Un modèle génératif pour la comparaison de métriques en classification de profils d 'expression de gènes

Alpha Diallo, Ahlame Douzal-Chouakria, Françoise Giroud

Un modèle génératif pour la comparaison de métriques en classification de profils d'expression de gènes

Alpha Diallo[1,2], Ahlame Douzal-Chouakria[1], Françoise Giroud[2]

[1] LIG AMA, Université Joseph Fourier
BP 53 - 38041 Grenoble cedex 9
(alpha.diallo, ahlame.douzal)@imag.fr

[2] TIMC-IMAG RFMQ (CNRS-UMR 5525), Université Joseph Fourier
F-38706 La Tronche Cedex, France
francoise.giroud@imag.fr

Résumé : La prolifération cellulaire traduit le processus cyclique de division des cellules responsable de la croissance des tissus. Un tissu tumoral peut alors être caractérisé par la présence de cellules cancéreuses qui présentent une prolifération anormale. Ce papier s'intéresse à la classification des gènes différentiellement activés au cours du processus de division cellulaire. Les gènes étudiés ici sont décrits par leurs profiles d'expression (c-à-d des données temporelles d'expression de gènes) tout au long de 3 cycles cellulaires successifs. Le travail présenté concerne l'évaluation de l'efficacité de 4 métriques majeures pour la classification et le classement des profiles d'expression de gènes. Il est basé sur la mise en œuvre d'un modèle périodique aléatoire pour la simulation de gènes d'expression cyclique au long du cycle de division cellulaire. Le modèle traduit les événements cycliques de régulation moléculaire du cycle cellulaire, avec des variations en décalage dans le temps des profiles d'expression de différents gènes. Le modèle rend compte également de phénomènes de désynchronisation cellulaire provoquant à la fois l'atténuation en amplitude des valeurs d'expression et des modifications de périodicité des cycles cellulaires successifs.

Mots-clés : Séries temporelles, métriques, classification, profils d'expression de gènes.

1. Introduction

Toutes les cellules de notre corps contiennent les mêmes gènes, mais tous n'interviennent pas dans chaque cellule : les gènes sont activés ou exprimés au besoin. De tels gènes spécifiques définissent le modèle moléculaire lié à une fonction spécifique d'une cellule et apparaissent dans la plupart des

cas comme organisés dans des réseaux de régulation moléculaire. Pour comprendre comment les cellules réalisent une telle spécialisation, il est nécessaire d'identifier quels gènes s'expriment dans chaque type de cellules (par exemple, des tissus cancéreux versus des tissus sains). La technologie des puces à ADN nous permet d'étudier simultanément les niveaux d'expression de plusieurs milliers de gènes, au cours de processus biologiques importants, pour déterminer ceux qui sont exprimés dans un type de cellule spécifique Eisen & Brown (1999). Les techniques de classification et de classement sont utilisées et se sont montrées particulièrement efficaces pour comprendre la fonction des gènes, des voies de régulation et des processus cellulaires (e.g., Liu *et al.* (2008), Park *et al.* (2008), Scrucca (2007)). Nous distinguons au moins deux principales approches de classification et de classement de profils ou de séries temporelles. D'une part, les approches paramétriques consistant à projeter les séries temporelles dans des espaces de fonctions correspondant, par exemple, aux polynômes d'un modèle ARIMA, aux transformées de Fourier, ou plus généralement aux paramètres d'un modèle approximant les séries temporelles. Des mesures conventionnelles peuvent ensuite être utilisées dans le nouvel espace de projection (e.g.,Bar-Joseph *et al.* (2003), Caiado *et al.* (2006), Garcia-Escudero & Gordaliza (2005)). D'autre part, on distingue les approches non-paramétriques dont l'objectif est la proposition de nouvelles mesures de proximités définies dans l'espace de description initial et intégrant la dimension temporelle des données (e.g.,Anagnostopoulos *et al.* (2006), Heckman & Zamar (2000), Keller & Wittfeld (2004)). Dans le cadre des approches non-paramétriques, nous proposons d'étudier l'efficacité de quatre métriques majeures pour la classification et le classement des profils temporels d'expression de gènes. Cette étude est basée sur la mise en œuvre d'un modèle périodique aléatoire pour la simulation de gènes d'expression cyclique. Ce modèle tient compte des caractérisques principalement observées sur les profils de gènes du cycle cellulaire : l'amplitude initiale du profil, la période du profil, l'atténuation des amplitudes dans la longueur du temps et les effets de tendance. La suite de l'article est organisé en quatre sections. La section suivante définit ce que sont les données d'expression de gènes et présente le problème biologique abordé. La section 3. présente les quatre principales métriques à évaluer et discute de leurs caractéristiques. La section 4. indique comment les mesures seront comparées par la classification et le classement des profils de gènes. Enfin, l'ensemble des méthodes d'évaluation basées sur le modèle utilisé et la discution des résultats obtenus sont présentés dans la section 5..

2. Identification des gènes exprimés au cours du cycle cellulaire

Le problème biologique d'intêret est l'analyse de la progression de l'expression des gènes durant le processus de la division cellulaire. La division cellulaire est le processus principal assurant la prolifération des cellules, et se décompose en quatre phases principales (G_1, S, G_2 et M) et trois transitions de phase (G_1/S, G_2/M et M/G_1)(Figure 1). Le processus de division commence à la phase G_1 pendant laquelle la cellule se prépare à la synthèse de l'ADN. Vient la phase S où l'ADN est dupliqué (c-à-d chaque chromosome est dupliqué), suivie par la phase G_2 pendant laquelle la cellule se prépare à sa division. Enfin vient la mitose M où la cellule se divise en deux cellules filles. Pendant ces quatre phases, certains gènes sont actifs (fortement exprimés) à des périodes spécifiques, d'autres pas. Un des objectifs consiste à identifier les gènes fortement exprimés et caractérisant chaque phase du cycle cellulaire. Ceci fournit des informations importantes, par exemple, pour comprendre comment le traitement hormonal peut induire la prolifération cellulaire par l'activation de gènes spécifiques. Afin d'accroître la compréhension de l'expression des gènes au cours du processus de la division cellulaire, des molécules d'ADN représentant les différents gènes sont placées sur des spots discrets régulièrement répartis en une matrice ligne/colonne (appelée puce à ADN). En déposant sur ces puces à ADN des extraits cellulaires on peut mesurer le niveau d'expression de chaque gène au sein des populations cellulaires étudiées. En échantillonnant au cours du temps une population cellulaire initialement synchronisée, chaque gène étudié peut être décrit par son profil d'expression observé au cours du temps sur un ou plusieurs cycles de la division cellulaire.

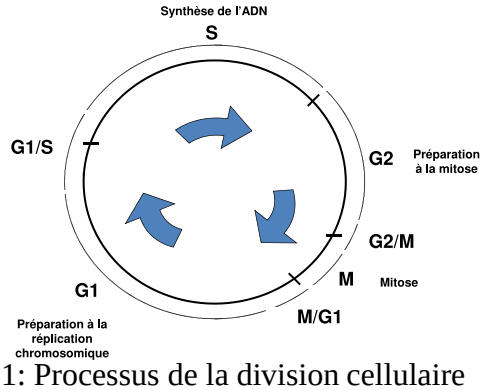

FIG. 1: Processus de la division cellulaire

3. Proximité entre profils d'expression de gènes

La classification et le classement des données d'expression implique le plus souvent la distance euclidienne ou le coefficient de corrélation de Pearson. Cette section introduit quatre métriques majeures pour l'analyse des gènes et leurs spécifications, en tenant compte des proximités en valeurs et en forme des profils de gènes. Soit $g_1 = (u_1, ..., u_p)$ et $g_2 = (v_1, ..., v_p)$ les niveaux d'expression de deux gènes observés aux instants $(t_1, ..., t_p)$.

3.1. La distance euclidienne

La distance euclidienne δ_E entre g_1 et g_2 est définie par :

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

Il ressort de cette définition, que la proximité entre deux gènes dépend de la proximité des valeurs de leurs niveaux d'expression, sans tenir compte de la forme de leurs profils. En d'autres termes, la distance euclidienne ignore la dépendance temporelle des données.

3.2. Le coefficient de corrélation de Pearson

De nombreux travaux utilisent le coefficient de corrélation de Pearson comme mesure de proximité en forme entre 2 séries temporelles. Sans perte de généralité, supposons que les valeurs de g_1 et g_2 évoluent dans $[0, N]$. Les gènes g_1 et g_2 sont dits de formes similaires si à chaque période d'observation $[t_i, t_{i+1}]$, ils croient ou décroient simultanément (monotonicité), avec un taux d'accroissement égal. En revanche, g_1 et g_2 sont de formes opposées si dans chaque période d'observation $[t_i, t_{i+1}]$ où g_1 croit , g_2 décroit et vice-versa avec le même taux d'accroissement en valeur absolue. Afin d'illustrer la limite du coefficient de corrélation à mesurer la proximité en forme des gènes, considérons son expression basée sur les différences entre les valeurs observées :

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

En impliquant les différences entre tous les couples d'observations (c-à-d, observées à tous les couples d'instants (i, i')), le coefficient de corrélation de Pearson fait l'hypothèse d'indépendance entre les données observées. Conséquence, ce coefficient peut surestimer la proximité en forme. Par exemple, la section 4. illustre le cas de données dotées d'un effet de tendance où deux gènes de formes opposées peuvent avoir un coefficient de corrélation de valeur positive forte.

3.3. Le coefficient de corrélation temporelle

Pour surmonter les limites du coefficient de corrélation de Pearson (Eq. (2)), le coefficient de corrélation temporelle est utilisé. Il réduit le coefficient de corrélation de Pearson aux différences de premier ordre :

$$\text{CORT}(g_1, g_2) = \frac{\sum_i (u_{(i+1)} - u_i)(v_{(i+1)} - v_i)}{\sqrt{\sum_i (u_{(i+1)} - u_i)^2}\sqrt{\sum_i (v_{(i+1)} - v_i)^2}}. \quad (3)$$

avec $\text{CORT}(g_1, g_2) \in [-1, 1]$. La valeur $\text{CORT}(g_1, g_2) = 1$ indique que g_1 et g_2 présentent une forme similaire. La valeur $\text{CORT}(g_1, g_2) = -1$ signifie que g_1 et g_2 sont de formes opposées. Enfin, $\text{CORT}(g_1, g_2) = 0$ exprime que les taux d'accroissement de g_1 et g_2 sont stochastiquement linéairement indépendants, identifiant ainsi des gènes de formes différentes (non similaires ni opposées).

3.4. Mesures de proximité alliant forme et valeurs

Pour une mesure de proximité couvrant simultanément les écarts en forme et en valeurs des niveaux d'expression, l'indice de dissimilarité D_k proposé dans Douzal-Chouakria *et al.* (2010, 2009) est considéré :

$$D_k(g_1, g_2) = f(\text{CORT}(g_1, g_2))\ \delta_E(g_1, g_2), \quad (4)$$

$$f(x) = \frac{2}{1 + exp(k\ x)},\ \ k \geq 0. \quad (5)$$

Cet indice couvre à la fois la distance euclidienne (Eq. (1)) pour la proximité en valeurs, et la corrélation temporelle (Eq. (3)) pour la proximité en forme. Il est basé sur une fonction de reglage $f(x)$ qui module la proximité en valeurs en fonction de la proximité en forme. Une fonction exponentielle $f(x)$ est préférable à une forme linéaire afin d'assurer un effet de réglage approximativement égal pour les valeurs extrêmes (i.e., CORT=-1, +1 et 0) et leurs

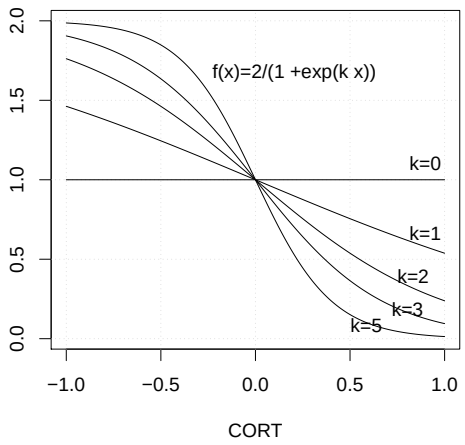

FIG. 2: L'effet du réglage en fonction de k

plus proches voisins. La figure 2 montre l'effet de réglage pour plusieurs valeurs du paramètre $k \geq 0$. Dans le cas de gènes de formes différentes (i.e., CORT voisin de 0), $f(x)$ est voisin de 1 quelles que soient les valeurs de k, et D_k est approximativemnt égal à δ_E. En revanche, si CORT $\neq 0$ (forme non différentes) le paramètre k module les contributions des deux types de proximité (valeurs et forme) à l'indice de dissimilarité D_k. Lorsque k augmente, la contribution de la proximité en forme $1 - 2/(1 + exp(k\ |\text{CORT}|))$ augmente, tandis que celle de la proximité en valeurs $2/(1+exp(k\ |\text{CORT}|))$ diminue. Par exemple, pour $k = 0$ et $|\text{CORT}| = 1$ (forme similaire ou opposée), la proximité en forme contribue 0% à D_k tandis que la proximité en valeurs contribue 100% à D_k (la valeur de D_k est totalement déterminée par δ_E). Pour $k = 2$ et $|\text{CORT}| = 1$, la proximité en forme contribue 76.2% à D_k tandis que celle en valeurs contribue 23.8% à D_k (23.8% de la valeur de D_k sont déterminés par δ_E et les 76.2% restants par CORT).

Notons que la dynamique time warping (e.g.,Kruskall & Liberman (1983), Shieh & Keogh (2008)), qui est largement utilisée, n'est pas abordée dans ce travail puisqu'elle n'est pas appropriée pour analyser des profils d'expressions de gènes au cours du cycle cellulaire. En effet, l'identification de gènes exprimés au cours cycle cellulaire repose principalement sur l'instant où les gènes sont fortement exprimés. Ainsi, les instants d'observation ne doivent pas subir de décalage lors de l'évaluation des proximités entre profils d'expression de gènes.

4. Comparison des métriques

Une étude de simulation est effectuée pour évaluer l'efficacité des métriques définies dans les équations (1) à (4). Pour la procédure de classification, nous proposons d'utiliser l'algorithme PAM (Partitioning Around Medoïds) afin de partitionner les gènes simulés en n classes (n étant le nombre de phases du cycle cellulaire ou de transitions de phases étudiées). L'algorithme PAM est préféré à l'approche classique des K-means pour plusieurs raisons. Il est plus robuste aux valeurs aberrantes qui sont nombreuses dans les données d'expression de gènes. PAM permet une analyse détaillée de la partition en fournissant des indices permettant d'apprécier la qualité des classes ainsi que celle des gènes. En effet, PAM mesure la *silhouette width* (sw) de chaque gène, qui est un indicateur de confiance quant à l'appartenance d'un gène à une classe. Pour plus de détails sur l'algorithme PAM voir Kaufman & Rousseeuw (1990). L'efficacité de chaque métrique est basée sur trois critères : *la silhouette width* moyenne d'une partition notée asw, le ratio standard $wbr = \frac{intra}{inter}$ et l'indice de Rand corrigé (RI). Pour la procédure de classement des gènes, l'algorithme 10-NN est utilisé, et les taux d'erreur de gènes mal classés sont retenus pour apprécier l'efficacité de chaque métrique.

5. Etude comparative

5.1. Modèle génératif de profils d'expression périodiques

Nous utilisons des profils simulés, générés sur la base du modèle de regression non-linéaire proposé par Liu *et al.* (2004). Ce modèle permet de simuler l'atténuation des amplitudes de l'expression des gènes liée aux variations stochastiques au cours des différentes phases du cycle cellulaire. La fonction sinusoïdale caractérisant la périodicité du profils d'expression d'un gène g au cours de différents cycles cellulaires est :

$$f(t, \theta_g) \quad = \quad a_g + b_g t + \frac{K_g}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} cos(\frac{2\pi t}{Texp(\sigma z)} + \Phi_g) exp(-\frac{z^2}{2}) dz \quad (6)$$

où $\theta_g = (K_g, T, \sigma, \Phi_g, a_g, b_g)$ est spécifique du gène g. Le paramètre K_g représente son amplitude initiale, T est la durée du cycle cellulaire. Le paramètre σ contrôle le taux d'atténuation des amplitudes au cours des différents cycles, Φ_g correspond à la phase du cycle cellulaire où le gène est le plus exprimé. Les paramètres a_g et b_g (l'ordonnée à l'origine et la pente, respectivement)

contrôlent les tendances des profils. La figure 3 illustre la progression des expressions de gènes au cours des 5 phases et transitions de phase G_1/S, S, G_2, G_2/M et M/G_1. Nous allons utiliser le terme phase de manière générique pour parler de "phase" et "transition de phase" dans tout ce qui suit.

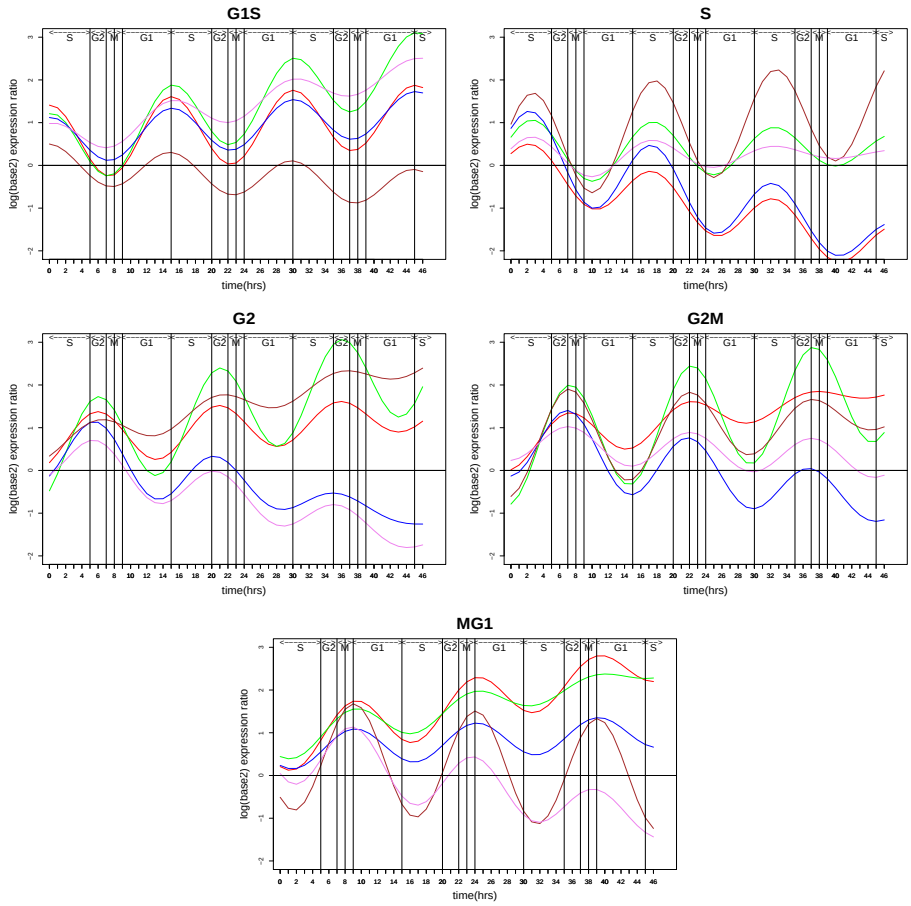

FIG. 3: Progression de l'expression des gènes durant les 5 phases G_1/S, S, G_2, G_2/M et M/G_1.

5.2. Protocole de simulation

Sur la base de ce modèle et des valeurs de paramètres spécifiées dans Liu *et al.* (2004), quatre expériences sont menées pour étudier la façon dont chaque métrique considère les variations d'expression de gènes. La première

expérience génère des profils avec une variation observée uniquement au niveau de l'amplitude initiale K_g variant dans $[0.34, 1.33]$. La seconde expérience inclut une atténuation des amplitudes σ évoluant dans $[0.054, 0.115]$. La troisième expérience inclut les effets de tendance $b_g \in [-0.05, 0.05]$ et $a_g \in [0, 0.8]$ et enlève les effets de σ. Enfin la quatrième expérience simule des profils avec une variation simultanée des paramètres K_g, σ, a_g, b_g dans les mêmes intervalles que précédemment. La valeur d'un paramètre est prise de manière aléatoire dans l'intervalle auquel il appartient. L'évolution des profils est suivi sur 3 cycles cellulaires, T est fixé à 15 heures pour toutes les simulations et Φ_g prend les valeurs 0, 5.190, 3.823, 3.278 et 2.459 pour la génération respective des 5 phases G_1/S, S, G_2, G_2/M et M/G_1. La figure 4 montre les variations produites dans les quatre expériences pour les gènes exprimés dans la phase G_1/S. La spécification des paramètres du modèle des quatre expériences est résumée dans le tableau 1. Pour chaque expérience $j \in \{1, ..., 4\}$, 10 échantillons S_{ij} $i \in \{1, ..., 10\}$ sont simulés. Chaque échantillon est composé de 500 profils d'expression (de longueur 47) de gènes avec 100 gènes pour chacune des 5 phases G_1/S, S, G_2, G_2/M et M/G_1. La comparaison est effectuée pour chaque expérience sur 5000 gènes simulés (c'est-à-dire 10 échantillons de 500 gènes chacune)

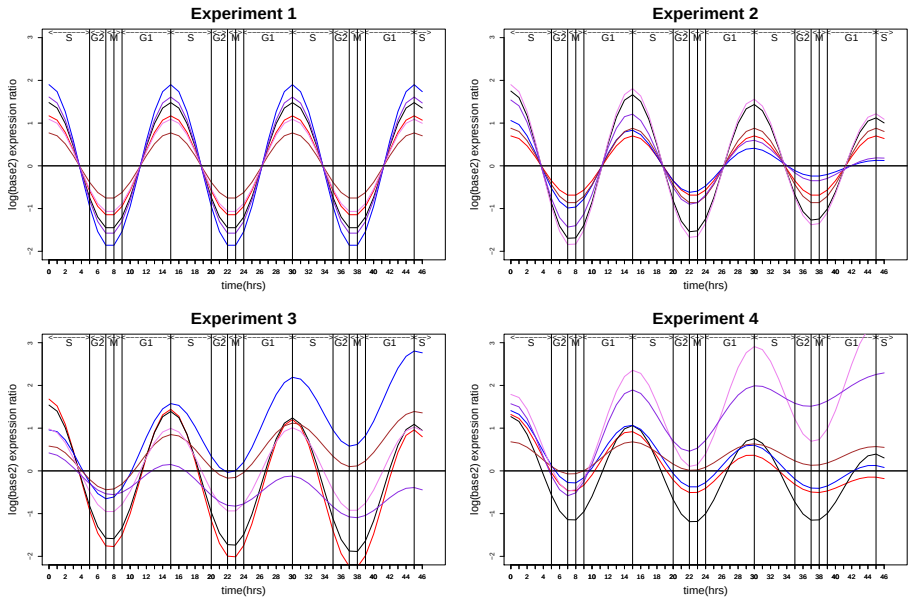

FIG. 4: Profils des gènes de la phase G_1/S suivant les quatre expériences.

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]

TAB. 1: Spécification des paramètres du modèle.

5.3. Evaluation de l'efficacité des métriques pour la classification des gènes

Pour chaque expérience et pour chaque métrique δ_E (Eq. (1)), COR (Eq. (2)), et CORT (Eq. (3)), nous partitionnons l'ensemble des profils de chaque échantillon S_{ij} en 5 classes (correspondant aux 5 phases). Par exemple, pour l'expérience j et la métrique δ_E, l'algorithme PAM est appliqué sur les 10 échantillons $S_{1j}, ..., S_{10j}$ afin d'extraire les 10 partitions $\mathcal{P}^{1j}_{\delta_E}$,..., $\mathcal{P}^{10j}_{\delta_E}$. Pour chaque partition $\mathcal{P}^{ij}_{\delta_E}$, les valeurs des trois critères asw, wbr, RI sont retenues. Ainsi, l'évaluation de l'efficacité de la métrique δ_E par rapport à l'expérience j est réalisée en considérant les valeurs moyennes des critères asw, RI et wbr sur les 10 partitions $\mathcal{P}^{1j}_{\delta_E}$,..., $\mathcal{P}^{10j}_{\delta_E}$. Une classification adaptative est appliquée pour l'indice de dissimilarité D_k (Eq. (4)). Elle consiste, sur un échantillon S_{ij}, à exécuter l'algorithme PAM pour k allant de 0 à 6 (avec un pas égal à 0.01). Ceci permet d'apprendre la valeur k^* qui fournit la partition optimale $\mathcal{P}^{ij}_{D_{k*}}$ selon les critères asw et wbr. Notons que k^* fournit la meilleure contribution des proximités en valeurs et en forme à D_{k^*}. Par conséquent, D_{k*} appris est considéré comme le meilleur pour le partitionnement de S_{ij}. Le tableau 2 donne pour chaque expérience, la moyenne et la variance $(\overline{k*}, var(k*))$ de $k*$. Comme dans le cas des métriques δ_E, COR et CORT, l'évaluation de l'efficacité de la métrique D_k par rapport à l'expérience j, est résumée par les valeurs moyennes asw, RI et wbr sur les 10 partitions $\mathcal{P}^{1j}_{D_{k*}}$,..., $\mathcal{P}^{10j}_{D_{k*}}$. La figure 5 montre (pour chaque métrique) la progression des valeurs moyennes des critères asw (en haut à gauche), wbr (en haut à droite) et RI (en bas) suivant les quatre expériences.

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)

TAB. 2: k*(moyenne, variance)

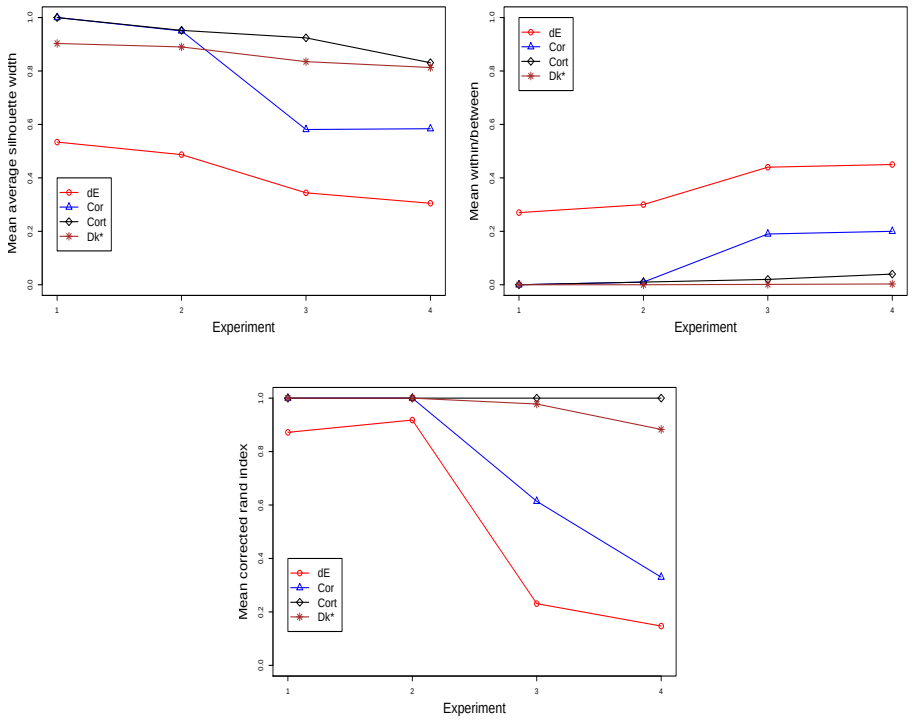

FIG. 5: Evaluation des métriques pour la classification des profils d'expression simulés. La progression des valeurs moyennes des critères asw (en haut à gauche), wbr (en haut à droite) et RI (en bas) est illustrée.

5.4. Evaluation de l'efficacité des métriques pour le classement des gènes

Pour chaque expérience et pour chaque métrique δ_E, COR et CORT, nous éxécutons l'algorithme 10-NN, pour chaque échantillon S_{ij}. Par exemple, pour l'expérience j et la métrique δ_E, l'algorithme 10-NN est appliqué sur les 10 échantillons $S_{1j}, ..., S_{10j}$ pour générer les 10 classes $\mathcal{C}_{\delta_E}^{1j}$,..., $\mathcal{C}_{\delta_E}^{10j}$. Pour chaque classe $\mathcal{C}_{\delta_E}^{ij}$, le taux de profils de gènes mal classifiés est retenu. L'évaluation de la métrique δ_E dans l'expérience j est résumée par le taux d'erreur moyen sur les 10 classes $\mathcal{C}_{\delta_E}^{1j}$,..., $\mathcal{C}_{\delta_E}^{10j}$. Pour l'indice de dissimilarité D_k, un classement adaptatif est appliqué. Il consiste à exécuter l'algorithme 10-NN sur l'échantillon S_{ij} avec des valeurs de k allant de 0 à 6 (avec un pas égal à 0.01). Ceci permet d'estimer la valeur k^* minimisant le taux d'erreur de profils mal classifiés de la classe $C_{D_k}^{ij}$. L'évaluation de la métrique D_k, pour

le classement des profils de gènes dans l'expérience j, se résume par le taux d'erreur moyen calculé sur les 10 classes $\mathcal{C}_{D_k}^{1j}$,..., $\mathcal{C}_{D_k}^{10j}$. La figure 6 montre la progression suivant les quatres expériences du taux d'erreur moyen lié à chaque métrique.

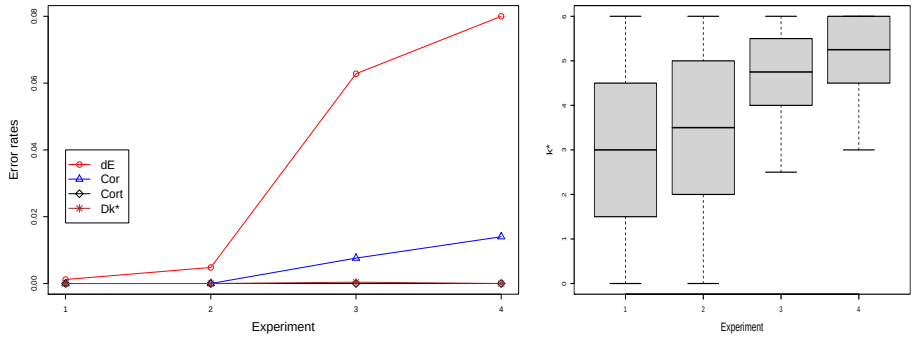

FIG. 6: Progression des taux d'erreur moyens des classements suivant les expériences (gauche). Distribution de $k*$ pour les classements adaptatifs suivant les expériences (droite)

5.5. Discussion

Nous discutons, d'abord, sur l'intérêt de considérer un modèle génératif pour l'évaluation de la classification ou du classement des gènes exprimés au cours du cycle cellulaire. Le modèle périodique aléatoire est d'une grande importance en biologie : il permet de simuler des trajectoires périodiques d'expressions de gènes semblables à celles observées expérimentalement. Ces trajectoires peuvent varier considérablement en forme d'un gène à un autre dans une même classe (i.e. pour les gènes exprimés dans la même phase du cycle cellulaire). Ce modèle permet aussi de simuler des variations observées expérimentalement : principalement, la reproduction des variations sur l'atténuation en amplitude des valeurs d'expression et sur la tendance des trajectoires par rapport à des modifications de périodicité des cycles cellulaires. Le modèle périodique aléatoire peut être utile pour étudier séparément et avec précision les effets de chaque condition expérimentale (i.e. variation) sur l'efficacité de métriques, de résultats de classification ou de classement. Il existe aussi d'autres modèles dans la littérature pour évaluer des métriques, des résultats de classification ou de classement de gènes exprimés au cours

du cycle cellulaire. On peut distinguer au moins deux principales techniques d'évaluation. D'une part, les données d'expression sont simulées à partir de modèles paramétriques, incluant : les modèles autorégressifs Ramoni *et al.* (2002), B-splines Luan & Li (2003), la décomposition en valeurs singulières (Alter *et al.* (2000), Holter *et al.* (2001)), ou les approches des moindres carrés partiels Johansson *et al.* (2003). Ces modèles fournissent une estimation assez bonne des trajectoires exprimées. Cependant, les paramètres estimés ne sont pas biologiquement interprétables et fournissent des périodicités anormales sur toute la durée du cycle cellulaire. D'autre part, les données d'expression sont extraites à partir de données réelles (e.g.,Spellman *et al.* (1998), Whitfield *et al.* (2002)) où les techniques d'évaluation sont a priori basées sur un petit ensemble de gènes connus dont les trajectoires sont cycliques (dits gènes de référence), et qui sont supposés caractéristiques des phases du cycle cellulaire. Ici, les évaluations sont fortement dépendantes des gènes de référence choisis, et la littérature ne fournit pas de consensus clair entre les biologistes sur l'ensemble approprié de gènes de référence à tenir en considération. Pour les raisons citées ci-dessus, nous avons opté pour l'utilisation d'un modèle génératif pour la simulation des données d'expression. Nous avons évalué des métriques avec des algorithmes classiques de classification (Kmeans, PAM, méthodes hiérarchiques, etc.) ou de classement (KNN, arbres de décisions, etc.). Ces métriques ont été évaluées par des critères de qualité : indice de Rand corrigé, taux de mauvais classement, etc.

Examinons d'abord les résultats de la classification. Notons quelques informations supplémentaires sur les critères en question. La valeur asw indique une forte structure (asw proche de 1) ou une faible structure ($asw < 0.5$) de classes. Le critère wbr mesure la compacité (variabilité au sein d'une classe) et la séparabilité (variabilité entre les classes) des classes. Une bonne partition est caractérisée par une faible valeur de wbr. Enfin, l'indice de Rand corrigé (RI) permet de comparer deux partitions. Une valeur $RI = 0$ correspond à une absence totale de correspondance entre la vraie partition et celle obtenue, alors qu'une valeur $RI = 1$ traduit une correspondance parfaite. La figure 5 montre que la classification basée sur δ_E donne, pour les expériences 1 à 4, les partitions les plus faibles comparée à celles fondées sur COR, CORT, ou D_k. En effet, les partitions fondées sur δ_E ont les plus faibles valeurs des critères asw et RI, et les valeurs les plus élevées pour wbr. En plus, les valeurs moyennes des critères asw, wbr et RI de la classification baseé sur δ_E se dégradent (diminution des asw et RI et augmentation de wbr) de l'expérience 1 à 4, montrant l'inadéquation de la distance euclidienne face aux variations com-

plexes des profils de gènes cyliques. La classification basée sur COR donne, pour les expériences 1 et 2, de bonnes structures de partitions avec de très bonnes valeurs des critères asw, wbr et RI. Toutefois, cette qualité diminue de façon drastique dans les expériences 3 et 4 (Figure 5). Comme expliqué dans la section 3, ces résultats affirment la limite du coefficient de corrélation de $Pearson$ face aux variations de tendance. Enfin, les meilleures classifications et les plus fortes structures de partitions sont produites par CORT et D_k sur toutes les quatre expériences, avec une asw variant dans $[0.8, 1]$, une valeur wbr autour de 0, RI dans $[0.83, 1]$. Notons que la qualité de la classification basée sur D_k est légèrement inférieure à celle qui est fondée sur CORT, révélant que les profils d'expression de gènes sont naturellement plus différenciés par leur forme que par leurs valeurs. Cette hypothèse est soutenue par les fortes valeurs de $k*$ (proche de 6, avec une variabilité de 0) obtenues dans la classification adaptative pour les quatre expériences (Tableau 2).
Considérons les résultats du classement, la figure 6 (gauche) montre que, pour les expériences 1 et 2, les quatre métriques sont toutes aussi efficaces, avec des taux d'erreurs de classement autour 0. Toutefois, pour les expériences 3 et 4, nous notons une forte augmentation du taux d'erreur pour les classements basés sur δ_E, une légère augmentation du taux d'erreur pour les classements fondés sur COR, une augmentation négligeable pour D_k. Le tableau 2 et la figure 6 (droite) illustrent la distribution des valeurs de $k*$ dans les classements adaptatifs. Pour les expériences 1 et 2, nous notons une distribution uniforme de $k*$ dans $[0, 6]$. Ce cas se présente lorsque un bon classement peut être obtenu avec une métrique fondée sur des valeurs ($k*$ proche de 0) et avec une métrique basée sur la forme ($k*$ proche de 6). En effet, dans les deux premières expériences, la figure 6 (gauche) montre que les quatre métriques sont toutes aussi efficaces pour le classement des gènes avec des taux d'erreur négligeables. Pour les expériences 3 et 4, $k*$ prend des valeurs plus élevées indiquant que les mesures fondées sur la forme (c-à-d CORT et D_k) sont les plus efficaces pour le classement des profils d'expression de gènes, avec de très faibles taux d'erreur (figure 6 (gauche)). Enfin, selon les résultats des quatre expériences, les mesures CORT et D_k peuvent être considérées comme les plus efficaces pour le classement des profils d'expression de gènes.

6. Conclusion

En conclusion, pour la classification ou le classement des gènes exprimés au cours du cycle cellulaire, il est souhaitable de considérer la corrélation tem-

porelle comme mesure de proximité. Toutefois, notons l'efficacité de la dissimilarité apprise D_k, qui fournit également une bonne classification et classement des gènes. La dissimilarité proposée D_k est particulièrement recommandée lorsque les instants d'observation ne doivent pas subir de décalage lors de l'évaluation des proximités (ce qui est le cas des profils d'expression de gènes du cycle cellulaire). Notons que le dissimilarité D_k généralise les métriques conventionnelles ; elle correspond à la corrélation temporelle pour k^* voisin de 6, à la distance euclidienne pour k^* voisin de 0 et plus généralement à une métrique couvrant à la fois des proximités portant sur les valeurs et sur les formes.

Références

ALTER O., BROWN P. & BOSTEIN D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. volume 97, p. 10101–10106.

ANAGNOSTOPOULOS A., VLACHOS M., HADJIELEFTHERIOU M., KEOGH E. & YU P. (2006). Global distance-based segmentation of trajectories. p. 34–43.

BAR-JOSEPH Z., GERBER G., GIFFORD D., JAAKKOLA T. & SIMON I. (2003). Continuous representations of time-series gene expression data. volume 10, p. 341–356.

CAIADO J., CRATO N. & PENA D. (2006). A periodogram-based metric for time series classification. volume 50, p. 2668–2684.

DOUZAL-CHOUAKRIA A., DIALLO A. & GIROUD F. (2009). Adaptive clustering for time series : application for identifying cell cycle expressed genes. volume 53, p. 1414–1426. Elsevier.

DOUZAL-CHOUAKRIA A., DIALLO A. & GIROUD F. (2010). A random-periods model for the comparison of a metrics efficiency to classify cell-cycle expressed genes. volume 31, p. 1601–1617. Elsevier.

EISEN M. & BROWN P. (1999). Dna arrays for analysis of gene expression. volume 303, p. 179–205.

GARCIA-ESCUDERO L. A. & GORDALIZA A. (2005). A proposal for robust curve clustering. volume 22, p. 185–201.

HECKMAN N. E. & ZAMAR R. (2000). Comparing the shapes of regression functions. volume 22, p. 135–144.

HOLTER N., MARITAN A., CIEPLAK M., FEDOROFF N. & BANAVAR J. (2001). Dynamic modeling of gene expression data. volume 98, p. 1693–1698.

JOHANSSON D., LINDGREN P. & BERGLUND A. (2003). A multivariate approach applied to microarray data for identification of genes with cell cycle-coupled transcription. volume 19, p. 467–473.

KAUFMAN L. & ROUSSEEUW P. (1990). Finding groups in data. an introduction to cluster analysis. New York : John Wiley and Sons.

KELLER K. & WITTFELD K. (2004). Distances of time series components by means of symbolic dynamics. volume 14, p. 693–704.

KRUSKALL J. & LIBERMAN M. (1983). The symmetric time warping algorithm : From continuous to discrete.

LIU D., UMBACH D., PEDDADA S., LI L., CROCKETT P. & WEINBERG C. (2004). A random-periods model for expression of cell-cycle genes. volume 101, p. 7240–7245.

LIU X., LEE S., CASELLA G. & PETER G. (2008). Assessing agreement of clustering methods with gene expression microarray data. volume 52, p. 5356–5366.

LUAN Y. & LI H. (2003). Clustering of time-course gene expression data using a mixed-effects model with b-splines. volume 19, p. 474–482.

PARK C., KOO J., KIM S., SOHN I. & LEE J. (2008). Classification of gene functions using support vector machine for time-course gene expression data. volume 52, p. 2578–2587.

RAMONI M. F., SEBASTIANI P. & KOHANE I. (2002). Cluster analysis of gene expression dynamics. volume 99, p. 9121–9126.

SCRUCCA L. (2007). Class prediction and gene selection for dna microarrays using regularized sliced inverse regression. volume 52, p. 438–451.

SHIEH J. & KEOGH E. (2008). isax : Indexing and mining terabyte sized time series. p. 623–631.

SPELLMAN P., SHERLOCK G., ZHANG M., IYER V., ANDERS K., EISEN M., BROWN P., BOTSTEIN D. & FUTCHER B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. volume 9, p. 3273–3297.

WHITFIELD M., SHERLOCK G., MURRAY J., BALL C., ALEXANDER K., MATESE J., PEROU C., HURT M., BROWN P. & BOTSTEIN D. (2002). Identification of genes periodically expressed in the human cell cycle and their expression in tumors molecular. volume 13, p. 1977–2000.