Which Distance for the Identification and the Differentiation of cell-cycle Expressed Genes ?

Alpha Diallo, Ahlame Douzal-Chouakria, Francoise Giroud

Laboratory TIMC-IMAG, CNRS UMR 5525,
Faculté de Médecine, 38706 LA TRONCHE Cedex, France
Université Joseph Fourier Grenoble 1, France
Lab. TIMC-IMAG, CNRS UMR 5525
Alpha.Diallo, Ahlame.Douzal, Francoise.Giroud@imag.fr

Abstract. This paper addresses the clustering and classification of active genes during the process of cell division. Cell division ensures the proliferation of cells, but becomes drastically aberrant in cancer cells. The studied genes are described by their expression profiles (i.e. time series) during the cell division cycle. This work focuses on evaluating the efficiency of four major metrics for clustering and classifying gene expression profiles. The study is based on a random-periods model for the expression of cell-cycle genes. The model accounts for the observed attenuation in cycle amplitude or duration, variations in the initial amplitude, and drift in the expression profiles.
Keywords: Time series, distance, clustering, classification, gene expression profiles

1 Introduction

Though most cells in our bodies contain the same genes, not all genes are active in every cell: genes are turned on (i.e. expressed) when needed. Expressed genes define the molecular pattern of a specific cell's function, and are organized into molecular-level regulation networks. To understand how cells achieve such specialization, it is necessary to identify which genes are involved in different types of cells. Moreover, it is helpful to know which genes are turned on or off in diseased versus healthy human tissues, and which genes are expressed differently in the two tissues, thus possibly causing the disease. DNA microarray technology allows us to monitor the expression levels of thousands of genes simultaneously during important biological processes to determine which ones are expressed in a specific cell type [7]. Clustering and classification techniques have proven helpful in understanding gene function, gene regulation, and cellular processes (e.g., [10], [17], [20], [21]). We distinguish at least two main approaches to clustering and classifying profiles or time series. First, the parametric approach consists of projecting time series into a given functional basis space, which corresponds to a polynomial ARIMA or a discrete Fourier transform approximation of the time series. Time series clustering and classification is then performed on the fitted coefficients (e.g., [2], [3], [8], [11], [18], [22]). The second approach

is non-parametric, and consists of clustering or classifying time series based on their initial temporal descriptions. Thus, the challenge in this approach is how to include information on dependency between measurements (e.g., [1], [9], [13], [19], [23], [6]). Within the context of the non-parametric approach, we propose to evaluate the efficiency of four major metrics for the clustering and classification of gene expression. This study is based on a random-periods model for the expression of cell-cycle genes. The model accounts for observed biological variations, such as attenuation in cycle amplitude, drift in the expression profiles, and variations in the initial amplitude or the cycle duration. The remainder of the paper is organized as follows. The next section clarifies what constitutes gene expression data and introduces the biological problem of interest. Section 3 defines the four major metrics to be evaluated, and discusses their specifications. Section 4 indicates how the metrics will be compared through the clustering and the classification of genes. Section 5 presents the overall methods of evaluation based on a random-periods model, and discusses the results obtained.

2 Identification of genes expressed in the cell cycle

The biological problem of interest is the analysis of the progression of gene expression during the cell division process. Cell division is the main process in cell proliferation, and it consists of four main phases (G_1, S, G_2, and M) and three inter-phases (the G_1/S, G_2/M, and M/G_1 transition phases) (Figure **??**). The division process begins at the G_1 phase, during which the cell prepares for duplication (DNA pre-synthesis). Then comes the S phase, during which DNA is replicated (i.e., each chromosome is duplicated); this is followed by the G_2 phase, during which the cell prepares for cell division (DNA post-synthesis). Finally comes the mitosis phase, which is also called the M phase, during which the cell is divided into two daughter cells. During these four phases, genes are turned on and off at specific times, so one important aim in understanding cell proliferation is to identify those genes that are highly expressed in, and characteristic of, each phase of the cell cycle. This can help, for instance, to understand how hormonal treatment can induce cell proliferation by activating specific genes. To better our understanding of gene expression during the cell division process, DNA molecules representing many genes are placed in discrete spots organized in a line or column matrix, which is called a DNA microarray. Microarray technology allows us to determine which gene is represented by each spot, and to measure its expression level at specific points in the cell division cycle. Finally, each gene of interest is analyzed for its expression profile observed during one or more cell division cycles.

3 Proximity between gene expression profiles

Let $g_1 = (u_1, ..., u_p)$ and $g_2 = (v_1, ..., v_p)$ be the expressions of two genes observed at time $(t_1, ..., t_p)$. The clustering and classification of gene expression data commonly involve Euclidean distance or the Pearson correlation coefficient.

The following section defines four major metrics for gene expression analysis and their specifications in accounting for proximity in values or behavior.

3.1 Euclidean distance

The Euclidean distance δ_E between g_1 and g_2 is defined as:

$$\delta_E(g_1, g_2) = \left(\sum_{i=1}^{p} (u_i - v_i)^2 \right)^{\frac{1}{2}}.$$

Based on the above definition, the closeness between two genes depends on the closeness of their values, regardless of their expression behavior. In other words, the Euclidean distance ignores the temporal dependence of the data.

3.2 Pearson correlation coefficient

Many works use the Pearson correlation coefficient as a behavior proximity measure. Without loss of generality, consider that g_1 and g_2 have values in $[0, N]$. The genes g_1 and g_2 exhibit similar behavior if over any observed period $[t_i, t_{i+1}]$, they increase or decrease simultaneously at the same rate. In contrast, g_1 and g_2 have opposite behavior if over any observed period $[t_i, t_{i+1}]$ where g_1 increases, g_2 decreases, and vice-versa, at the same rates (in absolute value). To illustrate the correlation coefficient specification, let us consider the following formula, based on the differences between the expression values:

$$\textsc{Cor}(g_1, g_2) = \frac{\sum_{i,i'} (u_i - u_{i'})(v_i - v_{i'})}{\sqrt{\sum_{i,i'} (u_i - u_{i'})^2} \sqrt{\sum_{i,i'} (v_i - v_{i'})^2}}.$$

We see that the correlation coefficient is based on the differences between all pairs of values (i.e. observed at all the pairs of time (i, i')), which implicitly assumes the independence of the observed data. Consequently, the correlation coefficient can overestimate behavior proximity. For instance, in the case of a high tendency effect, as shown in Section 4, two genes with opposite behavior may have a relatively high, positive correlation coefficient.

3.3 Temporal correlation coefficient: a behavior proximity measure

To overcome the limitations of the Pearson correlation coefficient, the temporal correlation coefficient introduced in [4] is considered, as it reduces the Pearson correlation coefficient to the first order differences:

$$\textsc{cort}(g_1, g_2) = \frac{\sum_i (u_{(i+1)} - u_i)(v_{(i+1)} - v_i)}{\sqrt{\sum_i (u_{(i+1)} - u_i)^2} \sqrt{\sum_i (v_{(i+1)} - v_i)^2}}.$$

, with $\textsc{cort}(g_1, g_2) \in [-1, 1]$. The value $\textsc{cort}(g_1, g_2) = 1$ indicates that g_1 and g_2 exhibit similar behavior. The value $\textsc{cort}(g_1, g_2) = -1$ indicates that g_1 and g_2 exhibit opposite behavior. Finally, $\textsc{cort}(g_1, g_2) = 0$ expresses that the growth rates g_1 and g_2 are stochastically linearly independent, thereby identifying genes with different behavior that are neither similar nor opposite.

3.4 Behavior and values proximity measure

For a proximity measure to cover both behavior and value proximities, the dissimilarity index D_k proposed in [5] is considered. It includes both the Euclidean distance, for proximity with respect to values, and the temporal correlation, for proximity with respect to behavior:

$$D_k(g_1, g_2) = f(\textsc{cort}(g_1, g_2))\ \delta_E(g_1, g_2), \quad \text{with } f(x) = \frac{2}{1 + exp(k\ x)}, \quad k \geq 0.$$

This index is based on a tuning function $f(x)$ that modulates the proximity with respect to values according to the proximity with respect to behavior. An exponential function $f(x)$ is preferred to a linear form to ensure a nearly equal modulating effect for extreme values (i.e., CORT=-1, +1 and 0) and their nearest neighbors. In the case of genes with different behavior (i.e., with CORT near 0), $f(x)$ is near 1 whenever the value of k, and D_k is approximately equal to δ_E. However, if CORT $\neq$ 0 (that is, non-different behavior), the parameter k modulates the contributions of both types of proximity, with respect to values and with respect to behavior, to the dissimilarity index D_k. As k increases, the contribution of proximity with respect to behavior, $1 - 2/(1 + exp(k\ |\textsc{cort}|))$, increases, whereas the contribution of proximity with respect to values, $2/(1 + exp(k\ |\textsc{cort}|))$, decreases. For instance, for $k = 0$ and $|\textsc{cort}| = 1$ (similar or opposite behavior), the behavior proximity contributes 0% to D_k whereas the value proximity contributes 100% to D_k (the value of D_k is totally determined by δ_E). For $k = 2$ and $|\textsc{cort}| = 1$, the behavior proximity contributes 76.2% to D_k whereas the value proximity contributes 23.8% to D_k (23.8% of the value of D_k is determined by δ_E, and the remaining 76.2% by CORT). Note that the widely-used dynamic time warping (see for instance [14], [15]) is not addressed in this work, as it is not appropriate for generating cell-cycle gene expression profiles. Indeed, the identification of genes expressed during the cell-cycle is mainly based on the time at which the genes are highly expressed. To best cluster or classify gene expression profiles, time should not be warped when evaluating proximities.

4 Metrics comparison

A simulation study is performed to evaluate the efficiency of the metrics defined in Section 3. For the clustering process, the PAM (Partitioning Around Medoids) approach is used to partition the simulated genes into n clusters, n being the number of cell-cycle phases or inter-phases of interest. The PAM algorithm is preferred to the classical K-means for many reasons. It is more robust with respect to outliers, which are numerous in gene expression data. It also allows a more detailed analysis of the partition by providing clustering characteristics; in particular, it indicates whether each gene is well classified (i.e. highly expressed in a cell-cycle phase) or whether it lies on the boundary of the cluster (i.e. it is involved in a transition phase). For more details about the PAM algorithm, see Kaufman and Rousseeuw [12]. The efficiency of each metric in clustering gene

expression profiles is evaluated through the goodness of the obtained partitions. Three criteria are measured: the average silhouette width, the within-between ratio, and the corrected Rand index. For the classification process, the 10-NN approach is used to classify gene expression profiles. The efficiency of each metric is evaluated through the estimated misclassification error rate.

5 Simulation study

5.1 Random-periods model for periodically expressed genes

We use gene expression profiles generated using the random-periods model proposed by Liu et al. [16] to study periodically expressed genes. This model allows us to simulate attenuation in the amplitude of periodic gene expression with regard to stochastic variations during the various phases of the cell-cycle, while also permitting us to estimate the phase of the cycle in which the gene is most frequently transcribed. The sinusoid function for characterizing the expected periodic expression of a cell-cycle gene g is

$$f(t, \theta_g) = a_g + b_g t + \frac{K_g}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} cos(\frac{2\pi t}{T exp(\sigma z)} + \Phi_g) exp(-\frac{z^2}{2}) dz$$

, where θ_g is explicitly $(K_g, T, \sigma, \Phi_g, a_g, b_g)$, specific to each gene g. Integration in the model computes the expected cosine across the lognormal distribution of periods, and thereby accounts for the aggregation of expression levels across a large number of cells. The parameter Φ_g corresponds to the cell-cycle phase during which the gene undergoes its peak level of transcription, with $\Phi_g = 0$ corresponding to the point when cells are first released to resume cycling. The parameter K_g is the initial amplitude of the periodic expression pattern. The parameters a_g and b_g account for any drift (intercepts and slopes, respectively) in a gene's background expression level, and T and σ are the parameters of the lognormal distribution of cell-cycle duration. The parameter σ governs the rate of attenuation in amplitude. If σ is zero, the duration of the cell-cycle does not vary, as cells remain synchronous through time, and the expression profile shows no attenuation in amplitude. Larger values of σ correspond to faster attenuation of the peak amplitude. Figure 1 illustrates the progression of gene expression during five cell-cycle phases.

5.2 Simulation protocol

Based on the above random model and on the parameters specification given in [16], four experiments are simulated to study how each metric accounts for gene variations. The first experiment generates genes with varying initial amplitudes K_g varying in $[0.34, 1.33]$. The second experiment simulates genes with amplitude attenuation, with governed by σ, varying in [0.054, 0.115]. The third experiment varies the drift, with slopes $b_g \in [-0.05, 0.05]$ and intercepts $a_g \in [0, 0.8]$. The last experiment simulates genes with simultaneous variations of initial amplitude, amplitude attenuation during the cell-cycle, and drift. Figure 2 shows the

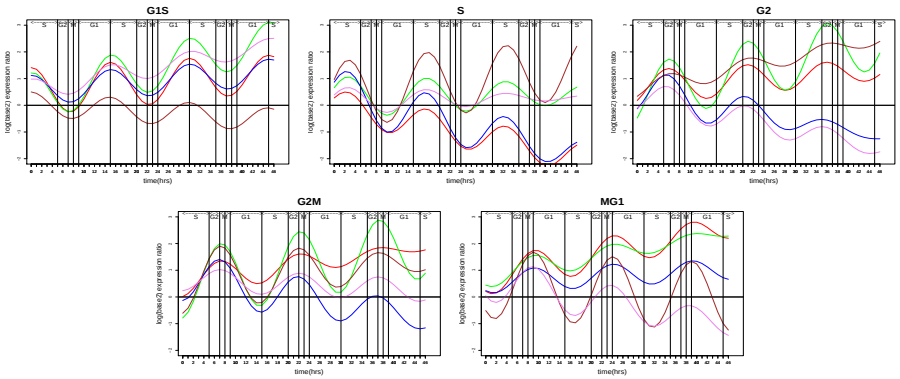

Fig. 1. Genes expression progression during five cell-cycle phases

variations generated across the four experiments for genes expressed in the G_1/S phases. The model parameter specifications of the four experiments are summarized in Table 1. For all simulations, T is fixed to 15, and Φ_g takes the values 0, 5.190, 3.823, 3.278, or 2.459 to simulate the expression profiles of the five classes G_1/S, S, G_2, G_2/M, or M/G_1, respectively. For each experiment $j \in \{1, ..., 4\}$, 10 samples S_{ij} $i \in \{1, ..., 10\}$ are simulated. Each sample S_{ij} is composed of 500 gene expression profiles with 100 genes for each of the five phases or inter-phases G_1/S, S, G_2, G_2/M, and M/G_1. The comparison of metrics is performed within each experiment through the clustering and the classification of 5000 simulated genes (i.e. 10 samples of 500 genes each).

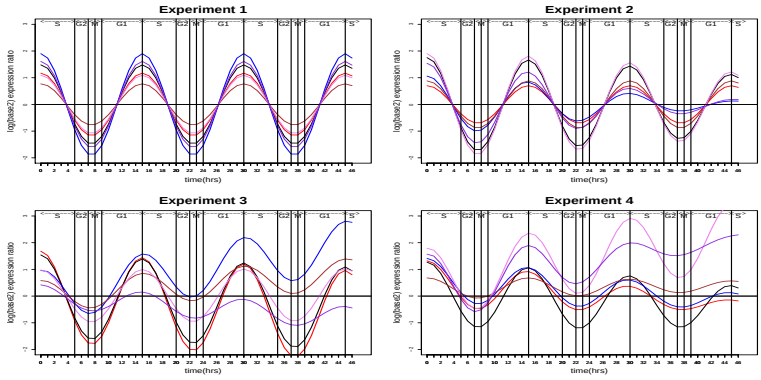

Fig. 2. G_1/S expression profiles through the four experiments

Table 1. Parameters specification

Experiment number	K_g	σ	b_g	a_g
1	[0.34, 1.33]	0	0	0
2	[0.34, 1.33]	[0, 0.115]	0	0
3	[0.34, 1.33]	0	[-0.05, 0.05]	[0, 0.8]
4	[0.34, 1.33]	[0, 0.115]	[-0.05, 0.05]	[0, 0.8]

5.3 Metrics efficiency for clustering gene expression profiles

For each experiment and for each metric δ_E, Cor, and Cort, a PAM algorithm is performed to partition each sample S_{ij} into 5 clusters (i.e. 5 cell-cycle phases and inter-phases). For instance, for the experiment j and for the metric δ_E, the PAM algorithm is applied to the 10 samples $S_{1j}, ..., S_{10j}$ to extract the 10 partitions $P_{\delta_E}^{1j}$,..., $P_{\delta_E}^{10j}$. For each partition, $P_{\delta_E}^{ij}$, three goodness criteria are measured: the average silhouette width (asw), the within/between ratio (wbr), and the corrected Rand index (RI). The corrected Rand index allows us to measure the proximity between $P_{\delta_E}^{ij}$ and the true partition (i.e. that defined by S_{ij}). Finally, the efficiency of the metric δ_E within the experiment j is summarized by the average values of the criteria asw, wbr, and RI of the 10 partitions $P_{\delta_E}^{1j}$,..., $P_{\delta_E}^{10j}$. For the dissimilarity index D_k, the adaptive clustering proposed in [5] is applied. The adaptive clustering of S_{ij} consists of performing the PAM algorithm for several values of k from 0 to 6 (per a lag of 0.01) to find the value k* that yields the optimal partition $P_{D_{k*}}^{ij}$, using as goodness criteria the average silhouette width and the within/between ratio. Note that $k*$ provides the best contribution of the proximity with respect to values and with respect to behavior to the dissimilarity index, thus the learned D_{k*} is identified as best clustering S_{ij}. Table 2 gives, for each experiment, the mean and the variance $(\overline{k*}, var(k*))$ of k*. As in the case of the metrics δ_E, Cor, and Cort, the efficiency of the metric D_k within the experiment j is summarized by the average values of the criteria asw, RI and wbr of the 10 partitions $P_{D_{k*}}^{1j}$,..., $P_{D_{k*}}^{10j}$. Figures 3, 4, and 5 depict, for each experiment and for each metric, the progression of the criteria asw, wbr, and RI across the 10 clustered samples $S_{1j}, ..., S_{10j}$. Figure 6 shows for each metric the progression, across the four experiments, of the average values of the criteria asw (top), wbr (middle) and RI (bottom)

Table 2. k* mean and variance

Adaptive	Exp1	Exp2	Exp3	Exp4
Clustering	(6,0)	(6,0)	(6,0)	(5.85,0.06)
Classification	(3,3.53)	(3,3.53)	(4.55,1.18)	(4.84,0.98)

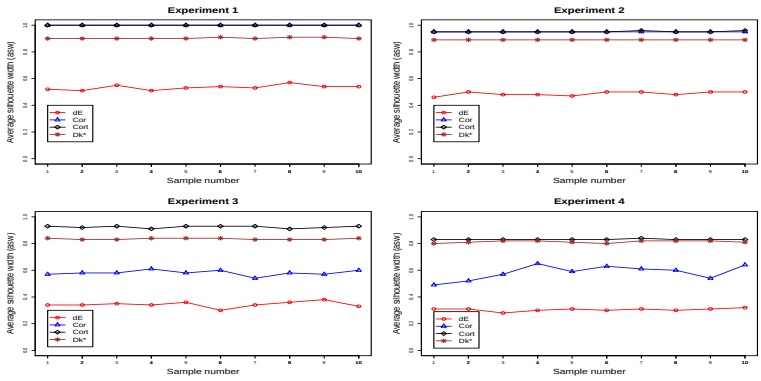

Fig. 3. Asw progression across the clustered samples

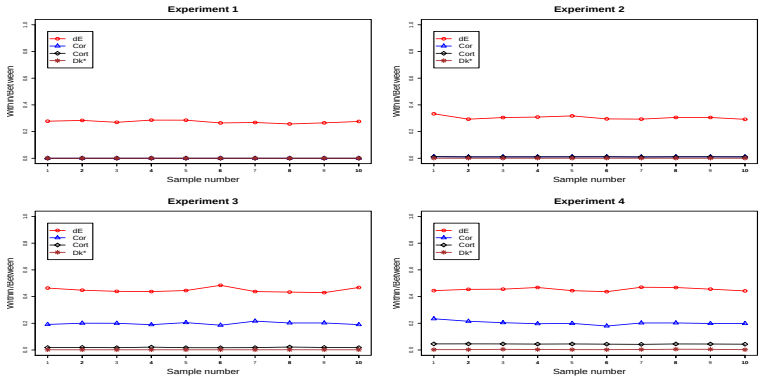

Fig. 4. Wbr progression across the clustered samples

5.4 Metrics efficiency for classifying gene expression profiles

For each experiment and for each metric δ_E, Cor, and Cort, a 10-NN algorithm is performed to classify each sample S_{ij}. For instance, for the experiment j and for the metric δ_E, the 10-NN algorithm is applied to the 10 samples $S_{1j}, ..., S_{10j}$ to generate the 10 classifications $C^{1j}_{\delta_E}$,..., $C^{10j}_{\delta_E}$. For each classification $C^{ij}_{\delta_E}$ the misclassification error rate is measured. The efficiency of the metric δ_E in classifying gene expression profiles within the experiment j, is summarized by the average misclassification error rates of the 10 classifications $C^{1j}_{\delta_E}$,..., $C^{10j}_{\delta_E}$. For the dissimilarity index D_k, an adaptive classification is performed. It consists of performing the 10-NN algorithm on S_{ij} for several values of k from 0 to 6 (in increments of 0.01) to find the $k*$ that minimizes the misclassification error rate of $C^{ij}_{D_k}$. The efficiency of the metric $D_k$in classifying gene expression profiles within the experiment j, is summarized by the average misclassification error rates of the 10 classifications $C^{1j}_{D_{k*}}$,..., $C^{10j}_{D_{k*}}$. Figure 7 depicts, for each of the four experiments, the progression of the misclassification error rates across the

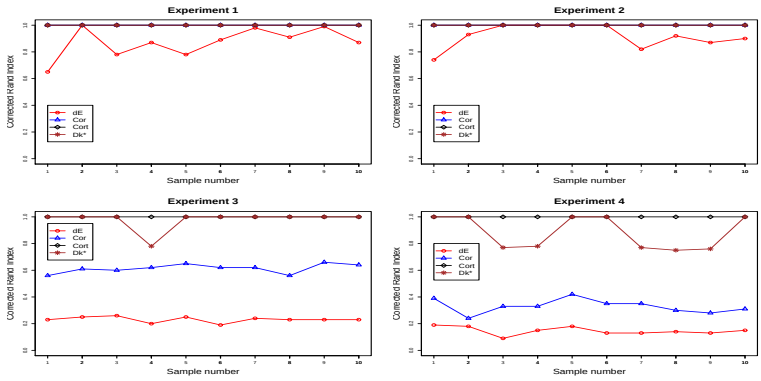

Fig. 5. RI progression across the clustered samples

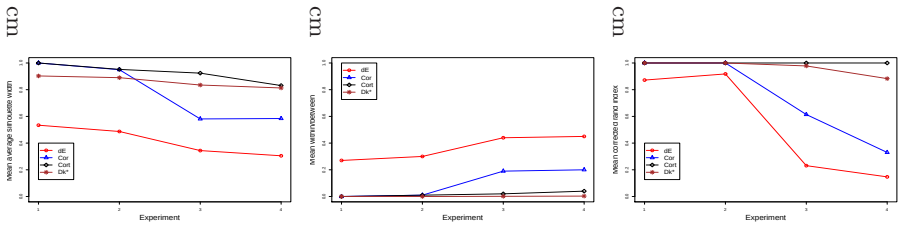

Fig. 6. Metric efficiency to cluster gene expression profiles

10 classified samples. Figure 8 shows, for each metric, the progression across the four experiments of the average misclassification error rates.

5.5 Discussion

We first discuss the clustering results. Let us give some additional information about the criteria in question. The asw indicates the strength (asw close to 1) or the weakness ($asw < 0.5$) of the obtained partitions, while wbr measures the compactness (i.e. within-cluster variability) and the separability (between-clusters variability) of the obtained clusters. A good partition is characterized by a lower within/between ratio. Finally, RI allows us to measure the similarity between the obtained partitions and the true ones ($RI = 1$ for a high level of similarity, and $RI = 0$ for non-similarity).
Figures 3, 4, and 5 show that the clustering based on δ_E gives, for experiments 1 to 4, weaker partitions than the ones based on COR, CORT, or D_k. Indeed, partitions based on δ_E have the lowest values for asw and RI and the highest values for wbr. Figure 6 shows that the average values of asw, wbr and RI of the clustering based δ_E decrease from the experiment 1 to 4, showing the inappropriateness of the Euclidean distance for cases with complex variations. Clustering based on COR gives strong partitions with the best values of asw, wbr, and RI, for the first two experiments. However, this quality decreases drastically

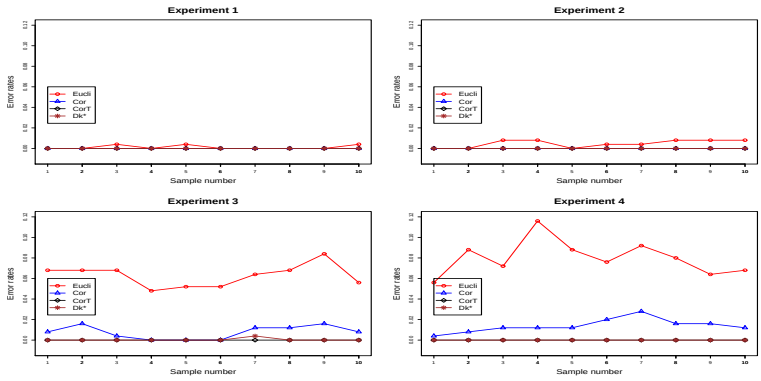

Fig. 7. Misclassification error rates progression through the classified samples

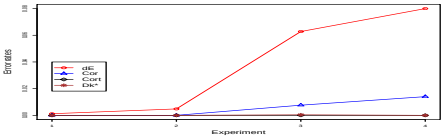

Fig. 8. Metric efficiency to classify gene expression profiles

in experiments 3 and 4 (Figures 3, 4, 5, and 6) showing the limitations of the Pearson correlation coefficient when faced with tendency variations, as explained in Subsection 3.2. Finally, the best clustering and the strongest partitions across the four experiments are given by CORT and D_k, with asw values varying in $[0.8, 1]$, wbr around 0, and RI varying in $[0.83, 1]$. Note that the quality of the clustering based on D_k is very slightly lower than that based on CORT, revealing that gene expression profiles are more naturally differentiated by their behaviors than by their values. This hypothesis is assessed by the higher values of $k*$ (near 6, with a variability of 0) obtained in the adaptive clustering across the four experiments (Table 2).

Let us now discuss the classification results. Figures 7 and 8 show that for experiments 1 and 2, the four metrics are equally efficient, with misclassification error rates around 0. However, for experiments 3 and 4, we note a drastic increase in the error rate for the partitions based on δ_E, a slight increase in the error rate for the partitions based on COR, and a negligible increase for D_k. Table 2 and Figure 9 indicate the distribution of k* in the adaptive classifications. For experiments 1 and 2, a uniform distribution of k* in $[0, 6]$ is noted. This case arises when a good classification can be obtained both with a metric based on values (k* near 0) and with a metric based on behavior (k* near 6). Indeed, Figures 7 and 8 show that the four metrics are equally efficient at classifying genes across the first two experiments. In experiments 3 and 4, k* takes higher values, indicating that the behavior-based metrics (i.e. CORT, D_k) are the most efficient for classifying gene expression profiles, as can easily be seen in Figures

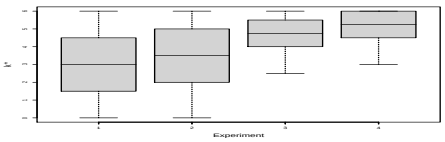

Fig. 9. k* distribution in the adaptive classification

7 and 8. Finally, according to the results of the four experiments, the metrics CORT and D_k can be said to be the most efficient at classifying gene expression profiles.

6 Conclusion

In conclusion, to cluster or classify cell-cycle gene expression profiles, it is advisable to consider the temporal correlation coefficient as a proximity measure. However, the effectiveness of the learned dissimilarity D_k, which also provides very good partitions and classifications, is worth noting. In general, when faced with data where time should not be warped for proximity evaluation (which is the case of cell-cycle gene expression profiles), the dissimilarity D_k proposed in this paper is recommended. The adaptive clustering or classification is used to learn the appropriate dissimilarity D_k to use in the next analysis task. The learned dissimilarity D_k can lead to the temporal correlation (for k* near 6), to the Euclidean distance (for k* near 0), or more generally to a metric covering both values and behavior proximities.

References

1. Anagnostopoulos, A., Vlachos, M., Hadjieleftheriou, M., Keogh, E.J., Yu, P.S. Global Distance-Based Segmentation of Trajectories. In Proc. of ACM SIGKDD, 34-43, 2006.
2. Bar-Joseph, Z., Gerber, G. K., Gifford, D.K., Jaakkola, T., Simon I.. Continuous Representations of Time-Series Gene Expression Data. Journal of Computational Biology. 10, 3, 341-356, 2003.
3. Caiado, J., Crato, N., Pena, D. A periodogram-based metric for time series classification. Computational Statistics and Data Analysis. 50, 2668-2684, 2006.
4. Douzal Chouakria, A., Nagabhushan, P.N. Adaptive dissimilarity index for measuring time series proximity. Advances in Data Analysis and Classification Journal. 1, 5-21, Springer, 2007.
5. Douzal-Chouakria, A., Diallo, A., Giroud, F. Adaptive clustering for time series: application for identifying cell-cycle expressed genes. Computational Statistics and Data Analysis 53 (4), 1414-1426, Elsevier, 2009.
6. Deroski, S., Gjorgjioski, V., Slavkov, I., Struyf, J. Analysis of time series data with predictive clustering trees. In S. Deroski and J. Struyf, editors, Knowledge Discovery in Inductive Databases, 5th International Workshop, KDID, Berlin, Germany, 2006.

7. Eisen, M.B., and Brown, P.O. DNA arrays for analysis of gene expression. Methods Enzymol. 303, 179-205, 1999.
8. Garcia-Escudero, L. A., Gordaliza, A. A proposal for robust curve clustering. Journal of Classification. 22, 185-201, 2005.
9. Heckman, N. E., Zamar, R. H. Comparing the shapes of regression functions. Biometrika. 22, 135-144, 2000.
10. He, Y., Pan, W., Lin, J. Cluster analysis using multivariate normal mixture models to detect differential gene expression with microarray data. Computational Statistics and Data Analysis. 51, 2, 641-658, 2006.
11. Kakizawa, Y., Shumway, R. H., Taniguchi, N. Discrimination and clustering for multivariate time series. Journal of the American Statistical Association. 93, 328-340, 1998.
12. Kaufman, L., Rousseeuw, P.J. Finding Groups in Data. An Introduction to Cluster Analysis. John Wiley & Sons, New York, 1990.
13. Keller, K., Wittfeld, K. Distances of time series components by means of symbolic dynamics. International Journal of Bifurcation Chaos. 14, 693-704, 2004.
14. Keogh, E.J., Pazzani, M.J. Scaling Up Dynamic Time Warping for Data Mining Applications. In Proc. of ACM SIGKDD, 285-289, 2000.
15. Kruskall, J.B., Liberman, M. The symmetric time warping algorithm: From continuous to discrete. In Time Warps, String Edits and Macromolecules. Addison-Wesley, 1983.
16. Liu, D., Umbach, D. M., Peddada, S. D., Li L., Crockett, P. W., Weinberg, C. R. A Random-Periods Model for Expression of Cell-Cycle Genes. Proc Natl Acad Sci USA. 101, 7240-7245, 2004.
17. Liu, X., Lee, S., Casella, G., Peter, GF. Assessing agreement of clustering methods with gene expression microarray data. Computational Statistics and Data Analysis. 52, 12, 5356-5366, 2008.
18. Maharaj, E. A. Cluster of time series. Journal of Classification. 17, 297-314, 2000.
19. Oates, T., Firoiou, L., Cohen, P. R. Clustering time series with Hidden Markov Models and Dynamic Time Warping. In: Proc. 6th IJCAI-99, Workshop on Neural, Symbolic and Reinforcement Learning Methods for Sequence Learning, Stockholm, 17-21, 1999.
20. Park, C., Koo, J., Kim, S., Sohn, I., Lee, J. W. Classification of gene functions using support vector machine for time-course gene expression data. Computational Statistics and Data Analysis. 52, 5, 2578-2587, 2008.
21. Scrucca, L. Class prediction and gene selection for DNA microarrays using regularized sliced inverse regression. Computational Statistics and Data Analysis. 52, 1, 438-451, 2007.
22. Serban, N., Wasserman, L. CATS: Cluster After Transformation and Smoothing. Journal of the American Statistical Association. 100, 990-999, 2004.
23. Shieh, J., and Keogh, E.J. iSAX: Indexing and Mining Terabyte Sized Time Series. In Proc. of ACM SIGKDD, 623-631, 2008.