Classification adaptative de séries temporelles : application à l'identification des gènes exprimés au cours du cycle cellulaire

Alpha Diallo*,**, Ahlame Douzal-Chouakria* et Françoise Giroud**

*TIMC-IMAG TIMB(CNRS-UMR 5525), Université Joseph Fourier Grenoble 1
F-38706 La Tronche Cedex, France
(alpha.diallo, ahlame.douzal)@imag.fr
http://www-timc.imag.fr/Ahlame.Douzal/
**TIMC-IMAG RFMQ (CNRS-UMR 5525), Université Joseph Fourier Grenoble 1
F-38706 La Tronche Cedex, France
francoise.giroud@imag.fr
http://www-timc.imag.fr/Francoise.Giroud/index.html

Résumé. Ce travail s'inscrit dans le cadre de l'étude de la division cellulaire assurant la prolifération des cellules. Une meilleure compréhension de ce phénomène biologique nécessite l'identification des gènes caractérisant chaque phase du cycle cellulaire. Le procédé d'identification est généralement basé sur un ensemble de gènes dits gènes de référence, sélectionnés expérimentalement et considérés comme caractérisant les phases du cycle cellulaire. Les niveaux d'expression des gènes étudiés sont mesurés durant le cycle de la division cellulaire et permettent de construire des profils d'expression. Chaque gène étudié est affecté à la phase du cycle cellulaire correspondant au groupe de gènes de référence le plus similaire. Cette approche classique souffre de deux limites. D'une part, les mesures de proximité les plus couramment utilisées entre profils d'expression de gènes sont basées sur les écarts en valeurs sans tenir compte de la forme des profils. D'autre part, dans la littérature il n'y a pas consensus quant à l'ensemble des gènes de référence à considérer. Dans cet article, notre but est de proposer une classification adaptative, basée sur un indice de dissimilarité incluant les proximités en valeurs et en forme des profils d'expression de gènes, permettant d'identifier les phases d'expression des gènes étudiés, et de présenter un nouvel ensemble de gènes de référence validé par une connaissance biologique.

1 Introduction

Les puces à ADN permettent de mesurer simultanément le niveau d'expression de plusieurs milliers de gènes dans un type cellulaire et un contexte physiologique et/ou pathologique particulier. Ces mesures du transcriptome permettent notamment d'étudier la cinétique des phénomènes cellulaires comme le cycle cellulaire (Spellman et al., 1998) pour ainsi défi-

nir des groupes présentant des cinétiques semblables. Les algorithmes de classification se sont montrés particulièrement efficaces pour comprendre la caractérisation de la fonction des gènes et des voies de régulation leur permettant de réaliser les processus biologiques dans lesquels ils sont impliqués. La plupart des cellules de notre corps contiennent les mêmes gènes, mais tous n'interviennent pas dans chaque cellule : les gènes sont activés et 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 savoir comment les cellules réalisent une telle spécialisation, les scientifiques ont besoin d'identifier quels gènes s'expriment dans chaque type de cellule. La technologie des puces à ADN nous permet maintenant de visualiser simultanément de nombreux gènes et de déterminer ceux qui sont exprimés dans un type de cellule spécifique par mesure du transcriptome (ensemble des ARN messagers transcrits) d'une cellule, reflet de la fonction particulière d'une cellule (Eisen et Brown, 1999). Les chercheurs utilisent cette technologie haut-débit pour détecter des gènes qui sont exprimés ou non exprimés dans des tissus humains sains en comparaison avec des tissus malades par exemple. Les gènes qui sont exprimés différemment dans deux tissus peuvent être impliqués dans la cause de la maladie. Dans cet article, nous nous intéressons à la progression dynamique du cycle de la division cellulaire à travers quatre phases distinctes G_1, S, G_2 et M. Les niveaux d'expression d'un ensemble de gènes étudiés sont alors observés à des moments spécifiques durant le cycle de la division cellulaire. À ce jour, l'identification de l'ensemble des gènes caractérisant chaque phase du cycle cellulaire est a priori basée sur un ensemble de gènes de référence. Chaque gène étudié est affecté à la phase du cycle cellulaire correspondant au groupe de gènes de référence le plus similaire. Cette approche classique souffre de deux limites. D'une part, les mesures de proximité les plus couramment utilisées entre profils d'expression de gènes sont basées sur les écarts en valeurs sans tenir compte de la forme des profils. D'autre part, dans la littérature il n'y a pas consensus quant à l'ensemble des gènes de référence à considérer. Dans cet article, notre but est de proposer une classification adaptative, basée sur un indice de dissimilarité incluant les proximités en valeurs et en forme des profils d'expression de gènes, permettant d'identifier les phases d'expression des gènes étudiés. Ensuite, proposer un nouvel ensemble de gènes de référence validé par une connaissance biologique. La suite de l'article s'articule comme suit : la section suivante définit ce que sont des données d'expression de gènes et présente le problème biologique abordé. La section 3 rappelle la définition et les propriétés de l'indice de dissimilarité utilisé. La section 4 présente l'application de l'approche adaptative proposée à l'étude de la prolifération des cellules humaines "Hela". Enfin, dans la section 5 nous procédons à l'analyse comparative et à la discussion des principaux résultats obtenus.

2 Identification des gènes exprimés du cycle cellulaire

Le problème biologique qui nous préoccupe 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 pour la prolifération des cellules et se décompose en quatre phases principales. Elle 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) qui est suivie par la phase G_2 pendant laquelle la cellule se prépare à la phase M pour achever sa division (séparation en deux cellules filles). Pendant ces quatre phases, certains gènes s'expriment ou

pas, et un but important consiste à identifier les gènes fortement exprimés et caractérisant chaque phase du cycle cellulaire. Pour cela, 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). Les puces à ADN offrent de nombreuses perspectives. Leur principale application est l'étude du niveau d'expression des gènes et les mécanismes génétiques qui leur sont associés au niveau cellulaire. De nombreuses études ont notamment été réalisées pour étudier la cinétique des phénomènes cellulaires comme la différenciation ou le cycle cellulaire (Spellman et al., 1998). Grâce à cette technologie, on mesure le niveau d'expression de chaque gène à des moments spécifiques du cycle de la division cellulaire en échantillonnant au cours du temps une population cellulaire initialement synchronisée. Chaque gène étudié peut alors être décrit par son profil d'expression observé au cours du temps sur un ou plusieurs cycles de la division cellulaire. La figure 1 montre l'expression du gène CCNE1 observé au cours des trois premiers cycles cellulaires après synchronisation de la population cellulaire.

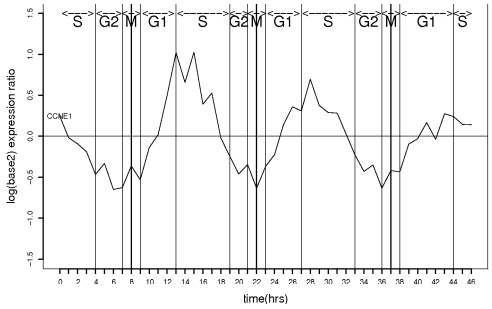

FIG. 1 – *Profil d'expression du gène CCNE1 observé sur une période de 46 heures après synchronisation correspondant à trois cycles cellulaires. Les traits verticaux gras représentent les mitoses des trois cycles. Chaque phase de cycle est délimitée par les traits verticaux et annotée par* G_1, S, G_2 *ou* M.

3 Mesure de proximité entre profils d'expression de gènes

Pour la classification d'un ensemble de profils d'expression de gènes évoluant dans le temps, le choix de la distance est crucial puisqu'il définit la mesure de ressemblance entre les profils de deux gènes. Considérons les niveaux d'expression de deux gènes $g_1 = (u_1, ..., u_p)$ et $g_2 = (v_1, ..., v_p)$ observés aux instants $(t_1, ..., t_p)$. La distance euclidienne δ_E, la plus fréquemment utilisée, entre g_1 et g_2 est définie par : $\delta_E(g_1, g_2) = \left(\sum_{i=1}^{p}(u_i - v_i)^2\right)^{\frac{1}{2}}$. Il ressort de cette définition, que la proximité dépend uniquement de l'écart entre les valeurs d'expression sans tenir compte de la forme des profils d'expression. En d'autres termes, deux profils d'expression de gènes sont dits proches au sens de δ_E si et seulement si les valeurs observées aux mêmes instants sont proches. Cette distance ignore l'information de dépendance entre les valeurs d'expression, elle est invariante à toutes permutations des instants d'observations. En réponse à ces limites, nous proposons d'utiliser un indice de dissimilarité couvrant la mesure

de proximité en valeurs et en forme des expressions de gènes proposé dans Douzal Chouakria et Nagabhushan (2007). Une étape préalable consiste à préciser la notion de proximité entre profils d'expression de gènes que tend à quantifier chaque mesure de distance, et indiquer les caractéristiques principales que doit vérifier cette dissimilarité .

3.1 Mesures de proximité entre formes

La proximité entre deux profils d'expression de gènes, fondée sur la forme, dépend de deux propriétés : la monotonicité mesurant la dépendance entre les tendances suivies à des périodes particulières et la proximité des taux d'accroissement observés. Sans perte de généralité, supposons que les valeurs de g_1 et g_2 évoluent dans $[0, D]$. g_1 et g_2 sont dits de formes similaires si à chaque période d'observation $[t_i, t_{i+1}]$, ils croissent ou décroissent simultanément (monotonicité), avec un taux d'accroissement égal. Ce concept de similarité peut être quantifié en considérant le coefficient de corrélation de Pearson classique, cependant cette corrélation mène à une surestimation dans le cas de données temporelles dépendantes. Pour plus de détails concernant les limites de la corrélation ainsi que des approches alternatives voir Douzal Chouakria et Nagabhushan (2007). Pour mesurer la proximité entre formes, nous proposons d'utiliser le coefficient de corrélation temporelle suivant :

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

$\text{CORT}(g_1, g_2)$ appartient à l'intervalle $[-1, 1]$. La valeur $\text{CORT}(g_1, g_2) = 1$ signifie que dans chaque période d'observation $[t_i, t_{i+1}]$, les expressions des gènes g_1 et g_2 croient ou décroient simultanément avec le même taux d'accroissement (formes similaires). Une valeur de $\text{CORT}(g_1, g_2) = -1$ exprime que dans chaque période d'observation $[t_i, t_{i+1}]$ g_1 croit, g_2 décroit ou vice-versa avec un même taux d'accroissement en valeur absolue (formes opposées). Enfin une valeur de $\text{CORT}(g_1, g_2) = 0$ signifie une absence de monotonicité entre les accroissements de g_1 et g_2 et leurs taux d'accroissement sont stochastiquement linéairement indépendants (formes différentes). Une étude détaillée de la corrélation temporelle est proposée par Chouakria Douzal (2003).

3.2 Indice de dissimilarité pour les profils d'expression de gènes

Le but est de fournir un indice de dissimilarité qui couvre la distance euclidienne δ_E et la corrélation temporelle CORT. Cet indice de dissimilarité devra moduler la proximité en valeurs en fonction de la proximité en forme. La fonction de modulation devra augmenter la proximité en valeurs à mesure que les formes sont opposées (la corrélation temporelle décroit de 0 à -1). À l'inverse, elle diminuera la proximité en valeurs à mesure que les formes sont similaires (la corrélation temporelle évolue de 0 à +1). La dissimilarité résultante correspond à la distance euclidienne si les formes sont différentes (corrélation temporelle nulle). Tenant compte de ces propriétés, nous proposons d'utiliser l'indice de dissimilarité D_k défini comme suit :

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

où $f(x)$ est une fonction de réglage exponentielle :

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

La figure 2 montre l'effet du réglage en fonction du paramètre k. Dans le cas de gènes de

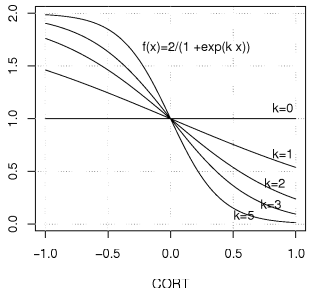

FIG. 2 – *L'effet du réglage en fonction de k.*

différentes formes (CORT $\rightsquigarrow$ 0), $f(x)$ est voisin de 1 quelle que soit la valeur du paramètre k et D_k tend vers δ_E. Dans le cas où CORT est différent de 0 (formes non différentes), le paramètre k module les contributions de la proximité en valeurs et en forme dans l'indice de dissimilarité D_k. La contribution de la proximité en forme $1 - 2/(1 + exp(k\ |\text{CORT}|))$ augmente quand k augmente et celle de la proximité en valeurs $2/(1 + exp(k\ |\text{CORT}|))$ diminue. Par exemple, pour $k = 0$ et |CORT| = 1, 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 |CORT| = 1, la proximité en forme contribue 76.2% à D_k tandis que celle en valeurs contribue 23.8% (23.8% de la valeur de D_k sont déterminés par δ_E et les 76.2% restantes par CORT). Le tableau 1 résume, dans le cas de formes similaires ou opposées (|CORT|=1), les contributions en formes et en valeurs à D_k. Remarquons que si $k = 0$, D_k pourrait être considéré comme

	Contribution en formes (%)	Contribution en valeurs (%)
$k = 0$	0	100
$k = 1$	46.2	53.7
$k = 2$	76.2	23.8
$k = 3$	90.5	9.4
$k \geq 5$	$\rightsquigarrow$ 100	$\rightsquigarrow$ 0

TAB. 1 – *Contribution de la proximité en valeurs et en formes à D_k fonction de k.*

une extension de δ_E aux mesures de proximité en valeurs et en forme. On note que si δ_E s'approche de 0 (i.e., les expressions de gènes sont proches en valeurs), CORT s'approche de 1 (i.e. les profils d'expression des gènes sont similaires en forme) alors D_k s'approche de 0. Nous pouvons vérifier que D_k vérifie les propriétés d'identité et de symétrie de la distance, mais pas d'inégalité triangulaire.

4 Classification non supervisée adaptative pour l'identification de gènes du cycle cellulaire

4.1 Description des données : cellule humaine Hela

Le cycle cellulaire, ou cycle de la division cellulaire, est la série d'événements entre une division cellulaire et la suivante. Le cycle cellulaire consiste en quatre phases successives : les phases G_1, S (Synthèse d'ADN ou réplication d'ADN), G_2 et M. Un système de surveillance moléculaire contrôle la progression des cellules pendant le cycle cellulaire. Entre ces différentes étapes, se situent des points de contrôle, qui ont pour but de vérifier l'intégrité de la transmission de l'ADN de la cellule mère vers les cellules filles. Ces points de restriction marquent les transitions d'interphases, G_1/S est la première d'entre elles. L'analyse des données d'expression des gènes pendant le cycle de la division cellulaire vise à déterminer ceux qui sont bien exprimés au cours des différentes phases du cycle cellulaire (Spellman et al. (1998), Oliva et al. (2005), Cho et al. (2001)). Étudier le transcriptome de la prolifération des cellules synchronisées mène à la construction de profils d'expression de gènes au cours du temps, c-à-d durant la progression du cycle cellulaire. Dans ce travail, nous nous limitons à l'analyse des données transcriptomiques concernant la prolifération des cellules humaines Hela publiée par Whitfield et al. (2002) (http ://genome-www.standford.edu/Human-CellCycle/Hela/). De manière plus spécifique, notre étude se concentrera sur les données, enregistrées dans la troisième expérimentation de l'application Hela. Seuls les 1099 gènes détectés comme présentant un évènement cyclique sont considérés dans notre étude. Ils sont décrits par leurs niveaux d'expression, pendant la progression du cycle cellulaire, tout au long des 46 heures qui suivent la synchronisation des cellules.

4.2 Identification conventionnelle des gènes du cycle cellulaire

Illustrons l'approche proposée par Whitfield et al. (2002) pour identifier les gènes du cycle cellulaire de l'application Hela. Les auteurs considèrent un ensemble de 20 gènes caractérisant les 5 phases et transitions du cycle cellulaire suivantes : G_1, G_1/S, G_2, G_2/M et M/G_1, nommées "phases" dans la suite de l'article. L'ensemble des 20 gènes est composé de 5 classes regroupant chacune 4 gènes de référence (tableau 2). La figure 3 montre, pour chaque phase du cycle cellulaire, les profils d'expression des 4 gènes de référence. Des auteurs ont discuté sur le choix des 20 gènes de référence par leurs pics d'expression dans chaque phase du cycle cellulaire. Chacun des 1099 gènes étudiés est affecté à la phase du cycle cellulaire correspondant au groupe de gènes de référence le plus similaire. La similarité utilisée est basée sur les valeurs d'expression sans tenir compte de la forme des profils. Si nous observons en détail les profils des 20 gènes sur la figure 3 nous trouvons quelques contradictions. Premièrement, les profils des gènes de référence CDC2, CCNF, CCNA2 caractérisant la phase G_2 ne culminent pas à la phase G_2 mais plutôt à la phase G_2/M. De même, les gènes de référence BUB1 et PLK de la phase G_2/M culminent à la phase M/G_1 au lieu de la phase G_2/M. Ces observations sont soutenues par les annotations de la base de données de Genecards (http ://www.genecards.org/) et la base de données de la voie moléculaire KEGG (http ://www.genome.ad.jp/kegg/kegg2.html).

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

TAB. 2 – *Les 20 gènes de référence de Whitfield et al.*

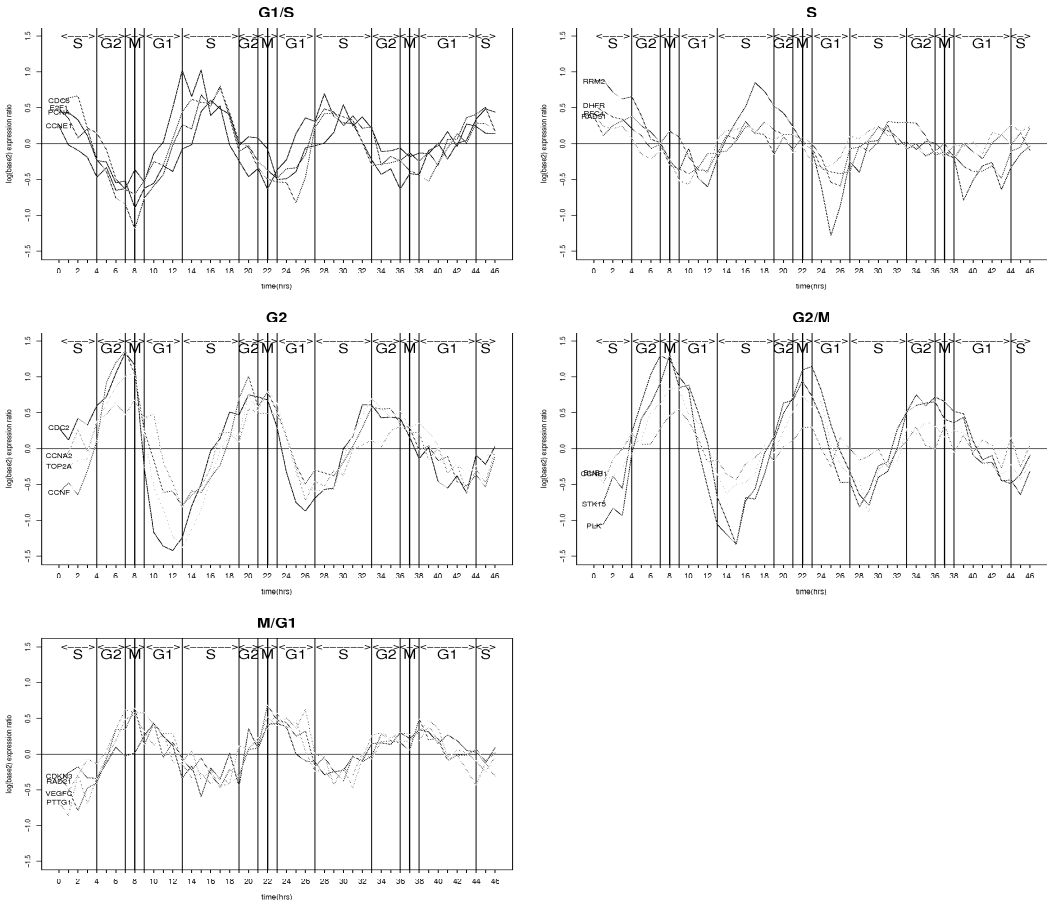

FIG. 3 – *Profils des 20 gènes de référence de Whitfield et al. dont les expressions culminent dans chacune des phases suivantes du cycle cellulaire : G_1/S, S, G_2, G_2/M et M/G_1.*

4.3 Classification non supervisée adaptative pour l'identification des phases du cycle cellulaire des gènes

Nous récapitulons le but de la classification non supervisée adaptative. Il permet d'extraire un ensemble de gènes caractérisant bien les phases du cycle cellulaire. L'approche de la classi-

fication est basée sur un indice de dissimilarité couvrant la proximité en valeurs et en forme. La classification adaptative aide à apprendre la contribution appropriée de la proximité en valeurs et en forme à l'indice de dissimilarité D_k.

Nous proposons d'utiliser l'algorithme PAM (Patitioning Around Medoids) pour partitionner l'ensemble des gènes étudiés en n classes (n étant le nombre de phases du cycle cellulaire 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. Il permet une analyse plus détaillée de la partition en fournissant des indices permettant d'apprécier la qualité des classes ainsi que des individus en mesurant leur valeur silhouette s_g définie comme suit :

$$s_g \quad = \quad \frac{b_g - a_g}{max(a_g, b_g)} \in [-1, 1]$$

où a_g représente la dissimilarité moyenne entre le gène g et les gènes de la même classe, b_g représente la dissimilarité moyenne entre le gène g et les gènes de la classe la plus proche (voisine du gène g). Pour une valeur de s_g proche de 1, le gène g est "bien classé" (bon représentant de la classe). Quand la valeur de s_g est voisine de 0, le gène g est dit frontalier (appartient aussi bien à sa classe d'appartenance qu'à la classe voisine). Enfin le gène g est "mal classé" si s_g est proche de -1. La largeur moyenne de la silhouette d'une classe est définie comme la moyenne des valeurs de la silhouette de tous les individus de la classe et la largeur moyenne de la silhouette (*lms*) est définie comme la moyenne des valeurs de la silhouette de tous les individus. Nous avons utilisé la *lms* pour estimer la qualité d'une partition. PAM fournit aussi un dispositif graphique représentant les silhouettes et permettant de comparer la qualité des classes. Pour plus de détails sur l'algorithme PAM voir Kaufman et Rousseeuw (1990).

Pour apprendre l'indice de dissimilarité le plus approprié pour nos données d'expression de gènes, nous exécutons l'algorithme PAM sur l'ensemble des 1099 gènes décrits précédemment pour plusieurs valeurs du paramètre k (k=0,...,6 avec un pas égal à 0.01). Soit $k*$ la valeur maximisant la lms et P_{k*} la partition correspondante. La valeur de $k*$ fournit la meilleure contribution des proximités en valeurs et en forme à l'indice de dissimilarité, et par conséquent l'indice de dissimilarité D_{k*} à utiliser pour les étapes suivantes. La seconde étape consiste à choisir, pour chaque classe, un ensemble de gènes noyau. Dans la littérature, nous avons trouvé 43 gènes (environ 10 gènes par phase) identifiés comme impliqués dans le processus de la division du cycle cellulaire (Whitfield et al., 2002). Pour cette raison nous avons extrait, de chaque classe de la partition P_{k*}, un ensemble noyau formé des 10 gènes ayant les valeurs de s_g les plus fortes et qui sont les 10 mieux classés de la classe. La figure 4 est le graphe de la silhouette associée. Nous visualisons sur la figure 5 les profils d'expression des gènes noyau et déterminons la phase du cycle cellulaire où culminent les expressions des gènes. L'observation de la progression des gènes noyau durant le cycle de la division cellulaire révèle que : les expressions des gènes noyau de la classe 1 culminent clairement à la phase S, celles de la classe 2 à la phase G_1/S, celles de la classe 3 à la phase G_2/M, celles de la classe 4 à la phase M/G_1 et finalement celles de la classe 5 à la phase G_1. Le tableau 3 donne pour chaque classe l'ensemble des gènes noyau (Type de Gène = K). Nous indiquons pour chaque gène noyau son nom, sa phase d'affectation par Whitfield et al. (2002), le numéro de la classe voisine et sa valeur s_g. Nous indiquons aussi l'ensemble des gènes de référence de Whitfield (tableau 2) appartenant à chaque classe (Type de Gène = R). Remarquons que dû à la désynchronisation des cellules, il est plus fiable de limiter nos interprétations aux premiers cycles cellulaires. Par

conséquent, chaque classe est affectée à la phase du cycle cellulaire de son ensemble noyau et chaque gène restant est affecté à la phase du cycle cellulaire de sa classe d'appartenance.

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

TAB. 3 – *Les 50 gènes noyaux (Type de Gène = K) caractérisant les phases : S, G_1/S, G_2/M, M/G_1 et G_1 avec la classification des 20 gènes de référence de Whitfield (Type de Gène = R) dans les 5 classes obtenues.*

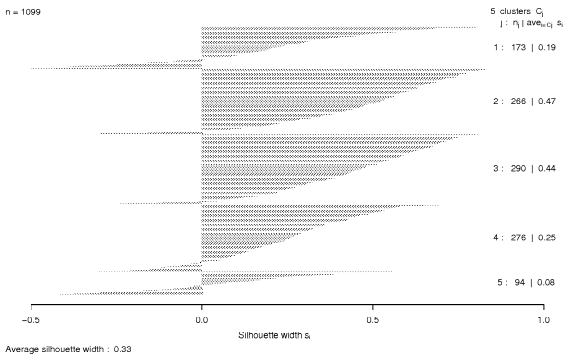

FIG. 4 – *Graphe de la silhouette de* $P_{k*=5.9}$

5 Analyse comparative et discussion

La partition optimale P_{k*} maximisant la lms est obtenue pour k*=5.9. Une valeur signifiant que les 5 principaux modèles de profils d'expression de gènes sont essentiellement distincts de par leurs formes (tableau 1). La figure 4 montre une lms de 0.33 qui indique que la structure de classe obtenue est raisonnable. Cependant, si on se limite aux 50 gènes des noyaux, on constate que la lms est égale à 0.67, ceci montre que les ensembles noyaux sont bien séparés les uns des autres. La figure 4 indique que la classe 2 (G_1/S) possède le plus grand coefficient de lms, par conséquent elle est bien séparée des autres. Par contre, avec une plus petite lms de 0.08, la classe 5 (G_1) n'est pas clairement séparées des autres classes de la partition.
On note que les gènes de référence CCNE1, CCNA2 et CCNB1 connus en tant que cyclines mitotiques, classés respectivement dans les phases G_1/S, G_2/M et M/G_1 apparaissent dans l'ordre temporel biologique attendu pendant le cycle de la division cellulaire (G_1, S, G_2 et M). On peut aussi noter que le gène E2F1, facteur de transcription connu comme un régulateur clé de la progression du cycle cellulaire impliqué dans le contrôle de la progression du cycle cellulaire de G_1 à S, est classé dans la phase G_1/S. Les gènes CCNE1 (s_g=0.786) et MCM6 (s_g=0.812) connus comme respectivement «activé» et «induit» par E2F1, sont également classés dans G_1/S. En accord avec les expériences qui ont montré que le gène CCNA2 favorise la transition G_2/M, notre approche classe bien CCNA2 dans la phase G_2/M (s_g=0.458), alors qu'il a été choisi comme gène de référence de la phase G_2 par Whitfield et al. (2002). Le gène UBE2C (s_g=0.779) appartenant à la classe G_2/M est bien évalué par la connaissance biologique : il représente une enzyme d'ubiquitination régulant la destruction des cyclines mitotiques en fin de mitose (transition G_2/M). Enfin, nous notons que tous les gènes de référence de Whitfield et al. marquant la phase G_2 sont affectés dans la classe G_2/M sauf le gène STK15, tous les gènes de référence marquant la phase G_2/M sont affectés dans la classe M/G_1. Ce qui corrobore clairement avec les contradictions discutées dans le paragraphe 4.2.

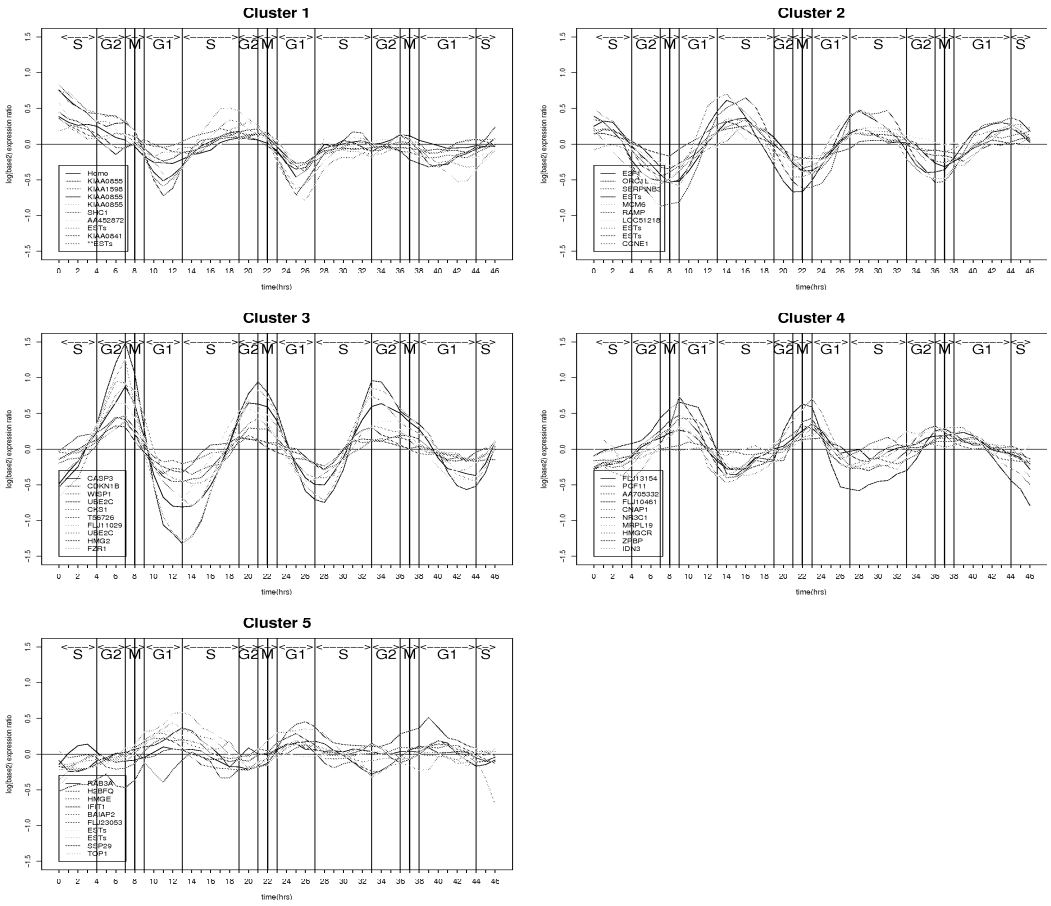

FIG. 5 – *Les profils d'expression des gènes noyau durant le cycle de la division cellualire.*

6 Conclusion

Dans cet article, nous présentons une méthode concurrente pour l'identification des gènes du cycle cellulaire. Cette méthode est fondée sur une classification adaptative basée sur un indice de dissimilarité pour l'analyse des profils d'expression de gènes incluant la proximité liée aux valeurs et aux formes. Cette procédure nous a permis d'abord d'identifier les phases du cycle cellulaire des gènes étudiés et enfin de proposer un nouvel ensemble de gènes de référence validé par une connaissance biologique publiée.

Références

Cho, R., M. Huang, M. Campbell, H. Dong, L. Steinmetz, L. Sapinoso, G. Hampton, S. Elledge, R. Davis, et D. Lockhart (2001). Transcriptional regulation and function during

the human cell cycle. *Nature Genetics 27*, 48–54.

Chouakria Douzal, A. (2003). Compression technique preserving correlations of a multivariate temporal sequence. In M. Berthold, H. Lenz, E. Bradley, R. Kruse, et C. Borgelt (Eds.), *Advances in Intelligent Data Analysis*, Volume V, pp. 566–577. Springer.

Douzal Chouakria, A. et P. Nagabhushan (2007). Adaptive dissimilarity index for measuring time series proximity. *Advances in Data Analysis and Classification Journal 1*, 5–21. Springer Berlin / Heidelberg.

Eisen, M. et P. Brown (1999). Dna arrays for analysis of gene expression. *Methods Enzymol 303*, 179–205.

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

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

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

Whitfield, M., G. Sherlock, J. Murray, C. Ball, K. Alexander, J. Matese, C. Perou, M. Hurt, P. Brown, et D. Botstein (2002). Identification of genes periodically expressed in the human cell cycle and their expression in tumors molecular. *Biology of the Cell 13*, 1977–2000.

Summary

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