Adaptive clustering of time series

Ahlame Douzal Chouakria, Alpha Diallo, Françoise Giroud

Adaptive clustering of time series

A. Douzal Chouakria [1], A. Diallo [1],[2] and F. Giroud [2]

[1] TIMC-IMAG TIMB (CNRS-UMR 5525), Université Joseph Fourier Grenoble 1, F-38706 La Tronche Cedex, France
[2] TIMC-IMAG RFMQ (CNRS-UMR 5525), Université Joseph Fourier Grenoble 1, F-38706 La Tronche Cedex, France

Keywords: Microarray technology, Gene expression data, proximity measure, classification, clustering.

abstract

This paper focuses on the cell division cycle insuring the proliferation of cells and which is drastically aberrant in cancer cells. The aim of this biological problem is the identification of genes characterizing each cell cycle phase. The identification process is commonly based on a prior set of well-characterized cell cycle genes, called reference genes. The expression levels of the studied genes are measured during the cell division cycle. Each studied gene is assigned a cell cycle phase by it's peak similarity to the reference genes. This classical approach suffers of two limitations. On the one hand, the most widely used proximity measures between gene expression profiles are based on the closeness of the values regardless to the similarity with respect to (w.r.t.) the genes expression behavior. On the other hand, many different ill-founded sets of reference genes are proposed in the literature, and biologists are not agree about those of genes best characterizing the observed cell cycle phases. Our aim in this paper is twice. We propose a new dissimilarity index for gene expression profiles to include both proximity measures w.r.t. values and w.r.t. behavior. An adaptive unsupervised classification, based on the proposed dissimilarity index, is then performed to identify the cell cycle phases of the studied genes. Finally, we propose a new set of reference genes, well-assessed by a biological knowledge.

1 Introduction

DNA microarray technology allows to monitor simultaneously the expression levels of thousands of genes during important biological processes and across collections of related experiments. Clustering and classification techniques have proved to be helpful to understand gene function, gene regulation, and cellular processes. Though most cells in our bodies contain the same genes, not all of them intervene in each cell: genes are turned on, or expressed when needed. Such specific genes define the molecular pattern related to a specific function of a cell and in most cases appear as organized in molecular regulation networks. To know how cells achieve such specialization, scientists need to identify which genes each type of cell expresses. Microarray technology now allows us to look at many genes at once and determine which are expressed in a specific cell type, i.e. which transcriptome (set of all mRNA or "transcripts") is characteristic of its particular function (Eisen et al.(1999)). Researchers are using this powerful technology to learn which genes are turned on or off in diseased versus healthy human tissues for example. The genes that are expressed differently in the two tissues may be involved in causing the disease. In this paper we will be interested in the dynamic progression of cell division cycle through the four distinct phases: G_1, S, G_2 and M phases. The expression levels of a set of studied genes are then observed at a specific instants of time during cell division cycle. The identification of the set of genes highly characterizing each cell cycle phase is generally based on a prior set of reference genes. Each studied gene is assigned a cell cycle phase by it's peak similarity to the reference genes. This classical approach suffer of two limitations. On the one hand, the most widely used proximity measures between gene expression

profiles are based on the closeness of the values regardless to the similarity with respect to (w.r.t.) the genes expression behavior. On the other hand, many different ill-founded sets of reference genes are proposed in the literature, and biologists are not agree about those of genes best characterizing the observed cell cycle phases. Our aim is twice. We propose a new dissimilarity index for gene expression profiles to include both proximity measures w.r.t. expressed values and w.r.t. genes expression behavior. An adaptive unsupervised classification, based on the proposed dissimilarity index, is then performed to identify the cell cycle phases of the studied genes. Finally and assessed by a biological knowledge, we propose a new well-justified set of reference genes.
The paper is organized as follows: the next section gives the definition and the properties of the new dissimilarity index. Section 3 presents the human HeLa cell line application, and gives the principal of the proposed adaptive unsupervised classification for cell cycle genes identification. Section 4 performs a comparative analysis and discuss the main obtained results.

2 Proximity measure between genes expression profile

For clustering or classifying a set of gene expression profiles evolving over time, the commonly used proximity measures are the euclidean distance. Let $g_1 = (u_1, ..., u_p)$ and $g_2 = (v_1, ..., v_p)$ be the expression levels of two genes g_1, g_2 observed at the instant of times $(t_1, ..., t_p)$. The Euclidean distance δ_E between g_1 and g_2 is defined as: $\delta_E(g_1, g_2) = \left(\sum_{i=1}^{p}(u_i - v_i)^2\right)^{\frac{1}{2}}$. It stems directly from the above definition that the closeness between two gene expression profiles depends on the closeness of the values regardless to the gene expression behavior. Our aim is to propose a dissimilarity index including both gene expression behavior and values proximity measures. A necessary prior step to the design of such a dissimilarity is to define what we mean about similar gene expression behaviors, and specify the main characteristics that the dissimilarity would measure. We distinguish at least two important characteristics of temporal data. On the one hand, the temporal data where only occurring events, and not their instants of time, are determinant for the proximity evaluation. For instance, in voice processing domain only the occurring syllables are used to identify words; the flow rate being specific to each person. On the other hand, the temporal data where both occurring events and their instants of time are determinant, for instance, ECG, delay response to a treatment, etc. The gene expression data lie within the scope of the latter case.

2.1 Behavior proximity measures

We define the similarity w.r.t gene expression behavior by considering two features. On the one hand, the strength of the monotonicity and, on the other hand, the closeness of the growth rates. Without loss of generality, assume that g_1 and g_2 values lie in $[0, D]$. g_1 and g_2 are similar w.r.t. behavior if at any observed period $[t_i, t_{i+1}]$ they increase or decrease simultaneously (monotonicity), with a growth rate (closeness of growth rates). One can quantify this similarity concept by considering the classical Pearson correlation coefficient, however this correlation leads to an overestimation when taking into account the temporal dependency between measurements. For more discussion about alternative approaches see Douzal Chouakria et al. (2007).
To overcome this problem we propose the following temporal correlation coefficient:

$$\text{CORT}(g_1, g_2) = \frac{\sum_{i=1}^{p-1}(u_{(i+1)} - u_i)(v_{(i+1)} - v_i)}{\sqrt{\sum_{i=1}^{p-1}(u_{(i+1)} - u_i)^2}\sqrt{\sum_{i=1}^{p-1}(v_{(i+1)} - v_i)^2}}$$

where $\text{CORT}(g_1, g_2)$ belongs to the interval $[-1, 1]$. The value $\text{CORT}(g_1, g_2) = 1$ signifies that in any observed period $[t_i, t_{i+1}]$, the genes g_1 and g_2 increase or decrease simultaneously with the same growth rate (similar behavior). The value $\text{CORT}(g_1, g_2) = -1$ means that in any observed period $[t_i, t_{i+1}]$ where g_1 increases, g_2 decreases and vice-versa with a same growth rate (in value; opposite behavior). Finally, $\text{CORT}(g_1, g_2) = 0$ expresses that there is no monotonicity between g_1 and g_2 and their growth rates are stochastically linearly independent (different behaviors). For

more details about temporal correlation see Chouakria Douzal (2003).
Now we present the new dissimilarity index based on the temporal correlation coefficient as a behavior proximity measure.

2.2 Dissimilarity index for gene expression profiles

The aim is to provide a new dissimilarity index which would cover both the euclidean distance for the proximity w.r.t. values and the temporal correlation for the proximity w.r.t. behavior.
Let us first describe the main specifications of the new dissimilarity index. The dissimilarity index should modulate the proximity w.r.t. values according to the proximity w.r.t. behavior. For the same proximity w.r.t. values, the dissimilarity measure should be dependent on the proximity w.r.t. behavior, and for the same proximity w.r.t. behavior, the dissimilarity should depend on the proximity w.r.t. values. The resulting dissimilarity measure should also allow to adjust the weight contribution between both quantities. The modulating function will increase conventional measure when the temporal correlation decreases from 0 to -1. The resultant dissimilarity should approach the conventional measure if the temporal correlation is zero. The modulating function should decrease the conventional measure when the temporal correlation increases from 0 to +1.
According to the specifications above, we propose a dissimilarity index D based on an automatic adaptive tuning function defined as follows:

$$D(g_1, g_2) = f(\text{CORT}(g_1, g_2)) \cdot \delta_E(g_1, g_2)$$

where $f(x)$ is an exponential adaptive tuning function:

$$f(x) = \frac{2}{1 + exp(k\ x)} \quad , \quad k \geq 0$$

with $f(0) = 1$. Figure 1 shows the adaptive tuning effect for several values of $k \geq 0$. In the case

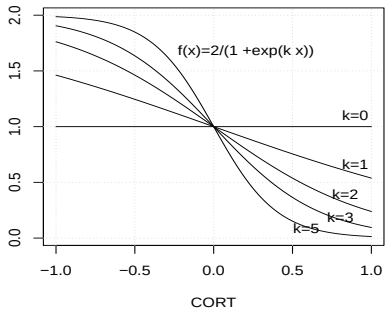

FIGURE 1. The adaptive tuning effect

of genes with different behavior (i.e. with CORT near 0), $f(x)$ is near 1 whatever the value of the weight k and D is approximately equal to δ_E. However, in the case of genes with an opposite or similar behavior (i.e. with $|\text{CORT}|$ near to 1), the parameter k modulates the contributions of the proximity w.r.t. values and w.r.t. behavior to the dissimilarity index D. As k increases, the contribution of the proximity w.r.t. behavior $1 - 2/(1 + exp(k\ |\text{CORT}|))$ increases, whereas the contribution of the proximity w.r.t. values $2/(1 + exp(k\ |\text{CORT}|))$ decreases. For instance, for $k = 0$, the proximity w.r.t. behavior contributes at 0% to D whereas the proximity w.r.t. values contributes at 100% to D (the value of D is totally determined by δ_E). For $k = 2$, the proximity w.r.t. behavior contributes at 76.2% to D whereas the proximity w.r.t. values contributes at 23.8% to D (23.8% of the value of D is determined by δ_E and the remaining 76.2% by CORT). Table 1 summarizes, in the case of similar or opposite behavior ($|\text{CORT}|$=1), the contributions of both, proximity w.r.t. behavior and w.r.t. values, to the dissimilarity index D. Let us add two remarks: First, if $k = 0$ the proposed dissimilarity index D is identical to δ_E; hence D could be considered as an extension of δ_E to both, behavior and value proximity measures. The second point is that if

	Proximity w.r.t. behavior Contribution (%)	Proximity w.r.t. values Contribution (%)
$k = 0$	0	100
$k = 1$	46.2	53.7
$k = 2$	76.2	23.8
$k = 3$	90.5	9.4
$k \geq 5$	$\rightsquigarrow$ 100	$\rightsquigarrow$ 0

TABLE 1. Contribution of the proximity w.r.t. behavior and w.r.t. values to D as a function of k

δ_E approaches 0 (i.e., the genes expression are close w.r.t. values), then CORT approaches 1 (i.e., the genes expression are similar w.r.t. behavior) and D approaches 0. Finally, we can check easily that D verifies the identity and the symmetry properties of a distance, but not the triangular inequality.

3 Adaptive unsupervised classification for identifying of cell cycle regulated genes

3.1 HeLa Cell line Data description

This paper focuses on a specific biological events occurring during cell proliferation, this process insuring the multiplication of cells, which is drastically aberrant in cancer cells. The cell cycle, or cell division cycle, is the series of events between one cell division and the next one. The cell cycle consists of progression along four distinct phases: G_1, S (DNA synthesis or DNA replication), G_2 and M phases. A molecular surveillance system monitors the cell's progress through the cell cycle and checkpoints ensure that a cell divides only when it has completed all of the molecular prerequisites for producing healthy daughter cells. These restriction points mark the inter-phases transitions, G_1/S is the first such transition. The genome-wide program of gene expression during the cell division cycle aims to determine the genes well expressed during studied cell cycles (Spellman et al (1998), Oliva et al. (2005) and Cho et al. (2001)). Looking at the transcriptome of proliferating synchronized cells leads to the construction of gene expression profiles along time, i.e. during cell cycle progress. This application is concerned with the analysis of experimental transcriptomic data from the human Hela cell line published in Whitfield et al. (2002) (http://genome-www.stanford.edu/Human-CellCycle/Hela/). Our study will focus on the 1099 genes, recorded in the third experimentation of the HeLa application. Genes are described by their expressing levels, during the cell-cycle progression, along 48 instants of times after cell synchronization.

3.2 Conventional identification of cell cycle genes

Let's illustrate the approach proposed by Whitfield et al. (2002) to identify the cell cycle genes of the HeLa application. Authors consider a set of 20 reference genes characterizing the following 5 cell cycle phases and transitions: S, G_1/S, G_2, G_2/M, and M/G_1. The set of 20 genes is composed of 5 classes of 4 reference genes per phase (table 2). Figure 4 gives, for each cell cycle phase, the expression profiles of the 4 reference genes. Authors argued the selection of the 20 reference genes by their peaks expression in each cell cycle phase. Each of the 1099 studied genes is then assigned a cell cycle phase of the most similar 4 reference genes. The used similarity is based on the expression values regardless to the gene expression behavior. If we observe in detail the 20 gene expression profiles given in figure 2 we find some contradictions. First, the reference genes CDC2, CCNF, CCNA2 characterizing the G_2 phase don't peak at G_2 but at G_2/M. Similarly, the G_2/M reference genes BUB1 and PLK peak at M/G_1 instead at G_2/M. These observations are supported by the annotations of Genecards database (http://www.genecards.org/) and KEGG molecular pathway database (http://www.genome.ad.jp/kegg/kegg2.html).

Phase	G_1/S	S	G_2	G_2/M	M/G_1
Name	CCNE1,E2F1 CDC6,PCNA	RFC4,DHFR RRM2, RAD51	CDC2, TOP2A CCNF, CCNA2	STK15,BUB1 CCNB1, PLK	PTTG1, RAD21 VEGFC, CDKN3

TABLE 2. The 20 reference genes.

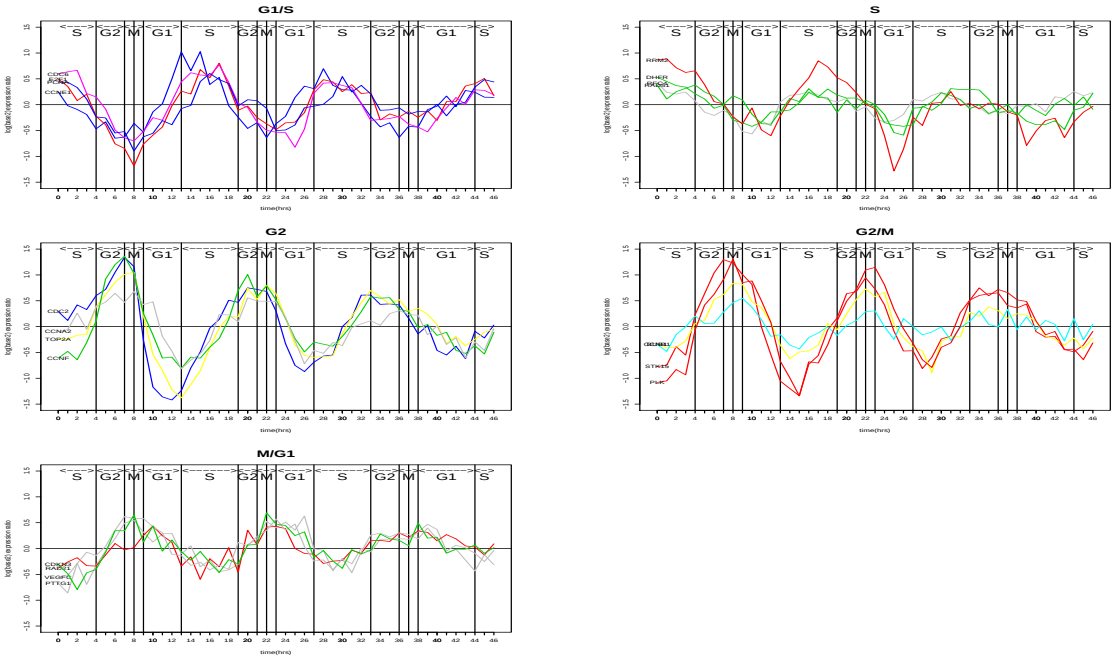

FIGURE 2. Gene expression profiles for the 20 reference genes whose expression peaks in each phase of the cell cycle : G_1/S, S, G_2, G_2/M and M/G_1. The double arrowed lines delimit the time duration for each cell cycle phase : G_1, S, G_2 and M.

3.3 Adaptive unsupervised classification for identifying genes cell cycle phases

Our purpose is first to identify the cell cycle phases of the studied genes; then to determine the set of genes well characterizing each cell cycle phase. For this, we propose to use the classical partitioning around medoids (PAM) method through an adaptive approach based on the new dissimilarity index. The idea of PAM method is that in order to classify objects into nb clusters, it selects nb objects called representative objects, the clusters are then obtained by assigning each remaining object to the nearest representative object. The clue is that the representative objects must be selected so that they minimize the average dissimilarity to all the other objects of the same cluster. In this work, PAM method is preferred to the familiar k-means approach for mainly two reasons. First, it is more robust with respect to outliers, which are numerous in genes expression data due to measurement errors. Secondly, it allows a more detailed analysis of the partition by providing clustering characteristics and a graphical display (a so-called silhouette plot). In particular, the extraction of the set of genes well characterizing each cluster (i.e. each cell cycle phase) is based on the silhouette width of each classified object indicating whether an object is well classified, misclassified or lies on the boundary of a cluster. The quality of a partition is estimated by the average of the silhouette width for all the classified objects. For more details about PAM see Kaufman and Rousseeuw (1990). The main idea of the adaptive approach is to perform, for several values of k (k=0,...,6 per a lag of 0.01), a PAM (number of clusters=5, for the 5 cell cycle phases in interest) method on the whole 1099 genes based on the dissimilarity index D_k to look for the best value of k maximizing the average silhouette width of the obtained partition. Let P_{k*} (k*=5.9) be such a partition and figure 3 the associated silhouette plot. In the literature, there are 43 genes (about 10 genes per phase) identified as involving in the cell cycle division process (Whitfield et al. (2002)); therefor we extract from each cluster of P_{k*} a

kernel set of the 10 well-classified genes maximizing the silhouette width. Table 3 gives for each cluster the set of kernel genes (Gene Type = K). We indicate for each kernel gene it's name, it's Whitfield assignment phase, the number of the neighbor cluster and it's silhouette width (sw) indicating if it is well-classified (sw close to 1) or misclassified (sw close to -1). We indicate also the set of Whitfield reference genes (table 2) belonging to each cluster (Gene Type = R). Figure 4, visualizes for each cluster the expression profiles of the associated kernel genes. The observation of the progression of the kernel genes during the cell division cycle reveals that: kernel genes of the cluster 1 peak clearly at S phase, kernel genes of cluster 2 peak at G_1/S, kernel genes of cluster 3 peak at G_2/M phase, kernel genes of cluster 4 peak at M/G_1 and finally, kernel genes of the cluster 5 peak at G_1 phase. Let's remark that due to the asynchronization of cells, it's more reliable that interpretations will be limited to the earlier cell cycles. According to that, each cluster is assigned the cell cycle phase of its kernel set genes (indicated in the last column of the table 3). Finally, each of the 1099 studied genes is then assigned the cell cycle phase of the cluster it belongs to.

Cluster Number	Gene Name	Whitfield Assignment	Gene Type	Neighbor Cluster	Silhouette width (sw)	High peaked phase
	Homo	S	K	2	0.806	
	KIAA0855	S	K	3	0.697	
	KIAA1598	S	K	2	0.688	
	KIAA0855	S	K	2	0.686	
	KIAA0855	S	K	3	0.681	
	SHC1	S	K	2	0.677	
1	AA452872	S	K	3	0.674	S
	ESTs	S	K	3	0.665	
	KIAA0841	S	K	2	0.658	
	**ESTs	S	K	3	0.635	
	RRM2	S	R	2	0.586	
	DHFR	S	R	2	0.315	
	RAD51	S	R	3	0.238	
	E2F1*	G_1/S	K	1	0.832	
	ORC1L	G_1/S	K	1	0.829	
	SERPINB3	G_1/S	K	1	0.82	
	ESTs	G_1/S	K	1	0.812	
	MCM6	G_1/S	K	1	0.812	
	RAMP	G_1/S	K	1	0.812	
	LOC51218	G_1/S	K	1	0.802	
2	ESTs	G_1/S	K	1	0.794	G_1/S
	ESTs	G_1/S	K	1	0.794	
	CCNE1	G_1/S	K/R	5	0.786	
	E2F1	G_1/S	R	1	0.775	
	CDC6	G_1/S	R	1	0.682	
	PCNA	G_1/S	R	1	0.625	
	RFC4	S	R	1	0.526	
	CASP3	G_2	K	4	0.811	
	CDKN1B	G_2	K	4	0.807	
	WISP1	G_2	K	4	0.799	
	UBE2C	G_2	K	4	0.788	
	CKS1	G_2	K	4	0.784	
	T56726	G_2	K	4	0.779	
	FLJ11029	G_2	K	1	0.779	
	UBE2C	G_2	K	4	0.779	
	HMG$_2$	G_2	K	4	0.768	
3	FZR1	G_2	K	4	0.765	G_2/M
	CCNF	G_2	R	4	0.757	
	TOP2A	G_2	R	4	0.669	
	CDC2	G_2	R	1	0.618	
	STK15	G_2/M	R	4	0.478	
	CCNA2	G_2	R	4	0.458	
	FLJ13154	M/G_1	K	3	0.737	
	PCF11	M/G_1	K	5	0.717	
	AA705332	G_2/M	K	5	0.695	
	FLJ10461	G_2/M	K	3	0.651	
	CNAP1	G_2/M	K	3	0.599	
	NR3C1	G_2	K	3	0.593	
	MRPL19	M/G_1	K	3	0.585	
	HMGCR	M/G_1	K	3	0.579	
4	ZPBP	M/G_1	K	3	0.578	M/G_1
	IDN3	G_2	K	3	0.576	
	RAD21	M/G_1	R	3	0.433	
	CDKN3	M/G_1	R	3	0.320	
	PTTG$_1$	M/G_1	R	5	0.282	
	BUB1	G_2/M	R	3	0.184	
	VEGFC	M/G_1	R	3	0.148	
	CCNB1	G_2/M	R	3	0.095	
	PLK	G_2/M	R	3	0.003	
	RAB3A	M/G_1	K	2	0.561	
	H2BFQ	M/G_1	K	2	0.502	
	HMGE	M/G_1	K	4	0.489	
	IFIT1	M/G_1	K	2	0.484	
5	BAIAP2	G_1/S	K	2	0.478	G_1
	FLJ23053	G_1/S	K	2	0.475	
	ESTs	M/G_1	K	4	0.429	
	ESTs	G_1/S	K	2	0.407	
	SSP29	G_2/M	K	4	0.398	
	TOP1	M/G_1	K	4	0.394	

TABLE 3. The 50 Kernel genes (Gene Type = K) well-characterized the cell cycle phases: S, G_1/S, G_2/M, M/G_1 and G_1, with the classification of the 20 Whitfield reference genes (Gene Type = R) through the 5 obtained clusters.

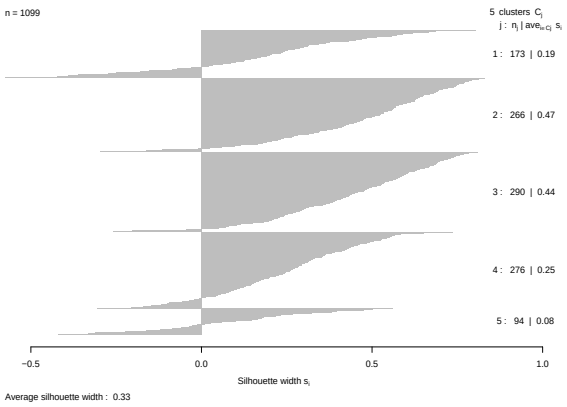

FIGURE 3. Silhouette width of $P_{k*=5.9}$

4 Comparative analysis and discussion

Let's first discuss the PAM obtained results. The optimal partition P_{k*} maximizing the average silhouette is obtained for k*=5.9. This value means that the 5 main patterns of gene expression profiles are distinctive essentially through their behaviors (table 1). Figure 3, reveals an average silhouette (sw) of 0.33 which indicates that the clustering structure is no better than reasonable. However, if limited to the 50 kernel genes, the average silhouette coefficient is about 0.67, which means that the kernel sets are well separated from each other. Figure 3 indicates that the cluster 2 (G_1/S) possesses the largest sw of 0.47, which means that this cluster is well separated from the other clusters, whereas the cluster 5 (G_1) possesses a rather very narrow sw of 0.08, which means that the cluster G_1 is not very clearly separated from the other clusters.

According to the biological knowledge, note that the well studied genes CCNE1, CCNA2, and CCNB1 known as mitotic cyclins, respectively classed into the clusters G_1/S, G_2/M and M/G_1, appear in the expected biological temporal order during the cell division cycle (G_1, S, G_2 and M). As a support to the obtained G_1/S cluster note that E2F1, a transcription factor known as a key regulator of cell cycle progression involved in the control of cell cycle progression from G_1 to S, exhibits the largest silhouette (sw=0.832). The genes CCNE1 (sw=0.786) and MCM6 (sw=0.812) known as activated and induced respectively by E2F1, are also classified into the G_1/S cluster. It's shown experimentally that the CCNA2 gene promotes G_2/M transition.

Our proposed approach classified well CCNA2 into the cluster G_2/M (sw=0.458), whereas it is selected as a G_2 Whitfield reference gene. The gene UBE2C (sw=0.779) belonging to the G_2/M cluster is well assessed by the biological knowledge: it represents an enzyme of the ubiquitin pathway regulating destruction of mitotic cyclins near the end of mitosis (G_2/M transition). Finally, note all Whitfield reference genes labeled as G_2 phase are classified into the cluster G_2/M and except STK15, all the reference genes labeled as G_2/M are classified into the cluster M/G_1, which clearly corroborate the contradictions discussed in the paragraph 3.2.

5 Conclusion

This paper focuses on an alternative adaptive clustering for the identification of the cell cycle genes. We propose a new dissimilarity index for gene expression profiles to include both proximity measures w.r.t. expressed values and w.r.t. genes expression behavior. An adaptive unsupervised classification, based on the proposed dissimilarity index, is then performed to identify the cell cycle phases of the studied genes. Finally, we propose a new well-justified set of reference genes, assessed by a published biological knowledge.

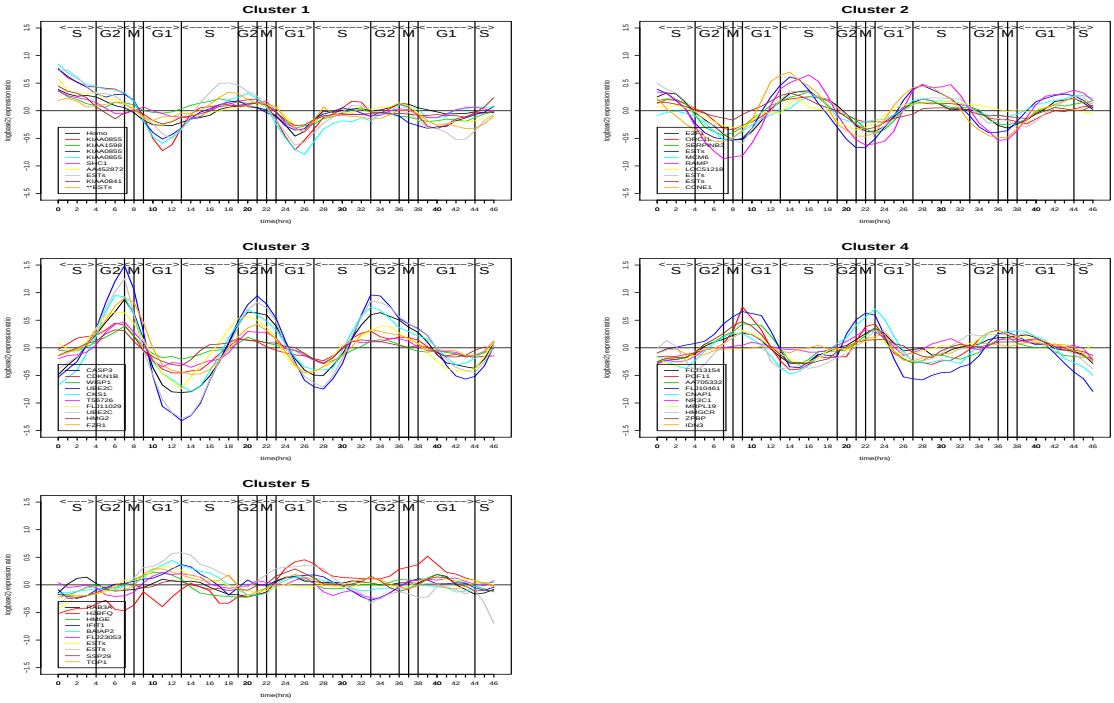

FIGURE 4. Kernel Gene expression profiles during cell division cycle

References

Cho, R.J., Huang, M., Campbell, M.J., Dong, H., Steinmetz, L., Sapinoso, L., Hampton, G. , Elledge, S.J., Davis, R.W., and Lockhart, D.J. (2001). Transcriptional regulation and function during the human cell cycle. *Nature Genetics*, Vol. **27**(1):48–54.

Chouakria Douzal A., (2003). Compression Technique Preserving Correlations of a Multivariate Temporal Sequence. In: M.R. Berthold, H-J Lenz, E. Bradley, R. Kruse, C. Borgelt (eds.) Advances in Intelligent Data Analysis, V, 566-577, Springer, ISBN: 3-540-40813-4.

Douzal Chouakria, A., Nagabhushan, P.N. (2007). Adaptive dissimilarity index for measuring time series proximity. *Advances in Data Analysis and Classification Journal*. Springer.

Eisen, M.B. , and Brown, P.O. (1999). DNA arrays for analysis of gene expression. *Methods Enzymol*. Vol. **303**, 179–205.

Kaufman L., and Rousseeuw P.J. (1990). Finding Groups in Data. An Introduction to Cluster Analysis. John Wiley & Sons, New York.

Oliva, A., Rosebrock, A., Ferrezuelo, F., Pyne, S., Chen, H., Skiena, S., Futcher, B., and Leatherwood, J. (2005). The cell cycle-regulated genes of Schizosaccharomyces pombe. *PLoS Biol*, Vol.**3**(7):e225.

Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell*, Vol. **9**, 3273–3297.

Whitfield, M.L., Sherlock, G., Murray, J.I., Ball, C.A., Alexander, K.A., Matese, J.C. , Perou, C.M., Hurt, M.M., Brown, P.O., Botstein, D. (2002). Identification of Genes Periodically Expressed in the Human Cell Cycle and Their Expression in Tumors Molecular. *Biology of the Cell*, Vol. **13**, 1977–2000.