UNIVERSITY
OF TRENTO

DEPARTMENT OF INFORMATION AND COMMUNICATION TECHNOLOGY

38050 Povo – Trento (Italy), Via Sommarive 14
http://www.dit.unitn.it

SIMPLE METHODS FOR PEAK AND VALLEY DETECTION IN TIME SERIES MICROARRAY DATA

A. Sboner, A. Romanel, A. Malossini, F. Ciocchetta, F. Demichelis, I. Azzini, E. Blanzieri, R. Dell'Anna

March 2005

Technical Report # DIT-05-015

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

A. Sboner(1,2), A. Romanel(2), A. Malossini(2), F. Ciocchetta(1,2), F. Demichelis(1,2), I. Azzini(1), E. Blanzieri(2), R. Dell'Anna(1)*.

(1) Bioinformatics Group, SRA, ITC-Irst, Via Sommarive 18, I-38050 POVO (TN) - Italy

(2)Department of Information and Communication Technology, Trento University, Via Sommarive 14, I-38050 POVO (TN) - Italy

+39 0461 882091

ABSTRACT

Given a set of gene expression time series obtained by a microarray experiment, this work proposes a novel quality control procedure, which exploits some analytical methods enabling the identification of genes which spike expressions within narrow time-windows and over a chosen amplitude threshold. The procedure automatically provides a list of genes and time points in which abrupt variations have been detected. The quality control has to be performed by a biologist, who assesses those spikes as bearing biology relevance or being artifacts. In the latter case spikes have to be substituted by a smoothing procedure. In particular, we focused on transcriptome of *Plasmodium falciparum* Intraerythrocytic Developmental Cycle. Assuming that spikes detected in this

---

*These authors share senior leadership.

set have been labeled as artifacts by a biologist, in the second part of this paper we discuss the effects of smoothing on subsequent different types of analyses.

General Terms

Algorithms, Measurement.

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; 2. the *P. falciparum* has evolved 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 *P. falciparum* biology [3,15], researchers' attention is mainly placed on the poor knowledge about 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) lacks sequence similarity to 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 to increase signal to noise ratio. However, given the experimental design of present dataset, namely a temporal series, it is possible to use this temporal information in order to further detect expression points that could be strongly affected by noise. Abrupt variations in the transcriptional profile can be assessed as artifacts or carrying relevant information. Among abrupt variation we were particularly interested in peaks and valleys, as they preserve signal periodicity, which is an IDC transcriptome characteristic, as highlighted by [2]. Usually the approach to time series analysis [1,5,6] is to approximate them with a continuous interpolating function. However, in this study we chose to preserve the actual information about each single point. In fact, our goal is to identify ORFs that present a relevant variation also with very short duration with respect to the overall length of IDC (48 hours). To achieve this goal, we built different simple methods based on the discrete derivative and integral operators. An additional method directly matches abrupt variations on transcriptional profiles. The panel of methods allows to perform the investigation of candidate genes in an automatic way, avoiding direct visual inspection of all available time series microarray data. The

presence of peaks and valleys in the signal as pointed out by our methods can be judged by the biologist as bearing biology relevance or being an artifact. In the former case further analysis for biological validation are required. In the latter case, peaks and valleys can be ruled out as artifacts that were not detected by conventional quality control procedures. These abrupt variations are therefore removed and substituted by a smoothing procedure.

The first part of the paper is devoted to the description of our procedure for peaks and valleys detection. Assuming that any biological relevance of the abrupt variations is ruled out by a biologist, in the second part of the paper we verify whether the smoothing procedure influences further analyses. Obtained results suggest that our proposed quality control procedure should be used whenever biologist judges detected time variations to be artifacts.

2. PRELIMINARY ANALYSIS

Given the intrinsic complexity of the experiments involving DNA microarray (see for example [10,17]), 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 [12] to analyze these images, and finally, 3. we verified the results of GenePixPro3.0 quality control algorithm. GenePixPro 3.0 [9] is the software used in [2] to 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]. 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. Moreover, this choice allows us to prove that our quality control procedure is really a specific one and therefore its results do not necessary resemble those of a conventional one. 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 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 values in Table 1 were imputed with the loess() function, provided by the `stats` package of R (version 2.01) [11]. Local weighting parameter was reduced to 12%.

Table 1. The data matrix obtained by QC_Dataset

In order to find within the **E** matrix gene expressions with rapid changes over time (in particular candidate peaks and valleys), we used an automated procedure exploiting six different methods (briefly Mi) concisely reported in Table 2. They can be split in three main classes: *Derivative Methods* (M1, M2, M3), *Integral Methods* (M4, M5), and *Other Methods* (M6). For each transcriptional profile, method Mi detects a time point $_i$ in which the maximum expression variation, measured by score So, occurs. The higher the score So the higher the probability to find a significant peak (or valley) with respect to average signal amplitude. The results of each method may confirm or complete the results of the other ones. M1 method proceeds as described in Figure 1: it simply computes (Step

2) for each oligo in **E** the maximum absolute value of the discrete derivative, obtained from Step 1. Similar procedure characterizes method M2. Method M3 characterizes ideal abrupt time variations as those for which expression So reported in Table 2 is zero. Differently from the other five methods, the smaller So the higher is therefore the probability to find a significant peak (or valley).

Figure 1. The derivative method M1.

For method M4 reported in Figure 2 (and similarly for M5) the score So is the fraction of area under a possible peak. Method M6 reported in Figure 3 looks for three-point structures in each gene profile, weighting their possible asymmetry and selecting that structure for which the area is maximum.

Figure 2. The integral method M4.

Figure 3. The method M6

Table 2. The methods used in this work.

Each method Mi is separately and iteratively applied to "clean" transcriptional profiles by filtering abrupt time variations larger than a chosen threshold (PEAK_VALUE). For each method Mi the iterative procedure may be schematized as follows: (i) For each gene expression the time point with maximum variation $_i$ is found; (ii) if the expression variation in $_i$ is larger than PEAK_VALUE, the expression value in $_i$ is substituted by applying the loess() function with local weighting parameter reduced to 15%; (iii) steps (i) and (ii) are repeated until no new $_i$ in which variation is larger than PEAK_VALUE is found. Each method ultimately provides a set of oligos for which at least a spike is found and the list of the corresponding time points..

It is worth noting that iterative procedure for each Mi method is completely automatic, i.e. it does not need any user intervention. The procedure is implemented in R [11].

The numbers of affected oligos and of detected time points depend on the threshold PEAK_VALUE chosen. By analyzing these numbers and performing a visual inspection of the expression profiles the most appropriate threshold value can be chosen, depending on the goal of the analysis. If the goal is quality control, the profiles in the original datasets are substituted by the smoothed profiles. In this paper, for each oligo, distinct time points detected by the methods were substituted by applying the loess() function with local weighting parameter reduced to 30%. Figure 4 presents two example of expression time series pointed out by the iterative procedure, performed with PEAK_VALUE=2. Profiles before and after the described smoothing are reported.

3.2 Evaluation of the detection methods

Let us assume that the detection methods presented above lead to detect artifacts, namely the application of the detection methods to the QC_Dataset leads to the creation of a new dataset QC_Dataset_smooth. It is now necessary to provide evaluation methods in order to assess the impact that the smoothing of the artifacts can have on further analysis. We considered a functional classification with Support Vector Machine and power spectrum.

3.2.1 Effect of spike smoothing on a MSVM functional classification

Support vector machine 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 Multiclass SVM classifier [13] provided by the package e1071 of R [11]. In particular, we adopted the pair wise 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, 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.

We first used the data set provided in TableS2 [2,4], hereinafter called “raw_dataset”. TableS2 describes the known functional classification of 530 genes belonging to the QC_Dataset. Afterwards the second dataset hereinafter called smooth_dataset, in which the same genes are extracted from QC_Dataset_smooth, was considered.

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

The effect of spike smoothing was assessed by considering its influence on the model selection and on the functional class prediction. 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 leave-one-out accuracy on the raw_dataset and smooth_dataset.

Then, as in the model selection procedure, we trained the MSVM on the raw_dataset and on the smooth_dataset and used the obtained models to predict, respectively, the genes without functional annotation in the QC_Dataset and in the QC_Dataset_smooth. Given a dataset, the parameter C maximizing leave-one-out accuracy was chosen to build the corresponding model. We then compared the prediction for the two models on the genes without functional annotation.

3.2.2 Effects on power spectrum

We assessed how our smoothing procedure affects the power spectrum used in [2] to select those 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 the number of oligos having at least one spike for different values of PEAK_VALUE.

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

PEAK_VALUE equals to 2 sensibly discriminate between irrelevant time variations (PEAK_VALUE < 2) and too stringent spike detection conditions (PEAK_VALUE > 2). This choice was confirmed by visually inspecting a number of selected expressionary profiles. Accordingly, a new dataset, "QC_Dataset_smooth", was obtained by substituting in the original QC_Dataset the 334 transcriptional profiles obtained by our iterative procedure with their smoothed version.

As reported in Table 3, methods Mi identified 334 oligos, each presenting abrupt expression variation in at least one time point. For sake of simplicity Table 4 only reports those 56 genes with the functional annotation. The complete list is available upon request (or as supplemental material).

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.

Figure 5. Peak temporal distribution.

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

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

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

From Table 5 it is evident that to always guarantee the maximal LOO accuracy, using the raw_dataset the best parameter should be C=0.1, instead using the smooth_dataset the chosen parameter should be C=1. Hence, despite of the very few modifications induced by the smoothing procedure (56 out of 530 genes of the training set) two different models would be obtained.

To evaluate the differences of these two models in terms of classification we predicted the functional class of genes without annotation in the QC_Dataset as well in the QC_Dataset_smooth as described in Section 3.2.1. Table 6 shows the confusion matrix we obtained. There were 970 off-diagonal elements, i.e. 970 elements were classified differently by the two classifier obtained by raw_dataset and smooth_dataset.

Table6. Confusion matrix regarding the prediction of functional expression of unknown genes between the two M-SVM models, selected respectively for C=0.1 and C=1.0

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, power spectrum shifts towards 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 work described in this paper can be divided into two conceptual distinct parts. In the first part we perform a quality control procedure by detecting in the gene expression time series anomalous rapid changes. Biologists have to assess if they represent artifacts or are instead biologically relevant. In the first case the spikes have to be smoothed. The detection is achieved by exploiting six different simple methods combined in an automatic iterative procedure. The choice of the PEAK_VALUE parameter permits to control the amplitude and number of detected spikes, therefore allowing the biologist to control the possible smoothing basing on his/her own personal knowledge of the expected dynamic of the temporal series. Moreover, our detection procedure preserves the data while smoothing only the points that are considered to be artifacts. In contrast a complete smoothing change the data completely.

In the case of the *P. falciparum* asexual cycle, if this peaks are artifacts we discuss the effects of their removal and substitution with smoothed values on a popular analysis technique such as supervised functional classification by means of MSVM. Evidence in favor of being artifacts is the bigger number of valleys with respect to peaks. In fact in case of low signals the relative noise is higher so it seems to be more reasonable to detect a valley. We found out that removing artifacts detected by our methods has an impact on

both the results of the model selection procedure and the functional classification of genes without annotation. In this last case, 970 genes are differently classified before and after the smoothing procedure.

Concerning power spectrum computation, smoothing result confirms and enhances the periodicity of expression profiles used for subsequent analysis in [2]. This result reflects the aim of our quality control procedure at preserving as much as possible signal periodicity. Therefore, while preserving periodicity, nonetheless our approach may have impact on functional analyses.

In the temporal distribution of spike positions as reported in Figure 5, the most populated channel is located in time steps 18. The result of Kolmogorv-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 one, suggesting that spikes, if considered artifacts, are not due to random experimental errors. This analysis may suggest to biologists, aware of performed experimental procedure, the possible causes of artifacts. In this way, improvements of experimental process could be achieved.

6. REFERENCES

[1] Bar-Joseph, Z., Analyzing time series gene expression data, *Bioinformatics*, 20(16):2493-2503, 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., Ozgur, O., Armbruster, D., Ferhatosmanoglu, H., Ray, W.C., A Time Series Analysis of Microarray Data *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., Zhi, J., Analysis techniques for microarray time-series data. *J Com Biol* 9(2):317-30, 2002.

[7] Furey, T. S., and et al. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10):906–914, 2000.

[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: 498–511*, 2002.

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

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

[11] Ihaka, R., Gentleman, R.. R: A Language for Data Analysis and Graphics. Journal of Computational and Graphical Statistics, 5(3):299-314, 1996.

[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, editors, Advances in Kernel methods–SV Learning*, pages 255-268, Cambridge, MA, 1999, MIT Press.

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

[15] 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*. 12;301(5639):1503-8, 2003. Epub 2003 Jul 31.

[16] 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*).

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

[18] Simek, K., and et al. Using SVD and SVM methods for selection, classification, clustering and modeling of DNA microarray data. Engineering application of Artificial Intelligence, 17(4):417–427, 2004.

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

[20] Ward G. (editor), *Monitoring Malaria: Genomic Activity of the Parasite in Human Blood Cells* Public Library of Science, Open-access article, PLoS Biol. 1(1):5-6, 2003.

APPENDIX

Table 2. 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)

Table 2. The methods 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	Figure 3.

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

Peak_Value	# 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 Table 6 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		

Table 5. LOO accuracy values of model selected 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

Table6. Confusion matrix regarding the prediction of functional expression of unknown genes between the two M-SVM models, selected respectively for C=0.1 and C=1.0

(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).

			M-SVM with smooth_dataset C=1.0												
		AM	CT	DR	DS	ER	GP	M	MI	OT	P	PG	RS	TC	TM
M-SVM with raw_dataset C=0.1	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
	M	0	0	9	1	0	0	25	6	0	1	0	0	21	0
	MI	62	46	7	0	5	1	3	654	4	16	2	1	16	0
	OT	0	10	54	1	0	1	10	0	349	8	0	15	15	0
	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

Given in input Matrix **E**,

Do ∀o,

{ **Step 1.** (Discrete Derivative). ∀ t∈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:

$$\max_t \left|\Delta E(o,t)\right| = S_0$$

}

Figure 1. The derivative method M1.

Given in input Matrix **E**,

Do ∀o,

{ **Step 1.** (Normalization). Compute:

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

Step 2. (Integral). Compute:

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

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:

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

Step 5. (Local Integral). Compute

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

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

}

Figure 2. The integral method M4.

Given in input Matrix **E**,

Do ∀o,

{ **Step 1.** (Spike Value Detection). ∀ t∈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:

$$\frac{\max_{t} SV(E(o,t))}{\sum_{t\in TP} SV(E(o,t))} = S_0$$

}

Figure 3. The method M6

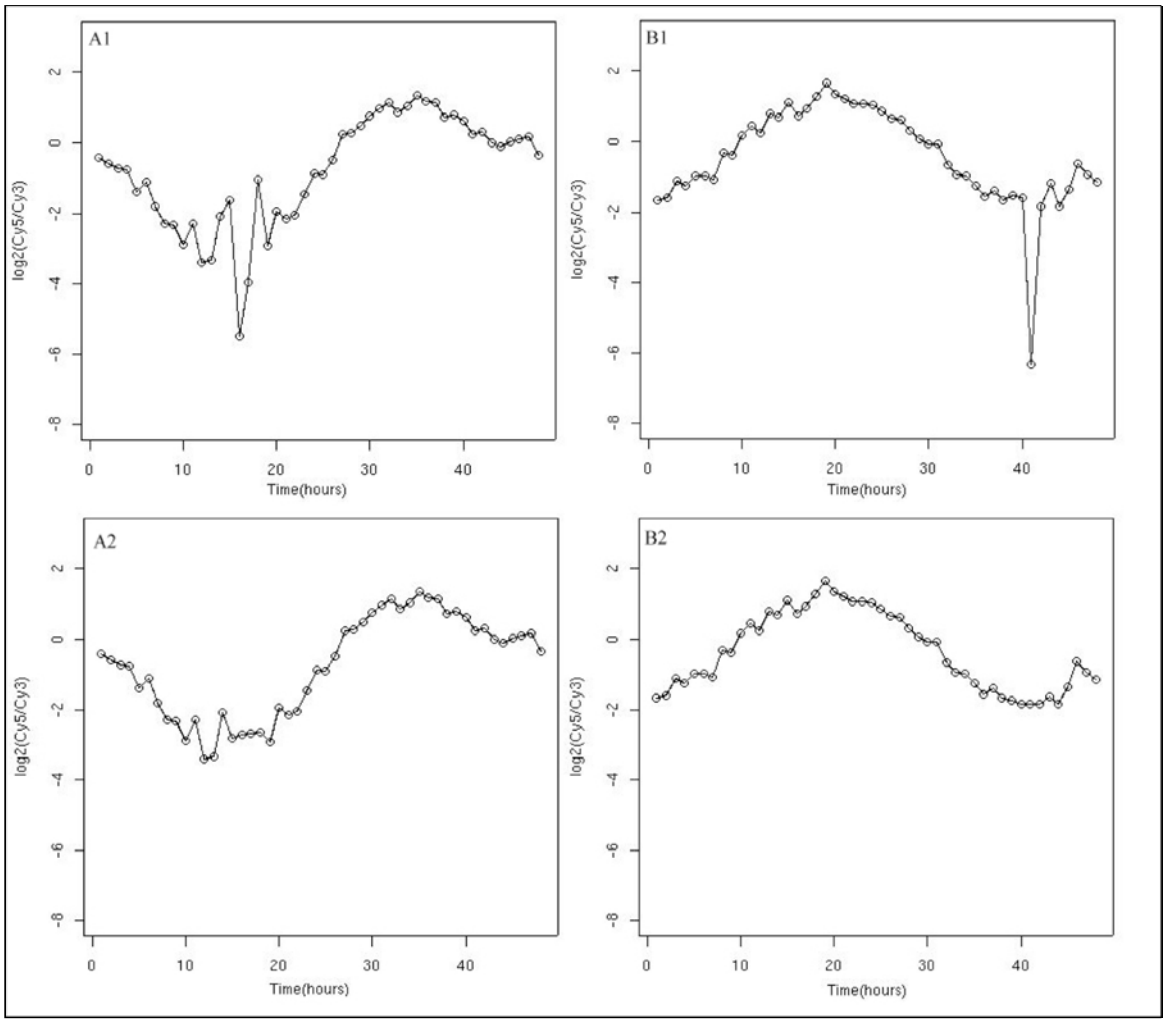

Figure 4. Example of expression time series detected by the iterative procedure, performed with PEAK_VALUE=2. Profiles before (A1 and B1 panels) and after (A2 and B2) the described smoothing are reported. Gene reported in A and B are respectively b541 (TP: 15, 16, 17, 18) and f71224_1 (TP detected =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

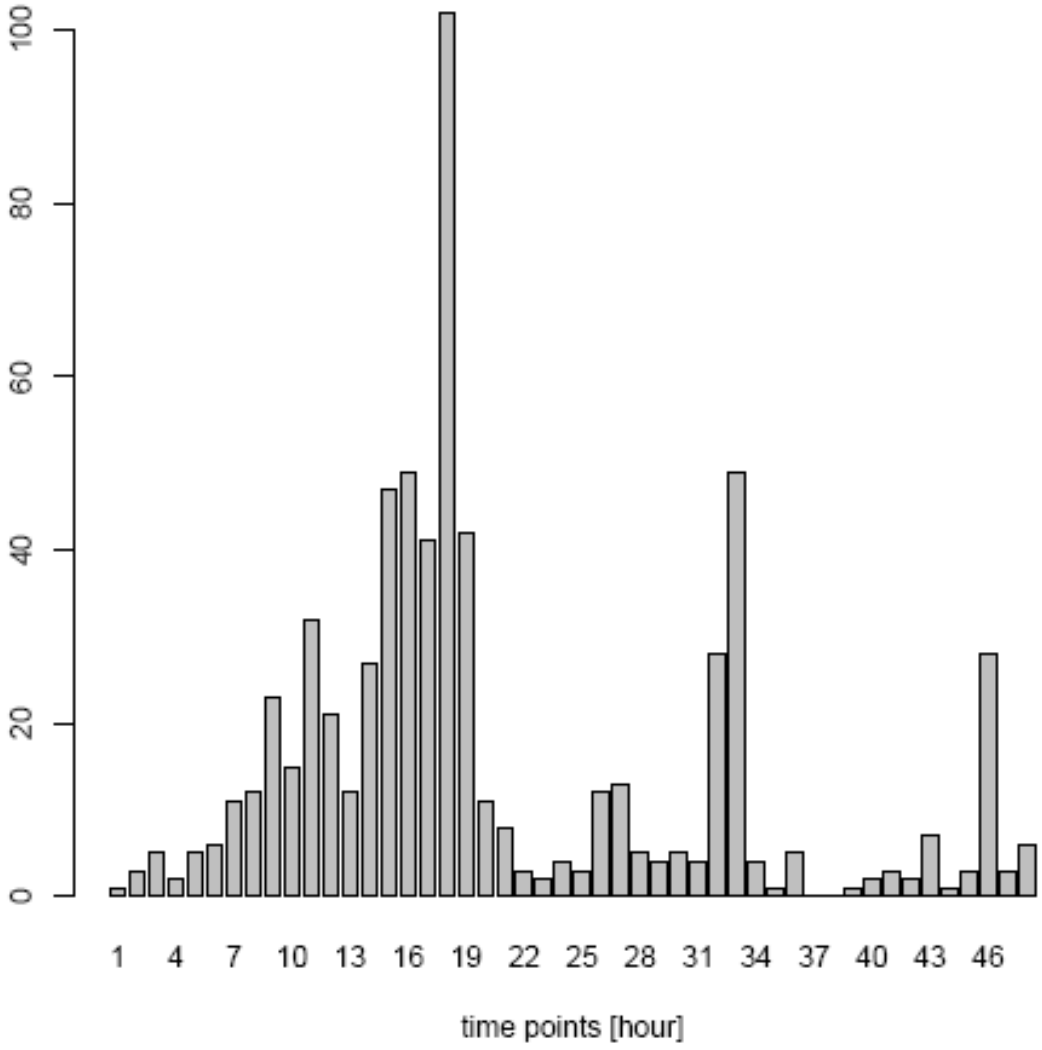

Figure 5. Peak temporal distribution.