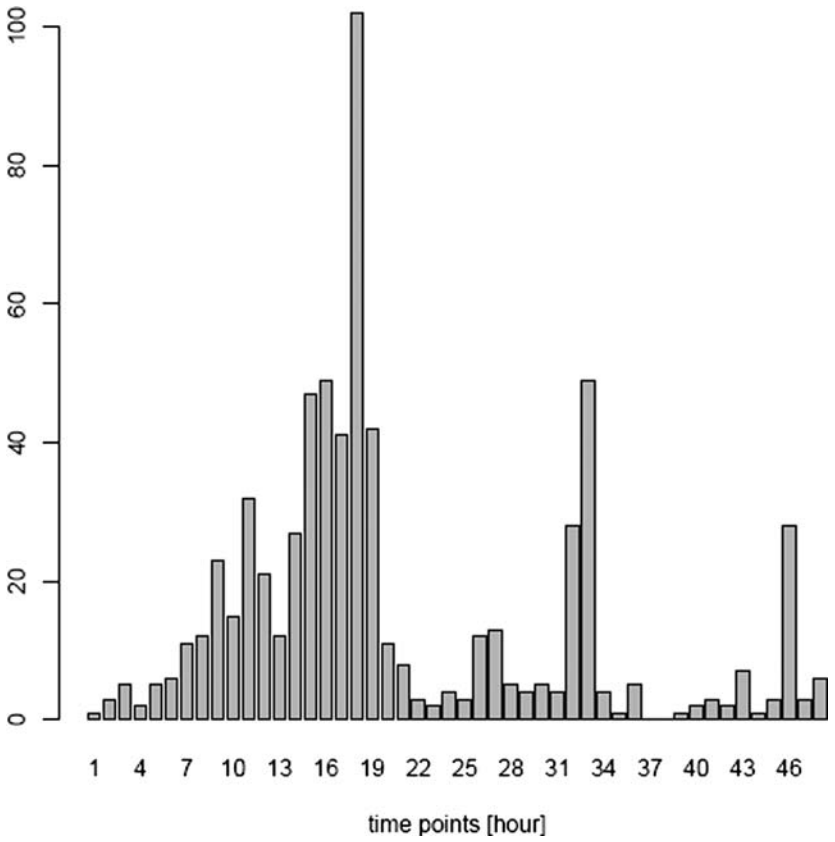

Figure 5. Spike temporal distribution.

Table 5. LOO accuracy values of MSVM for different values of cost parameter *C* using raw_dataset and smooth_dataset

C	LOO accuracy raw	LOO accuracy smoothed
0.001	56.4	56.8
0.01	69.4	69.6
0.1	72.5	71.9
1	72.1	72.5
10	69.2	69.4
100	67.4	67.4
1000	67.0	64.3
10000	64.7	66.8
100000	66.6	66.8

higher percentage. About 50 more genes have a power spectrum greater than 90% in the smoothed dataset. Concerning the cut-off value of 70%, which was used in [2] to select periodic genes, 12 more genes have a power spectrum greater than 70%.

5. DISCUSSION

The study described in this paper can be divided into two conceptually distinct parts. In the first part we perform an automated quality control procedure by detecting anomalously rapid changes in the gene expression time series. Biologists need to assess whether these represent artifacts or are biologically relevant. In the former case, such anomalous rapid changes have to be properly accounted. In the latter case, further biological investigation on those spikes and on the temporal distribution of their positions should be performed.

The detection of these spikes is achieved by exploiting six different simple methods in an automated and iterative way, and then suitably combining their results. The choice of the *pv* parameter permits to control the amplitude and number of detected spikes, therefore allowing the biologist to control the smoothing procedure based on his/her own personal knowledge of the expected dynamics of the temporal series.

In the case of the *P. falciparum* asexual cycle, assuming that these peaks are artifacts, we discuss the effects of their substitution with smoothed values on a popular analysis technique such as supervised functional classification by means of MSVM. The greater number of valleys with respect to peaks seems to indicate that they are artifacts. In fact, in the case of low signals the relative noise is higher, so it seems reasonable to detect more valleys than peaks. We found that removing artifacts detected by our methods affects both the results of the MSVM model selection procedure and the MSVM functional classification of genes without annotation. In the latter case, 970 genes are differently classified before and after the smoothing procedure. It is worth noting that we do not discuss the degree of reliability of either classification. Our aim is to show that our quality control procedure influences data analysis results. It is also worth noting that the smoothing procedure we propose is locally applied only to the temporal points in which artifacts occur. Therefore, it preserves the overall temporal profile. This strengthens the effectiveness of our quality control procedure.

Concerning power spectrum computation, the smoothing procedure confirms and enhances the periodicity of the expression profiles used for subsequent analysis in [2]. This result is consistent with the aim of our quality control procedure at preserving as much as possible signal periodicity. However, though preserving periodicity, our approach may affect functional analysis.

In the temporal distribution of spike positions as reported in Figure 5, the most crowded channel is located at time steps 18. The result of Kolmogorov–Smirnov test performed on this distribution allows us to state with a high level of confidence ($p < 0.0005$) that this spike position distribution does not come from a uniform distribution, suggesting that spikes, if considered artifacts, are not due to random experimental errors. This analysis may suggest to biologists, aware of the performed experimental procedure, the possible causes of artifacts. In this way, improvements of the experimental process could be achieved.

Supplemental table. Confusion matrix regarding the prediction of functional expression of unknown genes between the two MSVM models, selected for $C = 0.1$ and $C = 1.0$, respectively

		MSVM with smooth_dataset $C = 1.0$													
		AM	CT	DR	DS	ER	GP	M	MI	OT	P	PG	RS	TC	TM
	AM	4	0	0	0	0	0	0	0	0	0	0	0	0	0
	CT	0	1393	0	0	7	31	0	0	0	33	0	75	0	27
	DR	0	0	340	7	0	0	50	1	103	0	3	0	21	0
	DS	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	ER	1	27	0	0	181	0	0	0	0	0	0	2	0	4
	GP	0	14	0	0	0	51	0	0	1	29	0	20	0	1
MSVM	M	0	0	9	1	0	0	25	6	0	1	0	0	21	0
with	MI	62	46	7	0	5	1	3	654	4	16	2	1	16	0
raw_dataset	OT	0	10	54	1	0	1	10	0	349	8	0	15	15	0
$C = 0.1$	P	0	36	9	0	0	5	5	9	77	443	2	3	7	0
	PG	0	0	16	1	0	0	5	1	0	0	8	0	1	0
	RS	0	2	0	0	0	5	0	0	6	4	0	129	0	0
	TC	0	0	0	0	0	0	0	0	0	0	0	0	1	0
	TM	0	2	0	0	0	0	0	0	0	0	0	2	0	0

AM = Actin myosin motors, CT = Cytoplasmic Translation machinery, DR = DNA replication, DS = Deoxynucleotide synthesis, ER = Early ring transcripts, GP = Glycolytic pathway, M = Mitochondrial, MI = Merozoite Invasion, OT = Organellar Translation machinery, P = Proteasome, PG = Plastid genome, RS = Ribonucleotide synthesis, TC = TCA cycle, TM = Transcription machinery.

REFERENCES

[1] Bar-Joseph, Z., Analyzing time series gene expression data, *Bioinformatics*, **20**(16) (2004), 2493–2503.

[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.*, **1**(1) (2003 October), 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*: http://www.camda.duke.edu/camda04/datasets (last access 13/06/2005).

[5] Erdal, S., Ozgur, O., Armbruster, D., Ferhatosmanoglu, H., and Ray, W.C., A time series analysis of microarray data, in: *4th IEEE International Symposium on BioInformatics and BioEngineering (BIBE 2004)*, 19–21 March 2004, Taichung, Taiwan, IEEE Computer Society, 2004, ISBN 0-7695-2173-8.

[6] Filkov, V., Skiena, S., and Zhi, J., Analysis techniques for microarray time-series data, *J. Com. Biol.*, **9**(2) (2002), 317–330.

[7] Furey, T.S., Cristianini, N., Duffy, N., Bednarski, D.W., Schummer, M., and Haussler, D., Support vector machine classification and validation of cancer tissue samples using microarray expression data, *Bioinformatics*, **16**(10) (2000), 906–914.

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

[9] GenePix Pro, The image analysis software for microarrays, tissue arrays and cell arrays: http://www.axon.com (last access 13/06/2005).

[10] Griffiths, A.J.F., Gelbart, W.M., Miller, J.H., and Lewontin, R.C., *Modern Genetic Analysis*, W.H. Freeman & Co, New York, 1999.

[11] R Development Core Team, R: *A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2005, http://www.R-project.org (last access 13/06/2005).

[12] Institute for Genomics Research (TIGR), www.tigr.org.

[13] Kreßel, U., Pairwise classification and support vector machine, in: B. Schölkopf, C.J.C. Burges, and A.J. Smola, Eds., *Advances in Kernel Methods–SV Learning*, MIT Press, Cambridge, MA, 1999, pp. 255–268.

[14] Lee, Y. and Lee, C.-K., Classification of multiple cancer types by multicategory support vector machines using gene expression data, *Bioinformatics*, **19**(9) (2003), 1132–1139.

[15] Le Roch, K.G., Zhou, Y., Blair, P.L., Grainger, M., Moch, J.K., Haynes, J.D., De La Vega, P., Holder, A.A., Batalov, S., Carucci, D.J., and Winzeler, E.A., Discovery of gene function by expression profiling of the malaria parasite life cycle, *Science*, **12**(301) (5639) (2003), 1503–1508. Epub 2003 Jul 31.

[16] Molla, M., Waddell, M., Page, D., and Shavlik, J., Using machine learning to design and interpret gene-expression microarrays, *AI Magazine*, **25** (2004), 23–44.

[17] Sebastiani, P., Gussoni, E., Kohane, I.S., and Ramoni, M., Statistical challenges in functional genomics (with discussion), *Statistical Science*, **18** (2003), 33–70.

[18] Simek, K., Fujarewicz, K., Swierniak, A., Kimmel, M., Jarzab, B., Wiench, M., and Rzeszowska, J.J., Using SVD and SVM methods for selection, classification, clustering and modeling of DNA microarray data, *Engineering Application of Artificial Intelligence*, **17**(4) (2004), 417–427.

[19] Suh, K.N., Kain, K.C., and Keystone, J.S., Malaria, *CMAJ*, **170**(11) (2004 May 25), 1693–1702. DOI: 10.1053/cmaj.1030418.
[20] Ward, G., Ed., Monitoring Malaria: Genomic Activity of the Parasite in Human Blood Cells, Public Library of Science, Open-access article, *PLoS Biol.*, **1**(1) (2003), 5–6.
[21] Simon, R.M., Korn, E.L., McShane, L.M., Radmacher, M.D., Wright, G.W., and Zhao, Y., *Design and Analysis of DNA Microarray Investigations*, 1st ed. Springer, 2004.