

Classification de profils d'expression de gènes: application à l'étude de la régulation du cycle cellulaire chez les eucaryotes.

Alpha Diallo

▶ **To cite this version:**

Alpha Diallo. Classification de profils d'expression de gènes: application à l'étude de la régulation du cycle cellulaire chez les eucaryotes.. Sciences du Vivant [q-bio]. Université de Grenoble, 2010. Français. NNT : . tel-00516398v2

HAL Id: tel-00516398

https://theses.hal.science/tel-00516398v2

Submitted on 13 Sep 2010

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

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

Ecole Doctorale Ingénierie pour la Santé, la Cognition et l'Environnement (EDISCE)

Classification de profils d'expression de gènes :

application à l'étude de la régulation du cycle cellulaire chez les eucaryotes.

THÈSE

présentée et soutenue publiquement le 03 Juin 2010

pour l'obtention du

Doctorat de l'Université de Grenoble

Spécialité modèles, méthodes et algorithmes en biologie, santé et environnement

par

Alpha DIALLO

Directeur de thèse : Françoise GIROUD
Co-directeur de thèse : Ahlame DOUZAL-CHOUAKRIA

Composition du jury

Jean-Francois BOULICAUT	Président
Gérard GOVAERT	Rapporteur
Mohamed NADIF	Rapporteur
Françoise GIROUD	Examinateur
Ahlame DOUZAL-CHOUAKRIA	Examinateur

Laboratoire TIMC-IMAG — Université Joseph Fourier UMR 5525

Mis en page avec la classe thloria.

Remerciements

L'ensemble des travaux présentés dans ce rapport de thèse a été effectué au Laboratoire TIMC-IMAG, au sein des équipes TIMB (Traitement de l'Information et Modélisation en Bio-médecine) et RFMQ (Reconnaissance des Formes et de Microscopie Quantitative).

Je tiens tout particulièrement à remercier mes deux directrices de thèse, Mesdame Ahlame DOUZAL et Françoise GIROUD pour leur gentillesse et leur disponibilité. Elles n'ont ménagé aucun effort pour la réussite de cette thèse. Je n'aurais jamais pu me lancer sans elles dans l'aventure qu'a constitué cette thèse.

J'exprime ma profonde gratitude à Madame Ahlame DOUZAL, en sa qualalité de directrice de thèse, qui m'a guidé vers une plus grande maturité et une meilleure vision de la recherche, pour sa patience et son encouragement constant. Son investissement a été sans limite. Elle a toujours été là même les jours de plus grand doute.

Je tiens à faire part également de ma grande reconnaissance à l'égard de Madame Françoise GIROUD qui m'a guidé dans le monde de la biologie. Ses analyses, ses remarques pertinentes et ses nombreuses relectures m'ont permis d'améliorer ce rapport.

J'adresse mes sincères remerciements à Monsieur Gerard GOVAERT, Professeur à l'Université de Technologie de Compiègne et à Monsieur Mohamed NADIF, Professeur à l'Université Paris Descartes, pour l'intêret qu'ils ont bien voulu porter à ce travail en acceptant d'en être les rapporteurs.

Je remercie également Monsieur Jean-François BOULICAUT, Professeur à l'INSA Lyon d'avoir accepté de participer au jury de thèse.

Je remercie aussi tous les membres des équipes RFMQ et TIMB ainsi que du laboratoire TIMC-IMAG pour m'avoir aussi bien accueilli. Je tiens aussi à remercier mon collègue Cédric FRAMBOURG pour les discussions scientifiques qui m'ont éclairé sur certains aspects.

J'exprime ma reconnaissance à mes parents, mes frères et sœurs pour leur soutien inconditionnel.

Je dédie cette thèse
à ma mère.
qui fut la première de toutes.

Table des matières

Introduction générale **ix**

Table des figures **xiii**

Liste des tableaux **xvii**

Chapitre 1 Eléments de biologie et puces à ADN **1**

1.1 La cellule 1
1.2 Le cycle cellulaire 3
1.3 Le transcriptome 5
1.4 Les puces à ADN pour l'étude du transcriptome 6
1.4.1 Les sondes 7
1.4.2 Les cibles et l'hybridation 7
1.4.3 Acquisition des données 7
1.4.4 Normalisation des données et pré-traitement 8
1.5 Analyse des données du transcriptome 12

Chapitre 2 Classification des profils temporels d'expression de gènes : état de l'art **15**

2.1 Profils temporels d'expression de gènes 16
2.2 Techniques de classification non-paramétrique 18
2.2.1 Classification hiérarchique 18
2.2.2 Partitionnement par K-means 19
2.2.3 Classification par les réseaux de Kohonen (SOM) 20
2.2.4 Classification selon la théorie des graphes 21
2.2.5 Critères de validation du nombre de classes 22

2.3 Techniques de classification paramétrique 25
2.3.1 Classification par mélange de modèles gaussiens 25
2.3.2 Classification par modélisation des profils temporels . . . 26
2.4 Conclusion 29

Chapitre 3 Mesures de proximité entre des séries temporelles 31
3.1 Caractéristiques et prétraitements des séries temporelles 32
3.1.1 Régularité des séries temporelles 32
3.1.2 Séries temporelles de longueurs différentes 32
3.1.3 Décalage temporel 33
3.2 Mesures de proximités entre des séries temporelles 33
3.2.1 Notations pour une formalisation unifiée 34
3.2.2 Mesures de proximité fondées sur les valeurs 34
3.2.3 Mesures de proximité fondées sur la forme 37
3.2.4 Proximité entre formes : approches concurrentes 40
3.2.5 Mesures de proximité alliant forme et valeurs 44
3.3 Caractéristiques des profils d'expression de gènes au cours du cycle cellulaire 46
3.4 Mesures de proximité entre des profils d'expression de gènes . . 47
3.5 Conclusion 49

Chapitre 4 Analyse des gènes au cours du cycle cellulaire 51
4.1 Intêret de l'étude du cycle cellulaire 51
4.2 Approche conventionnelle 54
4.2.1 Classement des gènes cycliques 54
4.2.2 Recherche de profils types 55
4.3 Approche adaptative 55
4.4 Applications et analyses comparatives 56
4.4.1 Modèle aléatoire pour la comparaison de métriques . . . 57
4.4.2 Gènes périodiquement exprimés dans les cellules fibroblastiques humaines HeLa [106] 65
4.4.3 Gènes périodiquement exprimés dans les cellules du cancer du sein MCF-7 [63] 77
4.5 Conclusion 82

Chapitre 5 Conclusion **85**

Bibliographie **87**

Liste des publications **97**

Introduction générale

Avec le développement rapide de la technologie des puces à ADN depuis quelques années, il est aujourd'hui possible d'étudier simultanément l'expression de plusieurs milliers de gènes. Les données d'expression provenant de cette nouvelle technologie sont observées sous différentes conditions expérimentales. Nous distinguons principalement deux types de données d'expression. Le premier type de données porte sur l'expression d'un ensemble de gènes observés au travers d'un ensemble de conditions expérimentales. Ces conditions pouvant, par exemple, définir différents types de cellules, différents tissus (sains vs cancéreux). Le second type de données concerne l'évolution de l'expression des gènes au cours de processus biologiques, ce sont des données temporelles d'expression de gènes. Ces données sont généralement analysées pour des directions différentes. Certaines données d'expression de gènes sont spécifiques au cancer. Elles peuvent être utilisées pour inférer les gènes liés au cancer, afin d'identifier différents types de cancers sur la base de ces gènes (Liu [61]). D'autres données d'expression sont utilisées pour prédire les gènes impliqués dans certaines voies de signalisation cellulaire (Novak et Jain [72]). D'autres encore sont des données d'expression de gènes du cycle cellulaire, elles sont utilisées pour détecter les gènes impliqués dans les processus de régulation des différentes étapes du cycle cellulaire (Cho et al. [16], Spellman et al. [94]).

La grande quantité de gènes observés, combinée à la complexité des réseaux biologiques posent un vrai challenge pour la compréhension et l'interprétation des masses de résultats obtenus. Une étape préalable à l'exploitation des données produites par une expérience de puces à ADN est l'utilisation de techniques de classification. Elles permettent notamment de prédire la fonction de certains gènes ou d'extraire des profils types d'expression de gènes.

Des données d'expression de gènes, obtenues à partir de conditions physiologiques réalisées indépendamment, ont été étudiées par Hughes et al. [43]. Ces auteurs ont analysé des profils d'expression du génome de la levure *Saccharomyces cerevisea* dans 300 conditions différentes. Leurs travaux ont permis de regrouper les gènes de la levure en classes fonctionnelles et en particulier d'identifier de nouvelles protéines jouant un rôle dans la mitochondrie. Lorsque les conditions expérimentales ne sont plus indépendantes dans le sens où elles correspondent à la chronologie d'un évènement biologique (*e.g.* le cycle de la

division cellulaire), on a des données de profils temporels d'expression de gènes. Par hypothèse, les gènes de profils temporels similaires sont impliqués dans les mêmes processus biologiques. Ainsi, la classification non supervisée des profils d'expression de gènes pourrait permettre, entre autres, d'identifier la fonction de nouveaux gènes sur la base des fonctions déjà identifiées pour d'autres gènes. Les mesures de proximité classiques, telles que la distance euclidienne, la distance de Manhattan ou celle fondée sur le coefficient de corrélation de Pearson, souvent utilisées dans le contexte des profils temporels d'expression de gènes, ne semblent pas appropriées puisque leur utilisation suppose que les données observées pour chaque gène sont indépendantes entre elles. Le principal challenge pour l'évaluation de la proximité entre profils temporels de gène est alors l'intégration de l'information de dépendance temporelle. Afin de pallier ce problème nous nous intéresserons donc au développement de mesures de proximité entre profils temporels d'expression de gènes et démontrons leur efficacité par rapport aux méthodes plus conventionnelles. Ce travail sera réalisé dans le cadre de l'analyse, la classification et l'interprétation de profils d'expressions de gènes durant le processus de division cellulaire, et sera développé en 4 grandes parties.

Dans le chapitre 1, nous saisissons quelques notions et vocabulaires utiles, sur la biologie moléculaire, afin de mieux aborder la suite de notre étude. Nous présentons les différentes étapes permettant d'obtenir des données d'expression de gènes par les expériences des puces à ADN. Ces données sont toujours très bruitées, et pas assez fiables en raison de nombreux facteurs qui affectent les expériences. Une normalisation des données est donc nécessaire. Nous présentons les méthodes de normalisation les plus récentes et les plus utilisées dans ce domaine.

Le chapitre 2 dresse un état de l'art des techniques de classification les plus utilisées pour l'analyse des données du transcriptome. L'apprentissage non supervisé est la méthode la plus populaire actuellement utilisée dans la première étape d'analyse des expressions de gènes. De nombreuses techniques de classification non supervisée sont utilisées. D'une part, nous distinguons les techniques de classification non-paramétriques, basées essentiellement sur des mesures de proximité entre paires de gènes. Les algorithmes de classification hiérarchique, de partitionnement et ceux portant sur la théorie des graphes sont employés. D'autres part, au cours de ces dernières années, nous assistons à l'utilisation de plus en plus fréquente des techniques de classification paramétriques, modélisant les profils temporels d'expression de gènes (Bar-Joseph et al. [5], Wu et al. [108], Garcias-Frias [116]). L'un des principaux moteurs de cette approche est l'hypothèse que les profils temporels d'expression de gènes sont générés par des modèles dependant du temps. Ces méthodes tentent de quantifier cette hypothèse.

Les méthodes de classification non paramétrique se fondent souvent sur des

mesures de proximité. Cependant la pertinence de ces mesures est très discutable : soit elles ne mesurent pas la proximité en formes entre deux profils d'expression de gènes donnés (distance euclidienne), soit elles peuvent mal évaluer la proximité en forme (coefficient de corrélation de Pearson). Tenant compte de la structure temporelle des données d'expression, nous avons étudié dans le troisième chapitre trois familles de mesures de proximité. La première famille définit des mesures limitées à la comparaison des valeurs des expressions en ignorant la contrainte de dépendance temporelle des données. La seconde famille se limite à la comparaison des formes des profils de gènes. Bien que l'analyse des données d'expression de gènes s'intéresse principalement à la forme des profils de gènes, il nous semble intéressant de considérer simultanément les informations portant sur les écarts en valeurs et en forme des profils. Dans ce contexte, nous proposons et développons une troisième famille de mesures couvrant simultanément les aspects formes et valeurs des profils temporels d'expression de gènes. Ces mesures dépendent d'un facteur définissant les contributions relatives des composantes valeurs et forme. Une formalisation unifiée de ces mesures est proposée.

L'identification et la caractérisation des gènes impliqués dans les phases du cycle cellulaire sont souvent effectuées par des approches classiques. Ces approches, souvent supervisées, considèrent un ensemble de gènes de référence connus qui sont supposés caractéristiques des phases ou transitions de phase du cycle cellulaire considérées, et utilisent des mesures de proximité non justifiées. Au chapitre 4, nous développons une approche de classification non-supervisée adaptative dont le but est d'apprendre la mesure de proximité à utiliser pour les profils temporels d'expression de gènes, et de proposer un nouvel ensemble de gènes de référence validé biologiquement. Nous comparerons notre approche adaptative avec d'autres approches classiques sur un jeu de données simulées et sur deux jeux de données d'expression de gènes du cycle cellulaire réelles.

Table des figures

1.1 Configurations plane et spatiale de la structure de l'ADN. . . . 3

1.2 Processus de la division cellulaire 4

1.3 Dogme central de la biologie moléculaire (Source site ISIMA, auteur Vincent Barra). 6

1.4 Les étapes de fabrication d'une puce à ADN (Source site ENS, auteur Philippe Marc). 8

1.5 Une puce constituée de 16 blocs (4×4) illustrant l'image de la totalité du transcriptome de la levure (6300 gènes)(Soularue et Gidrol [92]). 9

1.6 Nuage de points pour une puce à ADN avant et après transformation log. Sur cette figure, on peut voir, à gauche sans aucune transformation, que les effectifs sont plus importants vers les faibles intensités tandis qu'à droite avec un passage au logarithme à base 2, les intensités sont distribuées de façon uniforme. 10

1.7 Nuage M vs A pour une puce à ADN. 11

1.8 Normalisation par *Lowess*. À gauche, le nuage M vs A avant normalisation. En rouge est représentée la courbe de régression *Lowess* épousant l'allure du nuage. À droite, le nuage M vs A après normalisation, la courbe de régression est une droite. . . . 12

2.1 Exemple de profil temporel d'expression de gène. Chaque puce correspond à un instant d'échantillonnage. Les niveaux d'expression mesurés sur chacune des puces sont inscrits sur la matrice d'expression de telle sorte que chaque colonne représente une puce. 17

2.2 Classification par la méthode hiérarchique ascendante de données temporelles (d'après Eisen et al. [26]). Chaque gène est représenté par une seule colonne de couleurs et chaque instant t par une seule ligne. Cinq classes distinctes, biologiquement intéressantes, sont indiquées par les cinq barres colorées (A,B,C,D et E) et par la coloration de la région correspondante du dendrogramme. 19

2.3 Principe des cartes auto-organisatrices. La position initiale des nœuds (en bleu) est sur une grille rectangulaire 3×2 et est indiquée par les lignes continues raccordant ces nœuds. Les lignes pointillées et les flèches représentent les trajectoires hypothétiques des nœuds lors de leur ajustement par itérations successives de l'algorithme SOM. Le résultat final est composé des six nœuds (en rouge) entourés par les points noirs représentant les profils d'expression de gènes (Tamayo et al. [97]). 21

3.1 Décalage temporel des séries S_1 et S_2. 33

3.2 Exemple de deux séries temporelles S_1 et S_2 (en haut). La recherche du chemin optimal entre les deux séries par la technique de la DTW (au milieu). Elle capture le décalage temporel le long de l'axe des temps (en bas). 35

3.3 La proximité définie par une distance euclidienne dépend uniquement de l'écart entre les observations prises aux mêmes instants. 36

3.4 Exemples de séries temporelles proches en valeurs et de formes opposées (S_1 et S_2), vs. distantes en valeurs et de formes similaires (S_1 et S_3). 37

3.5 Illustration des limites du coefficient de corrélation de Pearson face aux effets de tendance (a) et face aux valeurs extrêmes (b). Les valeurs enregistrées des coefficients sont significativement élevées 0.74 et 0.73 respectivement. Alors que les séries sont quasiment opposées. 38

3.6 Deux séries S_1 et S_2 de formes opposées, avec $D_q(S_1, S_2) = 0.33$, illutrant la limite de la distance de Todorovski. 41

3.7 À gauche, représentation de trois séries S_1, S_2 et S_3 avec $d_{STS}(S_1, S_3) = 6.4$ et $d_{STS}(S_1, S_2) = 5.3$. À droite, le passage à une standardisation des données donne $d_{STS}(S_1, S_2) = 8.7$ et $d_{STS}(S_1, S_3) = 1.8$. 42

3.8 L'effet du réglage en fonction de k. 45

4.1 Profil d'expression du gène cyclique CCNE1 (cycline E1) observé sur une période de 46 heures après synchronisation correspondant à trois cycles cellulaires. Ce gène s'exprime fortement pendant la transition G_1/S de l'interphase du cycle cellulaire. Chaque phase de cycle est délimitée par les traits verticaux et annotée par G_1, S, G_2 ou M. 53

4.2 Progression de l'expression des gènes durant les 5 phases G_1/S, S, G_2, G_2/M et M/G_1. 58

4.3 Profils des gènes de la phase G_1/S suivant les quatre expériences. Dans la première expérience (Experiment 1), les gènes sont différenciés par leur amplitude initiale. Ces amplitudes s'atténuent dans la longueur du temps, pour la seconde expérience (Experiment 2). La présence de tendance est observée pour les profils de la troisième expérience (Experiment 3), sans atténuation d'amplitude et avec atténuation d'amplitude pour la quatrième expérience (Experiment 3). 59
4.4 Pour chaque expérience et pour chaque mesure (δ_E, COR, CORT et D_{k*}), la progression des valeurs asw suivant les 10 échantillons est représentée. 61
4.5 Pour chaque expérience et pour chaque mesure (δ_E, COR, CORT et D_{k*}), la progression des valeurs wbr suivant les 10 échantillons est représentée. 61
4.6 Pour chaque expérience et pour chaque mesure (δ_E, COR, CORT et D_{k*}), la progression des valeurs RI suivant les 10 échantillons est représentée. 62
4.7 Evaluation des mesures pour la classification des profils d'expression simulés. La progression des valeurs moyennes des critères asw (gauche), wbr (droite) et RI (en bas) est illustrée. . . 62
4.8 Pour chaque expérience et pour chaque mesure (δ_E, COR, CORT et D_{k*}), la progression des taux d'erreur suivant les 10 échantillons est représentée. 63
4.9 Evaluation des mesures pour le classement des profils de gènes suivant les quatre expériences. La progression des valeurs moyennes des taux d'erreur est représentée. 64
4.10 La distribution des valeurs de $k*$ dans la classement adaptatif . 64
4.11 Profils des 20 gènes de référence (tableau 4.4) observés pendant trois cycles cellulaires, chaque gène s'exprimant fortement pendant l'une des périodes du cycle cellulaire. La durée de chacune des quatre phases du cycle cellulaire (G_1, S, G_2 et M) est délimitée par les lignes verticales. 67
4.12 Progression des valeurs du critère asw (gauche) et du critère wbr (droite) des partitions $\mathcal{P}_{N,k}$, N (Number of clusters) varie de 4 à 10 et k varie de 0 à 6. 68
4.13 Progression des ratio asw/wbr (gauche) en fonction du nombre de classes (variant de 4 à 10), le graphe de la silhouette associée à la partition $\mathcal{P}_{N^*=4,k^*=5.7}$ (droite). 69
4.14 Les profils d'expression des gènes noyaux au cours des 3 cycles cellulaires : classe 1(S), classe 2(G_1/S), classe 3(G_2/M) et classe 4(G_1). 69
4.15 La distribution des valeurs sw des profils simulés, classifiés par l'approche adaptative (gauche) et l'approche basée sur la corrélation (droite). 77

4.16 La progression des valeurs du critère *asw* (gauche) et celle du critère *wbr* (droite) des partitions $\mathcal{P}_{N,k}$, N (Number of clusters) varie de 2 à 15 et k varie de 0 à 6. 78
4.17 Le graphe de la silhouette associée à la partion $P_{N^*=8,k^*=6}$. . . . 79
4.18 Les profils d'expression des gènes noyaux des 8 classes issues de notre classification adaptative. Les 10 gènes noyau de chaque classe sont très similaires en forme entre eux et différents en forme des autres gènes noyaux des autres classes. 83

Liste des tableaux

2.1 Cinq expériences différentes pour l'étude du cycle cellulaire chez la levure. 17

3.1 Quatre séries temporelles illustrant les limites du coefficient de corrélation de Pearson face aux effets de tendance et aux valeurs extrêmes. 38

3.2 Définition de la fonction $Diff$. Pour deux séries S_1 et S_2, l'écart entre un accroissement et une décroissance est égal à 1, alors que l'écart entre un accroissement (ou une décroissance) et un état stable est égal à 0.5. 40

3.3 La description de deux séries opposées illustrant la limite de la distance de Todorovski à mesurer la proximité entre la forme des séries temporelles. 41

3.4 Trois séries temporelles montrant la limite de la distance d_{STS}. . 43

3.5 Contribution de la proximité en valeurs et en forme à D_k,pour $|Cort| = 1$. 46

3.6 Les principales mesures de proximité entre des séries temporelles, dans un cadre unifié. 47

4.1 Exemples d'études effectuées sur la levure et l'espèce humaine pour l'identification de gènes cycliques. 53

4.2 Spécification des paramètres du modèle. 60

4.3 Evaluation de l'efficacité des mesures pour la classification : chaque couple de valeurs représente la moyenne et la variance des dix valeurs de $k*$ trouvées pour les 10 partitions $\mathcal{P}^{1j}_{D_{k*}}$,..., $\mathcal{P}^{10j}_{D_{k*}}$. Evaluation de l'efficacité des mesures pour le classement : chaque couple de valeurs représente la moyenne et la variance des dix valeurs de $k*$ trouvées pour les 10 classements $\mathcal{C}^{1j}_{D_k}$,..., $\mathcal{C}^{10j}_{D_k}$ 60

4.4 Les 20 gènes de référence de Whitfield et al. [106]. 66

4.5 Les ensembles noyaux de la partition $\mathcal{P}_{N^*=4,k^*=5.7}$ 71

4.6 Affectation finale des gènes frontaliers ($sw \in [-0.05, 0.05]$) selon la continuité circulaire des phases du cycle cellulaire. 72

4.7 Liste des 39 gènes compilés depuis la littérature par Whitfield et al. [106]. Comparaison des phases publiées ($Publi - Assi$) avec celles issues de l'approche conventionnelle ($Whit - Assi$) et de la classification adaptative ($Ada - Assi$). 73
4.8 Les statistiques d'affectation des 700 profils simulés basés sur les classes noyaux (sous-tableau gauche) et les classes références (sous-tableau droit). Les nombres en gras correspondent aux nombres de profils de gènes simulés mal classifiés. 76
4.9 La structure de voisinage des 8 classes de la partition $P_{N^*=8,k^*=6}$. 79
4.10 Classification des 50 gènes identifiés par Peddada et al. [78] selon notre approche adaptative. Chacun de ces 50 gènes est décrit par son numéro d'identification (ID, dans la base de données UniGene en date du 11 février 2002), son numéro de classe ($Num - Class$), le numéro de sa classe voisine ($Num - Vois$), sa silhouette width (sw) obtenue par la méthode adaptative, et son affectation par rapport l'approche de Peddada et al. [78] ($Pedd - Aff$). 80

1

Eléments de biologie et puces à ADN

Sommaire

1.1 La cellule 1
1.2 Le cycle cellulaire 3
1.3 Le transcriptome 5
1.4 Les puces à ADN pour l'étude du transcriptome . 6
1.4.1 Les sondes 7
1.4.2 Les cibles et l'hybridation 7
1.4.3 Acquisition des données 7
1.4.4 Normalisation des données et pré-traitement 8
1.5 Analyse des données du transcriptome 12

Les méthodes d'analyse du transcriptome les plus utilisées reposent sur la technologie des puces à ADN. Ce chapitre nous permet d'appréhender quelques notions sur la biologie moléculaire et de dégager les principes de fabrication d'une puce à ADN. L'idée n'est pas de faire un cours sur les deux sujets mais simplement de fournir les éléments de base pour une bonne compréhension de cette technologie. Nous allons, dans une première étape, donner quelques notions de la biologie moléculaire utiles dans ce domaine. Ensuite dans un second temps nous abordons les différentes étapes permettant d'obtenir des données d'expression de gènes par les expériences des puces à ADN .

1.1 La cellule

La plus petite unité structurale et fonctionnelle de tous les êtres vivants est la cellule. Il existe une grande variété de cellules vivantes qui diffèrent par leur

taille, leur forme, leur fonction, leur comportement, et leur association. Chez les organismes les plus simples que sont les procaryotes (unicellulaires), le matériel génétique n'est pas compartimenté dans un noyau vrai mais est libre dans le cytoplasme, c'est le cas des bactéries. Par contre, pour les organismes plus complexes, les eucaryotes (uni ou pluricellulaires), l'information génétique est localisée dans un noyau. À la différence des procaryotes généralement pauvres en organites, les eucaryotes comprennent différents compartiments fonctionnels, membraneux (organites) ou dépourvus de membrane. L'Homme, les insectes, les plantes sont des organismes eucaryotes. On trouve environ 6×10^{23} cellules dans un humain, de 320 types différents (épidermique, musculaire, nerveuse...). La plupart de ces cellules sont capables de grossir et de se diviser. Elles ont un métabolisme, c'est à dire qu'elles importent des nutriments et les convertissent en molécules utiles et en énergie. Elles peuvent aussi réagir à leur environnement.

Il existe deux types fondementalement différents de molécules dans la cellule : de **petites molécules** et des **macromolécules**, qui tiennent des rôles essentiellement différents. Les petites molécules, par exemple les acides aminés et les nucléotides, servent surtout de substrats pour construire les macromolécules, et la cellule veille à s'approvisionner en toutes les espèces requises pour les synthèses des macromolécules. On regroupe les macromolécules en quatre grandes familles : les lipides, les glucides, les protéines et les acides nucléiques (ADN et ARN). Les principales caractéristiques et fonctions des protéines et des acides nucléiques sont décrites ci-dessous.

L'acide désoxyribonucléique ou ADN est une molécule présente dans le noyau de la cellule. Il joue un rôle central dans la vie cellulaire. En effet, il est à la fois détenteur de l'information génétique qui permet à la cellule de synthétiser continuellement ses protéines (par l'enchaînement séquentiel des processus de transcription, synthèse des ARN, et de traduction, synthèse des protéines), et il assure la transmission du message génétique au cours des générations cellulaires (par le processus de la réplication). Cette macromolécule a une structure double hélice (Watson et Crick) constituée de deux brins anti-parallèles (figure 1.1). Un brin simple de l'ADN (aussi appelé polynucléotide) est un polymère linéaire composé de quatre nucléotides. Un nucléotide comprend l'une des quatre bases : adénosine (A), cytosine (C), guanine (G) ou thymine (T), un désoxyribose et un groupement phosphate. Les couples $A-T$ et $G-C$ sont appelés paires de bases complémentaires, de ce fait les deux brins anti-parallèles vont s'associer par complémentarité des bases. Cette complémentarité des bases est respectée lors de la réplication, chaque brin de la molécule d'ADN servant de matrice pour la synthèse du brin complémentaire après ouverture de la double hélice. Elle permet également la réparation d'erreurs, donc la protection d'information liée à une lésion de l'ADN ; et à deux brins de se retrouver et de se réassocier dans un mélange moléculaire complexe.

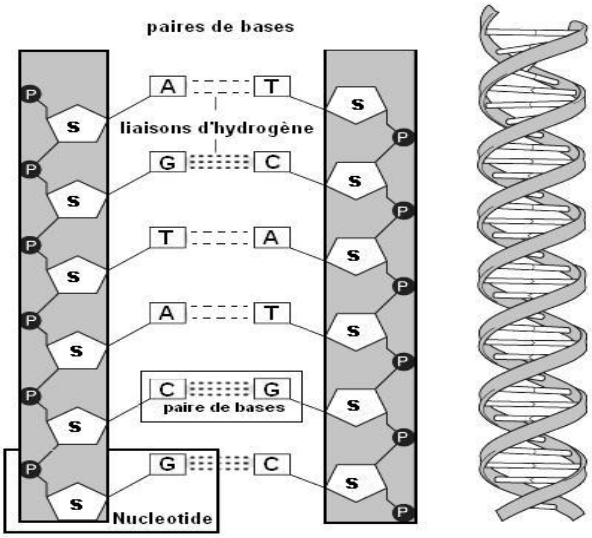


FIG. 1.1 – Configurations plane et spatiale de la structure de l'ADN.

Les acides ribonucléiques ou ARN sont des copies d'une région de l'un des brins de l'ADN. Le sucre de l'ADN (désoxyribose) est remplacé par un autre sucre (ribose) dans l'ARN et la thymine (T) de l'ADN est remplacée par l'uracile (U) dans l'ARN. L'ARN a de multiples fonctions cellulaires, les ARN ribosomaux (ARNr), de transfert (ARNt), messagers (ARNm) et de petits ARN tels les small nuclear ribosomes (snARN).

Les proteines, principales macromolécules de la cellule, sont composées par un ou plusieurs acides aminés. Elles sont exclusivement synthétisées dans le cytoplasme par l'intermédiaire de l'ARNm qui permet le transfert de l'information génétique du noyau vers le cytoplasme, lorsqu'un gène s'exprime. Les protéines sont donc le résultat final de l'expression d'un gène (*c.f.* § 1.3). Certaines sont des catalyseurs de réactions chimiques : elles permettent à des réactions chimiques de se dérouler rapidement dans les conditions de température et de pression conformes à la vie.

1.2 Le cycle cellulaire

Nous nous intéressons dans cette thèse à l'analyse de l'expression dynamique des gènes au cours du cycle cellulaire. Les données que nous allons étudier sont alors observées à des moments spécifiques durant le processus de la division cellulaire.

Chaque cellule eucaryote est née d'un processus de division nommé mitose. Elle peut être décrite selon la succession des étapes (ou phases) : G_1, S, G_2 et M communément référencée comme le cycle cellulaire ou cycle de la division

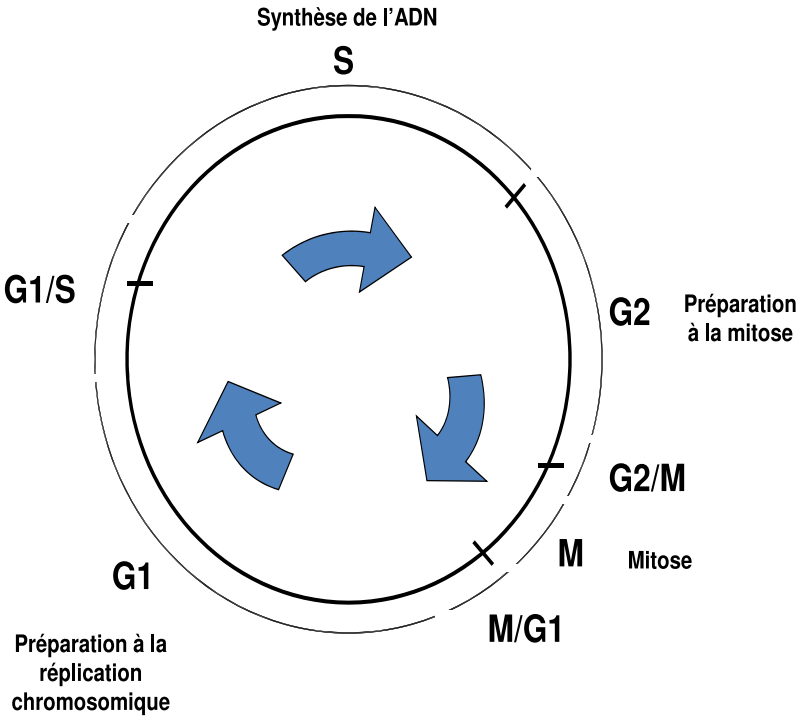


FIG. 1.2 – Processus de la division cellulaire

cellulaire (figure 1.2). Durant les phases G_1, S et G_2 (regroupées sous le nom de "interphase"), les chromosomes sont décondensés et distribués dans tout le noyau, la cellule grossit progressivement. La plupart d'entre elles doublent de taille entre le début de la phase G_1 et la phase G_2. La phase G_1 correspond à l'intervalle entre la fin de la phase M et le début de la réplication de l'ADN (phase S). Au cours de cette phase la cellule effectue son métabolisme habituel et grossit régulièrement, elle intègre les signaux mitogènes ou anti-mitogènes et se prépare pour effectuer correctement la phase S. Par contre, l'ADN n'est synthétisé que pendant une partie de l'interphase appelée phase S durant laquelle l'ADN chromosomique est répliqué. La phase G_2 est l'intervalle entre la fin de la réplication de l'ADN et la phase M. Durant cet intervalle la cellule continue de grossir et de synthétiser des protéines en vue de la phase M. Le début de la phase M est caractérisé par la désorganisation de l'enveloppe nucléaire (par phosphorylation des lamines), la compaction des chromosomes (par phosphorylation des condensines et des histones), l'organisation des chromosomes à la métaphase, la séparation des chromatides sœurs (par dégradation de cohésines) à l'anaphase, et enfin la cytodiérèse. La plupart des cellules des organismes pluricellulaires sont différenciées afin d'exercer des fonctions spécialisées et ne se divisent plus. Ces cellules sont considérées comme étant dans un compartiment spécial de la phase G_1, appelé phase G_0.

1.3 Le transcriptome

Le gène, unité de base du stockage de l'information génétique, est une petite séquence d'ADN. Il y a environ 6000 gènes chez la levure et 30000 chez l'homme. L'ensemble du matériel génétique d'un individu ou d'une espèce encodé dans son ADN est alors appelé **génome**. Aujourd'hui le séquençage du génome de diverses espèces (levure, souris, homme...) a été réalisé. Il s'agit de la description des séquences, ou gènes, constituant les molécules d'ADN au sein du noyau de chaque cellule.

En fonction de leurs besoins, les cellules utilisent à un instant donné une partie des gènes pour réaliser la synthèse des protéines nécessaires aux grandes fonctions cellulaires. Le passage du gène à la proteine s'effectue en deux grandes étapes de transcription et de traduction, à l'aide d'un intermédiaire essentiel : l'ARN messager (figure 1.3). Le gène est transcrit (synthèse de l'ARNm, c'est la transcription), l'ARNm est véhiculé hors du noyau dans le cytoplasme ou il va servir de matrice pour la synthèse des protéines (c'est la traduction). Pour réaliser la traduction, il y a besoin de ribosomes (contenant les ARNr, aussi synthétisés dans le noyau). Les sous-unités ribosomales sont des protéines ribosomales qui s'associent aux ARNr dans le noyau avant d'être véhiculées hors du noyau. Deux sous-unités (la petite et la grande) viennent former le ribosome sur l'extémité de l'ARNm. L'ARNt vient s'attacher au site du ribosome et la traduction en proteines peut commencer. À chaque codon (triplet de nucléotides) de l'ARNm va correspondre l'anticodon d'un ARNt. Cet anticodon est associé à un acide aminé qui est transité au niveau ribosome par l'ARNt. Le ribosome se déplace sur l'ARNm et la rencontre d'un codon qui ne correspond à aucun anticodon met fin à la traduction.

Sachant que plusieurs copies de différents ARNm peuvent être produites à des niveaux différents en fonction de l'activité de la cellule, le transcriptome reflétera donc le niveau d'expression de tous les gènes à un temps t pour une condition physiologique donnée. Il est le reflet instantané de l'activité cellulaire, et peut donc varier d'un type cellulaire à l'autre (neurone, cellule de la peau...), au cours du temps ou des différentes phases du cycle cellulaire, en fonction de conditions environnementales imposées aux cellules, ou en fonction de l'état sain ou pathologique de ces dernières. Le transcriptome, représentant l'ensemble des ARNm transcrits, participe clairement à la détermination de la fonction des gènes au sein d'une cellule. De manière générale, pouvoir comparer le transcriptome de différents types cellulaires, dans différentes conditions, ou pouvoir analyser l'ensemble du transcriptome d'une cellule à divers stades de son cycle cellulaire ou dans diverses conditions pathologiques, doit permettre d'une part de mieux comprendre le fonctionnement cellulaire sur le plan fondamental, et d'autre part d'offrir beaucoup d'intérêts en termes d'applications potentielles.

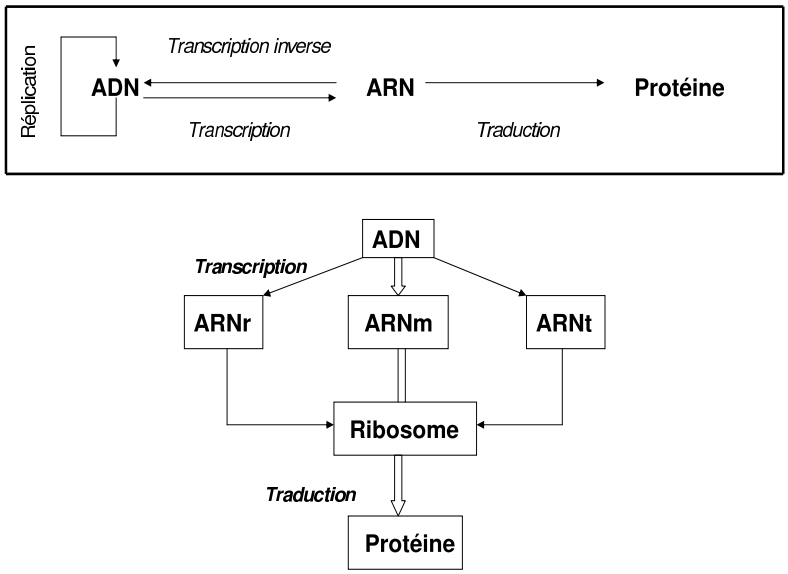


FIG. 1.3 – Dogme central de la biologie moléculaire (Source site ISIMA, auteur Vincent Barra).

1.4 Les puces à ADN pour l'étude du transcriptome

Le principe des puces à ADN repose sur les techniques d'hybridation de l'ADN simple brin à son brin complémentaire (Southtern blot [93]), ensuite adaptées pour l'étude des ARN (Northern blot) au niveau d'un seul gène. Ces techniques ont surtout posé le problème de l'étude au niveau du génome des processus biologiques complexes. Aujourd'hui, les puces à ADN offrent aux laboratoires la possibilité d'étudier des phénomènes à l'échelle du génome. Elles appartiennent à un ensemble de nouvelles technologies développées depuis quelques années à l'interface de nombreuses spécialités comme la biologie moléculaire, la chimie, l'informatique, l'électronique et la robotique.

Les puces à ADN connaissent depuis quelques années un essor considérable, car elles permettent l'analyse à haut débit de milieux biologiques, par exemple pour la détection de mutations associées à certaines maladies. Concrètement, elles permettent de visualiser simultanément le niveau d'expression de plusieurs milliers de gènes dans un type cellulaire et un contexte physiologique et/ou pathologique particulier. Elles ont d'abord été conçues sur de grandes membranes poreuses en nylon ou *macroarrays* (Gress et al. [35], Nguyen et al. [71], Pietu et al. [77]). La miniaturisation, rendue possible par les progrès de la robotique, a ensuite permis le développement des *microarrays*. Comme leur nom l'indique, ces puces sont de petites surfaces telles une lame de microscope (Schena et al. [85]) ou une petite membrane nylon (Jordan [48]). Elles

présentent l'avantage de pouvoir être de très haute densité et par conséquent susceptibles de recouvrir l'intégralité du génome humain. La fabrication d'une puce à ADN peut se décomposer en trois étapes : la production des sondes (fragments courts d'ADN de séquences connues) et leur dépot sur le support, la production et le marquage des cibles (fragments courts d'ADN que l'on cherche à identifier), enfin l'hybridation des sondes avec les cibles.

1.4.1 Les sondes

Pour l'étude du transcriptome, les sondes sont les séquences d'ADN double brin. Elles sont amplifiées par la technique PCR (polymerase chain reaction) et sont ensuite déposées sur la lame par un robot muni d'aiguilles sur laquelle elles sont fixées par un procédé chimique adéquat. La zone de dêpot du gène est appelée spot. L'ADN des sondes est dénaturé en simple brin, juste avant l'hybridation, pour permettre par la suite une hybridation avec les cibles marquées.

1.4.2 Les cibles et l'hybridation

La préparation des cibles consiste à extraire des ARNm de deux expériences de cultures cellulaires de même type dont on veut comparer l'expression des gènes (par exemple cellules saines versus cellules cancéreuses, cellules sauvages versus cellules mutées ou cellules témoins versus cellules traitées). Les cibles correspondant aux deux cultures sont marquées par deux fluorochromes, par exemple Cy5 (rouge) pour une culture et Cy3 (vert) pour l'autre, afin d'évaluer et de quantifier de manière comparative entre les deux cultures l'appariement sonde/cible. Ce marquage se fait lors d'une transcription inverse, où l'ARNm est retransformé en ADN complémentaire marqué ou ADNc. Les deux échantillons d'ADNc sont alors mélangés et hybridés sur la puce. Un brin d'ADNc fluorescent (cible) qui rencontre son complémentaire ADN simple brin (sonde) s'apparie avec lui pour redonner de l'ADN double brin. Cette hybridation, qui dure quelques heures en milieu liquide, est suivie d'un lavage du substrat qui permet d'éliminer les cibles non fixées, ou fixées non spécifiquement. La puce est alors séchée avant de passer au scanner pour détecter les hybridations.

1.4.3 Acquisition des données

L'acquisition des images se fait au moyen d'un scanner. Dans le cas du marquage avec deux fluorochromes, une image numérique est acquise pour l'échantillon marqué avec le Cy3 et une en Cy5. Un spot de couleur verte indique un gène dont le niveau d'expression est plus élevé dans l'échantillon marqué avec le Cy3 que celui marqué avec le Cy5, et inversement pour un spot de couleur rouge. Le spot apparait jaune lorsque le gène est exprimé de manière identique dans les deux échantillons comparés et le noir symbolise l'absence de signal. Il reste donc à analyser ces images en trois étapes. La localisation des spots

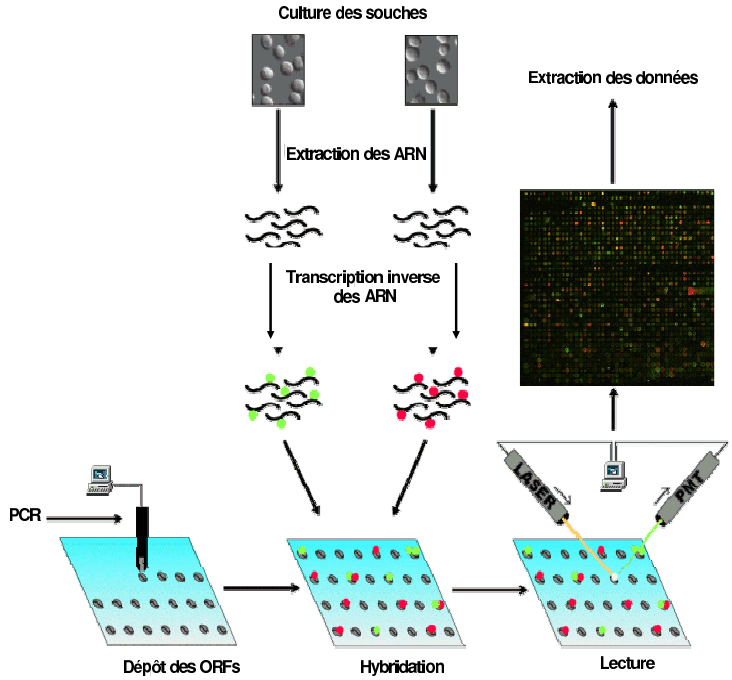


FIG. 1.4 – Les étapes de fabrication d'une puce à ADN (Source site ENS, auteur Philippe Marc).

qui permet de déterminer leurs coordonnées sur la puce, la segmentation qui classe, pour chaque spot, les pixels "signal" des pixels "bruit de fond" et enfin l'extraction des données consistant à définir les intensités des signaux Cy3 et Cy5. Pour chaque spot, on retient une valeur de l'intensité du signal rouge (notée R faisant référence au fluorochrome Cy5) et du signal vert (notée V pour le Cy3) ainsi que les valeurs du bruit de fond correspondant. L'intensité du signal est proportionnelle à l'intensité d'hybridation, donc à l'expression du gène ciblé. À l'issue de l'analyse d'image, les valeurs brutes des intensités obtenues sont ensuite filtrées et normalisées pour obtenir des valeurs relatives ou ratio d'expression pour chacun des gènes présent sur la puce.

1.4.4 Normalisation des données et pré-traitement

Avant de comparer de façon précise les deux échantillons, il est nécessaire d'effectuer une normalisation relative des deux intensités de fluorescence (Smyth and Speed [90]) ; afin de s'assurer que les différences observées dans les intensités sont dues à des différences réelles d'expression et non à des artefacts expérimentaux. En effet le rapport des intensités brutes R/V est biaisé pour plusieurs raisons bien connues. Lors de la fabrication de puces à ADN, les sources de variabilité sont nombreuses. Nous pouvons citer, entre autres, l'amplification des sondes par la technique PCR et leur positionnement sur la puce, l'hybridation sonde/cible, le nettoyage et le séchage des puces etc... Mais l'une

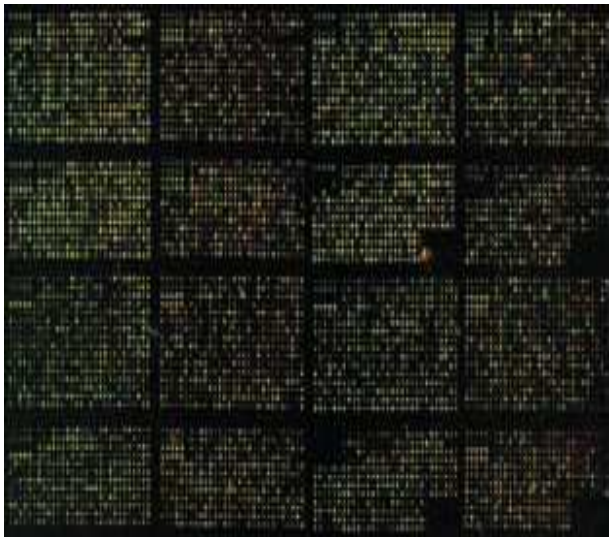


FIG. 1.5 – Une puce constituée de 16 blocs (4×4) illustrant l'image de la totalité du transcriptome de la levure (6300 gènes)(Soularue et Gidrol [92]).

des sources principales de variation intervient au moment de l'incorporation des fluorochromes Cy3 et Cy5. Le but de la normalisation est d'identifier les biais techniques, de les quantifier et de les soustraire du signal avant d'analyser l'expression des gènes. L'analyse biostatistique en aval recommande d'effectuer des replicats biologiques et techniques pour augmenter la fiabilité des résultats. Cependant les différentes expériences ne donnent jamais les mêmes résultats. Ces sources de variabilité et l'accumulation des erreurs expérimentales affectent de manière non négligeable les valeurs des intensités des signaux et ont tendance à bruiter les données d'expression de gènes. Bien souvent, on peut choisir de corriger l'intensité du signal mesuré par le bruit de fond (Kooperberg et al. [54]). Dans ce cas, chaque gène sera ainsi associé à un couple de valeurs corrigées (R_c,V_c).

Approches de normalisation

Des transformations préliminaires permettent de diminuer un peu la variabilité des mesures. En particulier, le remplacement des données brutes par leur logarithme à base 2 permet de réduire l'influence des valeurs aberrantes (figure 1.6).
Le principe de la normalisation repose sur l'hypothèse fondamentale que la plupart des gènes ont le même niveau d'expression dans les deux échantillons, ce qui signifie que peu de gènes sont différentiellement exprimés. Par la suite, on s'intéresse à la comparaison des expressions dans les deux fluorescences en visualisant les données sous la forme d'un nuage représentant les log_2-ratios

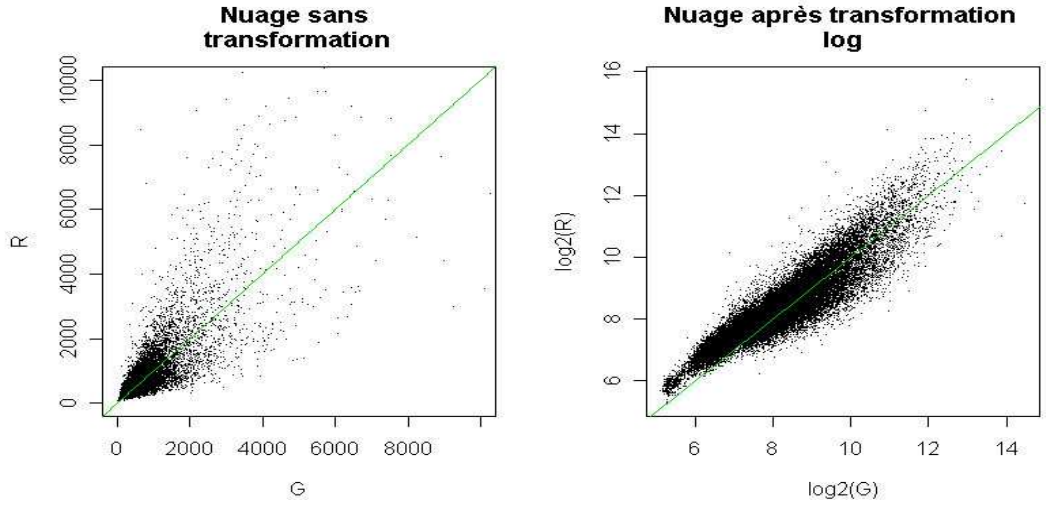


FIG. 1.6 – Nuage de points pour une puce à ADN avant et après transformation log. Sur cette figure, on peut voir, à gauche sans aucune transformation, que les effectifs sont plus importants vers les faibles intensités tandis qu'à droite avec un passage au logarithme à base 2, les intensités sont distribuées de façon uniforme.

notés M en fonction de la log_2-intensité globale moyenne sur le spot notée A. Ce nuage de points devrait alors se situer autour de l'axe des abscisses de façon symétrique. Or, nous observons une déformation de ce nuage que la normalisation va tenter de corriger (figure 1.7). Par définition :

$$A = log_2\sqrt{(R \times V)} = \frac{[log_2(R) + log_2(V)]}{2}, \tag{1.1}$$

$$M = log_2(R) - log_2(V) = log_2(R/V). \tag{1.2}$$

Il existe plusieurs types de normalisation des données (Park et al. [75], Wu et al. [109], Fujita et al. [29]). Les méthodes qui ont été utilisées sont de type global. Exemple, la normalisation par la moyenne ou la médiane : elle repose sur l'hypothèse que le rapport d'intensité entre R et V ne dépend pas du gène et reste constant. Ce qui équivaut à $M_i = K$ (K représente la moyenne ou la médiane des valeurs log_2-ratios), où $i = 1, ..., S$ et S le nombre de spots de la puce. On normalise en retranchant cette valeur à l'ensemble des log_2-ratios pour centrer l'ensemble des mesures sur un log_2-ratio à zéro. D'autre part, une normalisation par regression linéaire suppose que M et A sont liés par une fonction linéaire de type $M = a + b(A)$. Cependant ces méthodes sont limitées, car nous remarquons souvent que la nature des relations entre les intensités sont non linéaires. Une fonction qui prendrait la forme de l'allure du nuage M vs A serait plus adaptée. Dans ce contexte, Yang et al. [112] ont développé une

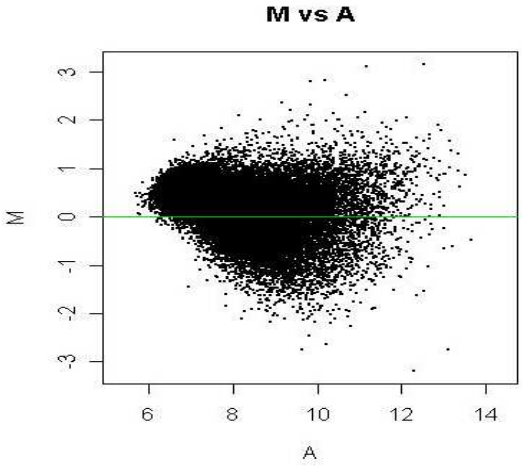


FIG. 1.7 – Nuage M vs A pour une puce à ADN.

technique trés utilisée et basée sur une méthode de régression locale pondérée, appelée *Lowess* (Locally Weighted Scatterplot Smoothing)(figure 1.8). Il s'agit donc d'une correction non-linéaire qui permet de prendre en compte la non-linéarité de la relation qui existe entre les intensités R et V et qui est définie par :

$$M_i = Lowess(A_i) \tag{1.3}$$

La correction lowess peut aussi être faite par bloc pour corriger principalement l'interaction gène×bloc×fluorochrome c'est à dire le biais spatial.

L'ajustement d'échelle

Il est nécessaire de procéder à un ajustement d'échelle. Ceci pour garantir la même dispersion des log_2-ratios au sein de chaque bloc. Ici, on fait l'hypothèse que les log_2-ratios suivent une distribution normale de moyenne nulle et de variance $\alpha_j^2\sigma^2$, où σ^2 est la variance des vrais log_2-ratios, α_j^2 est un facteur d'échelle pour le $j$ème bloc. Ce facteur est estimé par :

$$\hat{\alpha_j} = \frac{MAD_j}{(\prod_{j=1}^{J} MAD_j)^{\frac{1}{J}}} \tag{1.4}$$

où MAD est l'écart absolu médian (*median absolute deviation*) défini par :

$$MAD_j = median_i\{|M_{ij} - median_i(M_{ij})|\} \tag{1.5}$$

M_{ij} est le $i$ème log_2-ratio du $j$ème bloc, $i = 1, ..., I$; I est le nombre de gènes dans le $j$ème bloc et J est le nombre total de blocs. Une fois les valeurs $\hat{\alpha_j}$

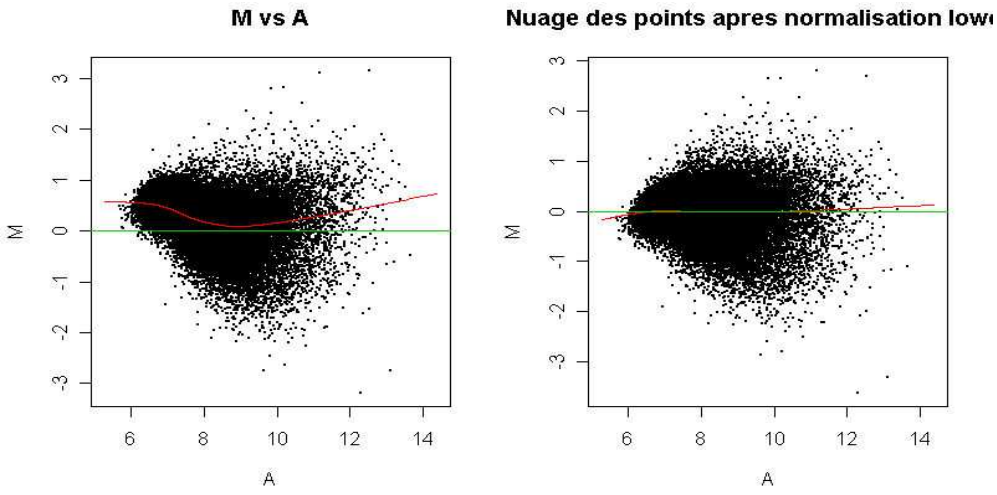


FIG. 1.8 – Normalisation par *Lowess*. À gauche, le nuage M vs A avant normalisation. En rouge est représentée la courbe de régression *Lowess* épousant l'allure du nuage. À droite, le nuage M vs A après normalisation, la courbe de régression est une droite.

estimées, on peut faire l'ajustement d'échelle :

$$M'_{ij} = \frac{M_{ij}}{\hat{\alpha}_j} \tag{1.6}$$

La normalisation *Lowess* par bloc avec un ajustement d'échelle entre les blocs est une approche de normalisation sensée dans le cadre des puces à ADN.

1.5 Analyse des données du transcriptome

La grande quantité de gènes observés (environ 6000 chez la levure, 30000 chez l'Homme) combinée à la complexité des réseaux biologiques posent un vrai challenge pour la compréhension et l'interprétation des masses de résultats obtenus. Une étape préalable à l'exploitation des données produites par cette nouvelle technologie est l'utilisation de techniques de classification. Elles permettent de réduire entre autres la dimensionnalité des données, d'extraire des profils types d'expression de gènes, ou de prédire la fonction de certains gènes.

Nous distinguons principalement deux types d'étude des données d'expression. La première porte sur l'expression d'un ensemble de gènes (individus) observés au travers d'un ensemble de conditions expérimentales (les dimensions). Ces conditions pouvant, par exemple, définir différents types de cellules, différents tissus (sains vs cancéreux), ou différentes expériences (Hughes et al. [43]). Un

second type d'analyse s'intéresse à l'étude de l'évolution de l'expression des gènes au cours de processus biologiques. Les gènes de profils temporels similaires, dits co-exprimés, sont considérés comme co-régulés, à savoir associés aux mêmes fonctions biologiques et impliqués dans les mêmes processus biologiques. Ainsi la classification non supervisée des profils d'expression de gènes, permet entre autres d'identifier la fonction de nouveaux gènes sur la base des fonctions déjà identifiées des gènes co-régulés (Eisen et al.[26], Iyer et al. [44], Gasch et al. [31], Storey et al. [95], Leng et Müller. [57]).

Ce travail de thèse s'inscrit dans ce dernier cadre d'analyse. Nous nous intéressons à l'analyse des profils temporels d'expression d'un ensemble de gènes au cours du processus de division cellulaire. La division cellulaire est le principal processus de prolifération des cellules, devenant aberrant dans le cas de cellules cancéreuses (Golub et al. [34]). L'enjeu principal de nos travaux se focalise sur des données de gènes exprimés au cours du cycle cellulaire. Dans toute la suite de ce rapport, nous utilisons "profils temporels d'expression de gènes" pour parler de la nature temporelle de ces données. Le chapitre suivant dresse un rapide état de l'art des principales techniques de classification non supervisée utilisées pour l'analyse de profils temporels d'expression de gènes. Nous utilisons simplement le terme classification à la place de classification non supervisée.

2

Classification des profils temporels d'expression de gènes : état de l'art

Sommaire

2.1 Profils temporels d'expression de gènes 16
2.2 Techniques de classification non-paramétrique . . 18
2.2.1 Classification hiérarchique 18
2.2.2 Partitionnement par K-means 19
2.2.3 Classification par les réseaux de Kohonen (SOM) . . 20
2.2.4 Classification selon la théorie des graphes 21
2.2.5 Critères de validation du nombre de classes 22
2.3 Techniques de classification paramétrique 25
2.3.1 Classification par mélange de modèles gaussiens . . . 25
2.3.2 Classification par modélisation des profils temporels 26
2.4 Conclusion 29

L'identification des gènes coexprimés est un des problèmes classiques de l'analyse du transcriptome. Ces dernières années, on assiste à l'émergence d'une littérature abondante portant sur l'utilisation des techniques de classification et d'analyse de données pour l'analyse de données d'expression de gènes (e.g., Yeung et al. [114], Jiang et al. [46], De Hoon et al. [21], Bréhélin [9], Chaiboonchoe et al. [12], etc...). On note en particulier les nombreux travaux dédiés à l'analyse de profils temporels d'expression de gènes (*e.g.*, Jiang et al. [47], Ernst at al. [27], Ma et al. [64], Kim et al. [52], Das et al. [20], etc...). Dans ce chapitre, nous dressons un rapide état de l'art des techniques de classification les plus utilisées pour l'analyse des données du transcriptome. Dans la première

section, nous donnons une description sommaire sur la génération de profils temporels d'expression de gènes. La section suivante rappelle les techniques non paramétriques qui, généralement, se fondent sur la définition d'une mesure de proximité. Nous aborderons le problème de détermination du nombre de classes à considérer et dégagerons certains critères qui les valident. Dans la dernière section, nous présentons les téchniques de classification paramétriques qui considèrent le problème sous un aspect statistique avant de dédager une conclusion.

2.1 Profils temporels d'expression de gènes

Les expériences des puces à ADN permettent de mesurer la quantité transcrite par un gène à un instant donné. En suivant l'évolution de l'expression d'un gène sur plusieurs instants successifs pour une durée déterminée, on obtient son profil temporel d'expression. Par exemple, pour une expérience sur n gènes que l'on souhaite observer sur p instants $(t_1, t_2, ..., t_p)$, on dispose de p puces correspondant à p échantillons. Sur chaque puce, on place le même gène i $(i = 1, 2, ..., n)$ en une même position. On dispose ensuite d'une matrice $n \times p$ dont chaque ligne représente un gène i et chaque colonne correspond à un instant t_k $(k = 1, 2, ..., p)$. À chaque instant t_k, la quantité d'expression x_{i,t_k} du gène i est prélevée et inscrite sur la ligne i et la colonne k de la matrice. À la fin de l'expérience, on obtient la matrice d'expression des gènes. Chaque ligne de la matrice représente un vecteur d'expression ou profil temporel d'expression du gène, que nous notons $g_i = (x_{i,t_1}, x_{i,t_2}, ..., x_{i,t_p})$ et chaque colonne de la matrice est un échantillon prélévé à l'instant t_k (figure 2.1).

Les profils temporels d'expression décrivent, par exemple, le niveau d'activation d'un gène au cours d'un processus biologique tel que la division cellulaire, le degré de réponse d'un gène à un traitement, ou plus généralement à la réponse aux conditions environnementales (Iyer et al. [44], Gasch et al. [31]). Le tableau 2.1 présente 5 expériences différentes réalisées pour étudier le cycle cellulaire chez la levure et produisant des profils temporels d'expression de gènes. Nous notons que les instants d'échantillonnage ne sont pas toujours uniformes et varient entre les différentes expériences. En plus, la durée du cycle cellulaire varie en fonction du type cellulaire, et pourrait varier aussi en fonction des conditions expérimentales pour un même type (Cho et al. [15], Spellman et al. [94], Zhu et al. [117], Pramila et al. [80]).

Souvent, la phase suivante pour l'analyse de la matrice d'expression est la classification des gènes qui présentent des profils temporels similaires. Plusieurs travaux ont été réalisés en utilisant des méthodes de classification non-paramétriques (e.g., Spellman et al. [94], Eisen et al [26], Tamayo et al. [97],

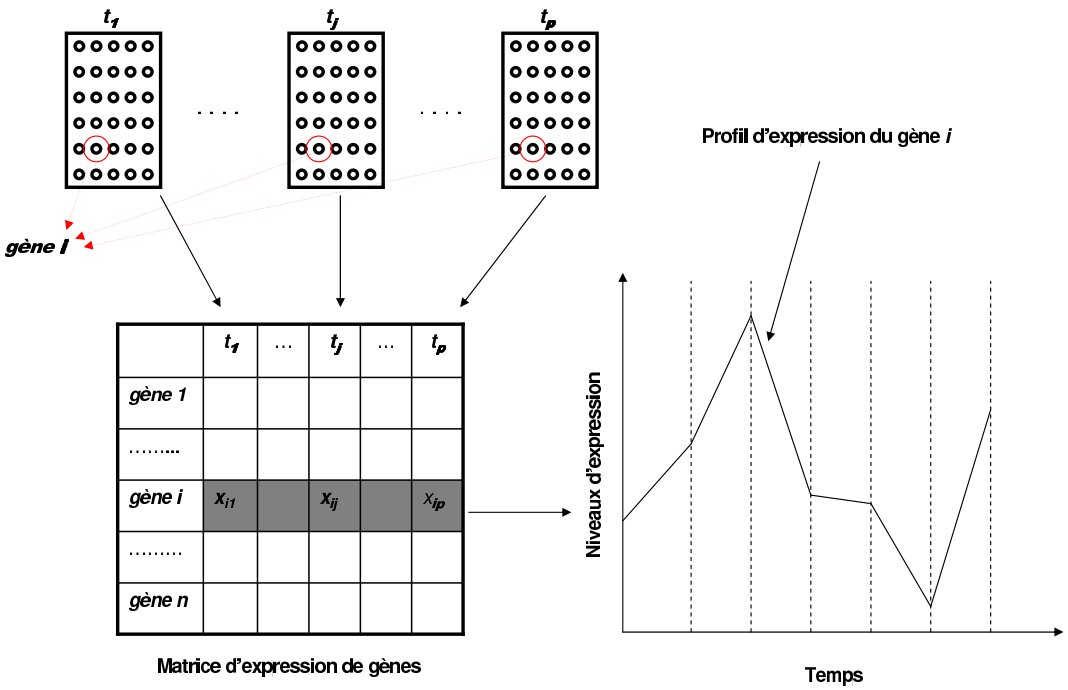


FIG. 2.1 – Exemple de profil temporel d'expression de gène. Chaque puce correspond à un instant d'échantillonnage. Les niveaux d'expression mesurés sur chacune des puces sont inscrits sur la matrice d'expression de telle sorte que chaque colonne représente une puce.

Nom	Début(min) échantillonnage	Fin (min) échantillonnage	Durée (min) du cyle cellulaire	Echantillonnage
Spellman et al. [94]	0	119	64	toutes les 7 min.
Spellman et al. [94]	10	290	112	toutes les 20 min de 0 à 70, toutes les 10 min de 70 à 250, toutes les 20 min de 250 à 290.
Cho et al. [15]	0	160	85	toutes les 10 min.
Zhu et al. [117]	0	210	105	toutes les 15 min de 0 à 165, puis 45 min de 165 à 210.
Pramila et al. [80])	0	120	60	toutes les 10 min.

TAB. 2.1 – Cinq expériences différentes pour l'étude du cycle cellulaire chez la levure.

Tavazoie et al. [100], Ben-Dor et al. [7]). D'autres, comme Peddada et al. [78], Bar-Joseph et al. [5], Ramoni et al. [82], Schliep et al. [86] se rapportent à des méthodes paramétriques. Nous présentons brièvement ces travaux dans la section suivante.

2.2 Techniques de classification non-paramétrique

Les techniques de classification non-paramétriques sont basées essentiellement sur des mesures de proximité entre paires d'objets. Les algorithmes de classification hiérarchique, de partitionnement (k-means et SOM) et ceux portant sur la théorie des graphes sont employés dans le cadre de l'analyse de profils temporels d'expression de gènes. Les résultats issus des ces techniques sont rigoureusement validés par des statistiques et des connaissances biologiques.

2.2.1 Classification hiérarchique

Les méthodes hiérarchiques sont les premières approches utilisées pour l'analyse des données de profils temporels d'expression (Eisen et al. [26], Spellman et al. [94], Wen et al. [105]). Eisen et al. [26] ont popularisé leur utilisation pour classer des profils temporels d'expression de gènes. Ils ont étudié l'évolution des niveaux d'expression de 2467 gènes de la levure au cours du cycle cellulaire. Une distance fondée sur le coefficient de corrélation est utilisée pour comparer les gènes. Le lien moyen, connu sous le nom de UPGMA (*Unweight Pair Group Method with Arithmetic mean*) est considéré comme distance entre classes. Par la suite, 10 groupes représentatifs sont extraits, contenant des gènes qui ont des profils temporels d'expression semblables. Les auteurs ont montré que ces groupes de gènes coexprimés partageaient une même fonction biologique. À partir de cette étude, ils ont montré que les gènes présentant un profil d'expression similaire (gènes de même classe) sont très souvent impliqués dans une même fonction biologique.

Ces auteurs ont développé un outil d'analyse graphique interactive et de visualisation des résultats (*Treeview*) permettant d'afficher les profils des gènes sous forme d'une carte d'expression colorée. Chaque gène est représenté par une seule colonne alternant les couleurs rouge (forte expression), verte (faible expression) et noire (expression nulle), et chaque ligne représente un instant t. L'association du dendrogramme avec la carte colorée (figure 2.2) permet une visualisation globale de la masse de données, met en évidence des groupes de gènes composant les différentes classes et donne une nouvelle lecture de ces classes en terme de niveau d'expression. La procédure de l'algorithme organise les gènes de manière hiérarchique dans un dendogramme, néanmoins, Gerstein et Jansen [32] notent qu'il n'y a aucune raison de croire qu'ils y sont naturellement organisés. D'autres part, ces méthodes sont mal adaptées aux grands ensembles de données car trop coûteuses en ressource mémoire et ne remettent jamais en cause les classes obtenues : si une mauvaise affectation est faite au début du processus, elle ne peut pas être corrigée et peut donc affecter le résultat final.

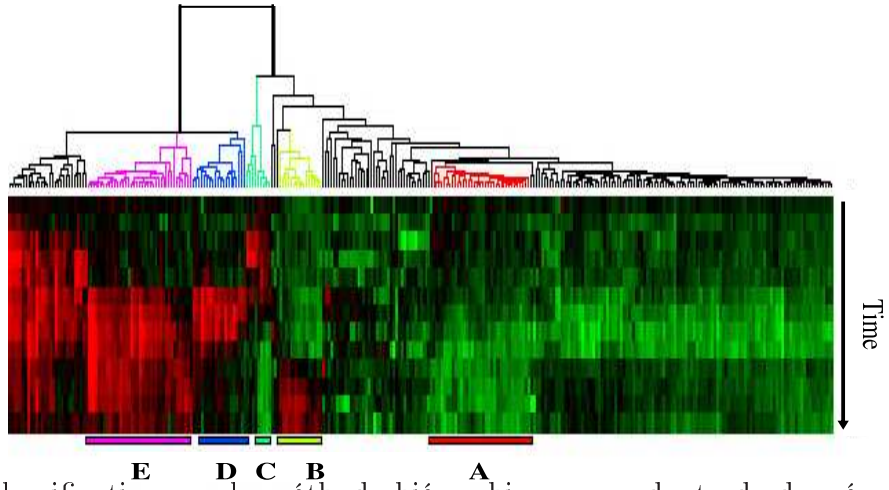


FIG. 2.2 – Classification par la méthode hiérarchique ascendante de données temporelles (d'après Eisen et al. [26]). Chaque gène est représenté par une seule colonne de couleurs et chaque instant t par une seule ligne. Cinq classes distinctes, biologiquement intéressantes, sont indiquées par les cinq barres colorées (A,B,C,D et E) et par la coloration de la région correspondante du dendrogramme.

2.2.2 Partitionnement par K-means

Une autre famille de méthodes de classification est celle des algorithmes de partitionnement. Ces méthodes nécessitent de spécifier préalablement le nombre de classes. L'une des plus utilisées est l'algorithme des K-means proposé par Mac Queen [67]. Sa grande popularité vient de sa simplicité conceptuelle, sa rapidité d'exécution et ses faibles exigences en taille mémoire. Il offre aussi la possibilité de partitionner un ensemble volumineux de données à moindre coût.

Les auteurs Tavazoie et al. [100] ont analysé 3000 profils de gènes de la levure (*Saccharomyces cerevisiae*), mesurés sur 15 instants d'échantillonnage et observés au cours de deux cycles cellulaires. Ils appliquent l'algorithme des K-means pour partitionner ces données successivement en 10, 30 et 60 classes. La similarité entre les gènes est mesurée par la distance euclidienne et le résultat en 30 classes est retenu. Ces travaux ont permis de montrer la présence, au sein des régions promotrices des gènes d'une même classe, d'une séquence commune correspondant au site de fixation d'un facteur de transcription, chaque classe étant ainsi associée à une séquence spécifique. Ces données suggèrent que l'expression des gènes d'une classe est co-régulée par un même facteur de transcription pour exercer une fonction biologique commune.

Cet algorithme est sensible aux valeurs extrêmes. Pour surmonter ce point, Kaufman et Rousseeuw [50] proposent d'utiliser l'algorithme PAM (Partitioning Around Medoids) où la médoïde, l'objet représentatif d'une classe, est la médiane de la classe. De plus, l'algorithme des K-means est une approche

non structurée : il produit un ensemble de classes non organisées (en terme de voisinage) dans lequel les relations entre classes sont difficiles à interpréter. Dans ce contexte, on cherche souvent à utiliser des méthodes qui ont une caractéristique particulière permettant d'imposer une structure partielle sur les classes.

2.2.3 Classification par les réseaux de Kohonen (SOM)

Les réseaux de Kohonen , appelés aussi cartes auto-organisatrices (Self Organisation Map - SOM), ont été développés par Kohonen [53]. C'est un algorithme qui permet à la fois un apprentissage non-supervisé rapide des gènes et leur représentation. Il utilise le réseau de neurone artificiel et respecte la topologie de l'espace des gènes en intégrant une notion de voisinage entre classes. Ainsi, deux gènes voisins dans l'espace d'entrée appartiennent à la même classe ou à des classes voisines à la sortie. Cette topologie est définie par des nœuds connectés entre eux, et le nombre de nœuds détermine le nombre de classes qui sont disposées sur une grille rectangulaire, par exemple (figure 2.3). Les nœuds sont caractérisés à la fois par leur position sur la grille et par un vecteur code représentant de chaque classe. La définition préalable de la configuration géométrique des classes, offrant une visualisation aisée des données, fait la distinction principalement entre SOM et les K-means.

Tamayo et al. [97] ont appliqué l'algorithme SOM, avec 30 nœuds définissant 30 classes disposés sur une grille 6×5, pour l'analyse du cycle cellulaire de la levure *Saccharomyces cerevisiae* en classant l'expression de 828 gènes. Les niveaux d'expression pour chaque gène sont recueillis à intervalles de 10 minutes sur deux cycles cellulaires (160 minutes). Les nœuds sont rapprochés aux gènes étudiés par l'utilisation de la distance euclidienne. Un processus itératif (entre 20000 et 50000 itérations) déplace les nœuds vers les groupes de gènes les plus proches.

Les points les plus remarquables de cet algorithme sont qu'il permet de réduire l'espace multidimensionnel des données en un espace 1D, 2D (qui sont les plus utilisés) ou 3D, et de faciliter leur visualisation et leur interprétation. Toutefois, il peut être difficile d'identifier les classes avec les outils de visualisation habituels de l'algorithme SOM. De plus, tout comme l'algorithme des K-means, la procédure de l'approche SOM exige un nombre de classes non connu *a priori*. Dans ce contexte, Fernandez et Balzarini [28] proposent la méthode RP-Q, appliquée à l'analyse des profils d'expression de gènes, qui semble intéressante pour donner une meilleure visualisation et estimation du nombre de classes dans SOM. Ghouila et al. [33] proposent l'algorithme de classification Multi-SOM, basé sur la méthode SOM avec plusieurs niveaux de grilles, afin de surmonter le problème de l'estimation du nombre de classes.

Dans le cas de profils d'expression de gènes, il est biologiquemt sensé de sup-

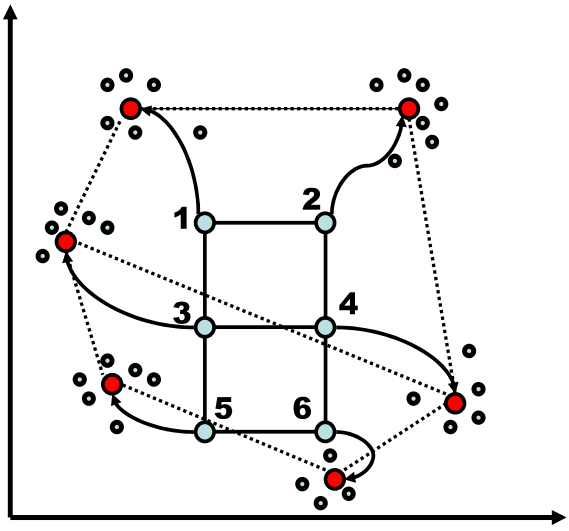


FIG. 2.3 – Principe des cartes auto-organisatrices. La position initiale des nœuds (en bleu) est sur une grille rectangulaire 3×2 et est indiquée par les lignes continues raccordant ces nœuds. Les lignes pointillées et les flèches représentent les trajectoires hypothétiques des nœuds lors de leur ajustement par itérations successives de l'algorithme SOM. Le résultat final est composé des six nœuds (en rouge) entourés par les points noirs représentant les profils d'expression de gènes (Tamayo et al. [97]).

poser que certaines vraies structures de classes n'existent que sous forme de graphes qui représentent la corrélation entre les différents gènes (Ben-Dor et al. [7]). Des approches de classification utilisent ainsi les graphes ou plus précisément la théorie des graphes pour partitionner un ensemble de gènes sans la nécessité de prédéfinir le nombre de classes.

2.2.4 Classification selon la théorie des graphes

Des techniques de classification fondées sur la théorie des graphes sont appliquées aux données d'expression de gènes et sont explicitement présentées en termes de graphe (Ben-Dor et al. [7], Shamir et Sharan [88], Hartuv et Shamir. [37], Das et al. [20]). Formellement, on représente l'ensemble des gènes sous forme d'un graphe appelé graphe de proximité $\mathcal{G}$ où chaque sommet correspond à un gène. Chaque paire de gènes est reliée par une arête qui peut être pondérée en fonction de la valeur de leur proximité (Shamir et Sharan [88]), ou en 0 et 1 en fonction d'une valeur seuil et dans ce cas seules les arêtes dont la pon-

dération est égale à 1 existent (Ben-Dor et al. [7], Hartuv et Shamir. [37]). Le problème consiste à chercher des coupes minimum ou des cliques maximales dans le graphe de proximité $\mathcal{G}$ pour partitionner l'ensemble des gènes. On distingue principalement deux algorithmes nommés CLICK (CLuster Identification via Connectivity Kernels) et CAST (Cluster Affinity Search Technique).

CLICK fait l'hypothèse que les valeurs de proximité entre les gènes sont normalement distribuées. Selon cette hypothèse, le poids w d'une arête est défini comme la probabilité que ses sommets soient dans la même classe. Les arêtes dont les poids ne dépassent pas un seuil de proximité prédéfini peuvent être omises du graphe. Le processus de classification trouve itérativement la coupe minimum dans le graphe et partage récursivement l'ensemble des données en un ensemble de composantes connexes de la coupe minimum. C'est le nombre de composantes connexes qui détermine le nombre de classes à considérer (Shamir et Sharan [88]). Dans Sharan et al. [89], les auteurs ont comparé les résultats produits par CLICK avec ceux obtenus par Tamayo et al. [97] et Eisen et al. [26] respectivement. Dans les deux cas, les classes obtenues par CLICK (respectivement 18 et 6) ont montré une meilleure qualité en termes d'homogénéité et de séparation. Toutefois, CLICK peut générer une partition qui sépare seulement quelques données atypiques des données restantes.

CAST (Ben-Dor et al. [7]) prend en entrée une matrice de proximité $\mathcal{P}$ ($\mathcal{P}(i,j) \in [0,1]$), un seuil d'affinité t et construit les classes une à une. On désigne par C_{open} la classe en cours de construction. L'affectation d'un gène g_i à la classe C_{open} est déterminée par une valeur a_{g_i} dite valeur d'affinité, qui représente la somme des valeurs de proximité entre g_i et tous les éléments contenus dans C_{open}. Si $a_{g_i} \geq t|C_{open}|$, le gène g_i est dit de forte affinité par rapport à C_{open}, et de faible affinité sinon. CAST alterne entre l'ajout des gènes de forte affinité, et la suppression des gènes de faible affinité dans C_{open}. Quand ce processus se stabilise, la formation de la classe C_{open} est terminée, et une nouvelle classe est démarrée. Une classe déjà construite n'intervient pas dans la formation des nouvelles classes. CAST repose sur le seuil d'affinité, t, une variable d'entrée définie par l'utilisateur. Sa détermination est cruciale parce que la taille et le nombre des classes produites par l'algorithme sont directement affectés par ce paramètre. E-CAST (Bellaachia et al. [1]) calcule le seuil de manière dynamique basé uniquement sur les gènes qui ne sont pas encore affectés à une classe.

2.2.5 Critères de validation du nombre de classes

La classification est un processus qui regroupe des données en classes sans aucune connaissance *a priori* sur leurs caractéristiques et leur nombre. Ce qui fait que le recouvrement de la structure naturelle des classes devient une tâche difficile et subjective. Supposons la classification de n profils d'expression de gènes $(g_1, g_2, ..., g_n)$ en un ensemble de N classes $\mathcal{C} = \{C_1, C_2, ..., C_N\}$. De nom-

breux algorithmes (e.g., K-means, SOM) exigent d'abord de définir ce nombre, et l'une des questions les plus importantes est comment l'évaluer afin de trouver la partition qui ajuste au mieux la structure des données. La détermination de cette valeur peut être une tâche particulièrement difficile pour les données de puces à ADN. En effet, les données issues de cette technologie peuvent être quelque peu uniformément distribuées dans l'espace des expressions de gènes, et donc peuvent ne pas avoir de solution présentant des caratéristiques d'isolement et de compacité de classes. Des indices de validité classiques tels que la silhouette width (Rousseeuw [83]), l'indice de Calinski et Harabasz (Calinski and Harabasz [11]) ou la statistique de Gap (Tibshirani et al. [101]) sont largement utilisés pour les techniques de classification fondées sur les métriques. D'autres approches, appliquées aux profils d'expression de gènes sont proposées (e.g., Yeung et al. [114], Datta and Datta [96]). Dans cette section, nous ne présentons que celles qui sont le plus souvent employées dans la littérature de l'analyse des expressions de gènes . L'approche classique consiste à exécuter un algorithme de classification sur les données pour plusieurs valeurs de N et d'obtenir des partitions différentes. Celle qui optimise la valeur de l'indice considéré sera retenue comme étant la plus appropriée. La validité du nombre de classes sera assurée par l'évaluation de la qualité des résultats issus de la classification. Plusieurs méthodes sont alors proposées.

Indice de Calinski et Harabasz

La qualité d'une classification peut être mesurée en termes d'homogénéité et de séparabilité des classes. Dans ce contexte, Calinski et Harabasz [11] ont proposé l'indice noté CH pour identifier des classes "homogènes et bien séparées". Il peut être formulé par :

$$CH(N) = \frac{BSS(N)/(N-1)}{WSS(N)/(n-N)} \tag{2.1}$$

où $BSS(.)$ et $WSS(.)$ représentent respectivement les dissimilarités inter-classes et intra-classe définies par :

$$BSS(N) = \frac{1}{2}\sum_{l=1}^{N}\sum_{g_i \notin C_l, g_j \in C_l} d^2(g_i, g_j), WSS(N) = \frac{1}{2}\sum_{l=1}^{N}\sum_{g_i, g_j \in C_l} d^2(g_i, g_j) \tag{2.2}$$

où d désigne la distance euclidienne. L'idée derrière l'indice de CH consiste à calculer la somme des carrés des écarts par paires de gènes (distances) entre les classes et les comparer à la somme des carrés des erreurs internes pour chaque classe. La valeur de N qui maximise $CH(N)$ est considérée comme le bon nombre de classes.

La statistique Gap de Tibshirani et al.

Tibshirani et al. [101] proposent la statistique de Gap qui consiste à calculer la distance moyenne intra-classe :

$$W(N) = \frac{1}{2}\sum_{l=1}^{N}\sum_{g_i,g_j\in C_l}\frac{1}{|C_l|}d(g_i,g_j) \tag{2.3}$$

et $E_n(log(W(N)))$ l'espérance de la statistique $log(W(N))$ sous une hypothèse nulle exprimant l'absence de structure du nombre de classes fixé. Ensuite, les deux courbes correspondant à $log(W(N))$ et $E_n(log(W(N)))$ en fonction du nombre de classes $l \in [1, N]$ sont construites et comparées. Le nombre de classes à retenir est la valeur N correspondant au plus grand écart entre les deux courbes.

$$Gap_n(N) = E_n(log(W(N))) - log(W(N)) \tag{2.4}$$

Notons que cette estimation est plus générale dans le sens où elle est applicable à toutes mesures de distance.

Figure Of Merit (FOM)

Un autre aspect de la validité du nombre de classes se focalise sur leur fiabilité. Yeung et al. [114], inspiré par l'approche de $Jackknife$ (Efron [25]), ont défini une *Figure Of Merit* (FOM) pour la validation du nombre de classes à considérer. Intuitivement, une méthode de classification a une signification biologique possible que si les gènes qui sont dans une même classe ont tendance à avoir des niveaux d'expression similaires dans d'autres expériences qui n'ont pas été utilisées pour former les classes. Supposons C_1, C_2,...,C_N, N classes issues des résultats d'une méthode de classification de profils de gènes considérés aux instants t_1,...,(t_{e-1}),(t_{e+1}),...,tp et t_e utilisé pour estimer le "pouvoir prédictif" de l'algorithme, la *Figure Of Merit* par rapport à t_e et N est définie par :

$$FOM(t_e, N) = \sqrt{\frac{1}{n}\sum_{k=1}^{K}\sum_{g_\alpha\in C_k}(R(g_\alpha, t_e) - \mu_{C_k}(t_e))} \tag{2.5}$$

où t_e est l'intant imputé, $R(g_\alpha, t_e)$ est le niveau d'expression du gène g_α à l'instant t_e et $\mu_{C_k}(t_e)$ est la moyenne des niveaux d'expression des gènes appartenant à la classe C_k à l'instant t_e. Chacun des instants d'échantillonnage peut être utilisé comme instant imputé et la *Figure Of Merit* globale est définie comme $FOM(N) = \sum_{t_e=1}^{p} FOM(t_e, N)$. Elle est censée évaluer une stabilité de la méthode de classification face à de petites perturbations. Une petite valeur de FOM indique un pouvoir prédictif élevé pour l'algorithme considéré, et donc une grande fiabilité des classes résultantes. Un inconvénient de FOM telle que définie est que la diminution de sa valeur en fonction du nombre de

classes peut être artificielle. Ceci est dû au fait que plus de classes signifie une taille moyenne plus faible pour les classes. La *FOM* ajustée est *Figure Of Merit* divisée par un facteur qui compense ce biais statistique avec un nombre de classes élevé :

$$FOM^a(N) = \frac{FOM(N)}{\sqrt{(n-N)/N}} \tag{2.6}$$

Il faut noter que la bonne performance d'un indice pour certaines données ne garantit pas le même comportement avec différentes données. Il est conseillé de faire la synthèse des résultats de plusieurs techniques pour sélectionner le bon nombre de classes. Et si une des limites des nombreuses méthodes de classification non paramétriques est leur incapacité à déterminer le nombre de classes, ceci est lié par le fait qu'il n'y a pas de définition claire de ce que c'est que une classe. Une solution alternative est l'utilisation de techniques paramétriques.

2.3 Techniques de classification paramétrique

Les approches fondées sur des modèles fournissent un cadre statistique pour modéliser la structure de classe. Des modèles statistiques tels le mélange de modèles gaussiens, les modèles auto-regressifs et les modèles de Markov cachés sont aujourd'hui largement utilisés dans le cadre de l'analyse des gènes (Liang et Wang [59], Dortet-Bernadet et Wicker [23], Hestilow et Huang [40]). Ils font l'hypothèse que les profils d'expression peuvent être caractérisés par un processus aléatoire paramétrique.

2.3.1 Classification par mélange de modèles gaussiens

Le mélange de modèles gaussiens prend en compte l'information probabiliste simultanément sur les observations et sur les classes. Le principe de ce type de méthode est fondé sur l'hypothèse que si un gène i $(i = 1, ..., n)$, représenté par son vecteur d'expression $g_i = (x_{i,t_1}, x_{i,t_2}, ..., x_{i,t_p})$, appartient à la classe C_k $(k = 1, ..., N)$, alors g_i suit $p_k(x)$ qui est la loi de distribution de la classe C_k. La fonction de densité des gènes est donc une loi de mélange :

$$p(x) = \sum_{k=1}^{N} \pi_k p_k(x) \tag{2.7}$$

la proportion π_k définit la probabilité *a priori* qu'un gène i appartienne à la classe C_k. Le but est d'estimer les paramètres des densités $p_k(x)$, les proportions π_k et le nombre de groupes N à partir de l'échantillon des gènes observés. Les gènes sont alors alloués aux classes en utilisant la règle de Bayes.

Biologiquement, il arrive que des gènes interviennent dans un processus de

transition, ou soient impliqués dans plusieurs processus biologiques. Ainsi, la méthode à base de mélange s'avère particulièrement intéressante puisqu'elle permet de modéliser cette propriété en permettant à un gène d'appartenir simultanément à plusieures classes différentes. En effet, contrairement aux méthodes décrites précédemment, celle-ci permet à chaque gène i d'avoir une probabilité *a posteriori* d'appartenir aux différentes classes. Dans leurs travaux, Yeung et al. [114] ont utilisé la classification par mélange de modèles gaussiens et l'ont comparé à d'autres algorithmes (e.g., CAST), le nombre de classes étant déterminé par le critère BIC (*Bayesian Information Criterion*). Ces auteurs ont eu de meilleurs résultats avec des données synthétisées, mais ils ont obtenu des résultats comparables sur de vraies données de profils d'expression de gènes. Le principal défaut de cette méthode est l'hypotèse que les données correspondent à une distribution gaussienne, ce qui n'est pas toujours le cas.

Notons que jusqu'ici, les approches que nous avons citées ignorent l'aspect temporel des données. Par exemple, ici, les densités sont essentiellement fondées sur les niveaux d'expression et ignorent la dépendance des instants d'échantillonnage. Afin de considérer cet aspect, il est nécessaire de modéliser les données.

2.3.2 Classification par modélisation des profils temporels

Le problème de modélisation des expressions de gènes a été étudié par plusieurs auteurs (Bar-Joseph et al. [5], Luan et Li [62], Möller-Levet et Yin. [69]). L'idée générale consiste à représenter les profils des expressions de gènes par des fonctions continues dépendant du temps. Par exemple, Luan et Li [62] ont mis en application le modèle à effets mixtes pour analyser les données temporelles d'expression de gènes en les modélisant par les *B-splines*. Tandis que Bar-Joseph et al. [5] représentent les profils temporels des gènes par des courbes continues avec les *cubic splines* (un ensemble de polynômes cubiques). La classification est effectuée avec un algorithme de type E-M et le nombre de classes est déterminé automatiquement dans la pénalisation du maximum de vraisemblance.

Les modèles auto-regressifs

L'approche basée sur le modèle dynamique tient explicitement compte de la nature dynamique des profils d'expression observés au cours du temps. Elle considère ces profils comme un ensemble de séries temporelles $S = \{S_1, S_2, ..., S_m\}$, réalisation d'une même mesure à des instants différents, non indépendantes et identiquement distribuées, et générées par un nombre inconnu de processus stochastiques. Son but est de partitionner l'ensemble des profils de gènes en classes disjointes, de sorte que ceux qui se trouvent dans la même classe soient

générés par le même processus. Une classe correspondrait à un processus, ce qui donnerait directement une information sur la structure temporelle des gènes groupés. Pour représenter les classes, la méthode emploie des processus auto-régressifs d'ordre p (AR(p)), chaque processus défini une classe. Deux grandes questions sont en jeu dans le choix du modèle AR(p) : l'ordre du modèle et la méthode d'estimation des paramètres.

Ramoni et al. [82] ont proposé un algorithme de classification basé sur le modèle Bayésien, où les profils de gènes sont représentés par des modèles auto-regressifs. Il consiste à chercher l'ensemble M_c des modèles ayant la plus forte probabilité *a posteriori* générant les profils de gènes observés. Mais, le nombre de possibilités augmente considérablement avec le nombre de gènes étudiés. L'espace des modèles est donc parcouru en utilisant une stratégie agglomérative sur la base d'une heuristique fusionnant les classes les plus similaires, selon la distance euclidienne, afin d'estimer le nombre de classes et les paramètres du modèle de chaque classe. Après plusieurs tests effectués, les auteurs trouvent que les meilleurs résultats de leur méthode sont fournis par les modèles auto-regressifs d'ordre $p = 1$.

Wu et al. [108] proposent une méthode de classification basée sur les modèles auto-regressifs pour les profils temporels de gènes. La méthode proposée considère l'ensemble des profils temporels de gènes comme un ensemble de séries temporelles, générées par un nombre de processus stochastiques. Chaque processus stochastique définit une classe et est décrit par un modèle auto-regressif. Un algorithme de réallocation itérative identifie les paramètres du modèle et les probabilités *a posteriori* sont employées pour affecter chaque gène à la classe appropriée. Les auteurs analysent trois jeux de données de profils temporels d'expression régulés au cours du cycle cellulaire et observés à intervalles de temps réguliers. Ces jeux de données sont respectivement constitués de 701 gènes dont les niveaux d'expression sont prélevés sur 18 instants, 789 gènes observés sur 14 instants et 1590 gènes avec 11 instants d'échantillonnage. Pour chaque jeu, la méthode de classification proposée avec différent ordre ($p = 1, 2, 3$) et l'algorithme de K-means, avec une mesure de distance fondée sur le coefficient de corrélation, sont appliqués. Une méthode bootstrap et l'indice Rand corrigé estiment la qualité des classifications. Les résultats montrent que l'augmentation de l'ordre des modèles auto-regressifs tend à diminuer la qualité de la classification proposée. Par conséquent, la meilleure qualité est fournie par le modèle auto-regressif d'ordre 1. L'application de la méthode proposée sur les trois jeux de données et avec les trois ordres a montré que la qualité de la classification est meilleure que celle produite par l'algorithme des K-means.

Le modèle auto-régressif est limité par l'exigence de stationnarité des séries. La structure temporelle est entièrement déterminée par le processus AR, mais le pas de l'échantillonnage n'est pas pris en compte. En effet, avec le modèle

$AR(1)$ la valeur x_{t_i} dépend de $x_{t_{i-1}}$ et d'un coefficient a, et donc la relation entre deux mesures consécutives reste toujours constante. Pour corriger ce problème, on peut ajouter des points par interpolation pour régulariser l'échantillonnage et pouvoir appliquer cet algorithme ou utiliser un algorithme basé sur un autre modèle que les séquences AR.

Les modèles de Markov cachés

Une autre façon d'exploiter les dépendances temporelles est l'utilisation de modèles de Markov cachés (HMM pour *Hidden Markov Models*). Un HMM peut être considéré comme un automate dont les états sont stochastiques et en nombre fini permettant de reconnaître des séquences d'états dits *cachés* à partir d'observations bruitées. Un HMM est formellement défini par les éléments suivants :

1. M le nombre d'états, $s_j^{t_k}$ l'état j à l'instant t_k appartenant au cadre de discernement $S^{t_k} = \{s_1^{t_k}, ..., s_M^{t_k}\}$.
2. La matrice de transition $A = [a_{jl}]_{1 \leq j,l \leq M}$ constituée de probabilités de transition entre les états associés. Dans cette matrice, l'élément a_{jl} représente la probabilité de passer de l'état $s_j^{t_k}$ à l'état $s_l^{t_k}$ et est défini par $a_{jl} = P(s_l^{t_k} | s_j^{t_{k-1}})$.
3. L'ensemble $B = \{b_j(x_{i,t_k}), 1 \leq j \leq M, 1 \leq k \leq p\}$ où $b_j(x_{i,tk}) = P(x_{i,tk} | s_j^{t_k})$ est la vraisemblance de l'état j conditionnelle à l'observation courante, à l'instant t_k.
4. $\pi = \{\pi_j, 1 \leq j \leq M\}$ où $\pi_j = P(\text{état initial} = s_j)$, la probabilité que s_j soit l'état initial.

L'ensemble de ces paramètres $\lambda = \{A, B, \pi\}$ constitue le modèle de Markov caché, qui est généralement désigné HMM λ.

Plusieurs auteurs ont utilisé les modèles $HMMs$ pour la classification des profils d'expression de gènes (Schliep et al. [86], Ji et al. [45], Zeng et Garcias-Frias [116]...). Par exemple, dans Schliep et al. [86], chaque classe est représentée par un HMM dans lequel une distribution de type gaussien est associée à chaque état. La méthode commence par un apprentissage partiellement supervivé. Un ensemble de 19 $HMMs$, représentant les classes, est choisi. Chacun des $HMMs$ a 9 états et englobe une forme typiquement qualitative ($up - down - regulated$). Les auteurs partent d'un nombre de classes fixé à 19 et utilisent un algorithme itératif pour ajuster ces modèles et affecter les gènes aux différentes classes de façon à maximiser la vraisemblance conjointe. La méthode est appliquée sur 1044 profils temporels de gènes de la levure, régulièrement échantillonnés sur deux cycles cellulaires, et possédant chacun 18 instants. Elle a permis aux auteurs d'utiliser de la connaissance *a priori* puisqu'elle est disponible dans de nombreux contextes biologiques où la réponse de certains gènes est déjà connue.

2.4 Conclusion

Dans ce chapitre, nous avons revu une série d'approches d'analyse de données de puces à ADN, en se focalisant en particulier sur les techniques de classification des gènes sur la base de leurs profils temporels d'expression. Le but est d'identifier les gènes qui sont coexprimés. Dans ce contexte, deux types d'approches sont adoptées : celle fondée sur des techniques non-paramétriques et celle portant sur des techniques paramétriques.

La première approche est purement descriptive et se fonde sur des mesures de proximité qui comparent directement les observations ou des caractéristiques extraites des données. Dans la seconde approche, on fait l'hypothèse que les données sont issues d'un mélange de modèles. Les classes sont décrites par des modèles probabilistes et les paramètres des modèles varient d'une classe à l'autre. L'un des principaux moteurs de cette approche est l'hypothèse que les profils temporels d'expression de gènes sont générés par des modèles dependant du temps, en ce sens que l'état actuel est fonction de l'état précédent de la cellule. Par conséquent, ces méthodes tentent de quantifier cette hypothèse. Les algorithmes de classification conventionnels utilisés dans le contexte non-paramétrique et revus ici, sont devenus très populaires mais généralement nécessitent de leur fournir le nombre de classes à générer *a priori*. Ils sont fondés souvent sur des mesures telles que la distance Euclidienne ou les distances basées sur les coefficients de corrélations. Dans ce cas, ils se reposent sur l'hypothèse que les niveaux d'expression, observés pour chacun des gènes, sont indépendants et identiquement distribués, donc inadaptés pour des données de profils temporels d'expression de gènes. Pour pallier cette limite, il est donc nécessaire d'utiliser des mesures de proximité qui considèrent la dépendance temporelle des données, ce travail est l'objet du chapitre suivant.

3

Mesures de proximité entre des séries temporelles

Sommaire

3.1 Caractéristiques et prétraitements des séries temporelles 32
3.1.1 Régularité des séries temporelles 32
3.1.2 Séries temporelles de longueurs différentes 32
3.1.3 Décalage temporel 33
3.2 Mesures de proximités entre des séries temporelles 33
3.2.1 Notations pour une formalisation unifiée 34
3.2.2 Mesures de proximité fondées sur les valeurs 34
3.2.3 Mesures de proximité fondées sur la forme 37
3.2.4 Proximité entre formes : approches concurrentes . . 40
3.2.5 Mesures de proximité alliant forme et valeurs 44
3.3 Caractéristiques des profils d'expression de gènes au cours du cycle cellulaire 46
3.4 Mesures de proximité entre des profils d'expression de gènes 47
3.5 Conclusion 49

Le principal challenge pour l'évaluation de la proximité entre des séries temporelles est l'intégration de l'information de dépendance temporelle. Nous distinguons dans la littérature au moins deux approches. 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., Garcia-Escudero and Gordaliza [30], Serban and Wasserman [87], Caiado et al. [10], Kakizawa et al. [49], Maharaj [49], Bar-Joseph et al. [5]). 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., Heckman and Zamar [38], Keller and Wittfeld [51], Oates et al. [73], Peddada et al. [78]).

Dans le cadre des approches non-paramétriques, ce chapitre introduit les principales mesures de proximités entre des séries temporelles. Préalablement, nous discutons en section 3.1 des principales approches de prétraitement des séries temporelles. Nous présentons ensuite deux catégories de mesures de proximité : celles fondées sur les valeurs et ignorant la forme des séries temporelles, et celles fondées sur la forme des séries. Enfin, nous proposons de nouveaux modèles alliant formes et valeurs. Après la définition des caractéristiques des profils temporels d'expression de gènes au cours du cycle cellulaire, nous proposons, dans la section 3.4, un choix de mesures pertinentes en vue de leur classification ou de leur classement .

3.1 Caractéristiques et prétraitements des séries temporelles

Une série temporelle est une suite d'observations numériques représentant l'évolution d'une quantité spécifique au cours du temps. Les séries temporelles à analyser sont souvent de caractéristiques différentes et dotées de variations diverses. Dans cette section nous présentons les principales caractéristiques rencontrées et prétraitements classiquement effectués.

3.1.1 Régularité des séries temporelles

Les instants d'observations d'une série peuvent être équidistants ou variés les uns des autres, on parlera de série régulière ou irrégulière. Certaines méthodes exigent la régularisation des séries temporelles. Ceci est effectué par la modélisation des séries ($B-splines$, ...) et interpolation des instants d'observations.

3.1.2 Séries temporelles de longueurs différentes

Soit $S_1 = (u_1, ..., u_p)$ et $S_2 = (v_1, ..., v_q)$ deux séries temporelles décrites respectivement par p et q observations effectuées aux instants $(t_1, ..., t_p)$ et $(t'_1, ..., t'_q)$. La différence de longueur des séries ($p \neq q$) est prise en compte de manière différente selon la mesure de proximité utilisée. En effet, certaines mesures de proximité se limitent aux valeurs observées aux même instants (e.g., distance euclidienne voir § 3.2.2), d'autres procèdent préalabement à l'alignement

des instants d'observations des deux séries (e.g., dynamic time warping voir § 3.2.2).

3.1.3 Décalage temporel

L'occurrence d'un événement au sein d'une série temporelle peut se produire de manière plus ou moins différée dans le temps, appelé ici décalage temporel (figure 3.1). Les mesures de proximité doivent prendre en compte la présence de décalages temporels dans les séries (e.g., dynamic time warping). De nom-

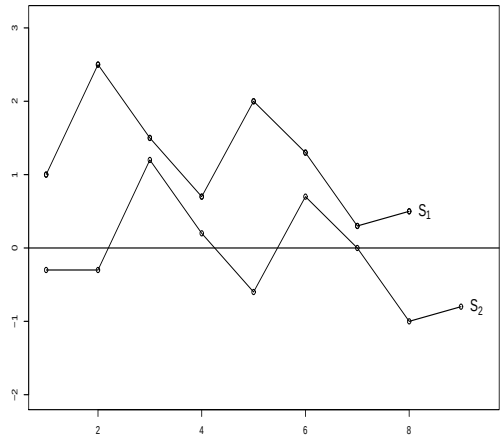


FIG. 3.1 – Décalage temporel des séries S_1 et S_2.

breux travaux ont porté, ces dernières années, sur les mesures de proximités entre séries temporelles. Ces propositions sont souvent liées à l'espace de description des données temporelles choisi (espace temps, espace fréquences,...), aux propriétés des séries temporelles à comparer (régularité ou non, égalité ou non des longueurs, égalité ou non des durées,...), et enfin à la manière d'aborder les problèmes de décalage temporel. Dans ce qui suit, nous présentons deux catégories de mesures de proximité : celles fondées sur les valeurs des séries et celles fondées sur la forme des séries.

3.2 Mesures de proximités entre des séries temporelles

Cette section présente, dans un formalisme unifié, les principales mesures de proximité entre des séries temporelles. Nous définissons d'abord, les mesures fondées sur les valeurs : la dynamic time warping, la distance Euclidienne, la distance de Manhattan et la distance de Fréchet. Nous présentons, ensuite, les mesures fondées sur la forme, principalement le coefficient de corrélation de

Pearson et le coefficient de corrélation temporelle. Enfin, nous définissons un modèle de mesure de proximité alliant forme et valeurs.

3.2.1 Notations pour une formalisation unifiée

Nous définissons un alignement r entre deux séries S_1 et S_2 par la séquence de m couples d'observations : $((u_{a_1}, v_{b_1}), (u_{a_2}, v_{b_2}), ..., (u_{a_m}, v_{b_m}))$, avec $a_i \in \{1, .., p\}$, $b_i \in \{1, ..q\}$ vérifiant, pour $i \in \{1, .., m-1\}$, les contraintes suivantes :

$$a_1 = 1,\ a_m = p,\ a_{i+1} = a_i \text{ ou } a_i + 1 \quad \text{, et} \quad b_1 = 1,\ b_m = q,\ b_{i+1} = b_i \text{ ou } b_i + 1.$$

Soit M l'ensemble des alignements, $R \subset M$ un sous-ensemble d'alignements vérifiant des contraintes supplémentaires et $c(r)$ une fonction coût d'un alignement ($r \in R$) mesurant l'écart entre les valeurs couplées dans r. Les mesures de proximité entre des séries temporelles peuvent être exprimées comme un problème de minimisation de la fonction de coût $c(r)$ sur un ensemble d'alignements R.

$$dUnif_{(c,R)}(S_1, S_2) \quad = \quad \min_{r \in R} c(r) \tag{3.1}$$

Sans perte de généralité, dans ce qui suit, nous désignerons par m la longueur de l'alignement considéré r.

3.2.2 Mesures de proximité fondées sur les valeurs

Nous introduisons ici quatre mesures de proximité conventionnelles fondées sur les écarts entre les valeurs des séries.

La dynamic time warping (DTW)

Soit $R \subset M$ un ensemble d'alignements et $c(r) = \sum_{i=1}^{m} |u_{a_i} - v_{b_i}|$ la fonction coût d'un alignement r. La dynamic time warping (Kruskall and Liberman 1983) est définie :

$$d_{Dtw}(S_1, S_2) \quad = \quad \min_{r \in R} c(r) = \min_{r \in R} \sum_{i=1}^{m} |u_{a_i} - v_{b_i}| \tag{3.2}$$

Ici, un alignement r entre deux séries S_1 et S_2 peut être représenté graphiquement par un chemin allant du point $h_1(1,1)$ au point $h_K(p,q)$ d'une grille de dimension $p \times q$ où la case (i,j) caractérise le couple (u_i, v_j) (figure 3.2). Ainsi, l'évaluation de la proximité entre S_1 et S_2 au sens de la DTW correspond à la recherche d'un chemin de longueur minimale.

Notons qu'à chaque nouvelle définition du couple (c, R), correspond une nouvelle variante de la dynamic time warping. Par exemple, en sont des variantes : les distances euclidienne, de Manhattan et de Fréchet définies ci-dessous.

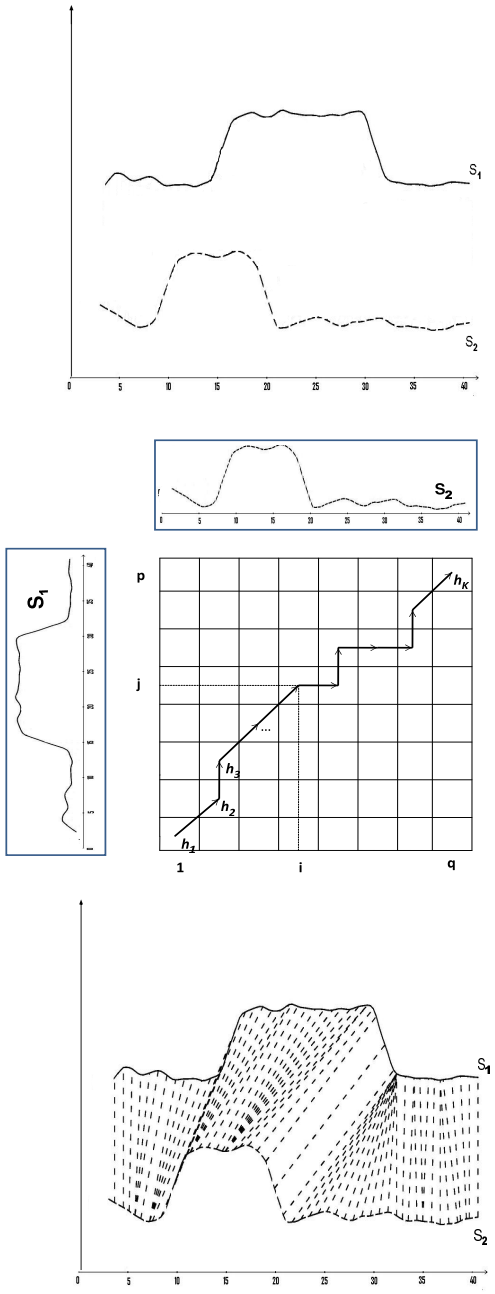


FIG. 3.2 – Exemple de deux séries temporelles S_1 et S_2 (en haut). La recherche du chemin optimal entre les deux séries par la technique de la DTW (au milieu). Elle capture le décalage temporel le long de l'axe des temps (en bas).

La distance de Fréchet

Pour un ensemble d'alignements $R \subset M$ et une fonction coût $c(r) = \max_{i=1}^{m} |u_{a_i} - v_{b_i}|$, la mesure de proximité $dUnif$ définit la distance de Fréchet (Fréchet

1906) :

$$d_F(S_1, S_2) \quad = \quad \min_{r \in R} c(r) = \min_{r \in R} \left(\max_{i=1}^{m} |u_{a_i} - v_{b_i}| \right) \tag{3.3}$$

Dans tout ce qui suit, nous considérons le cas particulier des séries temporelles de même longueur ($m = p = q$).

Les distances Euclidienne et de Manhattan

La minimisation de la fonction coût d'un alignement $c(r) = (\sum_{i=1}^{m}(u_i - v_i)^2)^{\frac{1}{2}}$ sur le singleton $R = \{r_0\}$, $r_0 = ((u_1, v_1), (u_2, v_2), ..., (u_m, v_m))$ définit la distance Euclidienne :

$$d_E(S_1, S_2) \quad = \quad \min_{r \in R} c(r) = c(r_0) = \left(\sum_{i=1}^{m} (u_i - v_i)^2 \right)^{\frac{1}{2}}, \tag{3.4}$$

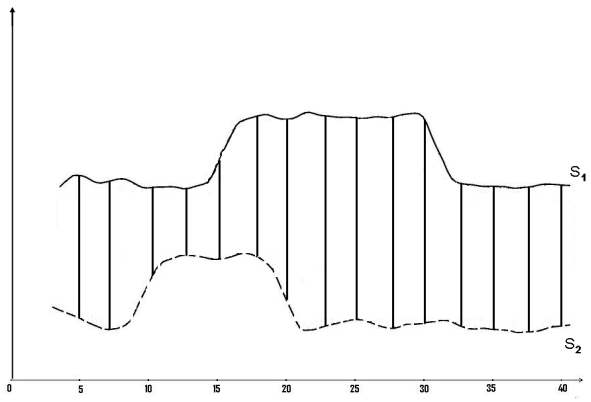


FIG. 3.3 – La proximité définie par une distance euclidienne dépend uniquement de l'écart entre les observations prises aux mêmes instants.

Minimiser la fonction coût $c(r) = \sum_{i=1}^{m} |u_i - v_i|$ sur $R = \{r_0\}$ définit la distance de Manhattan :

$$d_M(S_1, S_2) \quad = \quad \min_{r \in R} c(r) = c(r_0) = \sum_{i=1}^{m} |u_i - v_i| \tag{3.5}$$

Notons que les mesures conventionnelles ci-dessus impliquent des fonctions de coûts $c(r)$ fondées sur les écarts entre les valeurs couplées. Prenons, par exemple, les trois séries temporelles $S_1 = (0, 3, -1, 1)$, $S_2 = (2, 1, 0, -1)$, et $S_3 = (0, 8, 5, 8)$. S_1 et S_3 sont proches en forme mais distantes en valeurs, tandis que S_1 et S_2 sont de formes opposées et proches en valeurs (figure 3.4). La distance euclidienne et la dynamic time warping déterminent la série S_1 plus similaire à la série S_2 qu'à la série S_3 avec $d_E(S_1, S_2) = 3.60 < d_E(S_1, S_3) = 10.48 < d_E(S_2, S_3) = 12.60$, et $d_{Dtw}(S_1, S_2) = 7 < d_{Dtw}(S_1, S_3) = 18 < d_{Dtw}(S_2, S_3) = 23$.

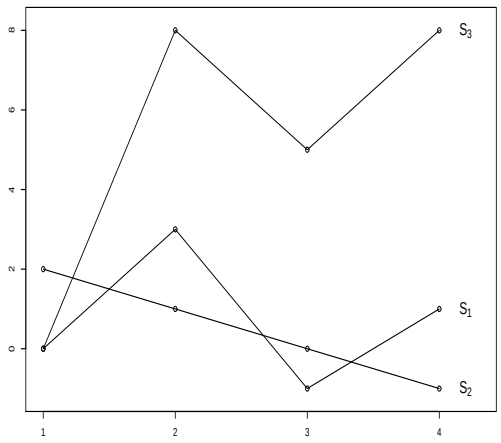


FIG. 3.4 – Exemples de séries temporelles proches en valeurs et de formes opposées (S_1 et S_2), vs. distantes en valeurs et de formes similaires (S_1 et S_3).

3.2.3 Mesures de proximité fondées sur la forme

Cette section introduit les principales mesures de proximité basées sur la forme des séries temporelles. Nous présentons d'abord les mesures impliquant les coefficients de corrélation de Pearson, de Spearman ou de corrélation temporelle.

Le coefficient de corrélation de Pearson

De nombreux travaux utilisent le coefficient de corrélation de Pearson comme mesure de proximité entre la forme des séries temporelles. Soit $S_1 = (u_1, ..., u_m)$ et $S_2 = (v_1, ..., v_m)$ deux séries temporelles observées aux intants $(t_1, ..., t_m)$. S_1 et S_2 sont de formes similaires si et seulement si dans chaque période d'observation $[t_i, t_{i+1}]$, S_1 et S_2 croissent ou décroissent simultanément avec le même taux d'accroissement. En revanche, S_1 et S_2 sont de formes opposées si dans chaque période d'observation $[t_i, t_{i+1}]$ où S_1 croit , S_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 entre la forme des séries, considérons son expression basée sur les différences entre les valeurs prises :

$$Cor(S_1, S_2) = \frac{\sum_{i,i'}(u_i - u_{i'})(v_i - v_{i'})}{\sqrt{\sum_{i,i'}(u_i - u_{i'})^2}\sqrt{\sum_{i,i'}(v_i - v_{i'})^2}}. \quad (3.6)$$

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. Comme première limite, le coefficient de corrélation peut surestimer la proximité entre les formes. En effet dans le cas de données dotées d'un effet de tendance, deux

séries temporelles de formes opposées peuvent avoir un coefficient de corrélation de valeur positive forte. La figure 3.5 (a) illustre ce cas avec un coefficient de corrélation de 0.74 pour les séries temporelles S_1 et S_2 de formes opposées. La deuxième limite du coefficient de corrélation est sa sensibilité face aux valeurs extrêmes. En effet, il peut être faussé par une unique valeur observée simultanément sur deux séries S_3 et S_4 et pouvant même inverser son signe. La figure 3.5 (b) illustre cet aspect, le coefficient de corrélation est de 0.73, en considérant toutes les observations, et chute à -0.95 après la suppression des observations de l'instant $t = 7$. Le tableau 3.1 donne les valeurs observées des séries S_1, S_2, S_3 et S_4 de la figure 3.5.

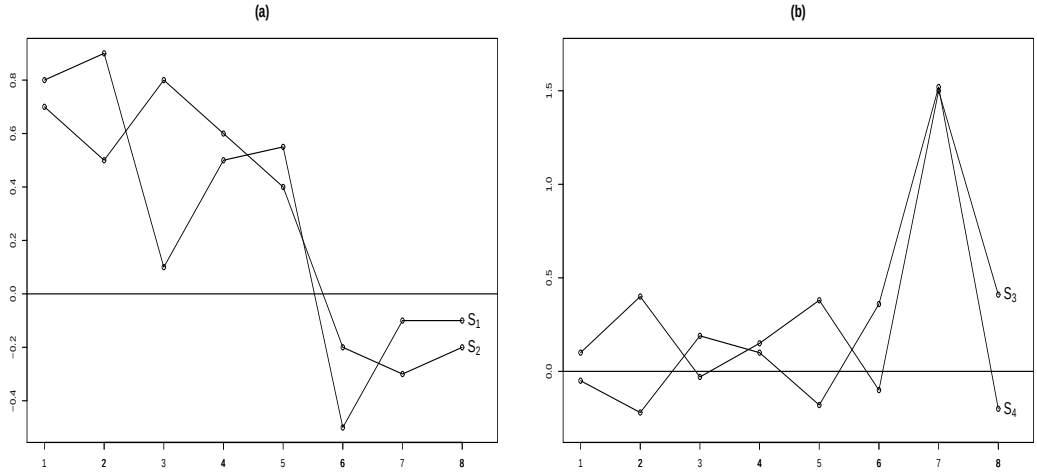


FIG. 3.5 – Illustration des limites du coefficient de corrélation de Pearson face aux effets de tendance (a) et face aux valeurs extrêmes (b). Les valeurs enregistrées des coefficients sont significativement élevées 0.74 et 0.73 respectivement. Alors que les séries sont quasiment opposées.

	t=1	t=2	t=3	t=4	t=5	t=6	t=7	t=8
S_1	0.80	0.90	0.10	0.50	0.55	-0.50	-0.10	-0.10
S_2	0.70	0.50	0.80	0.60	0.40	-0.20	-0.30	-0.20
S_3	0.10	0.40	-0.03	0.15	0.38	-0.10	1.50	0.41
S_4	-0.05	-0.22	0.19	0.10	-0.18	0.36	1.52	-0.20

TAB. 3.1 – Quatre séries temporelles illustrant les limites du coefficient de corrélation de Pearson face aux effets de tendance et aux valeurs extrêmes.

Notons que le problème des valeurs extrêmes, dans les séries temporelles, peut être résolu en utilisant le coefficient de $Jakknife$. Il est défini par :

$$J(S_1, S_2) \quad = \quad min\{\rho_{1,2}^{(1)}, ..., \rho_{1,2}^{(l)}, ..., \rho_{1,2}^{(p)}\} \tag{3.7}$$

où $\rho_{1,2}^{(l)}$ est le coefficient de Pearson de S_1 et S_2 avec la $l$ième mesure éliminée. L'utilisation de ce coefficient annulerait "l'effet de dominance" d'une valeur. Un coefficient de $Jakknife$ généralisé et plus robuste à plusieurs valeurs extrêmes, peut être obtenu. Cependant, c'est une méthode coûteuse en calcul et rarement utilisée (Laurie et al. [56]).

Le coefficient de corrélation temporelle

Pour surmonter les limites du coefficient de corrélation de Pearson, le coefficient de corrélation temporelle est utilisé, il réduit le coefficient de corrélation de Pearson aux différences de premier ordre :

$$Cort(S_1, S_2) = \frac{\sum_i (u_{(i+1)} - u_i)(v_{(i+1)} - v_i)}{\sqrt{\sum_i (u_{(i+1)} - u_i)^2} \sqrt{\sum_i (v_{(i+1)} - v_i)^2}}. \quad (3.8)$$

avec $Cort(S_1, S_2)$ appartenant à $[-1, 1]$. La valeur $Cort(S_1, S_2) = 1$ indique que S_1 et S_2 présentent une forme similaire. La valeur $Cort(S_1, S_2) = -1$ signifie que S_1 et S_2 ont des formes opposées. Enfin, $Cort(S_1, S_2) = 0$ exprime que les taux d'accroissement de S_1 et S_2 sont stochastiquement, linéairement indépendants, identifiant ainsi des séries temporelles de formes différentes (non similaires ni opposées). Notons que le coefficient de corrélation de Pearson et de la corrélation temporelle font l'hypothèse d'un alignement r_0 entre les séries temporelles (i.e., absence de décalage temporel). Pour un alignement donné $r = ((u_{a_1}, v_{b_1}), (u_{a_2}, v_{b_2}), ..., (u_{a_m}, v_{b_m}))$ appartenant à M, une définition plus générale du coefficient de corrélation de Pearson et de la corrélation temporelle est donnée :

$$Cor(S_1, S_2) = \frac{\sum_{i,i'} (u_{a_i} - u_{a'_i})(v_{b_i} - v_{b'_i})}{\sqrt{\sum_{i,i'} (u_{a_i} - u_{a'_i})^2} \sqrt{\sum_{i,i'} (v_{b_i} - v_{b'_i})^2}}, \quad (3.9)$$

$$Cort(S_1, S_2) = \frac{\sum_i (u_{a_i} - u_{a_{i+1}})(v_{b_i} - v_{b_{i+1}})}{\sqrt{\sum_i (u_{a_i} - u_{a_{i+1}})^2} \sqrt{\sum_i (v_{b_i} - v_{b_{i+1}})^2}}. \quad (3.10)$$

Les coefficients $Cor(S_1, S_2)$ et $Cort(S_1, S_2)$ sont tout simplement notée $Cor(r)$ et $Cort(r)$ dans la suite.

Coefficient des rangs de Spearman

Soit $S_1 = (u_1, ..., u_m)$ et $S_2 = (v_1, ..., v_m)$ deux séries temporelles, R_{S_1}=(ρ_{u_1},..., ρ_{u_m}) et R_{S_2}=(ρ_{v_1},..., ρ_{v_m}) deux séries décrivant respectivement la séquence des rangs des observations de S_1 et S_2 aux instants $(t_1, ..., t_m)$; avec ρ_{u_i} (resp. ρ_{v_i}) le rang de la valeur u_i dans S_1 (resp. v_i dans S_2). Considérons une fonction de coût $c(r) = 1 - \frac{6}{m(m^2-1)} \sum_{i=1}^{m} (\rho_{u_i} - \rho_{v_i})^2$ définie sur l'espace $r_0 = ((\rho_{u_1}, \rho_{v_1}), (\rho_{u_2}, \rho_{v_2}), ..., (\rho_{u_m}, \rho_{v_m}))$. La mesure de similarité fondée sur le coefficient de Spearman est définie :

$$d_{R(S_1,S_2)} = \min_{r \in \{r_0\}} c(r) = c(r_0) = 1 - \frac{6}{m(m^2-1)} \sum_{i=1}^{m} (\rho_{u_i} - \rho_{v_i})^2 \quad (3.11)$$

Introduisons maintenant trois mesures dédiées aux séries temporelles de petites longueurs. On définit d'abord la distance qualitative qui compare les états de variation proposée par Todorovski et al. [102], ensuite celle proposée par Möller-Levet et al. [68] utilisant les pentes pour comparer les formes entre séries et enfin celles reposant sur des coefficients de corrélation et d'indice proposées par de Son et Baek [91].

3.2.4 Proximité entre formes : approches concurrentes

La distance qualitative de Todorovski

Pour mesurer la proximité entre les formes des séries temporelles, Todorovski et al. [102] proposent une distance qualitative comparant les états de variation sur toutes les périodes $[t_i, t_j]$. Trois états sont considérés : accroissement ($u_i < u_j$), décroissance ($u_i > u_j$), et stable ($u_i \approx u_j$). Soit $q_1(u_i, u_j)$, $q_2(v_i, v_j)$ les états de variation sur la période $[t_i, t_j]$ dans S_1 et S_2 respectivement. La distance qualitative D_q mesurant l'écart entre les formes de S_1 et S_2 est définie :

$$D_q(S_1, S_2) = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \frac{2 \cdot Diff(q_1(u_i, u_j), q_2(v_i, v_j))}{m(m-1)} \tag{3.12}$$

avec $Diff(q_1, q_2)$ la fonction de dissimilarité mesurant l'écart entre deux états de variation (tableau 3.2).

	$Diff(q_1, q_2)$	q_1 accroissement	q_1 stable	q_1 décroissance
	accroissement	0	0.5	1
q_2	stable	0.5	0	0.5
	décroissance	1	0.5	0

TAB. 3.2 – Définition de la fonction $Diff$. Pour deux séries S_1 et S_2, l'écart entre un accroissement et une décroissance est égal à 1, alors que l'écart entre un accroissement (ou une décroissance) et un état stable est égal à 0.5.

Le facteur $\frac{2}{m(m-1)}$ normalise les valeurs de D_q dans l'intervalle $[0; 1]$. D_q compte le nombre de désaccords entre les états de variation de S_1 et S_2. Elle est égale à 0 si S_1 et S_2 ont les mêmes variations pour toutes les périodes $[t_i, t_j]$. La distance maximale de 1 est atteinte dans le cas où S_1 croit à chaque fois que S_2 décroit et vise-versa. Notons que D_q suppose l'indépendance des observations en impliquant toutes les paires d'instants (t_i, t_j). D_q souffre ainsi de la même limite que le coefficient de corrélation de Pearson. La figure 3.6, donne l'exemple de deux séries temporelles dont l'observation de S_1 diminue lorsque celle de S_2 augmente sauf dans la période $[t_5, t_6]$. Notons l'échec de D_q à déterminer les deux formes comme opposées avec une valeur $D_q(S_1, S_2) = 0.33$.

Remarquons que, dans le cas ou seules les paires d'instants consécutifs sont considérées, la valeur de la distance calculée est de 0.86, un résultat reflétant plus significativement l'écart entre les deux formes.

	t=1	t=2	t=3	t=4	t=5	t=6	t=7	t=8
S_1	0.80	0.90	0.10	0.50	0.55	-0.50	-0.10	-0.15
S_2	0.70	0.50	0.80	0.60	0.40	-0.20	-0.30	-0.20

TAB. 3.3 – La description de deux séries opposées illustrant la limite de la distance de Todorovski à mesurer la proximité entre la forme des séries temporelles.

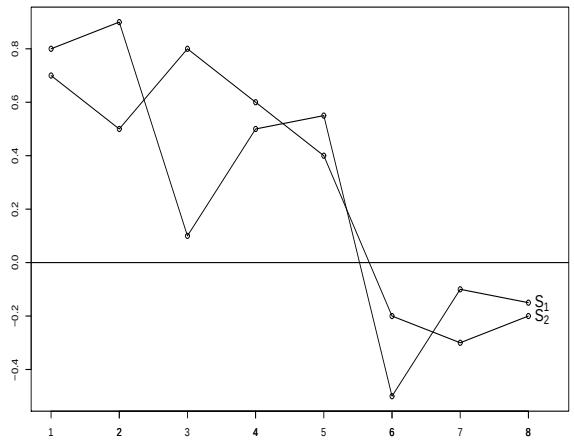


FIG. 3.6 – Deux séries S_1 et S_2 de formes opposées, avec $D_q(S_1, S_2) = 0.33$, illutrant la limite de la distance de Todorovski.

Proximité entre formes par comparaison des pentes.

Une série temporelle $S = (u_1, ..., u_m)$ observée aux instants $(t_1, ..., t_m)$ peut être considérée comme une fonction linéaire par morceaux d'intervalles $[t_i, t_{i+1}]$, $(i = 1, ..., m-1)$. La fonction linéaire entre les instants t_i et $t_{(i+1)}$ est définie par $u_i(t) = \alpha_i t + \beta_i$, $t_i \leq t \leq t_{(i+1)}$, où α_i est la pente et β_i est l'ordonnée à l'origine définies par :

$$\alpha_i = \frac{u_{t_{i+1}} - u_{t_i}}{t_{i+1} - t_i}, \tag{3.13}$$

$$\beta_i = \frac{t_{i+1} u_{t_i} - t_i u_{t_{i+1}}}{t_{i+1} - t_i}. \tag{3.14}$$

La proximité entre formes de deux séries $S_1 = (u_1, ..., u_m)$ et $S_2 = (v_1, ..., v_m)$ peut être estimée en comparant les pentes définies dans chaque intervalle

$[t_i, t_{i+1}]$. Möller-Levet et al. [68] déterminent la distance évaluant l'écart entre les formes des deux séries par :

$$d_{STS}(S_1, S_2) = \Big(\sum_{i=1}^{m-1} \Big(\frac{u_{t_{i+1}} - u_{t_i}}{t_{i+1} - t_i} - \frac{v_{t_{i+1}} - v_{t_i}}{t_{i+1} - t_i} \Big)^2 \Big)^{\frac{1}{2}} \tag{3.15}$$

Elle correspond à la somme des différences au carré des pentes entre deux instants consécutifs. Constatons que cette distance souffre d'un problème d'échelle. En effet, la figure 3.7 présente trois séries temporelles S_1, S_2 et S_3 expliquant ce problème. Les séries S_1 et S_3 présentent les mêmes variations avec des amplitudes différentes. Tandis que S_1 et S_2 sont de formes opposées. Or $d_{STS}(S_1, S_2) = 5.3$ et $d_{STS}(S_1, S_3) = 6.4$, ce résultat traduit la limite de d_{STS} à estimer la distance entre deux formes. Le passage à une standardisation des données permettra de résoudre ce problème (figure 3.7 droite), avec $d_{STS}(S_1, S_2) = 8.7$ et $d_{STS}(S_1, S_3) = 1.8$.

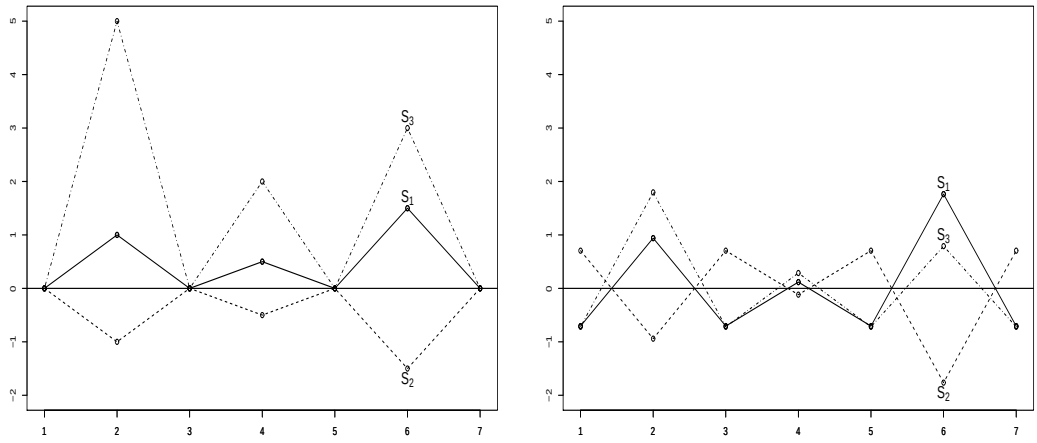


FIG. 3.7 – À gauche, représentation de trois séries S_1, S_2 et S_3 avec $d_{STS}(S_1, S_3) = 6.4$ et $d_{STS}(S_1, S_2) = 5.3$. À droite, le passage à une standardisation des données donne $d_{STS}(S_1, S_2) = 8.7$ et $d_{STS}(S_1, S_3) = 1.8$.

Une alternative pour éviter le problème d'échelle est de considérer l'angle correspondant à la pente (Chen et Dai [13]). À chaque série $S = (u_1, ..., u_m)$ ont fait correspondre le vecteur $\theta_S = (\theta_1, ..., \theta_{m-1})$, où θ_i défini l'angle de la droite reliant les deux observations u_i et u_{i+1} par rapport à l'axe du temps, c'est à

	t=1	t=2	t=3	t=4	t=5	t=6	t=7
S_1	0	1	0	0.5	0	1.5	0
S_2	0	-1	0	-0.5	0	-1.5	0
S_3	0	5	0	2	0	3	0

TAB. 3.4 – Trois séries temporelles montrant la limite de la distance d_{STS}.

dire :

$$\theta_i \quad = \quad arctan\Big(\frac{u_{i+1}-u_i}{t_{i+1}-t_i}\Big), \qquad i=1,2,...,m-1 \tag{3.16}$$

Approche de Son et Baek

Soit deux séries temporelles $S_1=(u_1,...,u_m)$ et $S_2=(v_1,...,v_m)$, pour décrire qualitativement leur évolution on définit le vecteur $L_{S_j,i}$ $(j=1,2)$ par :

$$L_{S_j,i}=\begin{cases} 1 & \alpha_i>0, \\ -1 & \alpha_i<0, \\ 0 & \alpha_i=0, \end{cases} \tag{3.17}$$

où α_i est la pente définie plus haut. Il permet de déterminer l'indice $A_{1,2}$ estimant la concordance en forme des séries S_1 et S_2 :

$$A_{1,2}=\sum_{i=1}^{m-1} I(L_{S_1,i}=L_{S_2,i})/(m-1) \tag{3.18}$$

où $I(D)$ est égale à 1 si D est vrai et 0 sinon. L'indice $A_{1,2}\in[0,1]$ et compte le nombre d'intervalles de temps $[t_i,\ t_{i+1}]$ dans lesquels S_1 et S_2 ont le même sens de variation. L'inconvénient avec cet indice est que l'écart entre les valeurs observées n'est pas considéré. Pour prendre en compte ce facteur, une alternative consiste à définir $A^*_{1,2}$ comme suit :

$$A^*_{1,2}=(Cor(d_1,d_2)+1)/2 \tag{3.19}$$

où Cor est la corrélation de Pearson, $d_1=(d_{1,1},...,d_{1,(m-1)})$, $d_2=(d_{2,1},...,d_{2,(m-1)})$, $d_{1,i}=u_{i+1}-u_i$ et $d_{2,i}=v_{i+1}-v_i$.

L'instant où les séries temporelles atteignent le maximum ou le minimum des mesures observées peut être capital dans certaines applications (e.g., étude du cycle cellulaire). Soient $T_{S_j}^{max}$ et $T_{S_j}^{min}$ $(j=1,2)$ représentant respectivement ces instants. Pour indiquer si les minimum et/ou maximum sont identifiés ou pas aux mêmes instants entre S_1 et S_2, un deuxième indice $M_{1,2}$ est défini :

$$M_{1,2}=\begin{cases} 1 & si\ T_{S_1}^{min}=T_{S_2}^{min}\ et\ T_{S_1}^{max}=T_{S_2}^{max}, \\ 0.5 & si\ T_{S_1}^{min}=T_{S_2}^{min}\ ou\ T_{S_1}^{min}=T_{S_2}^{min}, \\ 0 & si\ T_{S_1}^{min}\neq T_{S_2}^{min}\ et\ T_{S_1}^{max}\neq T_{S_2}^{max}. \end{cases} \tag{3.20}$$

De même, une alternative à $M_{1,2}$, exprimant le rapprochement en terme d'instants max/min des observations est $M^*_{1,2}$:

$$M^*_{1,2} = 1 - \frac{|T^{min}_{S_1} - T^{min}_{S_2}| + |T^{max}_{S_1} - T^{max}_{S_2}|}{2(m-1)} \tag{3.21}$$

Son et Baek [91] proposent quatre mesures de proximité, entre séries temporelles, qui reposent sur les coefficients de corrélation de Pearson, des rangs de Spearman et des indices définis ci-dessus. Les quatre mesures de proximité $Y^{R1}_{1,2}$, $Y^{R2}_{1,2}$, $Y^{S1}_{1,2}$, $Y^{S2}_{1,2}$ entre les séries S_1 et S_2 sont composées chacune d'un coefficient de corrélation et de deux indices :

$$Y^{R1}_{1,2} = \omega_1 R^*_{1,2} + \omega_2 A_{1,2} + \omega_3 M_{1,2}, \tag{3.22}$$
$$Y^{R2}_{1,2} = \omega_1 R^*_{1,2} + \omega_2 A^*_{1,2} + \omega_3 M^*_{1,2}, \tag{3.23}$$
$$Y^{S1}_{1,2} = \omega_1 S^*_{1,2} + \omega_2 A_{1,2} + \omega_3 M_{1,2}, \tag{3.24}$$
$$Y^{S2}_{1,2} = \omega_1 S^*_{1,2} + \omega_2 A^*_{1,2} + \omega_3 M^*_{1,2}, \tag{3.25}$$

avec $R^*_{1,2} = (R_{1,2} + 1)/2$ et $S^*_{1,2} = (S_{1,2} + 1)/2$, où $R_{1,2}$ et $S_{1,2}$ représentant respectivement les coefficients de corrélation de Pearson et de Spearman entre les deux séries S_1 et S_2 ; $\omega_k \in [0, 1]$ est le coefficient de pondération du k^{ieme} facteur composant la mesure avec $\sum_{k=1}^{3} \omega_k = 1$. Dans leurs travaux, Son et Baek ont montré que les mesures $Y^{R1}_{i,j}$ et $Y^{S1}_{i,j}$ sont préférables aux mesures $Y^{R2}_{i,j}$, $Y^{S2}_{i,j}$ et aux coefficients de corrélation conventionnels. Ils recommandent leur utilisation dans les applications.

3.2.5 Mesures de proximité alliant forme et valeurs

Selon les définitions ci-dessus (Sections 3.2.2, 3.2.3), à chaque alignement r correspond une fonction de coût fondée sur les valeurs $c(r)$ et une fonction de coût fondée sur la forme ($Cort(r)$ ou $Cor(r)$) notée $Co(r)$. Pour une mesure de proximité couvrant simultanément les écarts entre les formes et les valeurs des séries, on introduit la fonction de cout $c_k(r)$:

$$c_k(r) \quad = \quad f(Co(r)).c(r). \tag{3.26}$$

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

$$f(x) \quad = \quad \frac{2}{1 + exp(kx)}, \quad k > 0. \tag{3.27}$$

Le paramètre k définit les contributions relatives des composantes forme et valeurs à $c_k(r)$. Pour une fonction de coût $c_k(r)$ ne couvrant que la composante valeurs (c-à-d, en ignorant la composante forme), k est fixé à 0 et $c_{k=0}(r) = c(r)$. Pour $k \geq 6$, $c_{k=6}(r)$ comprend totalement la composante forme. Ainsi, si $Co(r) = 1$, alors $c_{k=6}(r) = 0$, signifiant que si deux séries temporelles sont de formes similaires, la fonction de coût est réduite à zéro, quelle que soit la

valeur de $c(r)$. Si $Co(r) = -1$, alors $c_{k=6}(r) = 2c(r)$, correspondant à une pénalisation de facteur 2 dans le cas de séries temporelles de formes opposées. Enfin, si $Co(r) = 0$, alors $c_{k=6}(r) = c(r)$, indiquant que dans le cas de séries temporelles de formes différentes (non similaires ni opposées), la fonction de coût $c_k(r)$ se résume à la seule information disponible l'écart en valeur $c(r)$. La figure 3.8 montre l'effet du réglage en fonction du paramètre k.

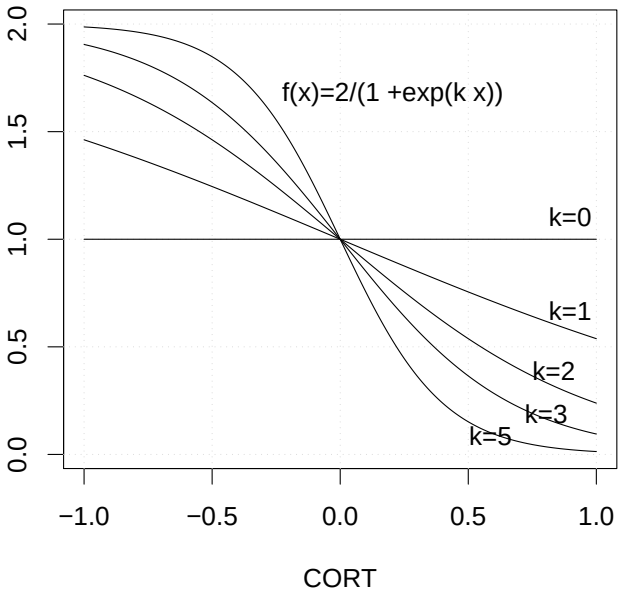


FIG. 3.8 – L'effet du réglage en fonction de k.

Basée sur $c_k(r)$ et sur un sous-ensemble d'alignements $R \subset M$, une forme générale de l'indice de dissimilarité adaptative couvrant simultanément les écarts entre les formes et les valeurs des séries est proposée :

$$D_k(S_1, S_2) \quad = \quad \min_{r \in R} c_k(r) = \min_{r \in R} \left(f(Co(r)).c(r) \right) \quad k > 0 \qquad (3.28)$$

Par exemple, pour $k = 0$, $Co(r) = Cort(r)$ et $|Cort| = 1$, la proximité entre forme contribue 0% à D_k tandis que la proximité en valeurs contribue 100% à D_k. Pour $k = 2$ et $|Cort| = 1$, 76.2% sont contribués par la proximité en forme contre 23.8% pour les valeurs. Le tableau 3.5 résume, dans le cas de formes similaires ou opposées mesurées par le coefficient de corrélation temporelle ($Cort$=1), les contributions en forme et en valeurs à D_k.

Distance Euclidienne étendue alliant forme et valeurs

Pour $R = \{r_0\}$, $Co(r) = Cort(r)$, et $c(r) = (\sum_{i=1}^{m}(u_i - v_i)^2)^{\frac{1}{2}}$, l'indice de dissimilarité adaptative D_k définit une extension de la distance Euclidienne, notée DE_k, couvrant les composantes forme et valeurs :

$$D_k(S_1, S_2) \quad = \quad DE_k = c_k(r_0) = f(Cort(r_0)).\Big(\sum_{i=1}^{m}(u_i - v_i)^2 \Big)^{\frac{1}{2}} \qquad (3.29)$$

	Contribution en forme (%)	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
$k \geq 6$	$\rightsquigarrow 100$	$\rightsquigarrow 0$

TAB. 3.5 – Contribution de la proximité en valeurs et en forme à D_k,pour $|Cort| = 1$.

Dynamic time warping étendue alliant forme et valeurs

Pour $R \subset M$, $Co(r) = Cort(r)$ et $c(r) = \sum_{i=1}^{m} |u_{a_i} - v_{b_i}|$, D_k donne une extension de la dynamic time warping, notée DTW_k, pour couvrir simultanément les composantes forme et valeurs :

$$D_k(S_1, S_2) \quad = \quad DTW_k = \min_{r \in R} \left(f(Cort(r)). \sum_{i=1}^{m} |u_{a_i} - v_{b_i}| \right) \qquad (3.30)$$

L'exemple donné dans la Figure 3.4 illustre le succès de $DE_{k=6}$ et $DTW_{k=6}$ à mesurer la proximité entre les formes des séries, en désignant S_1 plus proche de S_3 que de S_2, avec $DE_{k=6}(S_1, S_3) = 0.11 < DE_{k=6}(S_1, S_2) = 4.72 < DE_{k=6}(S_2, S_3) = 24.08$, et $DTW_{k=6}(S_1, S_3) = 0.20 < DTW_{k=6}(S_1, S_2) = 1.13 < DTW_{k=6}(S_2, S_3) = 12.52$.

Le tableau 3.6 résume, dans un cadre unifié, les principales mesures de proximité définies ci-dessus.

3.3 Caractéristiques des profils d'expression de gènes au cours du cycle cellulaire

L'analyse des séries temporelles issues des puces à ADN a pour objectif d'identifier des séquences d'évènements moléculaires qui s'étalent dans le temps. On s'intéresse en particulier à celles associées au cycle cellulaire. Les gènes qui y sont étudiés ont tous la même durée d'observation. Deux aspects rendent l'analyse de ces séries particulière. D'une part, les instants d'échantillonnage ne sont en général pas réguliers. D'autre part, ces séries ont peu d'observations. En effet, plus de 80% des séries ont une longueur inférieure ou égale à 8 (Ernst et al. [27]). À ces particularités, s'ajoutent, pour certains gènes observés, des comportements périodiques.

Sur la base des données observées au cours du cycle cellulaire, des méthodes d'estimation de la densité spectrale ont été employées pour identifier les gènes

Type	Ensemble des alignements : R	Valeurs fonct-coût : $c(r)$	Forme fonct-coût : $Co(r)$	Mesure de proximité
Valeurs	$R \subset M$	$\sum_{i=1}^{m} \lvert u_{a_i} - v_{b_i} \rvert$	-	$d_{Dtw} = \min_{r \in R} \sum_{i=1}^{m} \lvert u_{a_i} - v_{b_i} \rvert$
	$R = \{r_0\}$	$\Big(\sum_{i=1}^{m} (u_i - v_i)^2 \Big)^{\frac{1}{2}}$	-	$d_E = c(r_0) = \Big(\sum_{i=1}^{m} (u_i - v_i)^2 \Big)^{\frac{1}{2}}$
	$R = \{r_0\}$	$\sum_{i=1}^{m} \lvert u_i - v_i \rvert$	-	$d_M = c(r_0) = \sum_{i=1}^{m} \lvert u_i - v_i \rvert$
	$R \subset M$	$\max_{i=1}^{m} \lvert u_{a_i} - v_{b_i} \rvert$	-	$d_F = \min_{r \in R} \Big(\max_{i=1}^{m} \lvert u_{a_i} - v_{b_i} \rvert \Big)$
Forme	$R = \{r_0\}$	-	$1 - Cor(r)$	$d_{Cor} = Co(r_0) = 1 - Cor(r_0)$
	$R = \{r_0\}$	-	$1 - Cort(r)$	$d_{Cort} = Co(r_0) = 1 - Cort(r_0)$
	$R \subset M$	-	$1 - Cor(r)$	$d_{Cor} = \min_{r \in R}(1 - Cor(r))$
	$R \subset M$	-	$1 - Cort(r)$	$d_{Cort} = \min_{r \in R}(1 - Cort(r))$
Valeurs et forme	$R = \{r_0\}$	$\Big(\sum_{i=1}^{m} (u_i - v_i)^2 \Big)^{\frac{1}{2}}$	$Cor(r)$	$DE_k^{Cor} = f(Cor(r_0)).(\sum_{i=1}^{m} (u_i - v_i)^2)^{\frac{1}{2}}$
	$R = \{r_0\}$	$\Big(\sum_{i=1}^{m} (u_i - v_i)^2 \Big)^{\frac{1}{2}}$	$Cort(r)$	$DE_k^{Cort} = f(Cort(r_0)).(\sum_{i=1}^{m} (u_i - v_i)^2)^{\frac{1}{2}}$
	$R \subset M$	$\sum_{i=1}^{m} \lvert u_{a_i} - v_{b_i} \rvert$	$Cor(r)$	$DTW_k^{Cor} = \min_{r \in R} \Big(f(Cor(r)).\sum_{i=1}^{m} \lvert u_{a_i} - v_{b_i} \rvert \Big)$
	$R \subset M$	$\sum_{i=1}^{m} \lvert u_{a_i} - v_{b_i} \rvert$	$Cort(r)$	$DTW_k^{Cort} = \min_{r \in R} \Big(f(Cort(r)).\sum_{i=1}^{m} \lvert u_{a_i} - v_{b_i} \rvert \Big)$

TAB. 3.6 – Les principales mesures de proximité entre des séries temporelles, dans un cadre unifié.

qui s'expriment périodiquement. Spellman et al. [94] et Whitfield et al. [106] ont effectué une transformation de Fourier sur des données échantillonnées avec différentes méthodes de synchronisation, sous l'hypothèse que le signal du cycle cellulaire est une sinusoïde unique. Wichert et al. [107] ont appliqué le périodogramme moyenné et le test de Fisher pour détecter les phénomènes périodiques et les distinguer des processus purement aléatoires. Ahdesmäki et al. [2] ont mis en œuvre une procédure de test de périodicité en supposant que les bruits sont non gaussiens. La majorité de ces travaux ont porté sur des séries temporelles d'expression de gènes échantillonnées de façon régulière.

Pour l'étude du transcriptome du cycle cellulaire, les cellules sont synchronisées à partir d'une phase puis observées sur un, deux voir trois cycles cellulaires consécutifs. Au cours de l'observation des niveaux d'expression, certaines caractéristiques spécifiques peuvent varier pendant le processus de la division cellulaire : amplitude initiale fluctuant d'un gène à l'autre, atténuation d'amplitude au cours des cycles, effets de tendance, etc...

3.4 Mesures de proximité entre des profils d'expression de gènes

En général, l'analyse des gènes s'intéresse plus à la forme des profils d'expression. Celle-ci contient des informations sur la progression de l'expression du

gène. La proximité entre gènes peut être définie comme la proximité en forme de leurs profils associés, et peut être mesurée de différentes manières. Dans la section précédente, nous avons présenté différentes techniques pour apprécier la forme des profils et toutes ont été appliquées sur des données d'expression de gènes.

Dans Balasubramaniyan et al. [4], on s'interesse aux profils présentant des sous-séquences de formes similaires. La proximité est mesurée uniquement sur des sous-séquences et pas sur l'intégralité du profil. Dans cette situation, deux gènes sont proches s'ils présentent deux sous-séquences proches. Les auteurs proposent une mesure de proximité qui repose sur le coefficient de corrélation des rangs de Spearman. Toutefois, les méthodes fondées sur les rangs induisent une perte d'information qui peut entraîner une mauvaise interprétation des proximités entre gènes. Elles ne font pas de distinction entre les différences significatives des bruits : un petit accroissement dû à du bruit dans une courbe assez lisse peut avoir le même rang qu'un évènement significatif caractérisé par un pic. Un autre coefficient qui est utilisé est celui de Jackknife. Heyer et al. [41] l'ont introduit dans l'analyse de classification des données d'expression de gènes pour réduire l'influence des valeurs aberrantes. Mais, ce coefficient ne considère pas la structure temporelle des données puisqu'il repose sur le coefficient de Pearson. Dzeroski et al. [24] applique la distance qualitative de Todorovski, dans le cadre d'une classification prédictive qui combine classification (non supervisée) et des méthodes de prédiction pour profils temporels de gènes, afin de construire des groupes homogènes. Möller-Levet et al. [68] estiment la proximité des gènes par la méthode de comparaison des pentes. Cette méthode nécessite une standardisation des données qui est intéressante lorsque la forme des profils est la seule caractéristique importante pour différencier les gènes. Mais, son application ne permet pas de distinguer les gènes affectés par les bruits environnants des gènes qui varient beaucoup entre les différents instants. Signalons que toutes ces approches se focalisent uniquement sur la forme des profils en ignorant l'information sur les écarts en amplitude des niveaux d'expression.

D'autre part, les mesures de proximité ne considérant que les différences d'amplitude telles que la distance Euclidienne ou de Manhattan ne donnent aucune information sur la forme des profils de gènes. La forme intervient plus pour la DTW, bien qu'elle soit fondée sur les valeurs. Elle permet, en effet, de prendre en compte le décalage temporel qui peut rapprocher les formes similaires. Cependant, dans le cas de l'analyse des données d'expression de gènes au cours du cycle cellulaire, l'information cherchée est locale : on veut regrouper les gènes qui font un pic d'expression dans une phase donnée, ils sont alors caractéristiques de cette phase. L'utilisation de la DTW pourrait associer des gènes caractérisant des phases différentes.

Bien que l'amplitude d'expression ne soit pas un facteur critique pour dé-

terminer des groupes de gènes, son information est utile et doit être prise en compte lorsque l'on compare les formes des profils. Dans ce contexte, Kim et al.[52] opèrent une transformation des données tenant compte de l'information sur la forme par une nouvelle représentation dans un espace de plus grande dimension. Au lieu de travailler sur des profils avec n instants reliés entre eux, ils travaillent avec des points dans un espace de dimension $n(n+1)/2$ constitués par les écarts des nivaux d'expression entre les différents instants. Ils calculent alors une distance du Chi^2 par rapport à une estimation des paramètres de ces niveaux d'expression. Cette approche fournit de bons résultats mais nécessite une connaissance *a priori* sur la distribution des données.

3.5 Conclusion

Dans cette partie, nous avons présenté trois types de mesures de proximité pour les séries temporelles. Le premier type de mesures est fondé sur les valeurs des séries. Il est approprié d'utiliser la distance Euclidienne ou celle de Manhattan dans le cas de données n'incluant pas de décalage temporel. Par ailleurs, dans le cas de présence de décalage temporel, il est conseillé d'utiliser la DTW ou la distance de Fréchet. Le deuxième type de mesures repose sur la forme des séries et est principalement à utiliser pour l'analyse de profils temporels d'expression de gènes. Cependant, il nous semble intéressant d'introduire à la fois les informations sur les écarts en forme et en valeurs, afin d'avoir une analyse plus pertinente des données. Ceci constitue l'objet du troisième type de mesures que nous avons développé incluant simultanément la proximité en forme et en valeurs des séries temporelles. Ces mesures dépendent d'un paramètre qui définit les contributions relatives des composantes forme et valeurs. Dans le chapitre suivant, elles seront utilisées pour l'analyse des séries temporelles d'expression de gènes au cours du cycle cellulaire. Une étape d'apprentissage y sera introduite afin de saisir les proportions contribuées par les deux composantes afin d'estimer la mesure de proximité appropriée à utiliser.

4

Analyse des gènes au cours du cycle cellulaire

Sommaire

4.1 Intêret de l'étude du cycle cellulaire 51
4.2 Approche conventionnelle 54
4.2.1 Classement des gènes cycliques 54
4.2.2 Recherche de profils types 55
4.3 Approche adaptative 55
4.4 Applications et analyses comparatives 56
4.4.1 Modèle aléatoire pour la comparaison de métriques . 57
4.4.2 Gènes périodiquement exprimés dans les cellules fibroblastiques humaines HeLa [106] 65
4.4.3 Gènes périodiquement exprimés dans les cellules du cancer du sein MCF-7 [63] 77
4.5 Conclusion 82

4.1 Intêret de l'étude du cycle cellulaire

Grâce au mécanisme de la division cellulaire, les milliers de cellules qui disparaissent chaque jour, chez l'organisme d'un adulte, sont remplacées dans certains tissus par de nouvelles cellules. Ce mécanisme est un processus complexe, régulé par l'intervention de nombreuses protéines permettant la succession des différentes étapes du cycle cellulaire. Toutefois, la perte de régulation conduit à une prolifération anarchique des cellules (e.g., cancer), entraînant des altérations catastrophiques du gènome. Elle est provoquée par l'apparition des

tumeurs correspondant à une amplification de proto-oncogènes (gènes activateurs du cycle cellulaire) ou à la perte d'anti-oncogènes (gènes suppresseurs de tumeur qui freinent le cycle cellulaire).

La connaissance des gènes impliqués dans la régulation et leurs fonctions devrait permettre une meilleure compréhension du mécanisme de la division cellulaire. Elle constitue un point central pour l'étude de certaines maladies (héréditaires, cancéreuses,...) et permet, entre autre, de fournir des outils pour faire face à celles-ci. C'est pourquoi l'étude de la régulation du cycle cellulaire est nécessaire. Dans ce but, de nombreux chercheurs se sont intéressés à l'étude de l'expression des gènes au cours du cycle cellulaire (e.g., Cho et al. [15], Spelleman et al. [94], Zhu et al. [117], Whitfield et al. [106], Pramila et al. [80], Liu et al. [60], Oliva et al. [74], Bar-Joseph et al. [6], Côte et al. [18],...).

L'analyse des données d'expression pendant le cycle de division cellulaire vise à identifier, dans une première étape, les gènes présentant des profils d'expression périodiques (gènes cycliques). En effet, certains gènes s'expriment fortement pendant une phase spécifique du cycle cellulaire. Lorsque leurs niveaux d'expression sont suivis durant plusieurs cycles consécutifs, leurs profils d'expression sont périodiques. La figure 4.1 montre l'expression du gène cyclique CCNE1 (cycline E1) observé au cours des trois premiers cycles cellulaires après synchronisation d'une population cellulaire HeLa (Whitfield et al. [106]). Notons certaines caractéristiques spécifiques aux profils d'expression de gènes au cours du cycle cellulaire. L'amplitude initiale des niveaux d'expression varie d'un gène à un autre, des atténuations d'amplitude au cours des différents cycles et des effets de tendance sont observés sur les profils d'expression, en plus les périodes du cycle diffèrent selon les types de cellules considérés...

Les gènes cycliques sont, en général, impliqués dans le contrôle du cycle de la division cellulaire. Les premières études de l'évolution du transcriptome au cours du cycle cellulaire ont été réalisées chez la levure *Saccharomyces cerevisiae*. Cho et al. [15] identifient des gènes cycliques en visualisant les profils d'expression périodiques. Plus tard, Spellman et al. [94] utilisent trois méthodes de synchronisation différentes pour générer trois autres ensembles de données du cycle cellulaire. Ils déterminent des gènes cycliques sur la base de leur corrélation avec des gènes cycliques connus et sur la moyenne des résultats obtenus à partir de transformés de Fourier. Des études similaires ont ensuite été effectuées sur des fibroblastes humains (Cho et al. [16], Whitfield et al. [106]). Plus récemment, Bar-Joseph et al. [6] identifient des gènes cycliques dans les fibroblastes primaires de prépuce humain, Côte et al. [18] examinent des gènes cycliques dans des cultures de champignons pathogènes humains *Candida albicans*, en utilisant une approche fondée sur l'analyse spectrale. Les nombres de gènes étudiés et de gènes cycliques impliqués dans les exemples précités sont résumés dans le tableau 4.1. La détection des gènes cycliques est largement abordée dans la littérature. L'analyse qui nous préoccupe se focalisera donc

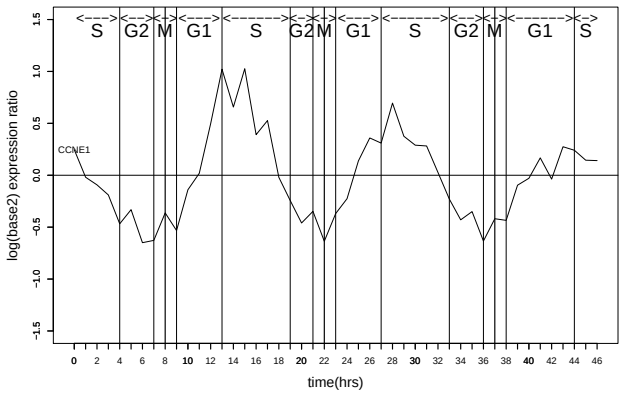


FIG. 4.1 – Profil d'expression du gène cyclique CCNE1 (cycline E1) observé sur une période de 46 heures après synchronisation correspondant à trois cycles cellulaires. Ce gène s'exprime fortement pendant la transition G_1/S de l'interphase du cycle cellulaire. Chaque phase de cycle est délimitée par les traits verticaux et annotée par G_1, S, G_2 ou M.

sur ces gènes cycliques, le but consistant à déterminer des groupes de gènes caractéristiques des différentes phases du cycle cellulaire (c-à-d regrouper des gènes avec des pics d'expression élévés dans une même phase du cycle cellulaire).

Auteurs	Espèce	Type cellulaire	Nbr de gènes étudiés	Nbr de gènes cycliques	Méthodes utilisées
Cho et al. [15]	levure	-	6220	416	visualisation
Spellman et al. [94]	levure	-	~ 6200	800	transformé de Fourier
Cho et al. [16]	humain	fibroblastes	40.000	731	transformé de Fourier
Whitfield et al. [106]	humain	fibroblastes	~ 29.621	850	transformé de Fourier
Bar-Joseph et al. [6]	humain	fibroblastes de prépuce	-	480	transformé de Fourier
Côte et al. [18]	humain	champignons pathogènes	50.000	494	analyse spectrale

TAB. 4.1 – Exemples d'études effectuées sur la levure et l'espèce humaine pour l'identification de gènes cycliques.

Par ailleurs, lors d'une expérience au cours du cycle cellulaire, les chercheurs se sont intéressés à sélectionner des gènes pertinents avec des profils de réponse-type, dont la forme peut être prédéfinie ou pas (Peddada et al. [78]).

Pour le regroupement des gènes cycliques et la sélection de gènes pertinents, des méthodes paramétriques et des méthodes à base de distance sont employées, afin de différencier les gènes selon les phases du cycle cellulaire ou de détecter des gènes de profils types. Dans le cadre des méthodes à base de distance, le regroupement ou la détection des gènes est fondé sur des profils choisis *a priori*. À ce jour, le regroupement de l'ensemble des gènes cycliques caractérisant chaque phase ou transition de phase du cycle cellulaire est *a priori* basé

sur un ensemble de gènes de référence. Chaque gène étudié est affecté à la phase ou transition de phase 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 sans fondement (distance euclidienne, distance fondée sur le coefficient de corrélation de Pearson,...). D'autre part, dans la littérature il n'y a pas de consensus quant à l'ensemble de gènes de référence à considérer. De même, pour la détection de profils de gènes pertinents, des profils types sont prédéfinis. La principale limite de cette approche est le très faible nombre de gènes sélectionnés (quelques dizaines parmi les milliers de gènes étudiés). Dans ce chapitre, notre but est de proposer une approche adaptative, basée sur la mesure de proximité développée dans le § 3.2.5 du chapitre précédent. Elle nous permet d'identifier les phases de pic d'expression des gènes étudiés ou de détecter des profils de gènes pertinents, ensuite proposer un nouvel ensemble de gènes de référence validé par une connaissance biologique. La suite de ce chapitre s'articule comme suit : la section suivante présente l'approche conventionnelle pour le regroupement et la sélection de gènes. Nous présentons ensuite les différentes étapes de l'approche adaptative, fondée sur une méthode de classification des gènes, dans la section 4.3. Enfin, l'application des approches et des études comparatives sont effectuées dans la section 4.4.

4.2 Approche conventionnelle

4.2.1 Classement des gènes cycliques

L'objectif du classement consiste à regrouper les gènes de même niveau d'expression à un instant donné du cycle cellulaire (phase ou transition de phase). Dans ce contexte, des approches paramétriques et des approches à base de distances sont adoptées.

La première de ces approches consiste à estimer l'angle correspondant à la phase de pic d'expression des gènes en modélisant leurs niveaux d'expression. Par exemple, dans Cheng et Li [14], le niveau d'expression x_t d'un gène observé à l'instant t ($t = 1, ..., p$) est formulé par :

$$x_t = \alpha cos\Big(2\pi\frac{t-\phi}{T}\Big) + \epsilon_t \tag{4.1}$$

où α est une contante positive, T la période du cycle cellulaire du gène et ϕ l'angle correspondant à la phase de pic d'expression du gène qui peut être estimée par la formule :

$$\hat{\phi} = arg\min_{\phi}\sum_{t=0}^{p}\Big(x_t cos\Big(2\pi\frac{t-\phi}{T}\Big)\Big) \tag{4.2}$$

La principale limite à estimer la phase de pic d'expression élévé d'un gène, par la méthode fondée sur un modèle, réside dans la non prise en compte de

certaines caractéristiques spécifiques aux profils d'expression de gènes au cours du cycle cellulaire évoquées plus haut.

La seconde approche est fondée sur un ensemble de gènes de référence exprimés dans les différentes phases ou transitions de phases du cycle cellulaire (e.g., Whitfield et al. [106], Bar-Joseph et al. [6]). Chaque gène cyclique est affecté à la phase du groupe de gènes de référence auquel il est le plus corrélé. Cette approche classique souffre de deux limites. D'une part, la mesure de proximité utilisée est souvent fondée sur le coefficient de corrélation de Pearson, inadapté pour la comparaison des profils temporels d'expression de gène (*c.f.* § 3.2.3). D'autre part, dans la littérature, il n'y a pas de consensus quant à l'ensemble des gènes de référence à considérer (*cf.* section 4.4.2), ainsi que le nombre de phases à utiliser. Dans Whitfield et al. [106], 5 phases et transitions de phases G_1, G_1/S, G_2, G_2/M et M/G_1 ont été considérées, tandis que Bar-Joseph et al. [6] n'en ont pris en compte que 4 (G_1/S, G_2, G_2/M et M/G_1) pour le même jeu de données. L'identification des phases d'expression des gènes dépend des gènes de référence initialement choisis.

4.2.2 Recherche de profils types

Peddada et al. [78] proposent une approche adhoc qui repose sur un ensemble de profils prédéfinis pour détecter des profils d'expression de gènes de réponse-type. Chaque profil prédéfini est représenté par un ensemble d'inégalités entre les valeurs d'expression estimées à des instants critiques. Par exemple, un profil présentant un acroissement monotone sur l'ensemble des instants $\{t_1, t_2,..., t_p\}$ est défini par :

$$\{\mu_{t_1} \leqslant \mu_{t_2} \leqslant ... \leqslant \mu_{t_p}\} \tag{4.3}$$

avec μ_{t_i} la valeur de l'expression estimée à l'instant t_i. Un profil prédéfini cyclique avec des minima aux instants t_1, t_j, et t_p et des maxima aux instants t_i et t_k est défini par :

$$\{\mu_{t_1} \leqslant ... \leqslant \mu_{t_i} \geqslant \mu_{t_{i+1}} \geqslant ... \geqslant \mu_{t_j} \leqslant \mu_{t_{j+1}} \leqslant ... \leqslant \mu_{t_k} \geqslant \mu_{t_{k+1}} \geqslant ... \geqslant \mu_{t_p}\} \tag{4.4}$$

Pour chaque profil prédéfini (i.e., chaque ensemble d'inégalités), une procédure d'inférence sous contrainte d'ordre est utilisée pour estimer les valeurs d'expression de chaque gène, afin que celles-ci vérifient les contraintes d'ordre définies par l'ensemble des inégalités. Chaque gène est alors affecté au profil prédéfini le plus convenable. Notons que cette procédure peut aboutir à des gènes de très petites valeurs d'expression sur tous les instants. Dans ce cas, ils sont retirés et ne seront pas classés.

4.3 Approche adaptative

L'identification des phases d'expression forte des gènes et la sélection de profils types au cours du cylce cellulaire sont réalisées par des approches para-

métriques et des approches à base de distances. Ces approches souffrent de certaines limites notamment le manque de considération de certaines caractéristiques spécifiques des profils temporels d'expression de gènes (pour les approches paramétriques), le choix sans consensus de gènes de référence et sur l'utilisation d'une mesure de similarité naïve (sans fondement).

Sous ce rapport, nous développons une classification adaptative des profils d'expression de gènes. Son but est d'apprendre d'abord la mesure de proximité à utiliser (D_k, présentée dans le § 3.2.5) incluant la forme et les valeurs des profils des gènes, ensuite sélectionner par une méthode non supervisée un ensemble de gènes de référence (ou un ensemble de gènes pertinents) validé par une connaissance biologique. Elle permet notamment d'identifier les phases de pic d'expression des gènes étudiés. Nous proposons d'utiliser l'algorithme PAM (Partitioning Around Medoïds) pour la procédure de classification. Il 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 des gènes. En effet, PAM mesure la *silhouette width* (sw) de chaque gène, c'est un indicateur de confiance sur l'appartenance d'un gène par rapport à sa classe. Pour plus de détails sur l'algorithme PAM voir Kaufman et Rousseuw [50]. Pour un nombre de classes fixé N, la classification adaptative fondée sur D_k consiste à exécuter l'algorithme PAM pour k allant de 0 (D_k totalement basé sur les valeurs) à 6 (D_k totalement basé sur la forme). Ceci permet d'apprendre la valeur k^* qui fournit la partition optimale $\mathcal{P}_{N,k^*}$ selon des critères prédéfinis. La valeur k^* donne la meilleure contribution des proximités en valeurs et en forme à D_{k^*}. Pour le nombre de classes variant de 1 à N, ce processus est répété pour chaque valeur du nombre de classes afin de déterminer $\mathcal{P}_{N^*,k^*}$.

Pour des fins d'analyse comparative avec des méthodes de classement reposant sur des gènes de référence, nous proposons notre technique de classement adaptatif fondée sur l'apprentissage d'une mesure appropriée D_k. L'algorithme des K plus proches voisins ou $K-NN$ (*K Nearest Neibghbor*) est utilisé pour classifier les gènes. Le classement adaptatif consiste à exécuter l'algorithme de classement pour k allant de 0 à 6 afin d'apprendre la mesure D_k appropriée selon un critère prédéfini.

4.4 Applications et analyses comparatives

Cette section présente trois applications portant sur la classification ou le classement de profils temporels d'expression de gènes au cours du cycle cellulaire. La première application présente un modèle aléatoire pour les profils temporels d'expression de gènes au cours du cycle cellulaire. Nous étudions l'efficacité de quatre mesures de proximité à classer ou classifier des gènes cycliques. Dans la

suite, nous confrontons des approches conventionnelles à notre approche adaptative fondée sur l'apprentissage de la métrique D_k développée en section 3.2.5. La seconde application est d'analyser des données transcriptomiques concernant la prolifération des cellules humaines HeLa, nous comparons nos résultats à l'approche conventionnelle fondée sur le coefficient de corrélation (Whitfield et al. [106]). La troisième application consiste à analyser des données transcriptomiques provenant de cellules cancéreuses du sein humain MCF-7 (Lobenhofer et al. [63]). L'approche adaptative de la classification est comparée à l'approche adhoc proposée par Peddada et al. [78]. Nous présentons également les discussions des principaux résultats obtenus.

4.4.1 Modèle aléatoire pour la comparaison de métriques

Dans cette partie, notre objectif est d'étudier l'efficacité de quatre mesures majeures : δ_E, COR, CORT et D_k pour la classification et le classement des profils temporels d'expression de gènes. Cette étude est menée sur un jeu de données temporelles simulé sur la base d'un modèle alèatoire "random-periods" (Liu et al.[60]). 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. L'analyse des résultats d'une classification adaptative est basée sur trois critères : la silhouette width moyenne d'une partition notée asw, le ratio standart $wbr = \frac{intra}{inter}$ et l'indice de Rand corrigé (RI) (Hubert et Arabie [42]). Les résultats des classements adaptatifs se reposent sur les taux d'erreur des gènes mal classifiés.

Simulation des profils d'expression périodiques

Nous utilisons des profils simulés, générés sur la base du modèle de regression non-linéaire proposé par Liu et al.[60]. 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 des expressions est :

$$f(t,\theta_g) \quad = \quad a_g + b_g t + \frac{K_g}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} cos(\frac{2\pi t}{Texp(\sigma z)} + \Phi_g) exp(-\frac{z^2}{2}) dz \quad (4.5)$$

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 4.2 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.

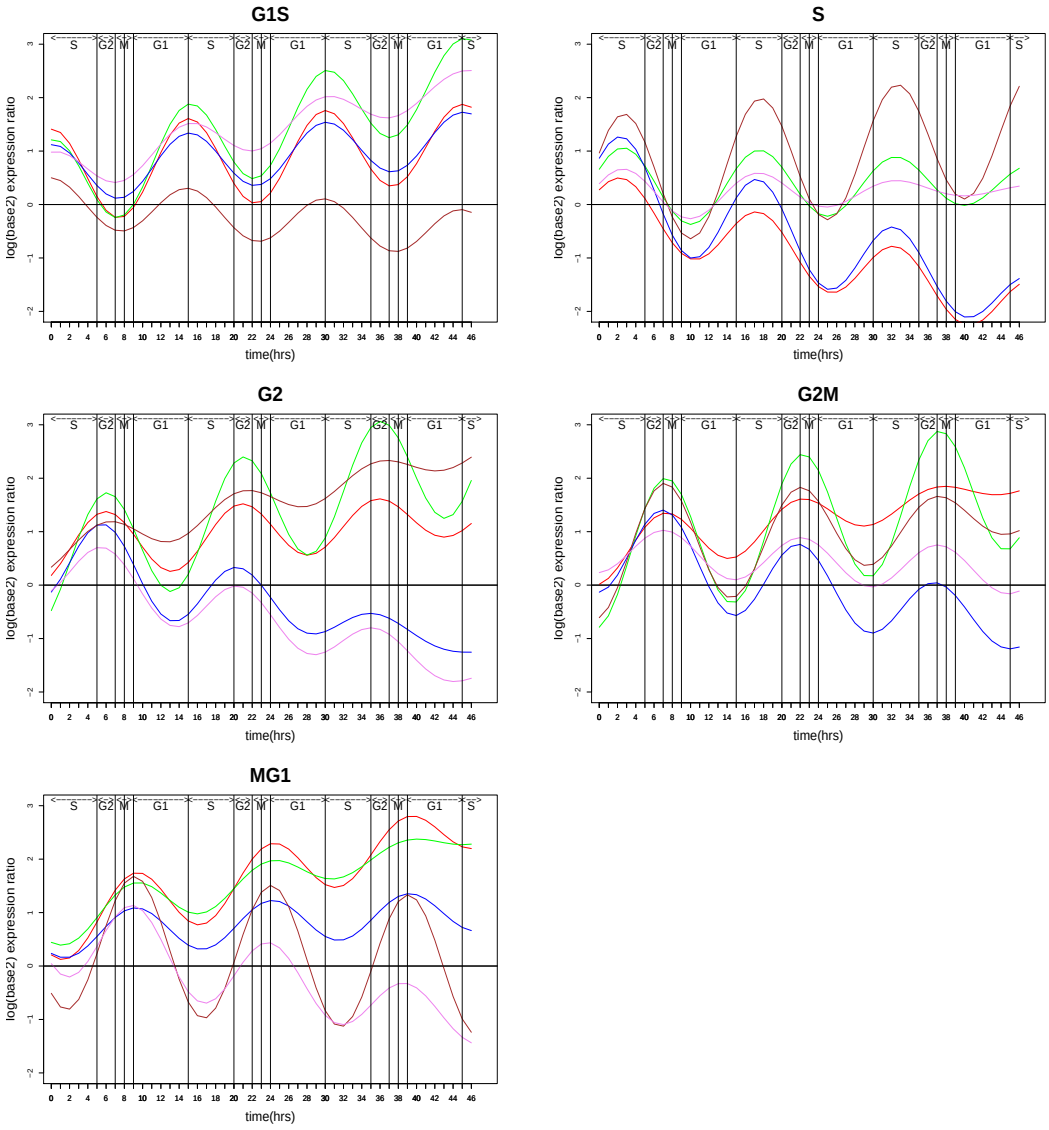


FIG. 4.2 – Progression de l'expression des gènes durant les 5 phases G_1/S, S, G_2, G_2/M et M/G_1.

Protocole de simulation

Basé sur ce modèle et les valeurs des paramètres spécifiées dans Liu et al.[60], quatre expériences sont menées pour étudier la façon dont chaque mesure de proximité 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 4.3 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 4.2. 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 pour chacune).

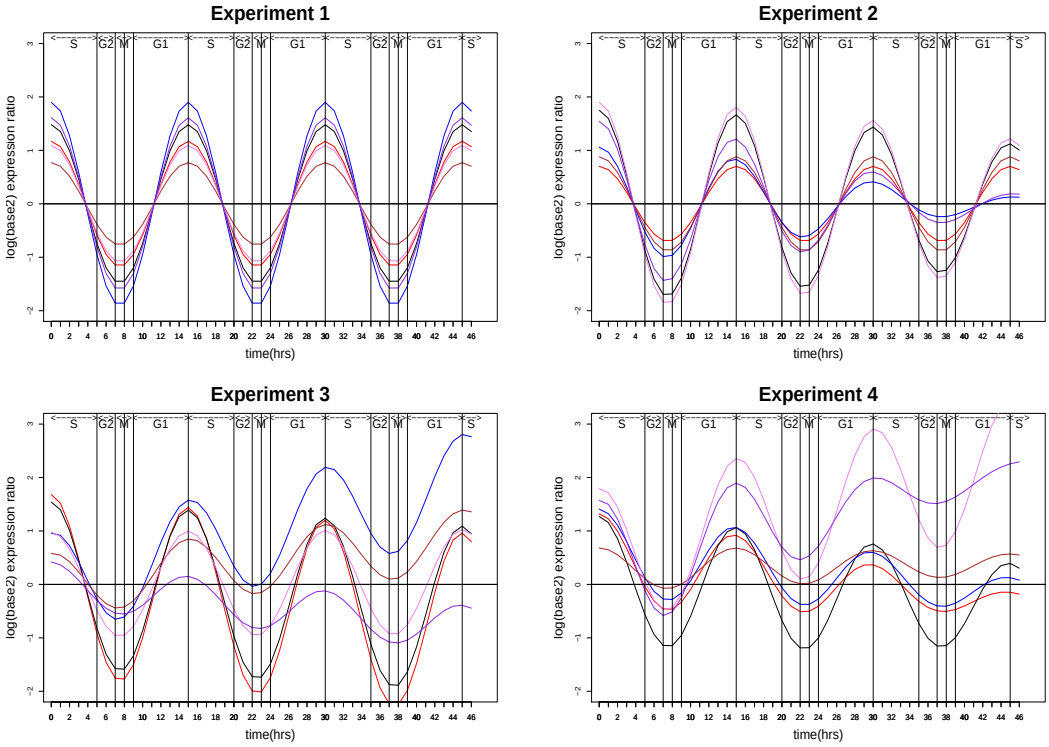


FIG. 4.3 – Profils des gènes de la phase G_1/S suivant les quatre expériences. Dans la première expérience (Experiment 1), les gènes sont différenciés par leur amplitude initiale. Ces amplitudes s'atténuent dans la longueur du temps, pour la seconde expérience (Experiment 2). La présence de tendance est observée pour les profils de la troisième expérience (Experiment 3), sans atténuation d'amplitude et avec atténuation d'amplitude pour la quatrième expérience (Experiment 3).

Numéro d'expérience	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. 4.2 – Spécification des paramètres du modèle.

Evaluation de l'efficacité des mesures pour la classification

Pour chaque expérience et pour chaque mesure δ_E, COR et CORT, nous partitionnons l'ensemble des profils d'expression générés de chaque échantillon S_{ij} en 5 classes (correspondant aux 5 phases). Par exemple, pour la mesure δ_E et pour l'expérience j, l'algorithme PAM est appliqué sur les 10 échantillons $S_{1j}, ..., S_{10j}$ afin d'extraire les 10 partitions $\mathcal{P}^{1j}_{\delta_E}$,..., $\mathcal{P}^{10j}_{\delta_E}$. Les valeurs des trois critères *asw*, *wbr*, *RI* (l'indice de Rand corrigé) sont retenues afin d'apprécier la qualité des classes. Ainsi, par exemple, l'évaluation de la mesure 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 $\mathcal{P}^{1j}_{\delta_E}$,..., $\mathcal{P}^{10j}_{\delta_E}$. Il en sera de même pour l'évaluation des mesures COR et CORT. Pour l'indice de dissimilarité D_k, notre méthode adaptative est appliquée. Le tableau 4.3 donne pour chaque expérience, la moyenne et la variance $(\overline{k*}, var(k*))$ de $k*$. Comme dans le cas des mesures δ_E, COR et CORT, nous considérons les valeurs moyennes des critères *asw*, *RI* et *wbr* sur les 10 partitions $\mathcal{P}^{1j}_{D_{k*}}$,..., $\mathcal{P}^{10j}_{D_{k*}}$ afin d'évaluer D_{k*} par rapport à l'expérience j. Les figures 4.4, 4.5, et 4.6 montrent, pour chacune des 4 expériences (tableau 4.2) et pour chaque mesure (δ_E, COR, CORT et D_{k*}) , la progression des critères *asw* (figure 4.4), *wbr* (figure 4.5), et *RI* (figure 4.6) sur les 10 échantillons $S_{1j}, ..., S_{10j}$. La figure 4.7 montre pour chaque mesure, sur l'ensemble des quatre expériences, la progression de la moyenne des valeurs des critères *asw* (gauche), *wbr* (droite) et *RI* (en bas).

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

TAB. 4.3 – Evaluation de l'efficacité des mesures pour la classification : chaque couple de valeurs représente la moyenne et la variance des dix valeurs de $k*$ trouvées pour les 10 partitions $\mathcal{P}^{1j}_{D_{k*}}$,..., $\mathcal{P}^{10j}_{D_{k*}}$. Evaluation de l'efficacité des mesures pour le classement : chaque couple de valeurs représente la moyenne et la variance des dix valeurs de $k*$ trouvées pour les 10 classements $\mathcal{C}^{1j}_{D_k}$,..., $\mathcal{C}^{10j}_{D_k}$

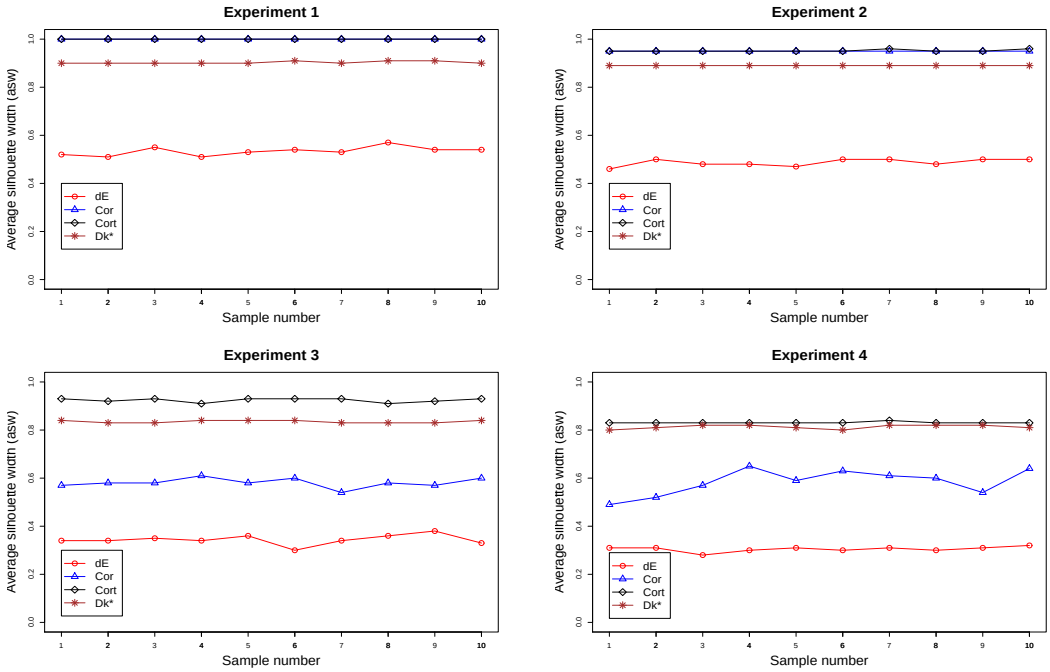


FIG. 4.4 – Pour chaque expérience et pour chaque mesure (δ_E, COR, CORT et D_{k*}), la progression des valeurs *asw* suivant les 10 échantillons est représentée.

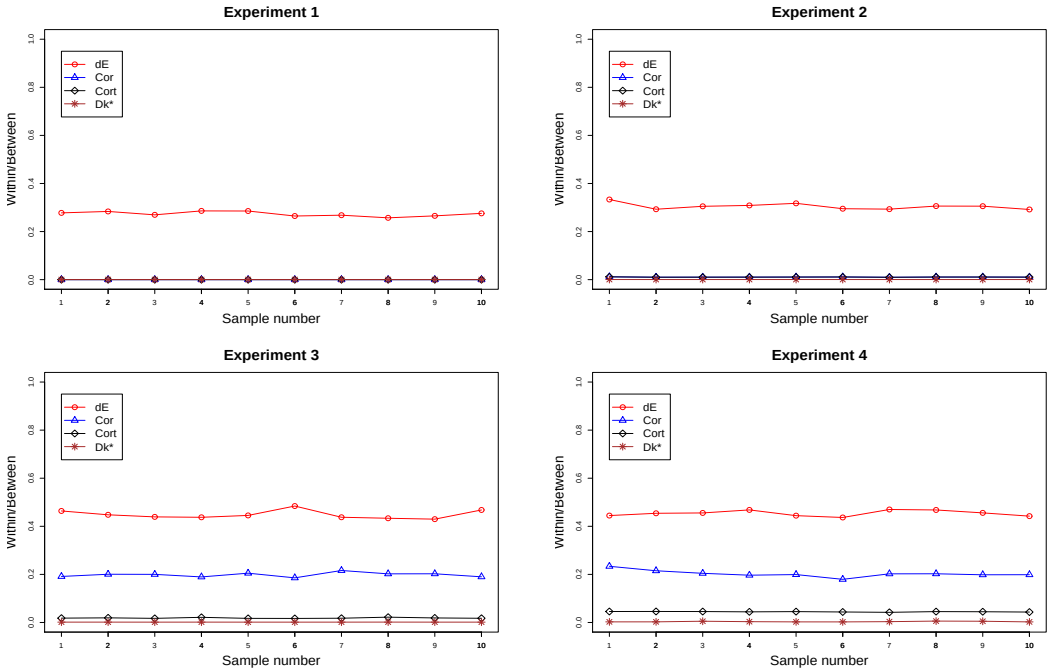


FIG. 4.5 – Pour chaque expérience et pour chaque mesure (δ_E, COR, CORT et D_{k*}), la progression des valeurs *wbr* suivant les 10 échantillons est représentée.

Evaluation de l'efficacité des mesures pour le classement

Pour chaque expérience et pour chaque mesure δ_E, COR et CORT, nous exécutons l'algorithme 10-NN, pour chaque échantillon S_{ij}. Par exemple, pour

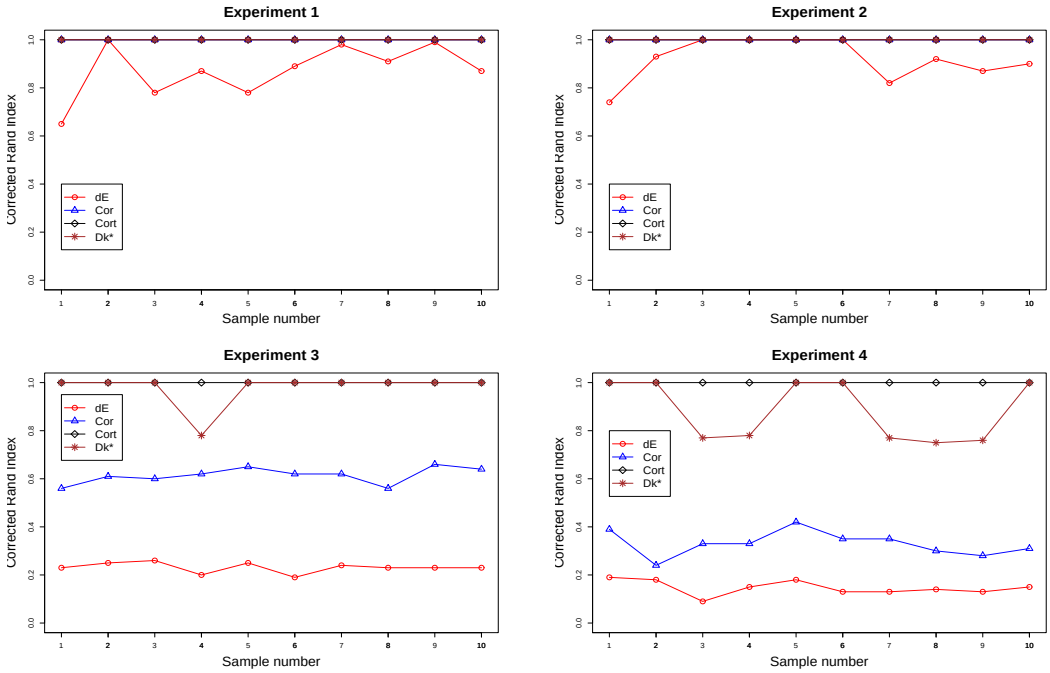


FIG. 4.6 – Pour chaque expérience et pour chaque mesure (δ_E, COR, CORT et D_{k*}), la progression des valeurs RI suivant les 10 échantillons est représentée.

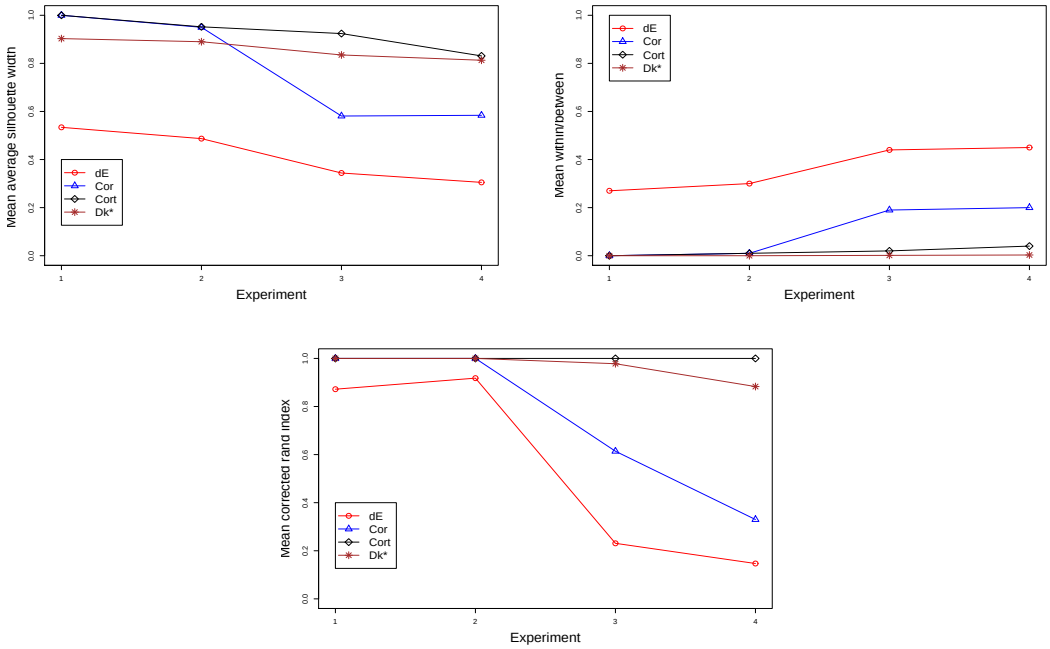


FIG. 4.7 – Evaluation des mesures pour la classification des profils d'expression simulés. La progression des valeurs moyennes des critères asw (gauche), wbr (droite) et RI (en bas) est illustrée.

la mesure δ_E et pour l'expérience j, l'algorithme 10-NN est appliqué sur les 10 échantillons $S_{1j}, ..., S_{10j}$ pour générer les 10 classements $\mathcal{C}_{\delta_E}^{1j}$,..., $\mathcal{C}_{\delta_E}^{10j}$. Pour chaque classement $\mathcal{C}_{\delta_E}^{ij}$, le taux de profils de gènes mal classifiés est retenu. L'évaluation de la mesure δ_E par rapport à l'expérience j est résumée par le taux d'erreur moyen des 10 classements $\mathcal{C}_{\delta_E}^{1j}$,..., $\mathcal{C}_{\delta_E}^{10j}$. Pour l'indice de dissimilarité D_k, notre procédure adaptative est appliquée. Son évaluation est basée sur le taux d'erreur moyen calculé sur les 10 classements $\mathcal{C}_{D_k}^{1j}$,..., $\mathcal{C}_{D_k}^{10j}$. La figure 4.8 montre, pour chacune des quatre expériences, la progression des taux d'erreur de profils de gènes mal classés suivant les échantillons classifiés. La figure 4.9 représente, pour chaque mesure, la progression de la moyenne des taux d'erreur suivant les quatre expériences.

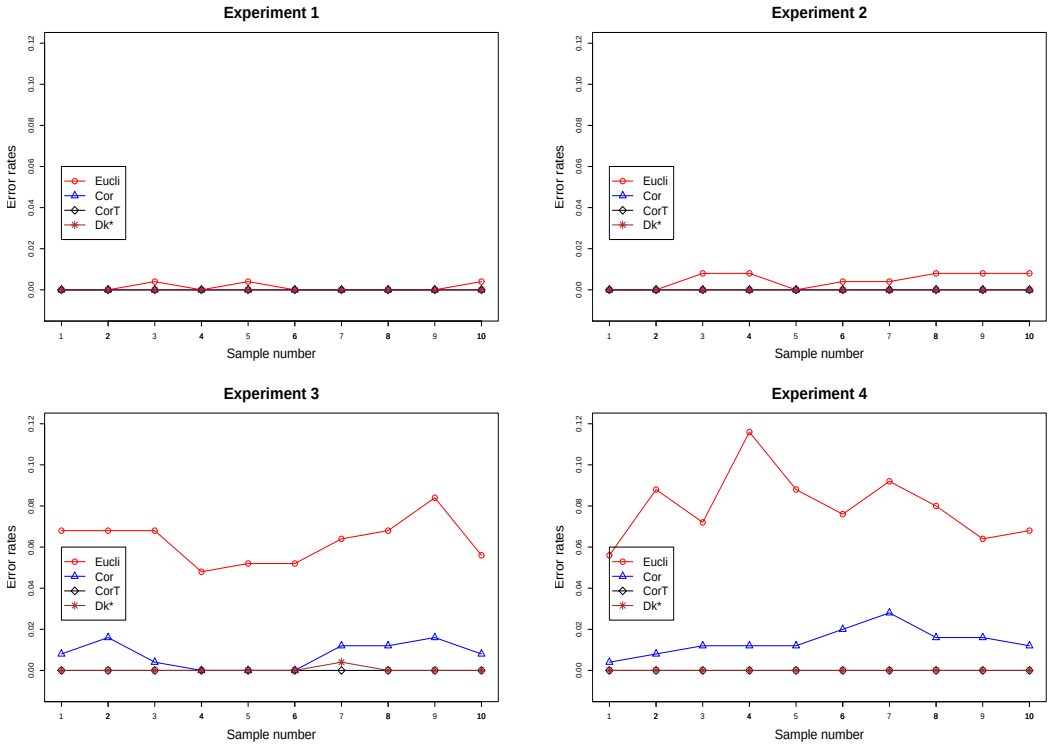


FIG. 4.8 – Pour chaque expérience et pour chaque mesure (δ_E, COR, CORT et D_{k*}), la progression des taux d'erreur suivant les 10 échantillons est représentée.

Discussion

Examinons dans un premier temps 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*. L'indice de

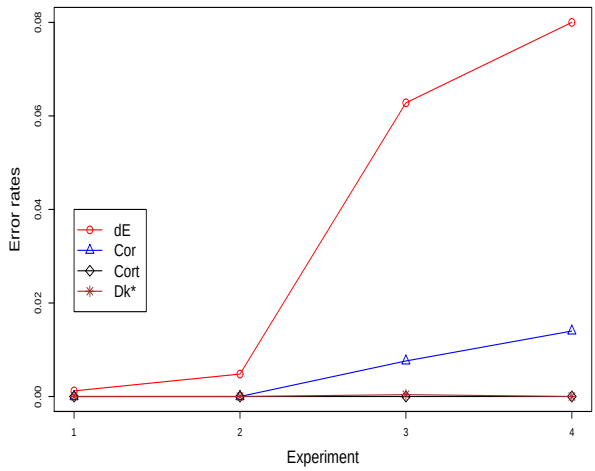


FIG. 4.9 – Evaluation des mesures pour le classement des profils de gènes suivant les quatre expériences. La progression des valeurs moyennes des taux d'erreur est représentée.

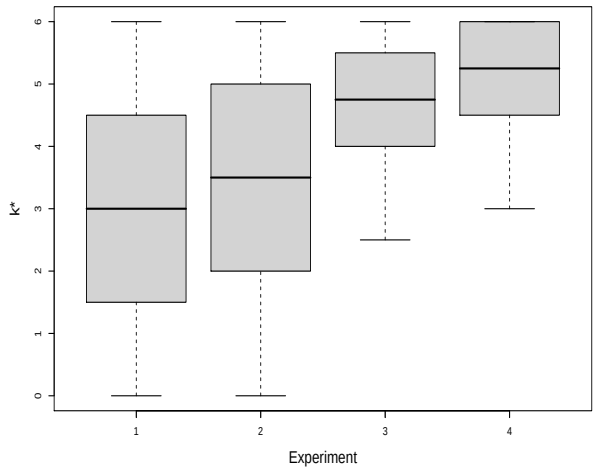


FIG. 4.10 – La distribution des valeurs de $k*$ dans la classement adaptatif

Rand corrigé (RI) permet de comparer deux partitions. Une valeur $RI = 0$ correspond à une absence totale de correspondance entre une vraie partition *a priori* et celle produite par un algorithme, alors qu'une valeur $RI = 1$ indique une correspondance parfaite. Les figures 4.4, 4.5, and 4.6 montrent que la classification basée sur δ_E donne, pour les expériences 1 à 4, les partitions les plus faibles comparé 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. La figure 4.7 montre que 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 complexes 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. En effet les figures 4.4, 4.5, 4.6, et 4.7 montrent les valeurs *asw* et *RI* qui s'affaiblissent et celles de *wbr* qui s'élèvent de façon significative pour les expériences 3 et 4. Ces résultats affirment la limite du coefficient de corrélation de *Pearson* face aux variations de tendance. Enfin, la meilleure classification et les structures de partitions les plus fortes sont données par CORT et D_k pour les quatre expériences, avec une *asw* variant dans $[0.8, 1]$ (figures 4.4 et 4.7), un *wbr* autour de 0 (figures 4.5 et 4.7), un *RI* évoluant dans $[0.83, 1]$(figures 4.6 et 4.7). 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 à travers les quatre expériences (Tableau 4.3).

Considérant les résultats sur le classement, les figures 4.8 et 4.9 montrent que pour les expériences 1 et 2, les quatre mesures 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 4.3 et la figure 4.10 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 quand un bon classement peut être obtenu avec une mesure fondée sur des valeurs ($k*$ proche de 0) et avec une mesure basée sur la forme ($k*$ proche de 6). En effet, dans les deux premières expériences, les figures 4.8 et 4.9 montrent que les quatre mesures 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 observés sur les figures 4.8 et 4.9. 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.

4.4.2 Gènes périodiquement exprimés dans les cellules fibroblastiques humaines HeLa [106]

Description des données

Les cellules HeLa, lignée cellulaire épithéliale dérivant du carcinome cervical humain, sont mises en cultures et bloquées en début de phase S du cycle cellu-

laire. Pour augmenter la fiabilité, deux blocages successifs sont réalisés sur les cultures en phase exponentielle de prolifération. Ainsi, après levée du blocage, les cellules repartent dans le cycle cellulaire de manière synchrone à partir du début de la phase S. Les gènes étudiés sont décrits par leurs niveaux d'expression, tout au long des 47 heures qui suivent la synchronisation des cellules, correspondant à trois cycles cellulaires. Nous nous limitons à l'analyse des données enregistrées dans la troisième expérimentation de l'application HeLa publiées par Whitfield et al. [106] (http ://genome-www.standford.edu/HumanCellCycle/Hela/), 1099 gènes cycliques sont considérés dans notre étude.

Analyse conventionnelle

Les auteurs Whitfield et al. [106] considèrent un ensemble de 20 gènes de référence. Ces gènes sont supposés être caractéristiques des cinq phases suivantes : $G1/S$, S, $G2$, $G2/M$ et $M/G1$, nommées classes dans la suite. Chaque classe est ainsi représentée par 4 gènes de référence (tableau 4.4). La figure 4.11 montre les profils d'expression des 4 gènes de référence pour chacune des classes. La

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. 4.4 – Les 20 gènes de référence de Whitfield et al. [106].

procédure d'affectation des gènes à l'une ou l'autre de ces classes est composée de deux étapes. D'abord, pour chacune des 5 classes, les niveaux d'expression des gènes de référence sont moyennés pour créer 5 profils moyens identifiant les classes $G1/S$, S, $G2$, $G2/M$ et $M/G1$. Ensuite, chacun des 1099 gènes étudiés est affecté à la classe du profil moyen le plus proche. La mesure de proximité est fondée sur le coefficient de corrélation de Pearson.

L'observation des 20 profils de référence sur la figure 4.11 révèle quelques contradictions. Nous remarquons que les niveaux d'expression des gènes CDC2, CCNF, CCNA2 n'atteignent pas leur maximum dans la classe G_2 mais plutôt dans G_2/M. De même, les gènes BUB1 et PLK sont plus exprimés dans la classe M/G_1 que dans la classe G_2/M. Ces contradictions sont soutenues par les annotations des bases de données Genecards (http ://www.genecards.org/) et KEGG (http ://www.genome.ad.jp/kegg/kegg2.html).

Classification adaptative

Nous effectuons la classification adaptative, décrite dans la section 4.3, sur les 1099 gènes évoluant de manière cyclique au cours du cycle cellulaire. Afin d'estimer le bon nombre de classes, nous faisons varier le nombre de classes

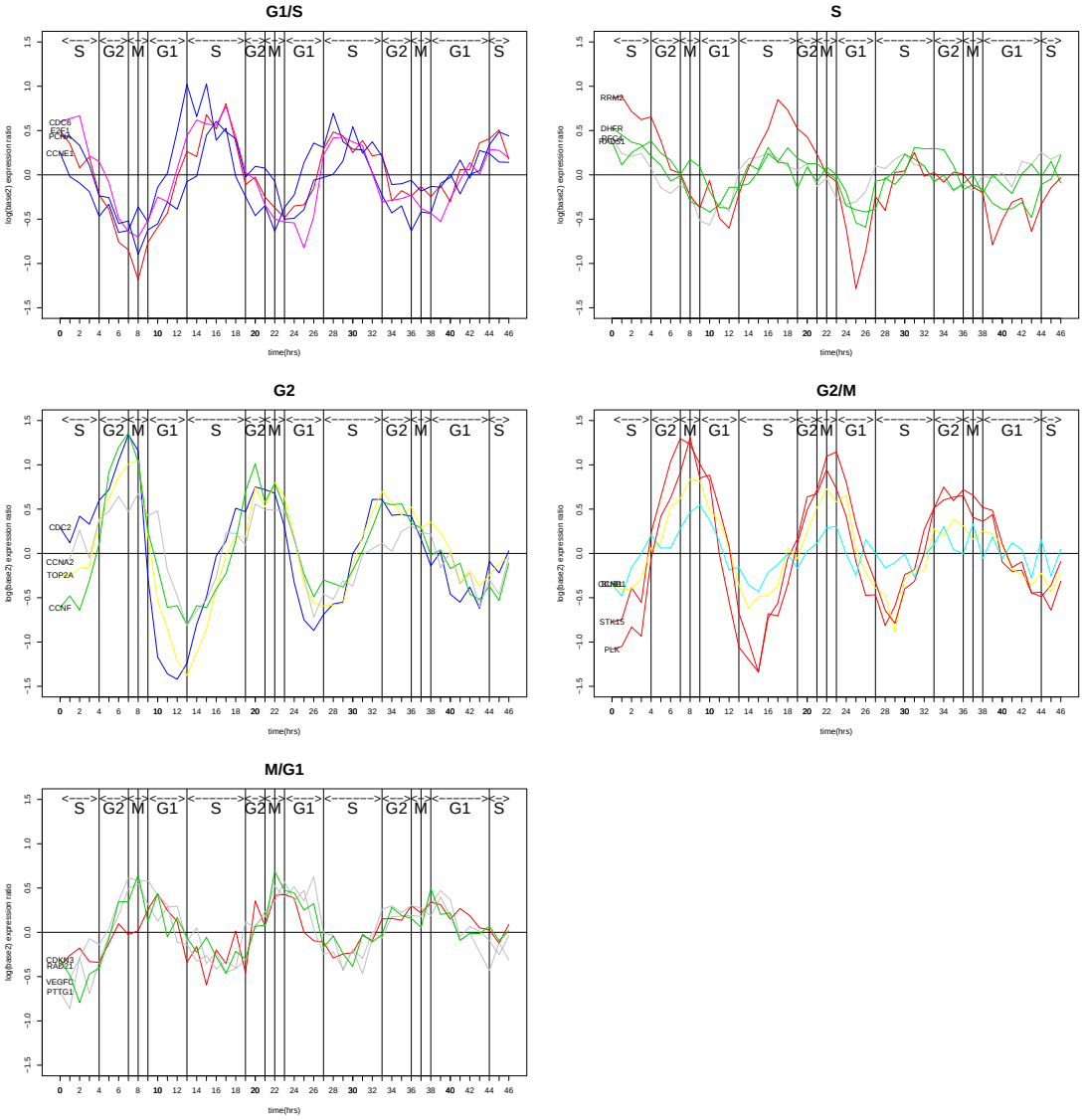


FIG. 4.11 – Profils des 20 gènes de référence (tableau 4.4) observés pendant trois cycles cellulaires, chaque gène s'exprimant fortement pendant l'une des périodes du cycle cellulaire. La durée de chacune des quatre phases du cycle cellulaire (G_1, S, G_2 et M) est délimitée par les lignes verticales.

N de 4 (pour les 4 phases du cycle cellulaire) à 10 (en incluant les phases du cycle cellulaire et leurs transitions). La figure 4.12 gauche indique, pour un nombre de classes N fixé, la progression des valeurs du critère asw en fonction du paramètre k. La figure 4.12 droite montre l'évolution des valeurs du critère wbr en fonction du nombre de classes, pour le paramètre k fixé dans $[0, 6]$. La partition optimale $\mathcal{P}_{N^*,k^*}$ est obtenue pour les paires de valeurs (N^*, k^*) maximisant la valeur asw, tout en minimisant la valeur wbr. Remarquons que, pour les deux critères asw et wbr, la qualité d'une partition $\mathcal{P}_{N,k}$ augmente (i.e., asw augmente, wbr diminue) lorsque k augmente (figure 4.12) ;

ceci montre que les classes de gènes sont plus séparées par la forme des profils d'expression que par leurs valeurs. Selon le critère *asw* (figure 4.12 gauche), la partition optimale est obtenue pour ($N = 4$, $k = 5.7$). Tandis que pour le *wbr* (figure 4.12 droite), l'optimalité est obtenue pour $k = 5.9$ et la valeur de N se situe dans le coude de la courbe correspondant à $k = 5.9$ ($N = 5$, 6, ou 7). Finalement, nous considérons une partition $\mathcal{P}_{N^*,k^*}$ comme optimale celle correspondant aux valeurs $k \in \{5.7, 5.9\}$ et $N \in \{4, ..., 10\}$ qui maximise le ratio asw/wbr. La figure 4.13 gauche permet de visualiser, pour $k = 5.7$ et $k = 5.9$ et pour N allant de 4 à 10, la progression du ratio asw/wbr. La partition optimale $\mathcal{P}_{N^*,k^*}$ est atteinte pour le couple de valeurs ($N^* = 4$, $k^* = 5.7$). La figure 4.13 droite donne le graphe de la silhouette de $\mathcal{P}_{N^*=4,k^*=5.7}$ et ses principales caractéristiques : la valeur *asw* de la partition, la taille et la valeur silhouette de chaque classe.

L'étape suivante consiste à choisir, dans chaque classe de la partition $\mathcal{P}_{N^*,k^*}$,

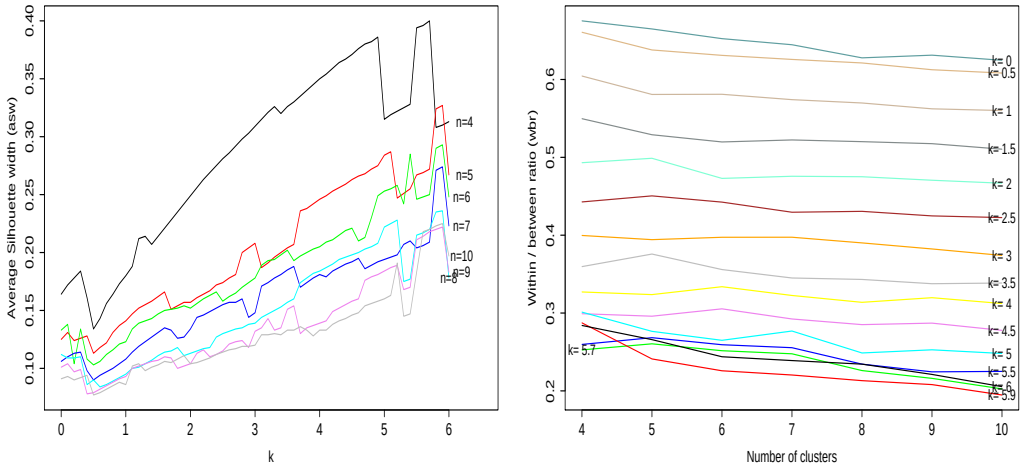


FIG. 4.12 – Progression des valeurs du critère *asw* (gauche) et du critère *wbr* (droite) des partitions $\mathcal{P}_{N,k}$, N (Number of clusters) varie de 4 à 10 et k varie de 0 à 6.

des gènes bien classés selons les critères, appelés gènes noyaux. Ils correspondent, pour chaque classe, aux 10 gènes différents et connus possédant les plus grandes valeurs *sw* supérieures à 0.4 (0.4 étant la *asw* de $\mathcal{P}_{N^*,k^*}$). Remarquons qu'un gène peut être représenté par plusieurs clones dans un ensemble noyau (par exemple, dans le classe 1, le gène $GOLGA8A$ est représenté par 5 clones), ou peut être un nouveau gène qui n'est pas encore étudié (gène non connu, c'est le cas du gène $LOC161527$ dans la même classe 1). La figure 4.14 permet de visualiser, pour chaque classe, les profils d'expression des gènes noyaux. Pour chaque ensemble noyau extrait, nous déterminons la phase pour laquelle ses gènes sont fortement exprimés. Chaque classe est alors affectée à la phase du cycle cellulaire de son ensemble noyau. L'analyse des profils des

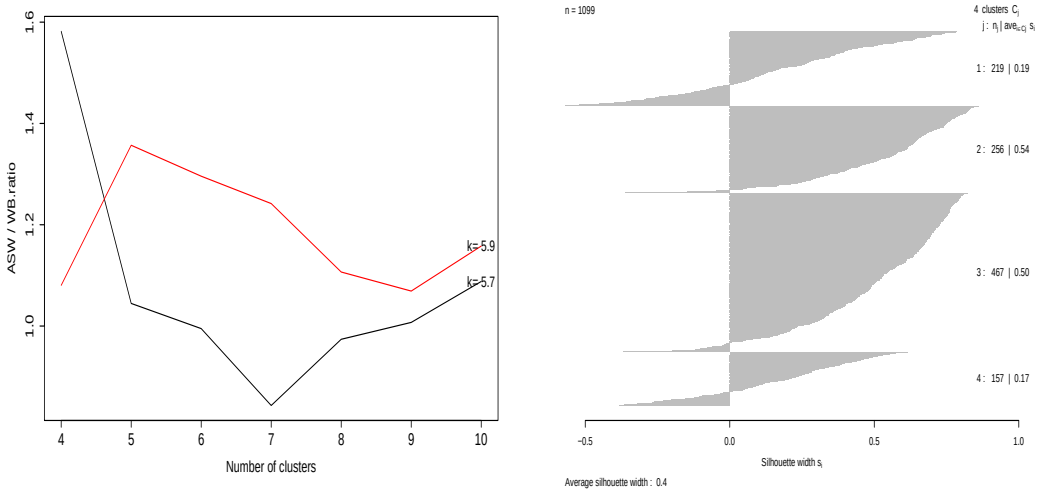


FIG. 4.13 – Progression des ratio asw/wbr (gauche) en fonction du nombre de classes (variant de 4 à 10), le graphe de la silhouette associée à la partition $\mathcal{P}_{N^*=4,k^*=5.7}$ (droite).

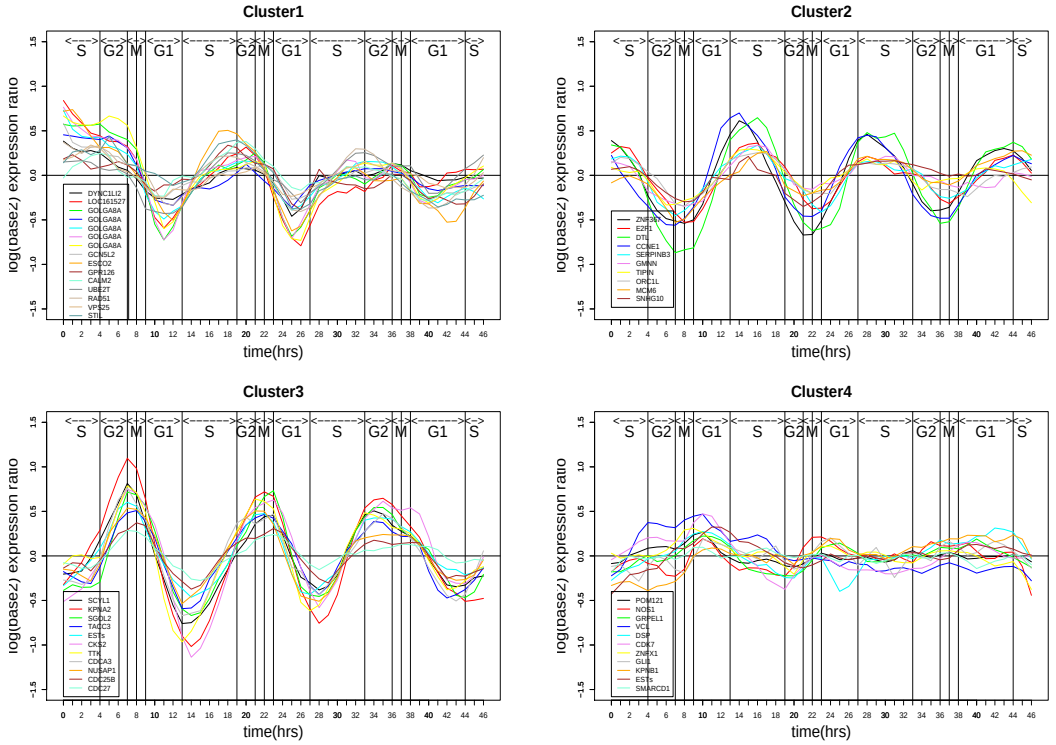


FIG. 4.14 – Les profils d'expression des gènes noyaux au cours des 3 cycles cellulaires : classe 1(S), classe 2(G_1/S), classe 3(G_2/M) et classe 4(G_1).

gènes noyaux révèle que : les niveaux d'expression des gènes noyaux de la classe 1 sont clairement plus exprimés à la phase S, ceux de la classe 2 à la phase G_1/S, ceux la classe 3 à la phase G_2/M et, enfin, ceux de la classe 4 à la phase

G_1. Remarquons que dû à la désynchronisation des cellules, il est plus fiable de limiter nos interprétations aux premiers cycles cellulaires.

Les ensembles noyaux choisis sont détaillés dans le tableau 4.5. Nous indiquons, pour chacun des gènes noyau (Type = N_{oy}) son nom, son numéro de classe ($Num - Class$), son numéro de classe voisine ($Num - Vois$), sa *silhouette width* (sw), sa phase d'affectation par notre classification adaptative ($Ada - Aff$) et enfin sa phase d'affectation par de l'approche de Whitfield et al. [106] ($Whit - Aff$). En outre, nous complétons cette liste avec l'ensemble des gènes de référence (Type = R) utilisés dans l'approche conventionnelle (tableau 4.4). Du fait de la continuité dans le temps du processus biologique étudié, et donc de la contiguïté circulaire des phases du cycle cellulaire dans le processus de division (figure 1.2), nous nous sommes également intéressés à l'ensemble des gènes situés à la frontière des classes (que nous appelons gènes frontaliers, avec des valeurs sw proches de 0) pour révéler des gènes impliqués dans la transition entre deux classes voisines. Ces gènes sont contenus dans le tableau 4.6 avec une valeur $sw \in [-0.05, 0.05]$. Pour chacun de ces gènes, une affectation finale est décidée selon la contiguïté circulaire *a priori* des phases du cycle cellulaire. Par exemple, le gène $BAIAP2$ appartenant à la classe 2 (classe G_1/S) et voisin de la classe 4 (classe G_1) est affecté à la classe G_1. De même, le gène $UBL3$ de la classe 1 (classe S) et voisin de la classe 4 (G_1) est finalement affecté à la classe G_1/S. Tandis que, le gène $KATNA1$ de la classe 1 (classe S) et voisin de la classe 3 (G_2/M) est affecté à la phase G_2, puisque $G2$ se positionne entre S et G_2/M. Les affectations finales de ces gènes sont dressées dans la dernière colonne du tableau 4.6. Dans Whitfield et al. [106], les auteurs compilent depuis la littérature une liste de 39 gènes et comparent les phases publiées avec celles obtenues par leur approche. Le tableau 4.7 présente ces résultats. Nous indiquons, pour chaque gène, sa phase publiée dans la littérature ($Publi - Assi$), celle obtenue par l'approche de ces auteurs ($Whit - Assi$), et enfin la phase d'affectation de notre approche adaptative ($Ada - Assi$).

Justification biologique de la classification adaptative

Examinons d'abord les résultats de la classification adaptative. La partition optimale obtenue $\mathcal{P}_{N^*=4,k^*=5.7}$ pour $asw = 0.4$, signifie que l'ensemble des 1099 profils d'expression peut être raisonnablement divisé en 4 classes basées essentiellement sur leurs formes (k^* voisin de 6). La figure 4.12 montre que, pour chaque nombre de classe, le critère asw croît avec le paramètre k, ce qui révèle que les profils d'expression des gènes sont plus différenciés par leurs formes que par leurs valeurs. La figure 4.13 indique que les classes 2 (G_1/S) et 3 (G_2/M) possèdent les valeurs asw les plus significatives 0.54 et 0.50 respectivement, ce qui signifie qu'elles sont bien séparées des autres classes (G_1 et S). En plus, nous remarquons que tous les gènes de référence de la classe M/G_1 utilisés dans l'approche conventionnelle (tableau 4.4) appartiennent à la classe 3 labé-

Nom	Type	Num-Class	Num-Vois	sw	Ada-Aff	Whit-Aff
DYNC1LI2	N_{oy}	1	1	0.784	S	S
LOC161527	N_{oy}	1	2	0.784	S	S
GOLGA8A	N_{oy}	1	3	0.781	S	S
GOLGA8A	N_{oy}	1	3	0.757	S	S
GOLGA8A	N_{oy}	1	2	0.751	S	S
GOLGA8A	N_{oy}	1	2	0.745	S	S
GOLGA8A	N_{oy}	1	3	0.744	S	S
GCN5L2	N_{oy}	1	2	0.738	S	S
ESCO2	N_{oy}	1	2	0.732	S	S
GPR126	N_{oy}	1	2	0.705	S	S
CALM2	N_{oy}	1	3	0.700	S	S
UBE2T	N_{oy}	1	2	0.697	S	S
RAD51	N_{oy}/R	1	3	0.697	S	S
VPS25	N_{oy}	1	3	0.696	S	G_2
STIL	N_{oy}	1	3	0.696	S	G_2
RRM2	R	1	2	0.635	S	S
CDC2	R	1	3	0.07	S	G_2
DHFR	R	1	2	0.058	S	S
ZNF367	N_{oy}	2	1	0.862	G_1/S	G_1/S
E2F1	N_{oy}	2	1	0.851	G_1/S	G_1/S
DTL	N_{oy}	2	1	0.848	G_1/S	G_1/S
CCNE1	N_{oy}/R	2	1	0.844	G_1/S	G_1/S
SERPINB3	N_{oy}	2	1	0.841	G_1/S	G_1/S
GMNN	N_{oy}	2	1	0.841	G_1/S	G_1/S
TIPIN	N_{oy}	2	1	0.834	G_1/S	G_1/S
ORC1L	N_{oy}	2	1	0.834	G_1/S	G_1/S
MCM6	N_{oy}	2	1	0.833	G_1/S	G_1/S
SNHG10	N_{oy}	2	1	0.829	G_1/S	G_1/S
E2F1	R	2	1	0.820	G_1/S	G_1/S
CDC6	R	2	1	0.753	G_1/S	G_1/S
PCNA	R	2	1	0.641	G_1/S	G_1/S
RFC4	R	2	1	0.618	G_1/S	S
SCYL1	N_{oy}	3	4	0.825	G_2/M	G_2
KPNA2	N_{oy}	3	4	0.815	G_2/M	G_2
SGOL2	N_{oy}	3	4	0.811	G_2/M	G_2/M
TACC3	N_{oy}	3	4	0.810	G_2/M	G_2/M
ESTs	N_{oy}	3	4	0.809	G_2/M	G_2/M
CKS2	N_{oy}	3	4	0.806	G_2/M	G_2/M
TTK	N_{oy}	3	4	0.805	G_2/M	G_2/M
CDCA3	N_{oy}	3	4	0.805	G_2/M	G_2/M
NUSAP1	N_{oy}	3	4	0.799	G_2/M	G_2/M
CDC25B	N_{oy}	3	4	0.799	G_2/M	G_2/M
CDC27	N_{oy}	3	4	0.799	G_2/M	G_2/M
STK15	R	3	4	0.780	G_2/M	G_2/M
BUB1	R	3	4	0.756	G_2/M	G_2/M
CCNF	R	3	1	0.725	G_2/M	G_2
CCNA2	R	3	4	0.720	G_2/M	G_2
PLK	R	3	4	0.710	G_2/M	G_2/M
TOP2A	R	3	1	0.709	G_2/M	G_2
VEGFC	R	3	4	0.675	G_2/M	M/G_1
CDKN3	R	3	4	0.634	G_2/M	M/G_1
PTTG1	R	3	4	0.565	G_2/M	M/G_1
RAD21	R	3	4	0.504	G_2/M	M/G_1
CCNB1	R	3	4	0.483	G_2/M	G_2/M
POM121	N_{oy}	4	3	0.616	G_1	G_2/M
NOS1	N_{oy}	4	3	0.594	G_1	M/G_1
GRPEL1	N_{oy}	4	3	0.583	G_1	M/G_1
VCL	N_{oy}	4	3	0.574	G_1	M/G_1
DSP	N_{oy}	4	3	0.562	G_1	M/G_1
CDK7	N_{oy}	4	3	0.553	G_1	M/G_1
ZNFX1	N_{oy}	4	3	0.539	G_1	M/G_1
GLI1	N_{oy}	4	3	0.538	G_1	G_2/M
KPNB1	N_{oy}	4	3	0.531	G_1	M/G_1
ESTs	N_{oy}	4	3	0.527	G_1	M/G_1
SMARCD1	N_{oy}	4	3	0.513	G_1	G_2/M

TAB. 4.5 – Les ensembles noyaux de la partition $\mathcal{P}_{N^*=4,k^*=5.7}$

Nom	Ada-Assi	Neig	sw	Final-Assi
MZF1	S	G_1/S	0.049	début de S
TncRNA	S	G_1/S	0.049	début de S
CAPS	S	G_1/S	0.041	début de S
AURKB	S	G_2/M	0.039	G_2
ZFX	S	G_2/M	0.038	G_2
KATNA1	S	G_2/M	0.028	G_2
KBTBD2	S	G_2/M	0.026	G_2
CDKL5	S	G_2/M	0.020	G_2
TTC31	S	G_1/S	0.013	début de S
LOC134121	S	G_2/M	0.012	G_2
UBL3	S	G_1	0.011	G_1/S
CDKN2C	S	G_2/M	0.000	G_2
REEP1	S	G_1/S	-0.012	début de S
TOP2A	S	G_2/M	-0.023	G_2
MICA/HCP5	S	G_2/M	-0.039	G_2
CDH24	S	G_1/S	-0.041	début de S
ABCC5	S	G_1/S	-0.044	début de S
RECQL4	G_1/S	S	0.047	début de S
SLC9A3	G_1/S	S	0.046	début de S
FLJ13231	G_1/S	S	0.028	début de S
ESTs	G_1/S	S	0.001	début de S
EST	G_1/S	G_1	-0.019	fin de G_1
HIST1H2AM	G_1/S	S	-0.045	début de S
BAIAP2	G_1/S	G_1	-0.045	fin de G_1
CRLF3	G_1/S	S	-0.050	début de S
NR5A2	G_2/M	G_1	0.050	M
HERPUD2	G_2/M	G_1	0.045	M
AMD1	G_2/M	G_1	0.035	M
NIPBL	G_2/M	G_1	0.012	M
NFIC	G_2/M	G_1	0.008	M
ESTs	G_2/M	G_1	0.006	M
ChGn	G_2/M	G_1	0.003	M
BCLAF1	G_2/M	S	0.001	G_2
WWC1	G_2/M	G_1	-0.003	M
HLA-DOA	G_2/M	G_1	-0.012	M
AGPAT3	G_2/M	G_1	-0.015	M
C20orf199	G_2/M	G_1	-0.017	M
SLC39A10	G_2/M	G_1	-0.020	M
LARP1	G_2/M	G_1	-0.024	M
ANP32B	G_2/M	G_1	-0.026	M
ABHD10	G_2/M	S	-0.029	G_2
FXR1	G_2/M	G_1	-0.032	M
ZNF207	G_1	G_2/M	0.050	M
HSPA2	G_1	G_2/M	0.048	M
PPP2CA	G_1	G_2/M	0.044	M
CEP350	G_1	G_2/M	0.017	M
OC146517	G_1	G_2/M	0.013	M
SAP30BP	G_1	S	0.009	G_1/S
DR1	G_1	G_2/M	0.007	M
TMEM132A	G_1	G_2/M	0.002	M
W85890	G_1	G_2/M	-0.007	M
PCF11	G_1	G_2/M	-0.021	M
DNAJA1	G_1	G_2/M	-0.022	M
TSC22	G_1	G_2/M	-0.023	M
ETS	G_1	G_2/M	-0.024	M
SRF	G_1	G_2/M	-0.034	M
TYSND1	G_1	G_2/M	-0.037	M
C4B	G_1	G_2/M	-0.048	M

TAB. 4.6 – Affectation finale des gènes frontaliers ($sw \in [-0.05, 0.05]$) selon la continuité circulaire des phases du cycle cellulaire.

lisée G_2/M. En effet, l'observation des profils d'expression sur la figure 4.11, permet de voir que les gènes, référencés à la classe M/G_1, sont plus exprimés

Nom	Publi-Assi	Whit-Assi	Ada-Assi	Neig	sw
E2F5	G_1	G_2/M	G_2/M	G_1	0.657
CCNE1	G_1/S	G_1/S	G_1/S	S	0.844
NPAT	G_1/S	G_1/S	G_1/S	S	0.823
SLBP	G_1/S	G_1/S	G_1/S	G_1	0.823
E2F1	G_1/S	G_1/S	G_1/S	S	0.820
MCM6	G_1/S	G_1/S	G_1/S	S	0.782
MCM2	G_1/S	G_1/S	G_1/S	S	0.781
CDC25A	G_1/S	G_1/S	G_1/S	S	0.768
CDC6	G_1/S	G_1/S	G_1/S	S	0.753
CCNE2	G_1/S	G_1/S	G_1/S	S	0.739
PCNA	G_1/S	G_1/S	G_1/S	S	0.641
CDKN3	G_1/S	M/G_1	G_2/M	G_1	0.634
CENPF	G_2	G_2/M	G_2/M	G_1	0.763
CCNF	G_2	G_2	G_2/M	S	0.725
CCNA2	G_2	G_2	G_2/M	G_1	0.720
TOP2A	G_2	G_2	G_2/M	S	0.709
CDC2	G_2/M	G_2	G_2	G_2/M	0.070
CKS2	G_2/M	G_2/M	G_2/M	G_1	0.788
STK15	G_2/M	G_2/M	G_2/M	G_1	0.780
CDC25C	G_2/M	G_2	G_2/M	S	0.773
CENPA	G_2/M	G_2/M	G_2/M	G_1	0.762
BUB1	G_2/M	G_2/M	G_2/M	G_1	0.756
BIRC5	G_2/M	G_2/M	G_2/M	G_1	0.741
PLK	G_2/M	G_2/M	G_2/M	G_1	0.710
CKS1	G_2/M	G_2	G_2/M	S	0.706
CCNB2	G_2/M	G_2/M	G_2/M	G_1	0.696
CDC20	G_2/M	G_2/M	G_2/M	G_1	0.694
BUB1B	G_2/M	G_2/M	G_2/M	G_1	0.690
CDC25B	G_2/M	G_2/M	G_2/M	G_1	0.552
CCNB1	G_2/M	G_2/M	G_2/M	G_1	0.483
CDKN2D	G_2/M	G_2/M	G_2/M	G_1	0.264
RRM2	S	S	S	G_1/S	0.635
RRM1	S	S	S	G_1/S	0.568
DHFR	S	S	S	G_1/S	0.058
CDKN2C	S	G_2	G_2	G_2/M	0.000
NASP	S	G_1/S	G_1/S	S	0.749
BRCA1	S	S	G_1/S	S	0.447
MSH2	S	G_1/S	G_1/S	S	0.404
TYMS	S	S	G_1/S	S	0.311

TAB. 4.7 – Liste des 39 gènes compilés depuis la littérature par Whitfield et al. [106]. Comparaison des phases publiées ($Publi - Assi$) avec celles issues de l'approche conventionnelle ($Whit - Assi$) et de la classification adaptative ($Ada - Assi$).

pendant la transition G_2/M que pendant la transition M/G_1, ce qui corrobore avec l'affectation produite par la classification adaptative et révèle un possible problème d'annotation des gènes de référence de la classe M/G_1. Enfin, le tableau 4.5 montre, à travers toutes les classes, que les gènes de référence (interprétés depuis l'expérimentation) caractérisent moins bien les phases du cycle cellulaire que les gènes noyaux (la valeur *sw* d'un gène de référence est inférieure à celle d'un gène noyau).

Evaluons biologiquement certaines affectations des gènes noyaux basées sur les annotations de GeneCards (http ://www.genecards. Org/ index.shtml). Le gène $CDK7$ de la classe 4 (G_1) correspond à une kinase participant à la formation d'un complexe trimérique avec la cycline H et un activateur, complexe

impliqué dans l'activation de la cycline D régulatrice de la phase G_1. Dans la classe 2 (G_1/S), le gène $E2F1$, un facteur de transcription connu comme régulateur majeur de la division du cycle cellulaire et impliqué dans le contrôle de la progression de la phase G_1 à la phase S, expose une valeur très significative de $sw = 0.851$. Les gènes connus $CCNE1$ ($sw = 0.844$), qui code pour la protéine cycline E1 impliquée dans la régulation de la transition G_1/S, et $MCM6$ ($sw = 0.833$), qui code pour une protéine impliquée dans la formation des complexes de reconnaissance des origines de réplication, activés et induits respectivement par $E2F1$ sont également classés en phase G_1/S. Dans la classe 3 (G_2/M), le gène $CCNA2$ (codant pour la cycline A2) est expérimentalement connu en tant que régulateur de la transition G_2/M ; le gène $SGOL2$, jouant un rôle central dans la cohésion des chromosomes au cours de la prophase, est actif à l'entrée de la phase M (ou G_2/M). Enfin, dans la classe 1 (S), le gène $UBE2T$, enzyme de conjugaison d'ubiquitine, est impliqué dans un processus de réparation des lésions de l'ADN au cours de la phase S. Nous remarquons que tous les gènes notés "Est" ou "Locxxx" ne sont ni annotés ni étudiés dans la littérature. L'approche proposée va aider les chercheurs à faire de nouvelles hypothèses sur leur fonction dans le contexte de la régulation de la division cellulaire.

Nous nous focalisons maintenant sur certaines affectations des gènes frontaliers indiqués dans le tableau 4.6. Par exemple, le gène $UBL3$ de la classe 1 (S) et voisin de la classe 4 (G_1) est rangé dans la classe G_1/S. Cette affectation est soutenue par le fait que $UBL3$ est connue comme une proteine ubiquitine ligase, protéine connue pour agir dans la dégradation des protéines régulatrices par la voie ubiquitine-protéasome à la transition G_1/S. Le gène $HIST1H2AM$ appartenant à la classe 2 (G_1/S) et voisin de la classe 1 (S) est impliqué dans la transition G_1/S et au cours de la phase S. En effet, la synthèse de la protéine histone $H2A1$ (un élément essentiel du nucléosome qui enveloppe et comprime l'ADN en chromatine) codée par le gène $HIST1H2AM$, commence à la transition G_1/S et s'exprime fortement au cours de la phase S avec un pic dés le début de phase S. Le gène $KATNA1$ de la classe 1 (S) et voisin de la classe 3 (G_2/M), est affecté à la phase G_2. Cette attribution semble cohérente avec la littérature, puisque que ce gène est impliqué dans la réorganisation des microtubules pendant la phase G_2 pour préparer la transition G_2/M. Le dernier exemple concerne le gène $NIPBL$, appartenant à la classe 3 (G_2/M) et voisin de la classe 4 (G_1), qui est considéré comme un gène de la phase M. En effet, il est connu que le produit du gène $NIPBL$ fonctionne comme une protéine de cohésion des chromatides sœurs composant les chromosomes métaphasiques, et est essentielle pour la ségrégation normale des chromosomes, survenant au cours de la phase M.

Comparaison d'un classement adaptatif et d'un classement conventionnel

Attachons nous à comparer l'efficacité du classement de Whitfield et al. [106] et le classement basé sur les gènes noyaux (tableau 4.5) et la métrique $D_{k*=5.7}$. Cette comparaison est effectuée sur des profils simulés par le modèle (4.5) présenté dans la première application. Les gènes de référence et les gènes noyaux ne couvrent pas toutes les phases et transitions de phase du cycle cellulaire, nous avons donc choisi de simuler des gènes qui caractérisent les 7 phases S, G_2, G_2/M, M, M/G_1, G_1 et G_1/S. Ceci permettra de savoir si les gènes simulés des phases représentées seront bien affectés aux mêmes phases et si ceux des phases non représentées seront affectés à la phase voisine. Nous simulons 700 profils de gènes, dont 100 par phase. $\theta_g = (K_g, 15, \sigma, \Phi_g, 0.1, 0.04)$ avec $K_g \in [0.3, 1.5]$, $\sigma \in \{0, 0.05, 0.075, 0.13\}$, et Φ_g prend les valeurs 5.214, 10.227, 9.625, 9.224, 8.823, 8.422, 7.019 respectivement pour les phases S, G_2, G_2/M, M, M/G_1, G_1, G_1/S.

Le processus de classement est le suivant : nous disposons d'une part de 5 classes de gènes de référence et d'autre part de 4 classes de gènes noyaux. Chaque gène simulé est affecté à la classe des gènes noyaux la plus proche (i.e., la classe des gènes noyaux qui minimise la distance moyenne, avec $D_{k*=5.7}$ comme mesure de proximité). De même, selon l'approche fondée sur le coefficient de corrélation, chaque gène simulé est affecté à la classe de référence la plus corrélée (*cf.* § 4.4.2). Les résultats du classement des 700 profils simulés sont résumés dans le tableau 4.8. Les 7 classes des gènes simulés sont représentées en lignes et les classes des gènes noyaux et des gènes de référence sont en colonnes. Par exemple, la première ligne du tableau représente la classe S des gènes simulés. Tous les 100 gènes simulés de la classe S sont affectés à la même classe S par le classement fondé sur les gènes noyaux et de la dissimilarité apprise $D_{k*=5.7}$. Tandis que l'approche fondée sur le coefficient de corrélation et sur les gènes de référence affecte 29 gènes de la classe simulée S à la même classe S et 71 gènes simulés de la classe S à la phase G_1/S. Un profil de gène simulé est supposé être bien classifié, s'il est affecté à la classe qui porte la même étiquette que lui ou s'il est affecté à la classe voisine (s'il n'existe pas de classe même étiquette), et mal classifié dans le cas contraire (indiqué en gras dans le tableau 4.8).

Le tableau 4.8 (sous-tableau gauche) donne les statistiques d'affectation des profils simulés par rapport aux classes des gènes noyaux. Nous notons que les gènes simulés des classes S, G_2/M et G_1/S sont tous bien affectés aux classes des gènes noyaux de même étiquette. Les gènes simulés des classes $G2$, M et $M/G1$ sont tous affectés à la classe de gènes noyaux voisine G_2/M, puisqu'il n'existe pas de classes de gènes noyaux d'étiquettes G_2, M et M/G_1. Enfin, le mauvais classement concerne 27% des gènes simulés de la classe G_1, mal affectés à la classe de gènes noyau G_2/M. Le tableau 4.8 (sous-tableau droit)

Classes simulées	Classes des gènes noyaux				Classes des gènes de référence				
	S	G_2/M	G_1	G_1/S	S	G_2	G_2/M	M/G_1	G_1/S
S	100	0	0	0	29	0	0	0	**71**
G_2	0	100	0	0	0	65	0	**35**	0
G_2/M	0	100	0	0	0	0	0	**100**	0
M	0	100	0	0	0	0	0	100	0
M/G_1	0	100	0	0	0	0	0	100	0
G_1	0	**27**	73	0	0	0	0	100	0
G_1/S	0	0	0	100	0	0	0	0	100

TAB. 4.8 – Les statistiques d'affectation des 700 profils simulés basés sur les classes noyaux (sous-tableau gauche) et les classes références (sous-tableau droit). Les nombres en gras correspondent aux nombres de profils de gènes simulés mal classifiés.

donne les statistiques d'affectation des gènes simulés par rapport aux classes de gènes de référence. Nous remarquons que les gènes simulés des classes M/G_1 et G_1/S sont tous bien affectés à des classes de gènes de référence de même étiquette. Les gènes simulés des classes M et G_1 sont tous affectés à la classe de gènes de référence voisine M/G_1, puisqu'il n'existe pas de classes de gènes de référence d'étiquettes M et G_1. La mauvaise classification concerne 71% des gènes de la classe S, 35% des gènes de G_2, et 100% des gènes de G_2/M.

Nous avons calculé la valeur sw de chaque gène pour indiquer s'il est bien classifié, mal classifié ou s'il peut être aussi affecté à sa classe voisine. La figure 4.15 illustre la distribution des valeurs sw des profils simulés, classifiés par les gènes noyaux et la métrique $D_{k*=5.7}$ d'une part et d'autre part par les gènes de référence et la distance fondé sur le coefficient de corrélation. La figure 4.15 (gauche) montre que presque tous les gènes simulés sont très bien classifiés par rapport aux classes des gènes noyaux (médiane $sw > 0.6$). Par exemple, le troisième boxplot à partir de la gauche, indique que les 100 gènes de la classe G_2/M sont très bien classifiés dans la classe des gènes noyaux G_2/M étant donné que leurs sw sont dans l'intervalle [0.918, 0.969]. Enfin, sur la figure 4.15 (droite), nous pouvons noter que les gènes simulés sont moins bien classifiés par rapport aux classes de références qu'aux classes noyaux, compte tenu de leurs très faibles valeurs sw.

En résumé, cette étude montre que le classement des gènes simulés fondé sur les gènes noyaux extraits de manière non supervisée et sur la dissimilarité apprise $D_{k*=5.7}$ donne de meilleurs résultats que l'approche fondée sur les gènes de référence et le coefficient de corrélation. Le taux d'erreur de classement est d'environ 0.038 ($\frac{27}{700}$) pour la première approche (gènes noyaux) contre 0.294 ($\frac{71+35+100}{700}$) pour la seconde (gènes de référence). D'autre part, les valeurs sw

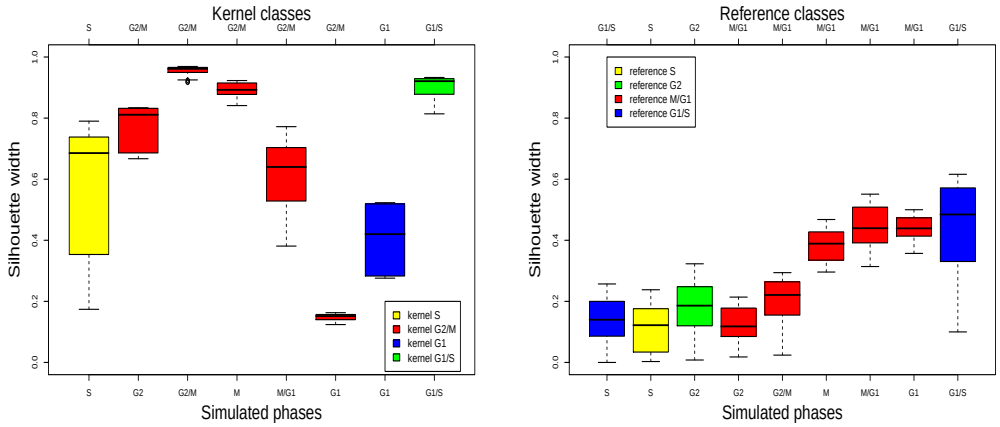


FIG. 4.15 – La distribution des valeurs *sw* des profils simulés, classifiés par l'approche adaptative (gauche) et l'approche basée sur la corrélation (droite).

(figure 4.15) des gènes simulés montrent clairement qu'ils sont mieux classifiés par les gènes noyaux que par les gènes de référence.

4.4.3 Gènes périodiquement exprimés dans les cellules du cancer du sein MCF-7 [63]

Description des données

Le but de cette application est l'analyse de données transcriptomiques d'une lignée de cellules cancéreuses du sein MCF-7, étudiée dans Lobenhofer et al. [63]. Dans cette étude, les cellules ont été stimulées avec de l'œstrogène (17β-estradiol) à des doses mitogéniques et observées à différents intervalles de temps (1, 4, 12, 24, et 48 heures) pendant un cycle cellulaire. L'objetctif est d' analyser les profils des gènes en fonction de leurs réponses aux œstrogènes. Notre étude se concentre sur les 1901 gènes de la première expérience d'hybridation. Selon l'expérimentation de Lobenhofer, à l'instant $t = 1$ heure la plupart des cellules sont en début de G_1, à l'instant $t = 4$ heures les cellules progressent vers le point de contrôle G_1/S, l'instant $t = 12$ heures correspond à une entrée en phase S avec une réplication du génome de 24 à 36 heures, et la fin du cycle cellulaire est atteinte à $t = 48$ heures. Les détails sur le traitement de ces données et les hybridations réalisées peuvent être trouvés dans Lobenhofer et al. [63] (http ://dir.niehs.nih.gov/microarray/chips.htm).

Analyse conventionnelle

Dans Peddada et al. [78], les auteurs proposent une approche supervisée qui repose sur un ensemble de profils de référence prédéfinis. Chaque profil de référence est exprimé par un ensemble d'inégalités entre les valeurs d'expression estimées aux différents instants. 10 profils de référence sont considérés : 1 profil avec une décroissance monotone ("dec"), 1 profil avec un accroissement

monotone ("acc"), 4 profils qui croissent puis décroissent ("acc-dec") avec des maximas aux instants 4, 12, 24, 36 heures, respectivement, et 4 profils qui décroissent puis croissent ("dec-acc") avec des minimas aux instants 4, 12, 24, 36 heures, respectivement. Après l'application de l'algorithme de Peddada et al. [78] sur les 1901 les gènes, seuls 50 ont été sélectionnés et affectés à 9 des 10 profils de référence prédéfinis. Un gène est sélctionné lorsqu'il respecter les critères basées sur une méthode de bootstrap. Ce qui veut dire que tous les gènes ne sont pas affectés à un profil de référence.

Classification adaptive

Notre classification adaptive est appliquée aux 1901 gènes pour identifier ceux qui présentent des profils d'expression pertinents. La figure 4.16 montre la progression des valeurs du critère *asw* et celle du critère *wbr* de la partition $P_{N,k}$ en fonction du paramètre k et du nombre de classes N ($N = 2$ à 15) respectivement. La partition optimale P_{N^*,k^*} est obtenue pour le couple de valeurs (N^*, k^*) maximisant la *asw* et minimisant le *wbr* à la fois. Pour les deux critères, la valeur optimale de k est de 6, tandis que le nombre optimal de classes est de 4 selon le critère *asw* (figure 4.16 gauche), et 8 (coude de la courbe $k = 6$) selon le critère *wbr* (figure 4.16 droite). En considérant simultanément les deux critère *asw* et *wbr*, la partition optimale est obtenue pour $N^* = 8$ et $k^* = 6$ (maximisant le rapport *asw*/*wbr* comme dans l'exemple précédent). La figure 4.17 illustre le graphe de la silhouette correspondant à la partition $P_{N^*=8,k^*=6}$ et ses principales caractéristiques : la silhouette moyenne de la partition (*asw*), la taille et la silhouette moyenne de chaque classe.

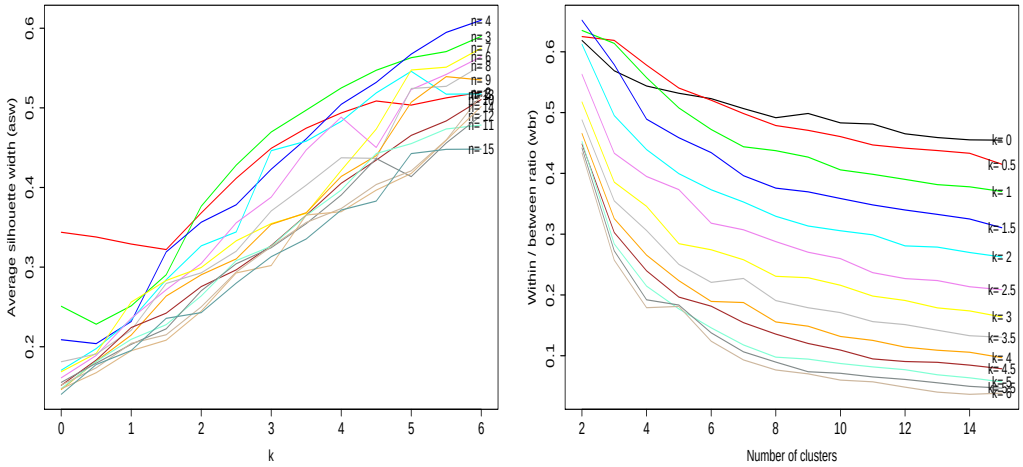


FIG. 4.16 – La progression des valeurs du critère *asw* (gauche) et celle du critère *wbr* (droite) des partitions $\mathcal{P}_{N,k}$, N (Number of clusters) varie de 2 à 15 et k varie de 0 à 6.

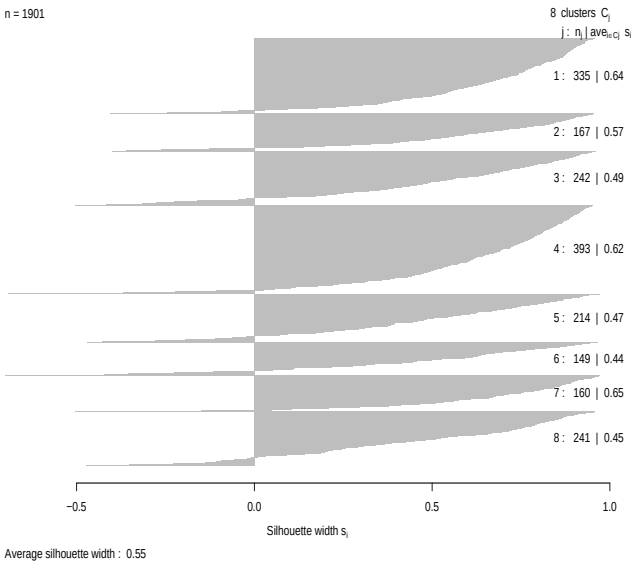


FIG. 4.17 – Le graphe de la silhouette associée à la partion $P_{N^*=8,k^*=6}$.

L'étape suivante consiste à sélectionner un ensemble de gènes caractérisant chaque classe de la partition $P_{N^*=8,k^*=6}$ (i.e., les profils les plus représentatifs). Ils correspondent, pour chaque classe, aux 10 gènes possédant les plus grandes valeurs de silhouette width (sw) supérieure à 0.55 (0.55 étant la valeur du critère asw de la partition $P_{N*=8,k*=6}$). Les profils des 8 ensembles noyaux extraits sont visualisés dans la figure 4.18. Le tableau 4.9 présente la structure de voisinage des classes obtenues. Pour chaque classe, nous indiquons ses quatres plus proches voisines (Vois 1 à Vois 4), en indiquant entre parenthèses le nombre de gènes correspondants. Par exemple, la première ligne indique que 194 gènes de la classe 1 sont voisins de la classe 8, 132 gènes de la classe 1 sont voisins de la classe 4, etc. Enfin, le tableau 4.10 fournit les affectations des 50 gènes sélectionnés par l'approche de Peddada et al. [78] aux 8 classes de la partition $P_{N^*=8,k*=6}$. Chacun de ces 50 gènes est décrit par son numéro d'identification (ID) dans la base de données UniGene en date du 11 février 2002, son numéro de classe ($Num - Class$), le numéro de sa classe voisine ($Num - Vois$), sa silhouette width (sw) obtenue par la méthode adaptative, et son affectation par rapport l'approche conventionnelle ($Pedd - Aff$).

Num-Class	Vois 1	Vois 2	Vois 3	Vois 4
1	8 (194)	4 (132)	7 (16)	2 (10)
2	4 (94)	7 (30)	5 (27)	6 (9)
3	4 (116)	5 (60)	8 (45)	1 (14)
4	1 (140)	2 (125)	3 (120)	6 (5)
5	3 (63)	6 (62)	2 (41)	7 (35)
6	5 (61)	4 (34)	2 (29)	7 (11)
7	8 (55)	2 (37)	1 (24)	5 (22)
8	1 (150)	7 (46)	3 (35)	5 (10)

TAB. 4.9 – La structure de voisinage des 8 classes de la partition $P_{N^*=8,k^*=6}$.

Num-Class	Num-Vois	ID	Pedd-Aff	sw
1	4	359191	dec-acc (Min 24)	0.888
1	8	248613	acc-dec (Max 4)	0.497
3	8	180789	dec-acc (Min 24)	0.711
3	5	365147	dec-acc (Min 36)	0.698
3	4	162479	dec-acc (Min 24)	0.434
3	1	430235	dec-acc (Min 24)	0.354
3	5	145093	dec-acc (Min 12)	0.016
4	6	470480	acc-dec (Max 36)	0.422
4	3	268652	dec-acc (Min 24)	0.16
4	3	361381	dec-acc (Min 12)	0.102
5	3	29682	dec-acc (Min 36)	0.84
5	6	339075	acc-dec (Max 24)	0.563
5	7	366842	acc-dec (Max 24)	0.47
5	2	485875	dec-acc (Min 12)	0.385
5	3	417226	dec (Max 1)	0.314
5	6	509614	acc-dec (Max 36)	0.301
6	5	229509	acc-dec (Max 24)	0.957
6	5	196676	acc-dec (Max 12)	0.943
6	5	49950	acc-dec (Max 24)	0.91
6	2	563187	acc-dec (Max 12)	0.902
6	5	129140	acc-dec (Max 24)	0.881
6	2	264117	acc-dec (Max 36)	0.863
6	2	49940	acc-dec (Max 24)	0.855
6	5	198205	acc-dec (Max 24)	0.766
6	5	487407	dec-acc (Min 4)	0.743
6	5	488059	acc-dec (Max 24)	0.694
6	5	346838	acc-dec (Max 24)	0.646
6	7	242182	acc-dec (Max 36)	0.643
6	2	136609	acc-dec (Max 24)	0.63
6	5	321207	acc-dec (Max 12)	0.448
6	7	563809	acc-dec (Max 24)	0.312
6	5	487757	acc-dec (Max 24)	0.306
6	5	32713	acc-dec (Max 24)	0.278
6	7	200573	acc-dec (Max 24)	0.0.15
6	2	359465	acc-dec (Max 24)	0.098
6	1	362059	acc-dec (Max 4)	-0.001
7	8	293274	acc-dec (Max 24)	0.955
7	8	248008	acc-dec (Max 24)	0.899
7	6	359119	acc-dec (Max 24)	0.867
7	6	489092	acc-dec (Max 24)	0.826
7	1	150163	acc-dec (Max 36)	0.794
7	6	510595	acc-dec (Max 36)	0.712
7	3	415639	acc-dec (Max 24)	0.56
8	3	428733	acc-dec (Max 4)	0.78
8	1	417503	acc-dec (Max 4)	0.705
8	1	110022	acc-dec (Max 4)	0.426
8	3	34821	dec-acc (Min 12)	0.293
8	7	285427	acc-dec (Max 24)	0.207
8	1	238545	acc-dec (Max 36)	0.07
8	3	545242	dec-acc (Min 24)	0.033

TAB. 4.10 – Classification des 50 gènes identifiés par Peddada et al. [78] selon notre approche adaptative. Chacun de ces 50 gènes est décrit par son numéro d'identification (ID, dans la base de données UniGene en date du 11 février 2002), son numéro de classe ($Num - Class$), le numéro de sa classe voisine ($Num - Vois$), sa silhouette width (sw) obtenue par la méthode adaptative, et son affectation par rapport l'approche de Peddada et al. [78] ($Pedd - Aff$).

Discussion

Comme dans l'application précédente, le critère de qualité de la partition $P_{N,k}$ augmente (i.e., augmentation de la valeur asw et diminution de la valeur wbr)

lorsque k augmente (figure 4.16), révèlant encore que les profils d'expression des gènes sont mieux séparés par leur forme que par leurs valeurs. La partition optimale $P_{N^*=8,k^*=6}$, obtenue pour une valeur $asw = 0.55$, montre qu'il est tout à fait raisonnable de diviser l'ensemble des 1901 gènes en 8 classes bien distinctes. En effet, la figure 4.18 illustre bien 8 ensembles noyaux représentant 8 classes de formes différentes, avec des profils très proches en forme au sein de chaque ensemble noyau. Notons que la méthode de Peddada et al. [78] est basée sur une approche supervisée qui tente à identifier des profils de gènes similaires aux profils de références de formes prédéfinies, tandis que la classification adaptive est une approche non supervisée, qui vise à déterminer ces gènes de référence (i.e., les gènes noyaux). Selon les formes des gènes noyaux obtenus, il semble que les 10 profils prédéfinis sont trop simplistes. Cela s'explique par le fait que de nombreux profils simplistes prédéfinis considérés par Peddada et al. [78] appartiennent à une même classe générée par l'approche adaptative (tableau 4.10). Par exemple, les gènes 359191 (dec-acc (min 24)) et 248613 (acc-dec (max 4)) sont tous deux affectés à la classe 1, et comme illustré sur la figure 4.18, le profil de la classe 1 peut être considéré comme le regroupement de deux profils simplistes : un acc-dec (max 4) et un dec-acc (min 24). D'autre part, parmi les 50 gènes sélectionnés, seuls 3 font partie des ensembles noyaux, et les 47 restants caractérisent moins bien les principaux profils que les gènes noyaux (sw des 47 gènes restants $\leq sw$ d'un gène noyau).

Discutons biologiquement sur les principaux résultats obtenus. On sait que chaque famille de cyclines, se lie à une protéine kinase spécifique Cdk, l'active à une phase spécifique du cycle cellulaire et contribue à l'activation du complexe cycline/Cdk suivant. Par exemple, l'activation des complexes cycline D/Cdk4-6, marque le début du cycle cellulaire (phase G_1), et conduit à l'activation de plusieurs autres cyclines. Dans cette expérience, le 17β-estradiol induit l'expression des trois cyclines D1, D2 et D3 appartenant aux classes 8, 1 et 3, respectivement. En outre, l'activation les cyclines D est toujours suivie par l'activation des cyclines E (nécessaire pour la transition de G_1 à S) et B (qui augmente au cours du cycle cellulaire jusqu'à la mitose) appartenant respectivement aux classes 7 et 6. Par conséquent, les gènes des classes 8, 1 et 3 sont activés pendant la première phase du cycle cellulaire (G_1), les gènes de la classe 7 sont activés pendant la transition G_1/S et les gènes de la classe 6 sont activés de la phase S à la transition G_2/M.

Selon la contiguïté circulaire des phases du cycle cellulaire et la structure de voisinage donnée dans le tableau 4.9, nous pouvons organiser séquentiellement les classes par rapport à leurs principaux instants d'activation : 4, 1, 8, 7, 2, 6. En effet, la figure 4.18 montre que les gènes noyau de la classe 4 sont fortement exprimés à l'instant $t = 1$ heure puis à l'instant $t = 48$ heures, ceux de la classe 1 à $t = 4$ heures puis à $t = 48$ heures, ceux de la classe 8 à $t = 4$ heures puis diminuent, ceux de la classe 7 à $t = 24$ heures, ceux de la classe 2 à $t = 48$ heures et enfin ceux de la classe 6 à $t = 12$ heures suivi d'une faible diminu-

tion. Chose intéressante, la classe 5 contient le facteur de transcription $E2F$ qui joue un rôle central dans la régulation de la prolifération cellulaire. Ceci explique probablement les profils lisses caractérisant la classe 5. Enfin, nous remarquons que la structure de contiguïté des 8 classes proposée pourraient être d'un grand intérêt pour la conception de réseaux moléculaires.

4.5 Conclusion

Pour identifier les gènes exprimés au cours du cycle cellulaire, nous proposons une nouvelle approche de classification adaptative qui consiste en deux points principaux. Premièrement, elle permet aux utilisateurs de choisir de manière non supervisée un ensemble de gènes de référence bien justifié, afin de confirmer ceux extraits expérimentalement. Deuxièmement, elle permet aux chercheurs de connaître la mesure de proximité appropriée à utiliser pour des données temporelles d'expression de gènes et qui couvre à la fois les proximités en valeurs et en forme.

Les résultats basés sur des données simulées nous permettent de conclure que, pour la classification des profils d'expression de gènes, il est souhaitable de considérer la corrélation temporelle comme mesure de proximité. Toutefois, l'efficacité de la dissimilarité apprise D_k, qui fournit également de très bonnes partitions, est à noter. En général, lorsqu'on est confronté à des données temporelles (ce qui est le cas des profils d'expression de gènes du cycle cellulaire), la dissimilarité proposée D_k est recommandée. Elle peut tendre vers la corrélation temporelle (pour $k* = 6$), vers la distance euclidienne (pour $k* = 0$), ou, plus généralement, vers une mesure couvrant à la fois les proximités en valeurs et en forme. Par rapport aux données réelles, les résultats obtenus, évalués par les connaissances biologiques et l'étude de simulation ont apporté de nouveaux arguments d'explication et de comparaison aux résultats publiés et d'en tirer des hypothèses sur de nouveaux gènes qui ne sont pas encore annotés.

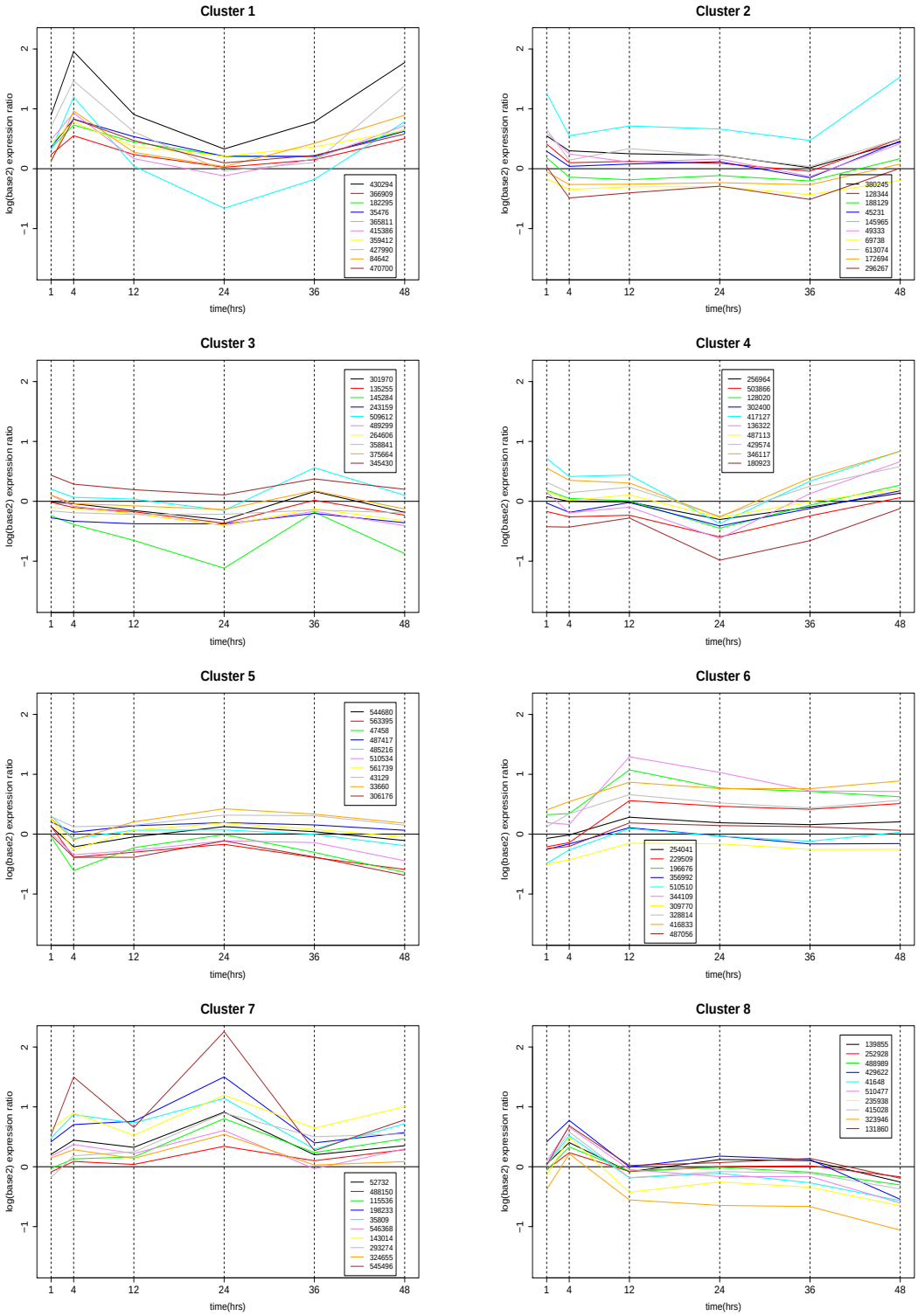


FIG. 4.18 – Les profils d'expression des gènes noyaux des 8 classes issues de notre classification adaptative. Les 10 gènes noyau de chaque classe sont très similaires en forme entre eux et différents en forme des autres gènes noyaux des autres classes.

5

Conclusion

Compte tenu des caractéristique atypiques des données temporelles d'expression de gènes (dépendance des observations pour chaque gène, longueur courte des profils de gènes considérés comme séries temporelles, données bruitées), l'utilisation de nouvelles méthodes d'analyse est incontournable. Nous avons assité, ces dernières années, à l'émergence d'une littérature abondante portant sur l'utilisation des méthodes d'analyse de données pour l'analyse de données temporelles d'expression de gènes. Les méthodes de classification sont les plus utilisées pour identifier des gènes fonctionnellement liés. À cet effet, nous avons adopté une méthode de classification non paramétrique basée sur une mesure de proximité. Nous nous sommes intéressés en particulier à développer une mesure de proximité alliant la forme et les valeurs des profils d'expression de gènes. À ce stade de notre travail, nous avons pu conclure que les profils temporels de gènes sont plus différenciés par leur forme que par leurs valeurs d'expression.

D'autre part, nous avons développé une méthode concurrente pour l'identification et la caractérisation de gènes impliqués dans les phases du cycle cellulaire. Les méthodes classiques considèrent un ensemble de gènes de référence censés caractéristiques des phases ou transitions de phase du cycle cellulaire, fixent le nombre de phases ou transitions de phase à considérer et utilisent des distances non fondées. Notre méthode repose sur une classification adaptative non supervisée biologiquement justifiée, basée sur une mesure incluant la proximité liée aux valeurs et aux formes des profils de gènes. Cette procédure nous a permis d'abord d'identifier les phases du cycle cellulaire des gènes étudiés, de tirer des hypothèses sur de nouveaux gènes non connus et enfin proposer un nouvel ensemble de gènes de référence validé par une connaissance biologique publiée. En outre, notre méthode nous a permis de proposer une structuration de contiguïté de classes qui pourraient être d'un grand intérêt pour la conception de réseaux moléculaires.

Notre travail futur concernera l'élargissement de la classification adaptative sur d'autres expériences pour l'élaboration de réseaux de régulation moléculaire, et appliquer la démarche proposée sur des profils de non cycliques. Les

perspectives de ce travail se situent essentiellement sur la représentation synthétique des résultats obtenus en relation avec le caractère continu du phénomène biologique étudié et sur l'intégration de la connaissance sur les fonctions des gènes (bases de données Bio) : interprétations biologiques des classes et des gènes aux frontières, hypothèses sur la fonction de gènes inconnus. Il serait aussi intéressant d'affiner la démarche pour la reconnaissance de l'ensemble des phases et transitions de phase.

Bibliographie

[1] Abdelghani Bellaachia et. al. *E-CAST : A Data Mining Algorithm for Gene Expression Data.* Proc. BIOKDD02 : Workshop on Data Mining in Bioinformatics (With SIGKDD02 conference) pp. 49-54, (2002).

[2] Ahdesmäki, M., Lähdesmäki, H., Pearson, R., Huttunen, H., and Yli-Harja, O. *Robust detection of periodic time series measured from biological systems* BMC Bioinformatics, vol. 6, article 117, pp. 118, (2005)

[3] Androulakis, I.P., Yang, E., Almon, R.R. *Analysis of Time-Series Gene Expression Data : Methodes, Challenges and Opportunities.* Annu. Rev. Biomed. Eng. 9 :3.13.24 (2007)

[4] Balasubramaniyan, R., Hüllermeier, E., Weskamp, N., Kämper, J. *Clustering of gene expression data using a local shape-based similarity measure* Bioinformatics 21(7) :1069-1077, (2005)

[5] Bar-Joseph ,Z., Gerber,G.K., Gifford,D.K., Jaakkola T.S., Simon,I. *Continuous Representations of Time-Series Gene Expression Data.* JOURNAL OF COMPUTATIONAL BIOLOGY Volume 10, Numbers 34, © Mary Ann Liebert, Inc.Pp. 341356 (2003)

[6] Bar-Joseph, Z., Siegfried, Z., Brandeis, M., Brors, B., Lu, Y., Eils, R., Dynlacht, B.D., Simon I. *Genome-wide transcriptional analysis of the human cell cycle identifies genes differentially regulated in normal and cancer cells.* PNAS, vol. 105, n°. 3, pp. 955-960, (2008).

[7] Ben-Dor, A., Shamir, R., and Yakhini, Z. *Clustering gene expression patterns.* Journal of Computational Biology, 6(3/4) :281297, (1999).

[8] Brazma A. and Vilo J. *Gene Expression Data.* FEBS Letterss 480, 17-24 (2000).

[9] Brehelin, L. *Une Approche Bayésienne pour la Classification de Cinétiques d'Expression de Gènes.* JOBIM'04 : Journées Ouvertes Biologie, Informatique, Mathématiques, Montreal, Canada (2004)

[10] Caiado, J., Crato, N., Pena, D. *A periodogram-based metric for time series classification.* Computational Statistics and Data Analysis 50, 26682684 (2006)

[11] Calinski, T., Harabasz, J. *A dendrite method for cluster analysis.* Comm Statist, 3 :1-27, 1974.

[12] Chaiboonchoe, A. , Samarasinghe,S. and Kulasiri,D. *Using emergent clustering methods to analyse short time series gene expression data from childhood leukemia treated with glucocorticoids.* 18th World IMACS / MODSIM Congress, Cairns, Australia 13-17 July (2009)

[13] Chen, G. and Dai, Y. *A New Distance Measurement for Clustering Time-Course Gene Expression Data.* Proceedings of the 26th Annual International Conference of the IEEE EMBS San Francisco, CA, USA. September 1-5, 2004.

[14] Cheng, C. and Li, L.M. *Systematic identification of cell cycle regulated transcription factors from microarray time series data* BMC Genomics, 9 :116, (2008).

[15] Cho, R.J., Campbell, M.J., Winzeler, E.A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T.G., Gabrielian, A.E., Landsman, D., Lockhart, D.J., Davis, R.W. *A genome-wide transcriptional analysis of the mitotic cell cycle.* Mol Cell. Jul ;2(1) :65-73, (1998)

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

[17] Cooper, G. *La cellule une approche moléculaire. Chapitre 14 : Le cycle cellulaire.* Eds. De Boeck Université, Paris, Bruxelles, 561-597(1999).

[18] Côte, P., Hogues, H., Whiteway, M. *Transcriptional analysis of the Candida albicans cell cycle* Molecular Biology of the Cell, vol. 20, pp. 3363-3373, (2009).

[19] Das, R., Kalita, J.K. and Bhattacharyya, D.K. *An effective dissimilarity measure for clustering gene expression time series data.* BIOT, Colorado, 2007.

[20] Das, R., Kalita, J.K. and Bhattacharyya, D.K. *A New Approach for clustering gene expression time series data.* Int. J. Bioinformatics Research and Applications, Vol. 5, no. 3 (2009).

[21] De Hoon, M.J.L.,Imota, S., Kobayashi, K., Ogasawara, N., Miyano, S. *Inferring Gene Regulatory Networks From Time-Ordered Gene Expression Data Of Bacillus Subtilis Using Differential Equations.* Pacific Symposium on Biocomputing 8 :17-28(2003)

[22] Do, J.H., Choi, D.K. *Clustering Approaches To Identifying Gene Expression Patterns from DNA Microarray Data.* Molecules and cells, Vol. 25, No. 2, pp. 1-10, (2008)

[23] Dortet-Bernadet, J.L. and Wicker, N. *Model-based clustering on the unit sphere with an illustration using gene expression profiles* Biostatistics 9, 1, pp. 66-80 (2008).

[24] Dzeroski, S., Slavkov, I., Gjorgjioski, V. and Struyf, J. *Analysis of Time Series Data with Predictive Clustering Trees.* ECML/PKDD06 Workshop Notes, pages : 47-58 (2006).

[25] Efron, B. *The Jackknife, the Bootstrap, and Other Resampling Plans.* CBMS-NSF Regional Conference Series in Applied Mathematics, 38, (1982)

[26] Eisen, M.B., Spellman, P.L., Brown, P.O., Botstein, D. *Cluster analysis and display of genome-wide expression patterns.* Proc. Natl. Acad. Sci. USA. Vol. 95, pp. 14863-14868 (1998).

[27] Ernst , J., Nau, G., Bar-Joseph, Z. *Clustering short time series gene expression data.* Bioinformatics 21, i159i168 (2005).

[28] Fernandez, E.A., Balzarini, M. *Improving cluster visualization in self-organizing in gene expression data analysis.* Computers in Biology and Medicine 37, 1677-1689 (2007)

[29] Fujita, A., Sato,J.R., Rodrigues,L.O., Ferreira,C.E., Sogayar, M.C. *Evaluating different methods of microarray data normalization.* BMC Bioinformatics 7 : 469, 2006.

[30] Garcia-Escudero, L.A., Gordaliza, A. *A proposal for robust curve clustering.* Journal of Classification 22, 185201 (2005)

[31] Gasch, A.P. et al. *Genomic expression programs in the response of yeast cells to environmentals changes.* Mol Biol Cell 11 : 4241-4257, (2000)

[32] Gerstein M, Jansen R. *The current excitement in bioinformatics-analysis of whole-genome expression data : how does it relate to protein structure and function ?* Curr Opin Struct Biol Oct ;10(5) :574-84 (2000)

[33] Ghouila, A.,Yahia, S.B., Malouche, D., Jmel, H., Laouini D., Guerfali, F.Z., Abdelhak, S. *Application of Multi-SOM clustering approach to macrophage gene expression analysis.* Infect. Genet. Evol.(2008).

[34] Golub, T.R. et al. *Molecular classification of cancer : class discovery and class prediction by gene expression monitoring.* Science 286 : 531-537, (1999).

[35] Gress, T.M., Hoheisel J.D., Lennon G.G., et al. *Hybridization fingerprinting of high-density cDNA-library arrays with cDNA pools derived from whole tissues.* Mamm Genome ; 3 :609-19, 1992.

[36] Hardin, J., Mitani, A., Hicks, L., and VanKoten, B. *A robust measure of correlation between two genes on a microarray.* BMC Bioinformatics 8, 220 (2007).

[37] Hartuv, E. and Shamir, R. *A clustering algorithm based on graph connectivity.* Information Processing Letters, 76(46) :175181, (2000).

[38] Heckman, N.E., Zamar, R.H. *Comparing the shapes of regression functions.* Biometrika 22, 135144 (2000)

[39] Herrero, J., Valencia, A., Dopazo, J. *A hierachical unsupervised growing neural network for clustering gene expression patterns.* Bioinformatics Vol. 17, no. 2, Pages 126-136 (2001).

[40] Hestilow, T.J. and Huang, Y. *Clustering of Gene Expression Data Based on Shape Similarity* Journal on Bioinformatics and Systems Biology, Volume 2009, Article ID 195712, 12 pages (2009).

[41] Heyer, L., Kruglyak, S., Yooseph, S. *Exploring expression data : Identification and analysis of coexpressed genes.* Genome Research 9, 11061115 (1999).

[42] Hubert, L., Arabie P. *Comparing partitions.* J. of Classification, pages 193218, (1985).

[43] Hughes, T.R., Marton, M.J, Jones, A.R, Roberts, C.J, Stoughton, R., Armour, C.D., Bennett, H.A., Coffey, E., Dai, H., He, Y.D., Kidd, M.J., King, A.M., Meyer, M.R., Slade, D., Lum, P.Y., Stepaniants, S.B., Shoemaker, D.D., Gachotte, D., Chakraburtty, K., Simon, J., Bard, M., Friend, S.H. *Functional discovery via a compendium of expression profiles.* Cell 102 : 1, 109-26, Jul, (2000).

[44] Iyer, V.R. et al. *The transcriptional program in the response of human fibroblasts to serum.* Science 283 : 83-87, (1999).

[45] Ji, X., Li-Ling, J., Sun, Z. *Mining gene expression data using a novel approach based on hidden Markov models.* FEBS 542, 125-131 (2003)

[46] Jiang, D., Pei, J., Zhang, A. *DHC : a density-based hierarchical clustering method for time series gene expression data.* Bioinformatics and Bioengineering, On page(s) : 393- 400. Proceedings. Third IEEE Symposium, (2003)

[47] Jiang, D.,Tang, C.,Zhang, A. *Cluster Analysis for Expression Data : A Survey.* IEEE Transactions on Knowledge and Data Endineering, vol. 16, no. 11, pp. 1370-1386,(2004).

[48] Jordan B.R. *Large-scale expression measurement by hybridization methods : from highdensity membranes to "DNA chips".* J. Biochemistry 124 : 251-258, 1998.

[49] Kakizawa, Y., Shumway,R.H., Taniguchi, N. *Discrimination and clustering for multivariate time series.* Journal of the American Statistical Association 93, 328340 (1998)

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

[51] Keller, K., Wittfeld, K. *Distances of time series components by means of symbolic dynamics.* International Journal of Bifurcation Chaos 14, 693704 (2004)

[52] Kim, K., Zhang, S., Jiang, K., Cai, L., Lee, I.-B., Feldman, L.J., Huang, H. *Measuring similarities between gene expression profiles through new data transformation.* BMC Bioinformatics 8, 29. (2007)

[53] Kohonen T. *Self-Organization and Associative Memory.* Spring-Verlag, Berlin (1984).

[54] Kooperberg,C.,Fazzio,T.G., Delrow, J.J., Toshio, T. *Improved Background Correction for Spotted DNA Microarrays.* Journal of Computational Biology, 9(1), 5566, 2002.

[55] Kustra, R., Zagdanski, A. *Data-Fusion in Clustering Microarray Data : Balancing Discovery and Interpretability.* IEEE/ACM Transactions on Computational Biology and Bioinformatics, 15 Nov. 2007.

[56] Laurie J.H., Semyon K., Shibu Y. *Exploring Expression Data : Identification and Analysis of Coexpression Genes.* Spring Harbor Laboratory Press. Pages 1106-1115 (1999).

[57] Leng, X. and Müller, H. *Classification using functional data analysis for temporal gene expression data.* Bioinformatics, vol.22, no.1, pages 68-76, (2006).

[58] Li, G.G., Wang, Z.Z. *Evaluation of similarity measures for gene expression data and their correspondent combined measures.* Interdisciplinary Sciences : Computational Life Sciences 1 : 72-80 (2009).

[59] Liang, F., Wang, N. *Dynamic agglomerative clustering of gene expression profiles* Pattern Recognition Leters 28, pp. 1062-1076, (2007).

[60] 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).

[61] Liu, E.T. *Classification of cancers by expression profiling.* Curr Opin Genet Dev. Feb ; 13(1) :97-103, 2003.

[62] Luan, Y. and Li, H. *Clustering of time-course gene expression datausing a mixed-effects model with B-splines.* Bioinformatics, Vol.19 no. 4, pages 474482 (2003).

[63] Lobenhofer, E.K., Bennett,L., Cable,P., Li,L., Bushel,P.R., and Afshari, C.A. *Regulation of DNA Replication Fork Genes by 17β-Estradiol* Molecular Endocrinology 16 (6) : 1215-1229, (2002).

[64] Ma, P., Castillo-Davis, C.I., Zhong, W., and Liu, J.S. *A data-driven clustering method for time course gene expression data.* Nucleic Acids Research, Vol. 34, No. 4 12611269, 2006.

[65] Magni, P., Ferrazzi, F., Sacchi, L., Bellazzi, R. *TimeClust : a clustering tool for gene expression time series.* Bioinformatics Advance Access published December 6, (2007).

[66] Maharaj, E.A. *Cluster of time series.* Journal of Classification 17, 297314 (2000)

[67] MCQueen J.B. *Some methods of classification and analysis of multivariate observations.* In Proc. of 5th Berkeley Symp. on Mathematical Statistics and Probability, pages 281-297 (1967).

[68] Möller-Levet, C.S., Klawonn, F., Cho, K.H., and Wolkenhauer, O. *Fuzzy clustering of short time series and unevenly distributed sampling points.* In : BertholdMR, Lenz HJ, Bradley E, Kruse R, Borgelt C(eds) : Advances in Intelligent Data Analysis. Springer, Berlin Heidelberg, pp 330-340 (2003).

[69] Möller-Levet, C. S., and Yin, H. *Modeling and analysis of gene expression time-series based on co-Expression.* International Journal of Neural Systems 15(4),311-322, (2005).

[70] Nettleton,D., Recknor,J., and Reecy, J.M. *Identification of differentially expressed gene categories in microarray studies using nonparametric multivariate analysis* Bioinformatics, Vol. 24 no. 2, pages 192201, (2008).

[71] Nguyen, C., Rocha,D., Granjeaud S., Baldit,M., Bernard,K., Naquet,P. and Jordan, B. R. *Differential gene expression in the murine thymus assayed by quantitative hybridization of arrayed cDNA clones.* Genomics 29 :207-216, (1995).

[72] Novak, B.A., Jain, A.N. *Pathway recognition and augmentation by computational analysis of microarray expression data.* Bioinformatics. Jan 15 ; 22(2) :233-41, (2006).

[73] Oates, T., Firoiou, L., Cohen, P.R. *Clustering time series with Hidden Markov Models and Dynamic Time Warping.* In : Proc. 6th IJCAI 1999, Workshop on Neural, Symbolic and Reinforcement Learning Methods for Sequence Learning, Stockholm, pp. 1721 (1999).

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

[75] Park,T., Yi, S.G, Kang,S.H., Lee,S.Y., Lee,Y.S. and Simon,R. *Evaluation of normalization methods for microarray data.* BMC Bioinformatics 4 :33, 2003.

[76] Pauling, L., Corey, R. B., Branson, H. R. *The structure of proteins : two hydrogen-bonded helical configurations of the polypeptide chain.* Proc. Natl. Acad. Sci. USA 37, 205-211, 1951.

[77] Pietu, G., Alibert, O., Guichard, V., Lamy, B., Bois, F., Leroy, E., Mariage-Sampson, R., Houlgatte, R., Soularue, P., and Auffray, C. *Novel gene transcripts preferentially expressed in human muscles revealed by quantitative hybridization of a high density cDNA array.* Genome Res 6, 492-503, 1996.

[78] Peddada, S.D., Lobenhofer, E.K., Li, L., Afshari, C.A., Weinberg, C.R., Umbach, D.M. *Gene selection and clustering for time-course and dose-response microarray experiments using order-restricted inference.* Bioinformatics 19, 834841, (2003).

[79] Pollard,T.D. et al. *Biologie Cellulaire : le cycle cellulaire.* Eds. Elseveir SAS, Paris, 721-836,(2004).

[80] Pramila, T., Miles, S., GuhaThakurta, D., Jemilo, D., Breeden, L. L. *Conserved homeodomain proteins interact with MADS box protein Mcm1 to restrict ECB-dependent transcription to the M/G1 phase of the cell cycle.* Genes Dev. 16 :30343045, (2002).

[81] Ramasso, E., Rombaut, M., Pellerin, D. *Modèles de Markov Cachés Crédibilistes en présence d'informations non distinctes.* LFA 2008, Lens : France (2008).

[82] Ramoni, M.F., Sebastiani, P., Kohane, I.S. *Cluster analysis of gene expression dynamics.* Proc. Natl. Acad. Sci. USA 99(14) :912126, (2002).

[83] Rousseeuw P.J. *Silhouettes : a graphical aid to the interpretation and validation of cluster analysis.* J. Comp App. Math, Vol. 20, pp. 53-65, (1987).

[84] Sankoff, D., Kruskal, J.B. *Time warps, string edits, and macromolecules : the theory and practice of sequence comparaison.* Addison-Wesley, Reading (1983).

[85] Schena, M., Renu A.H., Thomas P.T., Konrad K., Lachenmeier E., Ronald W.D. *Microarrays : biotechnology's discovery platform for functional genomics.* Trends in Biotechnology, Volume 16, Issue 7, 1 July , Pages 301-306, (1998).

[86] Schliep, A.,Schönhuth, A.,Steinhoff, C. *Using hidden Markov models to analyze gene expression time course.* Bioformatics, 19 (Suppl. 1), i255-i263, (2003).

[87] Serban, N., Wasserman, L. *CATS : Cluster After Transformation and Smoothing.* Journal of the American Statistical Association 100, 990999 (2004)

[88] Shamir R. and Sharan R. *Click : A clustering algorithm for gene expression analysis.* In In Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB 00). AAAI Press., (2000).

[89] Sharan, R., Maron-Katz, A., Shamir, R. *CLICK and EXPANDER : a system for clustering and visualizing gene expression data.* Bioinformatics 19 : 1787-99, (2003).

[90] Smyth, G.K., Speed, T. *Normalisation of cDNA Microarray Data.* (2003).

[91] Son, Y.S., Baek, J. *A* modified correlation coefficient based similarity measure for clustering time-course gene expression data. Pattern Recognition Letters archive Volume 29, Issue 3,Pages 232-242,ISSN :0167-8655. Elsevier Science Inc. New York, NY, USA. (2008)

[92] Soularue, P., Gidrol, X. *Puces à ADN* Techniques de l'Ingénieur, (2002).

[93] Southern E. *Detection of specific sequences among DNA fragments separated by gel electrophoresis.* Journal of Molecular Biology, 98 :503-517, (1975).

[94] Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., Futcher, B. *Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization.* Mol. Biol. Cell 9, 32733297 (1998).

[95] Storey, J.D., Xiao, W., Leek, J.T., Tompkins R.G., Davis, R.W. *Significance analysis of time course microarray experiments.* PNAS September 6, vol. 102 , no. 36, 12837-12842, (2005).

[96] Susmita Datta and Somnath Datta *Comparaisons and validation of statistical clustering techniques for microarray gene expression data.* Bioinformatics, vol. 19, n°4, pages 459-466 (2003).

[97] Tamayo P., Solni D., Mesirov J., Zhu Q., Kitareewan S., Dmitrovsky E., Lander E.S. and Golub T.R. *Interpreting patterns of gene expression with self-organizing maps : Methods and application to hematopoietic differentiation.* Proc. Natl. Acad. Sci. USA, Vol. 96(6) :29072912, March (1999).

[98] Tang, C. and Zhang, A.. *An iterative strategy for pattern discovery in high-dimensional data sets.* In Proceeding of 11th in International Conferene on Information and Knowledge Management (IKM 02), McLean, VA, November 4-9 (2002).

[99] Tang, C., Zhang, L., Zhang, A., Ramanathan M. *Interrelated two-way clutering : An unsupervised approach for gene expression data analysis.* In Proceeding of BIBE2001 : 2nd IEEE International Symposium on Bioinformatics and Bioengineering, pages 41-48, Bethesda, Maryland, November 4-5 (2001).

[100] Tavazoie,S.,Hughes J.D., Campbell M.J., Cho, R.J., Church G.M. *Systematic determination of genetic netwrk architecture.* Nature gnetics, vol. 22, (1999).

[101] Tibshirani,R., Walther, G., Hastie, T. *Estimating the numbers of cluster in a dataset via the Gap statistic.* URL : http ://www-stat.stanford.edu/ tibs/ftp/gap.pdf, (2000).

[102] Todorovski,L., Cestnik,B., Kline,M., Lavrac, S., and Dzeroski, S. *Qualitative clustering of short time-series : A case study of firms reputation data.* In ECML/PKDD02 workshop on Integration and Collaboration Aspects of Data Mining, Decision Support and Meta-Learning, pages 141149. Helsinki University Printing House, August (2002).

[103] Tseng, V.S.M. and Chen, Y.L. *An Effective Approach for Mining Time-Series Gene Expression Profile.* Studies in Computational Intelligence, Springer Berlin / Heidelberg, Volume 9, Pages 363-376, (2006).

[104] Van Der Laan, M.J.,Pollard, K.S.,Bryan, J. *A new partitioning around medoids algorithm.* Journal of statistical computation and simulation , vol. 73, no8, pp. 575-584, (2003).

[105] Wen, X., Fuhrman, S., Michaels, G. S., Carr, D. B., Smith, S. Barker, J. L. and Somogyi, R. *Large-scale temporal gene expression mapping of central nervous system development.* Proc. Natl. Acad. Sci. USA 95, pp. 334339, (1998).

[106] Whitfield, M.L., Sherlock, G.,Saldanha, A.J.,Murray J.I.,Ball,C.A., Alexander K.E., Matese J.C., Perou C.M., Hurt,M.M., Brown, P.O., and Botstein, D. *Identification of genes periodically expressed in the human cell cycle and their expression in tumors.* Molecular Biology of the Cell, Vol.13, 1977-2000, (2002).

[107] Wichert, S., Fonkianos, K., and Strimmer,K. *Identifying periodically expressed trascripts in microarry time series data* Bioinformatics, vol. 20, no. 1, pp. 520, (2004)

[108] Wu, F.X., Zhang, W.J., Kusalik, A.J. *Dynamic model-based clustering for time-course gene expression data.* Journal of Bioinformatics and Computational Biology , 3(4) :821-836 (2005).

[109] Wu,W., Xing,E.P., Myers,C., Mian,I.S., and Bissell,M.J. *Evaluation of normalization methods for cDNA microarray data by k-NN classification.* BMC Bioinformatics 6 :191, (2005).

[110] Yang, F.Z., Zhu, Y.Y. *An efficient method for similarity search on quantitative transaction data.* Journal of Computer Research and Development (in Chinese) 41, 361368 (2004).

[111] Yang, Jiong, Wang, Wei, Wang, Haixun and Yu, Philip S. *Capturing Subspace correlation in a Large Data Set.* In Proceeding sof 18th in International Conference on Data Engineering (ICDE 2002), pages 517-528, 2002.

[112] Yang, Y., Dudoit, S., Luu, P., Lin, D., Peng, V., Ngai, J., Speed, T. *Normalization for cDNA microarray data : A robust composite method addressing single and multiple slide systematic variation.* Nucleid Acids Research, vol. 30, (2002).

[113] Yang,Y.H., Xiao,Y., Segal,M.R. *Identifying differentially expressed genes from microarray experiments via statistic synthesis.* Center for Bioinformatics and Molecular Biostatistics July 2, (2004).

[114] Yeung, K.Y., Fraley, C, Murua, A., Raftery, AE., Ruzz WL. *Model-based clustering and data transformations for gene expression data.* Bioinformatics, 17 :977987,(2001).

[115] Yuan,Y., Li,C.T.,Wilson,R. *Partial mixture model for tight clustering of gene expression time-course.* BMC Bioinformatics 9 :287, (2008).

[116] Zeng, Y., Garcias-Frias, J. *A novel HMM-based clustering algorithm for the analysis of gene expression time-course data.* Computational Statistics Data Analysis Volume 50, Issue 9, Pages 2472-2494, 1 May (2006).

[117] Zhu, G., Spellman, P ;T., Volpe, T., Brown, P.O., Botstein, D., Davis, T.N., Futcher, B. *Two yeast forkhead genes regulate the cell cycle and pseudohyphal growth.* Nature 406 :90-94, (2000)

Liste des publications

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. *Pattern Recognition Letters*, volume 31, issue 12, 1 Septembre 2010, pages 1608-1617.

Alpha Diallo, Ahlame Douzal-Chouakria, Francoise Giroud (2009). Comparaison et évaluation de métriques pour la classification de profils d'expression de gènes. *XVIèmes Rencontres de la Société Francophone de Classification* (SFC09).

Alpha Diallo, Ahlame Douzal-Chouakria and Francoise Giroud (2009). Which Distance for the Identification and the Differentiation of Cell-Cycle Expressed Genes ? N. Adams et al. (Eds.) : *Advances in Intelligent Data Analysis VIII* (IDA09), LNCS 5772, pp. 273-284. Springer Berlin Heidelberg.

A. Douzal-Chouakria, A. Diallo, F. Giroud (2009). Adaptive clustering for time series : application for identifying cell cycle expressed genes. *Computational Statistics and Data Analysis* (CSDA) 53 (4), 1414-1426. Elsevier.

A. Diallo, A. Douzal-Chouakria, F.Giroud (2008). Classification adaptative de séries temporelles : application à l'identification des gènes exprimés au cours du cycle cellulaire. *Revue des Nouvelles Technologies de l'Information* (RNTI-E-11, ECG), 487-498, Cépaduès.

Ahlame Douzal Chouakria, A. Diallo, F. Giroud (2007). Adaptive dissimilarity index for Gene Expression Profiles Classification. *In : Selected Contributions in Data Analysis and Classification*, Series : Studies in Classification, Data Analysis, and Knowledge Organization, Brito, P., Bertrand, P., Cucumel, G., De Carvalho, F. (Eds.). XIII, 483-494, Springer Berlin Heidelberg.

A. Douzal Chouakria, A. Diallo and F. Giroud (2007). Adaptive clustering for time series. *International Association for Statistical Computing* (IASC), Statistics for Data Mining, Learning and Knowledge Extraction, Aveiro, Portugal.

A. Douzal Chouakria, A. Diallo and F. Giroud. (2006). Adaptive dissimila-

rity index for gene expression profiles classification. *Integrative Post Genomics Conference*. Lyon, France.

Résumé

La technologie des puces à ADN a rendu aujourd'hui possible de mesurer les niveaux d'expression de milliers de gènes durant des processus biologiques importants. Analyser des profils d'expression de multiples gènes offre la possibilité d'éclairer certains aspects de la génomique fonctionnelle. Ce travail porte sur l'analyse, la classification et l'interprétation de profils d'expressions de gènes durant le processus de division cellulaire. La division cellulaire est le processus biologique de prolifération des cellules qui devient drastiquement aberrant dans le cas de cellules cancéreuses. Tenant compte de la structure temporelle des données d'expression, nous avons étudié trois familles de mesures de proximités. La première famille définit des mesures limitées à la comparaison des valeurs des expressions en ignorant la contrainte de dépendance temporelle des données. La seconde famille se limite à la comparaison des formes des expressions. Enfin, la troisième famille de mesures couvre simultanément les aspects formes et valeurs. Une formalisation unifiée de ces mesures est proposée. Une classification adaptative de milliers de gènes est appliquée afin d'apprendre la mesure de proximité à considérer pour l'identification et la caractérisation de gènes impliqués dans les phases du cycle cellulaire.

Mots-clés: analyse de données temporelles, classification, analyse de profils d'expression de gènes, transcriptome.

Abstract

DNA microarray technology allows us to monitor the expression levels of thousands of genes simultaneously during important biological processes. Analyzing the expression profiles of multiple genes provides the opportunity to clarify certain aspects of functional genomics. This work focuses on the analysis, clustering and interpretation of gene expression patterns during cell proliferation, this process insuring the multiplication of cells, which is drastically aberrant in cancer cells. Taking into account the temporal dependency of expression data, we studied three families of proximity measures. The first family defines limited measures to compare genes expression values by ignoring the constraint of time-dependent data. The second family is limited to compar genes expression behaviors. Finally, the third group of measures covers both genes expression temporal behavior and values. A unified formalization of these measures is proposed. An adaptive clustering of thousands of genes is used to learn the proximity measure to be considered for the identification and characterization of genes involved in cell cycle phases.

Keywords: Temporal data analysis, clustering, analysis of gene expression profiles, transcriptome.