



HAL
open science

Adaptive dissimilarity index for Gene Expression Profiles Classification

Ahlame Douzal Chouakria, Alpha Diallo, Françoise Giroud

► **To cite this version:**

Ahlame Douzal Chouakria, Alpha Diallo, Françoise Giroud. Adaptive dissimilarity index for Gene Expression Profiles Classification. Brito, P., Bertrand, P., Cucumel, G., De Carvalho, F. Selected Contributions in Data Analysis and Classification, Springer Berlin Heidelberg, pp.483-494, 2007, Studies in Classification, Data Analysis, and Knowledge Organization. hal-00360479

HAL Id: hal-00360479

<https://hal.science/hal-00360479>

Submitted on 11 Feb 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Adaptive dissimilarity index for gene expression profiles classification

Ahlame Douzal Chouakria¹, Alpha Diallo², and Francoise Giroud³

¹ TIMC-IMAG TIMB (CNRS UMR 5525), Université Joseph Fourier Grenoble 1, F-38706 LA TRONCHE Cedex, France, *Ahlame.Douzal@imag.fr*

² TIMC-IMAG RFMQ (CNRS UMR 5525), Université Joseph Fourier Grenoble 1, F-38706 LA TRONCHE Cedex, France, *Alpha.Diallo@imag.fr*

³ TIMC-IMAG RFMQ (CNRS UMR 5525), Université Joseph Fourier Grenoble 1, F-38706 LA TRONCHE Cedex, France, *Francoise.Giroud@imag.fr*

Abstract. 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. However the conventional proximity measures between genes expression data, used for clustering or classification purpose, do not fit gene expression specifications as they are based on the closeness of the expression magnitudes regardless of the overall gene expression profile (shape). We propose in this paper an adaptive dissimilarity index which would cover both values and behavior proximity. The effectiveness of the adaptive dissimilarity index is illustrated through a classification process for identification of genes cell cycle phases.

1 Introduction to microarray technology

Though most cells in our bodies contain the same genes, not all of the genes are used in each cell. Some 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 a molecular regulation network. To know how cells achieve such specialization, scientists need a way 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 particular cell type (Eisen and Brown (1999)). DNA molecules representing many genes are placed in discrete spots regularly organized in a line/column matrix on a microscope slide. This is called a DNA microarray. Thousands of individual genes (clones) can be spotted on a single square inch slide surface. Next, total messenger RNA (the working copies of genes within cells, indicators of which genes are being used) is purified from cells. The RNA molecules are then labeled by attaching a fluorescent dye and spread over the DNA dots on the microarray. Due to a phenomenon termed base-pairing, RNA will stick to the gene it came from (this is the hybridization process). After washing away all of the unstuck RNA, we can look at the microarray under a microscope and see which RNA remains stuck to the DNA spots. Fluorescent mea-

measurements are performed using specific scanners and related spot fluorescent values are extracted from images (<http://genomewww.stanford.edu/Human-CellCycle/HeLa/>). Since we know which gene each spot represents, we can determine which genes are turned on in the cells. Some 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 other experiments time-course DNA microarray analysis are necessary to determine temporal genomic expression profiles relative to the dynamic progression of a specific biological process or to response at stimulation or treatment. In this paper we will be interested in the dynamic progression of cell division cycle. Additionally, in order to take in account systematic biases in the measured expression levels related to experimental factors, two-channel array experiments are usually performed. It consists in using a reference material in parallel to the tested material. For example: normal cells used as references versus pathological ones being the tested cells. Both materials are labeled using two different colors (green and red) and are mixed in equal proportion prior to hybridization. The final expression measured is given as $\log(\text{base}2)\text{ratio}$ between the tested material against the reference one.

The purpose of clustering or classification tasks is to determine co-expressed genes which indicate co-function and co-regulation. Because different genes are usually functionally implied in a same regulation network, users of microarrays data may not only be interested in clustering or classifying genes, but also be interested in the relationship between these clusters (e.g. which clusters are most close to each other), and the relationship between the genes within the same cluster (e.g. which gene can be considered as the representative of the cluster and which ones are at the boundary area of the cluster).

2 Proximity measure between genes expression data

For clustering or classifying a set of genes expression profiles evolving over time, the commonly used proximity measures are the euclidean distance or the person's correlation coefficient. Let $g_1 = (u_1, \dots, u_p)$ and $g_2 = (v_1, \dots, v_p)$ be the expressions levels of two genes g_1, g_2 observed at the instant of times (t_1, \dots, t_p) . On the one hand, 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 expression profiles depends on the closeness of the values observed at corresponding points of time. δ_E ignores the information of interdependence among the observed values. However, for genes expression data, the overall shapes of gene expression patterns are of greater interest than the individual magnitudes at corresponding instants of time.

2.1 Shape proximity measures

The alternate conventional measure to estimate the similarity between gene expression shapes is Pearson's coefficient correlation (called classical correlation). Unfortunately, we will illustrate in the following that the classical correlation do not score well for proximity between shapes either. For shape proximity measure, we propose the temporal correlation coefficient introduced in Chouakria Douzal (2003), Chouakria Douzal and Nagabhushan (2006) and defined as follows:

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

The temporal correlation coefficient $\text{CORT} \in [-1, 1]$ presents an interesting property, it allows to estimate the linear dependency between the growths of two gene expression profiles, observed at corresponding times. A value of $\text{CORT} = 1$ means that the growths (positive or negative) observed on both expression profiles, at any corresponding instant of time, are similar in direction and rate (similar behavior). On the contrary a value of -1 means that the growths observed on both expression profiles, at any corresponding instant of time, are similar in rate but opposite in direction (opposite behavior). Finally, a value of 0 expresses that the growths observed on both expression patterns are stochastically linearly independent (different behaviors).

2.2 Adaptive dissimilarity index for gene expression proximity

Our aim is to provide a new dissimilarity index model D which would cover both proximity on values $\delta_E(g_1, g_2)$ and on behavior $\text{cort}(g_1, g_2)$. The model would allow to adjust the weights of behavior (shape) or values components. The proposed model is based on an adaptive tuning function which modulates the proximity on values according to the proximity on behavior. The modulating function will increase the proximity on values if the proximity on behavior (i.e the temporal correlation) decreases from 0 to -1. The resultant dissimilarity D approaches the proximity on values if the temporal correlation is zero (different behaviors). Finally, the modulating function will decrease the proximity on values if the proximity on behavior (i.e temporal correlation) increases from 0 to +1. The formulation to compute the resultant dissimilarity index D is:

$$D(S_1, S_2) = f(\text{cort}(S_1, S_2)) \cdot \delta_E(S_1, S_2)$$

where $f(x)$ is an exponential adaptive tuning function fitting the above properties:

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

Figure 1 shows the adaptive tuning function effect for several values of k . The parameter k defines the weights, in the dissimilarity index D , of both behavior and values components as summarized in the Table 1. For instance for $k=5$, $D \rightsquigarrow 2 \delta_E$ when $cort \rightsquigarrow -1$ and decreases until $D \rightsquigarrow 0$ when $cort \rightsquigarrow 1$, finally when $cort \rightsquigarrow 0$ $D \rightsquigarrow \delta_E$. Figure 1 illustrates that higher is the value of k , higher will be the temporal correlation weight and lower will be δ_E weight.

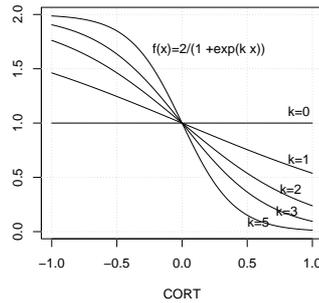


Fig. 1. The adaptive tuning effect

| | Behavior weight (%) | Values weight (%) |
|------------|--------------------------|------------------------|
| $k=0$ | 0% | 100 % |
| $k=1$ | 50% | 50% |
| $k=2$ | 80% | 20% |
| $k=3$ | 90% | 10% |
| $k \geq 5$ | $\rightsquigarrow 100\%$ | $\rightsquigarrow 0\%$ |

Table 1. Behavior ($cort$) and Values (δ_E) weights according to k

3 Classification for genes expression profiles

We propose to compare the adaptive dissimilarity index with the classical correlation through a genes classification (assignment) approach. For the genes classification purpose, we define first two conventionally used genes assignment approaches: a supervised and an unsupervised approaches. Two genes samples are considered: a learning sample based on a set of well-studied genes, and a test sample based on a set of published genes compiled from the literature. Let's give briefly the algorithmic details of these assignment approaches, first in the case of the classical correlation Cor as a genes proximity measure, then in the case of the adaptive dissimilarity index D .

3.1 Supervised and unsupervised assignment approaches based on the classical correlation Cor

The supervised assignment approach based on the classical correlation noted ($SupAss - Cor$) consists to assign each gene to the most similar prior class (Average-Link, Centroid-Link,...) of the well-studied genes. The assessment step consists to evaluate the rand index between the obtained and the prior classes of the published genes. The unsupervised assignment approach based on $1 - Cor$ noted ($UnsupAss - cor$) consists first to perform an hierarchical clustering of the whole genes to classify, then each obtained cluster is assigned to the most similar prior class as detailed in the following:

```

1 Begin UnsupAss-Cor
  % assignment part %
2 - Perform an Hierarchical clustering (Average-Link) of the whole genes
3 - Extract the Nb clusters partition,
4 - Estimate the proximity between each obtained cluster
5   and the Nb prior classes of the well-studied genes,
6 - Assign each cluster to the most similar class,
7 - Assign each gene to the cluster's class it belongs in.
  %assesement part%
8 - Evaluate the rand index between the obtained and prior classes
9   of the published genes
10 End

```

3.2 Supervised and unsupervised assignment approaches based on the adaptive dissimilarity index D

The main idea of the assignment approach based on the adaptive D , is to learn the weights of both values and behavior components of D to fit best the prior partition of the well-studied genes. Let's give the algorithmic steps of the supervised assignment approaches based on D and noted $SupAss - D$.

```

1 Begin SupAss-D based on the adaptive D
  % assignment part %
2 - For each value of k from 0 to 6 per 0.1  %(61 values)
3   - Assign each gene to the most similar class
4     (Average-Link, Centroid-link,...)
5   - Evaluate the rand index between the obtained
6     and the prior classes of the well-studied genes.
7 - End For
8 - Let k* be the value of k maximizing the rand index
9   and Pk* the corresponding obtained assignments
  % assesement part%
10 - Evaluate the rand index between the obtained and prior
11   classes of the published genes
12 End

```

Similarly the unsupervised assignment approach based on D noted $UnSupAss - D$ algorithmic steps are:

```

1 Begin UnsupAss-D
  % assignment part %
2 - For each value ok k from 0 to 6 per 0.1  %(61 values)
3   - Perform an Hierarchical clustering (Average-Link) of the whole genes
4   - Extract the Nb clusters partition,
5   - Estimate the proximity between each obtained cluster
6     and the Nb prior classes of the well-studied genes,

```

```

7     - Assign each cluster to the most similar class,
8     - Assign each gene to the cluster's class it belongs in.
9 - End For
10 - Let k* be the value of k maximizing the obtained rand
11   index and Pk* the corresponding obtained assignments
    %assessement part%
12 - Evaluate the rand index between the obtained and prior classes
13   of the published genes 19 End

```

4 Application and results

4.1 Data description

In this paper we will focus on the specific biological events occurring during cell proliferation, this process insuring the multiplication or reproduction of cells and 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: G1 phase, S phase (DNA synthesis or DNA replication), G2 phase and M phase. A molecular surveillance system monitors the cell's progress through the cell cycle and checkpoints help to 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 transition from one phase to another : the transition from G1 to S phase is the first such transition (G1/S). According to that, we will focus on the G1/S, S, G2, G2/M and M/G1 phases and transitions we will short cut named "cell cycle phases" in the text. The genome-wide program of gene expression during the cell division cycle has been investigated in a wide range of organisms Spellman et al. (1998), Cho et al. (2001), Oliva et al. (2005), using DNA microarrays. In this paper we will focus on a set of genes expression data recorded in the third experimentation of Whitfield et al. published data Whitfield et al. (2002) (<http://genome-www.stanford.edu/Human-CellCycle/Hela/>). The dataset describes 1099 genes, periodically expressed in the human cell cycle. RNA was isolated from HeLa cells et 1 hour intervals after release from a synchronous arrest in S phase. Two lists of genes are considered respectively for learning and assessment steps. On the one hand a list of 20 well-studied genes composed of 4 referenced genes for each of the 5 phases is used for learning step (Table 2, Figure 4). On the other hand, and for assessment step, a list of 39 genes was compiled Whitfield et al. (2002) from the literature that had been shown to be cell cycle regulated by traditional bio-molecular methods (Table 5).

| Phase | G1/S | S | G2 | G2/M | M/G1 |
|-------|-------------------------|--------------------------|----------------------------|--------------------------|------------------------------|
| Name | CCNE1,E2F1 CDC6,PCNA | RFC4,DHFR RRM2, RAD51 | CDC2, TOP2A CCNF, CCNA2 | STK15,BUB1 CCNB1, PLK | PTTG1, RAD21 VEGFC, CDKN3 |

Table 2. List of the 20 genes assigned in Whitfield et al. (2002) to the 5 cell cycle phases.

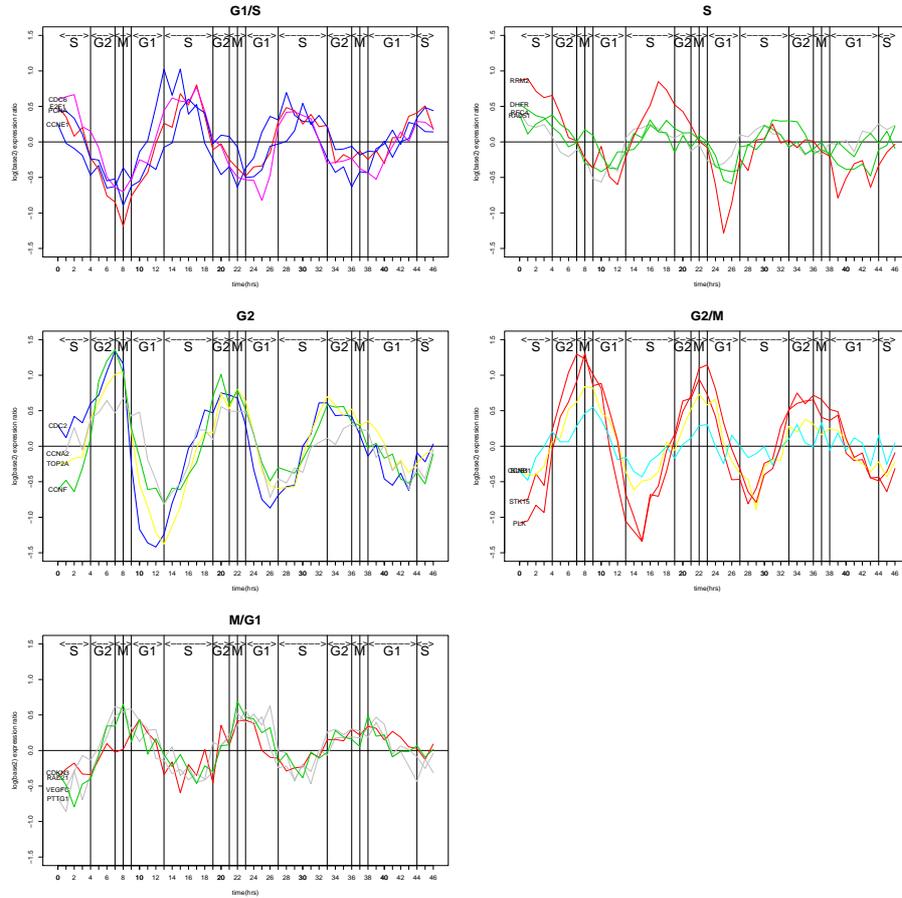


Fig. 2. Gene expression profiles for the 20 well-characterized cell cycle genes whose expression peaks in each phase of the cell cycle : G1/S, S, G2, G2/M and M/G1. The double arrowed lines delimit the time duration for each cell cycle phase : G1, S, G2 and M.

4.2 Identification of genes cell cycle phases results

To illustrate the efficiency of the adaptive dissimilarity index D against the classical correlation, we compare their effectiveness to identify the cell cycle phases of the 39 published genes, through the supervised and unsupervised assignment approaches. On the one hand, we have performed a supervised assignment *SupAss* for centroid and average link. We have then compared the obtained assignments when the supervised approach is based on the classical correlation and on the adaptive D . The obtained results of the assignments of the 39 published genes and the corresponding corrected rand index are reported in the Table 5 at the columns 3-4 for average-link and 5-6 for

centroid-link. On the other hand, we have performed the unsupervised assignment approach *UnsupAss* based respectively on the classical correlation and on the proposed dissimilarity index D . First an hierarchical clustering is performed on the whole 1099 genes based respectively on $1 - Cor$ and Dk . A 5 clusters partition is then extracted. For each extracted cluster we estimate it's dissimilarity to each of the well-referenced phases. The dissimilarity values between the 5 clusters and the 5 phases are reported in the Tables 3 and 4. The obtained dendrograms illustrating the 5 obtained clusters and the identified cell cycle phases are given in the Figure 3. Each gene is then assigned to the cluster's phase it belongs in. The assignments of the 39 published genes obtained through *UnsupAss - Cor* and *UnsupAss - Dk* are

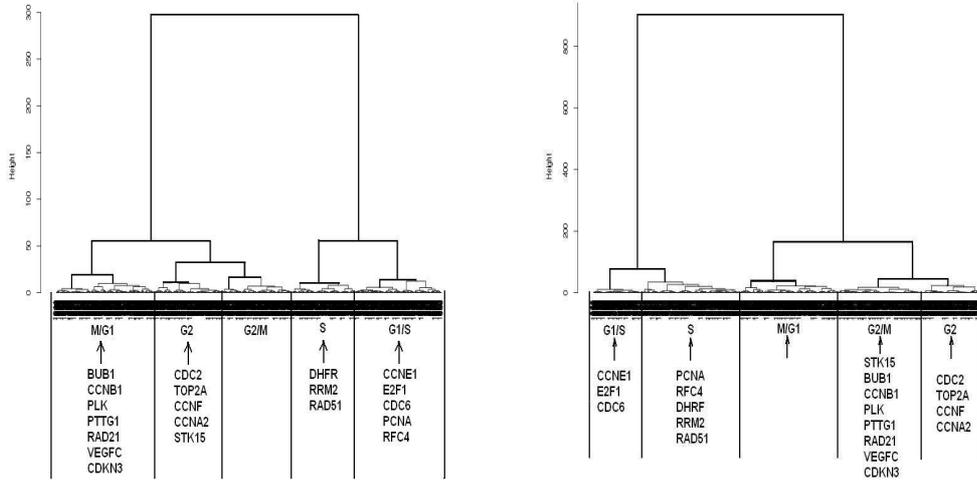


Fig. 3. The unsupervised approach: the dendrograms of the 1099 genes and there phases identification based on Cor(left) and D(right).

| | G1/S | S | G2 | G2/M | M/G1 |
|-----------|--------------|--------------|--------------|--------------|--------------|
| Cluster 1 | 0.755 | 0.416 | 0.806 | 1.236 | 1.461 |
| Cluster 2 | 0.404 | 0.632 | 1.314 | 1.589 | 1.512 |
| Cluster 3 | 1.461 | 0.976 | 0.345 | 0.451 | 0.663 |
| Cluster 4 | 1.500 | 1.056 | 0.540 | 0.475 | 0.694 |
| Cluster 5 | 1.494 | 1.411 | 0.737 | 0.457 | 0.426 |

Table 3. Unsupervised approach based on Cor: similarity between the 5 extracted clusters and the 5 well-referenced phases

| | G1/S | S | G2 | G2/M | M/G1 |
|-----------|--------------|--------------|--------------|--------------|--------------|
| Cluster 1 | 1.520 | 0.762 | 3.284 | 5.016 | 4.123 |
| Cluster 2 | 0.502 | 1.180 | 6.374 | 7.238 | 4.936 |
| Cluster 3 | 5.709 | 2.194 | 0.761 | 0.907 | 1.161 |
| Cluster 4 | 6.688 | 3.565 | 0.989 | 0.464 | 0.521 |
| Cluster 5 | 4.264 | 3.025 | 2.873 | 1.755 | 1.158 |

Table 4. Unsupervised approach based on D_{k^*} ($k^*=3.9$): similarity between the 5 extracted clusters and the 5 well-referenced phases

| Name | Published Phase | Supervised Average-Link | | Supervised Centroid-Link | | UnSupervised | |
|------------|-----------------|-------------------------|-----------|--------------------------|-----------|--------------|-----------|
| | | COR | D_{k^*} | COR | D_{k^*} | COR | D_{k^*} |
| E2F5 | G1 | G2/M | M/G1 | G2/M | M/G1 | G2/M | G2/M |
| CCNE1 | G1/S | G1/S | G1/S | G1/S | G1/S | G1/S | G1/S |
| CCNE2 | G1/S | G1/S | G1/S | G1/S | G1/S | G1/S | G1/S |
| CDC25A | G1/S | G1/S | G1/S | G1/S | G1/S | G1/S | G1/S |
| CDC6 | G1/S | G1/S | G1/S | G1/S | G1/S | G1/S | G1/S |
| CDKN3 | G1/S | M/G1 | M/G1 | M/G1 | M/G1 | M/G1 | G2/M |
| E2F1 | G1/S | G1/S | G1/S | G1/S | G1/S | G1/S | G1/S |
| MCM2 | G1/S | G1/S | G1/S | G1/S | G1/S | G1/S | S |
| MCM6 | G1/S | G1/S | G1/S | G1/S | G1/S | G1/S | G1/S |
| NPAT | G1/S | G1/S | G1/S | G1/S | G1/S | G1/S | G1/S |
| PCNA | G1/S | G1/S | G1/S | G1/S | G1/S | G1/S | S |
| SLBP | G1/S | G1/S | G1/S | G1/S | G1/S | G1/S | G1/S |
| BRCA1 | S | S | S | S | S | G1/S | S |
| CDKN2C | S | G2 | S | G2 | S | G2 | G2 |
| DHFR | S | S | S | S | S | S | S |
| MSH2 | S | G1/S | S | G1/S | S | G1/S | S |
| NASP | S | G1/S | G1/S | G1/S | G1/S | G1/S | G1/S |
| RRM1 | S | S | S | S | S | S | S |
| RRM2 | S | S | S | S | S | S | S |
| TYMS | S | S | S | S | S | S | S |
| CCNA2 | G2 | G2 | G2 | G2 | G2 | G2 | G2 |
| CCNF | G2 | G2 | G2 | G2 | G2 | G2 | G2 |
| CENPF | G2 | G2/M | G2/M | G2/M | G2/M | M/G1 | G2/M |
| TOP2A | G2 | G2 | G2 | G2 | G2 | G2 | G2 |
| BIRC5 | G2/M | G2/M | M/G1 | G2/M | M/G1 | M/G1 | G2/M |
| BUB1 | G2/M | G2/M | G2/M | G2/M | G2/M | M/G1 | G2/M |
| BUB1B | G2/M | G2/M | G2/M | G2/M | G2/M | G2/M | G2/M |
| CCNB1 | G2/M | G2/M | M/G1 | G2/M | M/G1 | M/G1 | G2/M |
| CCNB2 | G2/M | G2/M | M/G1 | G2/M | G2/M | M/G1 | G2/M |
| CDC2 | G2/M | G2 | G2 | G2 | G2 | G2 | G2 |
| CDC20 | G2/M | G2/M | M/G1 | G2/M | M/G1 | M/G1 | G2/M |
| CDC25B | G2/M | G2/M | M/G1 | G2/M | M/G1 | M/G1 | G2/M |
| CDC25C | G2/M | G2 | M/G1 | G2 | M/G1 | G2/M | G2 |
| CDKN2D | G2/M | M/G1 | M/G1 | G2/M | M/G1 | M/G1 | M/G1 |
| CENPA | G2/M | G2 | G2 | G2/M | G2 | G2 | G2/M |
| CKS1 | G2/M | G2 | G2 | G2 | G2 | G2 | G2 |
| CKS2 | G2/M | G2/M | G2/M | G2/M | G2/M | M/G1 | G2/M |
| PLK | G2/M | G2/M | G2/M | G2/M | G2/M | M/G1 | G2/M |
| STK15 | G2/M | G2/M | G2/M | G2/M | G2/M | G2 | G2/M |
| Rand Index | | 0.760 | 0.830 | 0.790 | 0.818 | 0.757 | 0.771 |

Table 5. The assignment cell cycle phases of the 39 published genes

5 Discussion and future scope

5.1 Comparative analysis

Let's note that the assignments obtained in Whitfield et al. (2002) corresponds to the centroid-linkage $SupAss-Cor$ (3rd column). We can first show, that whatever is the considered variant of the supervised approach, the rand index of $SupAss-D$ is greater than the one obtained through $SupAss-Cor$, as illustrated at the last row of the Table 5. Hence, the genes cell cycle

phases of the 39 published genes are better identified through the adaptive dissimilarity index D than through the classical correlation. Through the both assignment approaches $UnsupAss - Cor$ and $UnsupAss - Dk$, each cluster is assigned to a distinguish phases. However the 20 referenced genes are not well distributed through the 5 extracted clusters. Indeed, through $UnsupAss - Cor$, 7 referenced genes 4 from M/G1 and 3 from G2/M are merged in a same cluster labeled as M/G1, with one cluster including no referenced genes (Figure 3 on left) and labeled as G2/M. A nearly similar distribution is obtained through $UnsupAss - D$, 8 referenced genes 4 from M/G1 and 4 from G2/M are merged in a same cluster labeled as G2/M, with one cluster including non referenced genes (Figure 3 on right) and labeled as M/G1. Finally, all the obtained assignment results show that whatever is the assignment approach (supervised or unsupervised) the identification of the genes cell cycle phases is better through the adaptive dissimilarity index than through th classical correlation.

5.2 The unsupervised classification: a promising tool for better understanding of dynamic cell cycle events

Considering the actual fast progression in the acquisition of new biological data, mainly due to recent developments in high throughput experimental methods (such as DNA microarrays), biological concepts and knowledge are undergoing drastic and rapid evolution. Keeping this in mind it appears quite reasonable to expect some invaluable assistance from unsupervised classification methods rather than supervised ones to help in understanding the complexity of life. The results obtained in this specific study, dedicated to better understanding cell cycle progression and regulation, bring some support to such an expectation. For example, considering results obtained by the unsupervised classification associated to D (Fig.5 right) it's possible to draw-back three interesting and encouraging remarks. Note first, the classification process lead to the 5 expected cell cycle phases, then the PCNA gene which has been chosen by Whitfield et al. (2002) as representative of G1/S phase has been classified by the $UnsupASS - D$ approach in the S phase. And effectively it's quite well established that PCNA is a DNA polymerase expressed at the highest levels in the S-phase. Indeed if PCNA is first expressed in mid-G1, PCNA expression peaks in S phase and continues to be weakly expressed in G2 and M phases of the cell cycle. Finally, among the four misclassified M/G1 genes as G2/M by the $UnsupASS - D$ approach we will just discuss, as an example, on the PTTG1 gene. It has been recently demonstrated, by classical molecular biology methods, that the PTTG1 expression peaks at the S-G2 transition and declined thereafter Vlotides et al. (2006). According to that, it makes sense to work out PTTG1 gene as classified in the G2/M cluster rather than in the M/G1 one. On the basis of all these encouraging remarks our future works will focus on the biological processes related to the

different genes obtained in the 5 clusters including new biological knowledge and the genes implication in regulation cell cycle molecular network.

6 Conclusion

This paper focuses on a new application domain of the microarrays and genes expression profile analysis. We introduce the microarrays technology, discuss main challenges of genes expression profile analysis and the great need of clustering and classification techniques. For genes expression profile classification, we propose an adaptive dissimilarity index which would cover both values and behavior proximity. We show it's effectiveness for genes identification cell cycle phases, whatever is the considered assignment approach.

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* 27(1), 48-54.
- CHOUAKRIA DOUZAL, A. (2003): Compression technique preserving correlations of a multivariate temporal sequence. In: Berthold M R, Lenz H J, Bradley E, Kruse R, Borgelt C (eds): *Advances in Intelligent Data Analysis*. Springer, Berlin Heidelberg, 566-577.
- CHOUAKRIA DOUZAL, A. and NAGABHUSHAN, P. (2006): Improved Fréchet Distance for Time Series. In: V. Batagelj, H.-H. Bock, A. Ferligoj, A. Ziberna (Eds): *Data Science and Classification*, Springer, 29-38.
- EISEN, M. B. and BROWN, P. O. (1999): DNA arrays for analysis of gene expression. *Methods Enzymol* 303, 179-205.
- JAVIER, A., BATA-CSORGO, Z., ELLIS, C.N., KANG, S., VOORHEES, J.J. and COOPER, K.D. (1997): Rapamycin (Sirolimus) Inhibits Proliferating Cell Nuclear Antigen Expression and Blocks Cell Cycle in the G1 Phase in Human Keratinocyte Stem Cells. *J. Clin. Invest* 99(9), 2094-2099.
- 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*, 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* 9, 3273-3297.
- VLOTIDES, G., CRUZ-SOTO, M., RUBINEK, T., EIGLER, T., AUERNHAMMER, C.J. and MELMED, S. (2006): Mechanisms for growth factor-induced Pituitary Tumor Transforming Gene-1 (PTTG1) expression in pituitary folliculostellate TtT/GF cells, *Molecular Endocrinology*. First published ahead of print September 7, as doi:10.1210/me.2006-0280.
- 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.

and BOTSTEIN, D. (2002): Identification of Genes Periodically Expressed in the Human Cell Cycle and Their Expression in Tumors Molecular. *Biology of the Cell* 13, 1977-2000.