

Analyse Statistique de Données de Puces à ADN

D. Dembélé*

IGBMC, 1 rue Laurent Fries, BP 10142
67404 Illkirch Cedex, France
Doulaye.Dembele@titus.u-strasbg.fr

Résumé

La technologie dite de puces à ADN permet d'analyser les fonctions de milliers de gènes à la fois. Les données issues des puces à ADN sont organisées dans un tableau où les lignes correspondent aux gènes (des milliers) alors que les colonnes correspondent aux différentes expériences (des dizaines) réalisées. Chaque cellule du tableau correspond au ratio ou au niveau d'expression d'un gène dans une expérience.

Dans ce papier nous présentons de façon synthétique les méthodes statistiques utilisées pour analyser ces données. Nous présentons d'abord les méthodes basées sur des tests d'hypothèse puis les méthodes basées sur la statistique exploratoire. Le lecteur ou la lectrice intéressé(e) trouvera plus de détails sur les applications des puces à ADN et sur les méthodes présentées dans les références citées.

1 Introduction

Toutes les cellules vivantes contiennent des chromosomes qui sont des grandes pièces d'Acide DésoxyriboNucléique (ADN). L'ADN contient des centaines ou des milliers de gènes et tout système biologique est contrôlé par un nombre déterminé de gènes (près de 30 000 pour le système humain). Les gènes sont codés pour donner des protéines qui accomplissent les fonctions cellulaires. Le gène est d'abord transcrit en Acide RiboNucléique messenger (ARNm) qui va produire la protéine son par expression. Les protéines constituent le cheval de bataille des molécules de la cellule et sont responsables par exemple de la structure et de la reproduction cellulaire, elles produisent aussi l'énergie et des biomolécules comme l'ADN. À première vue, toutes les cellules d'un même organisme ont le même nombre de chromosomes et donc contiennent le même répertoire de protéines. Cependant les cellules de tissus différents (oeil, cheveux, ...) ont des propriétés différentes. En général les niveaux d'expression de l'ARNm reflètent l'abondance des protéines correspondantes dans la cellule. Il existe un lien logique entre l'état d'une cellule, les détails de ses protéines et la composition de l'ARNm. Les perturbations dans l'environnement cellulaire par des facteurs tels que la radiation, la mutation, conduisent à l'altération de l'expression d'un groupe spécifique de gènes.

Le but de la génomique fonctionnelle est d'expliquer les technologies permettant d'identifier parmi un grand nombre de gènes, les quelques uns qui sont associés à un changement moléculaire pour un prototype défini. L'identification de ces gènes peut nous aider à mieux diagnostiquer une maladie, à identifier les voies thérapeutiques d'intervention ou simplement à comprendre l'origine d'un phénomène biologique donné. Les puces à ADN qui sont utilisées pour mesurer simultanément l'expression de milliers de gènes dans une population cellulaire constituent un outil de base de la génomique fonctionnelle. Ces puces ont été utilisées pour trouver des gènes malades [5, 8, 15] ou pour classer des gènes tumoraux [5, 17].

*Ce travail a été fait avec le soutien de l'INSERM, du CNRS, de l'Hopital Universitaire de Strasbourg et du Centre National de Recherche en Génétique. Il a aussi bénéficié d'un soutien financier du Groupement d'Intérêt Public – Hoechst-Marion-Roussel

Une puce à ADN est une membrane de nylon ou une lame de verre chimiquement traitée et sur laquelle sont déposés des ADN complémentaires (ADNc) ou des oligonucléotides de séquences connues. Les oligonucléotides peuvent être aussi directement synthétisés sur la lame (c'est le cas par exemple des puces Affymetrix). Après hybridation de la puce avec des sondes spécifiques d'échantillons biologiques, suivie d'une quantification de chacun des spots de l'image scannée, voir [29, 30], nous obtenons des données caractérisant les niveaux d'expression des gènes ou transcription qui est la première étape vers la fonction biologique d'un gène.

Une étude complète peut nécessiter plusieurs puces. Par exemple pour comparer les gènes de trois tissus sains à ceux de quatre tissus malades, il faut au moins sept puces. Les données générées par les puces peuvent être organisées dans un tableau où les lignes correspondent aux gènes (des milliers) alors que les colonnes correspondent aux différentes expériences (des dizaines) représentant le nombre de puces de l'étude, voir tableau 1. Les données du tableau 1 sont obtenues après des ajustements

TAB. 1 – Données générées par les puces à ADN. Le nombre réel x_{ij} est le ratio (puce par dépôt) ou le signal (puce Affymetrix) de l'expression du gène i dans l'expérience j .

	exp_1	exp_2	exp_3	\dots	exp_p
$gene_1$	x_{11}	x_{12}	x_{13}	\dots	x_{1p}
$gene_2$	x_{21}	x_{22}	x_{23}	\dots	x_{2p}
$gene_3$	x_{31}	x_{32}	x_{33}	\dots	x_{3p}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
$gene_n$	x_{n1}	x_{n2}	x_{n3}	\dots	x_{np}

(normalisations intra-lame et inter-lame) des mesures brutes. Ces normalisations, [7, 21, 2, 32, 34, 26], non traitées ici permettent d'avoir des valeurs cohérentes pour les mesures d'une lame (expérience) d'une part, et d'autre part d'avoir des valeurs comparables pour les mesures des différentes expériences. Nous supposons aussi que les données ont été transformées, voir [45], afin de satisfaire les conditions d'utilisabilité des tests statistiques présentées par la suite.

Les expériences de puces à ADN sont réalisées pour répondre à des questions posées. Par exemple il peut s'agir de retrouver les gènes résistants à un traitement ou de retrouver les gènes permettant de diagnostiquer une maladie. Pour répondre à ce type de questions à partir des données d'expression, on utilise des méthodes statistiques qui font l'objet de ce papier. Nous faisons une présentation synthétique des méthodes statistiques qui sont utilisées pour analyser les données issues de puces à ADN. Deux familles de méthodes sont utilisées : la première basée sur des tests d'hypothèses (paramétrique ou non) permet entre autre de trouver des gènes qui sont exprimés différemment dans des groupes d'expériences, la deuxième famille de méthodes peut être utilisée pour réduire les dimensions des données ou pour regrouper les gènes et/ou les expériences.

La suite de ce papier est organisé comme suit. Nous présentons dans le paragraphe 2, les méthodes basées sur des tests d'hypothèses. Ces méthodes sont utilisées pour sélectionner les gènes qui sont analysés avec les méthodes de type statistique exploratoire que nous présentons dans le paragraphe 3. Dans le paragraphe 4, une conclusion est donnée.

2 Méthodes inférentielles

Il y a deux types de méthodes basées sur les tests d'hypothèses : paramétriques et non paramétriques. La première catégorie de méthodes utilise les données mesurées alors que la seconde catégorie de méthodes utilise indirectement les mesures.

Considérons une étude composée de deux groupes d'expériences, le premier constitué d'expériences sur des tissus sains et le second constitué d'expériences sur des tissus malades. Il est intéressant de

savoir s'il y a une différence significative entre les niveaux moyens d'expression d'un gène dans les deux groupes. Pour cela, on utilise le test de Student ou le test de Wilcoxon. Quand on a plus de deux groupes d'expériences (on considère par exemple plusieurs stades de développement d'une maladie), il est possible d'utiliser toutes les combinaisons deux à deux des groupes avec le test de Student pour retrouver les gènes ayant une différence significative entre les niveaux moyens d'expression. Au lieu de cela, on utilise une analyse de variance (ANOVA = ANalysis Of VAriance) ou un test de Kruskal-Wallis. Ces différents tests sont résumés dans les paragraphes suivants.

2.1 Tests paramétriques

Ces tests utilisent les mesures réellement relevées (valeurs x_{ij} du tableau 1). Le test de Student, encore appelé *test-T*, est utilisé quand on a deux groupes d'expériences et l'on cherche à savoir s'il y a une différence significative entre les niveaux moyens d'expression de ceux-ci.

2.1.1 Test de Student

Pour un gène, on calcule la moyenne et l'écart type des valeurs d'expression des expériences de chaque groupe. Les deux moyennes et les deux écarts types associés sont utilisés pour calculer la statistique T [20]. Puis une table statistique est utilisée pour décider l'acceptation de l'hypothèse nulle qui suppose l'égalité entre les deux moyennes ou le rejet de celle-ci si l'écart entre les moyennes des deux groupes est jugé important. Le rejet ou l'acceptation de l'hypothèse se fait avec une erreur dont le seuil est fixé *a priori*. Cette erreur est de type 1 par opposition à l'erreur que l'on commet en acceptant l'hypothèse nulle alors cela ne devrait pas être le cas. Dans ce dernier cas l'erreur est dite de type 2.

Par exemple si les données du tableau 1 sont issues de deux groupes d'expériences, le test de Student peut être fait pour chaque gène. Cela permet de retrouver les gènes qui se comportent différemment dans les deux groupes de tissus, rejet de l'hypothèse nulle. Nous supposons une cohérence entre les mesures relevées dans chaque groupe.

2.1.2 ANOVA

L'analyse ANOVA est une extension du test de Student au cas de données comportant plus de deux groupes. L'hypothèse nulle est acceptée si tous les groupes ont des valeurs moyennes égales à une erreur près. Il suffit qu'un groupe ait une expression moyenne différente des autres pour justifier le rejet de l'hypothèse nulle au profit de l'hypothèse alternative. Une description détaillée de l'analyse ANOVA est donnée dans [20].

2.2 Tests non paramétriques

Les tests non paramétriques utilisent indirectement les données mesurées, ils sont conseillés quand le nombre d'expériences par groupe est faible. Ces méthodes sont insensibles aux mesures aberrantes. Une mesure est jugée aberrante si sa valeur est très grande ou très faible comparée aux valeurs des autres mesures du même groupe.

2.2.1 Test de Wilcoxon

Ce test est l'équivalent non paramétrique du test de Student. Il consiste à ordonner toutes les valeurs des mesures par ordre croissant indépendamment du groupe auquel elles appartiennent (on note cependant le groupe de chaque mesure) puis à calculer la somme des rangs des mesures de chaque groupe. Intuitivement, les deux sommes calculées seront proches si les deux groupes proviennent de la même population (acceptation de l'hypothèse nulle). Si toutes les valeurs des mesures d'un groupe ont tendance à être plus grandes ou plus petites que celles de l'autre groupe, l'écart entre les deux sommes

sera important et il y aura peu de chance que les deux groupes proviennent de la même population (rejet de l'hypothèse nulle au profit de l'hypothèse alternative). Au lieu du test de Wilcoxon, on peut utiliser le test de Mann-Whitney avec lequel on obtient des résultats similaires.

2.2.2 Test de Kruskal-Wallis

Ce test est l'équivalent non paramétrique de l'analyse ANOVA à un facteur. Il consiste, comme pour le test de Wilcoxon, à ordonner toutes les valeurs des mesures par ordre croissant indépendamment du groupe auquel elles appartiennent puis à calculer la somme des rangs des mesures de chaque groupe. Ces sommes sont utilisées pour calculer une statistique H qui suit une loi de χ^2 avec un nombre de degré de liberté égal au nombre de groupe diminué de un [20]. Une table statistique est enfin utilisée pour prendre la décision sur l'acceptation de l'hypothèse nulle ou de son rejet.

Une méthode inférentielle peut être suffisante, dans certains cas celle-ci doit être complétée avec une méthode exploratoire.

3 Méthodes exploratoires

On peut diviser ces méthodes en deux catégories : les méthodes factorielles ou géométriques et les méthodes de classification.

3.1 Méthodes factorielles ou géométriques

Comme leur nom l'indique, les méthodes géométriques traitent l'aspect "représentation graphique" des données. Cela consiste à trouver un espace de dimension réduit (1 ou 2) dans lequel les données sont projetées pour leur visualisation et leur interprétation. La méthode la plus utilisée est l'Analyse en Composantes Principales (ACP).

3.1.1 Analyse en Composantes Principales et autres méthodes voisines

L'ACP permet de réduire la dimension des données. L'outil utilisé pour cela est aussi appelé SVD (Singular Value Decomposition) par les mathématiciens ou expansion de Karhunen-Loève par les traiteurs d'images. Soit X la matrice des données du tableau 1. L'ACP permet de mettre X sous la forme de produit de trois matrices : $X = UDV^T$, où D est une matrice diagonale contenant des valeurs réelles non négatives et ordonnées par ordre décroissant alors que U et V sont des matrices orthogonales. On peut montrer que $X = d_1 u_1 v_1^T + d_2 u_2 v_2^T + \dots + d_p u_p v_p^T$, où d_i est l'élément de la i^{eme} diagonale de la matrice D alors que u_i et v_i sont respectivement les vecteurs correspondants aux i^{eme} colonnes des matrices U et V . Dans cette décomposition de X en une somme de matrices de rang un chacun, le poids des termes est décroissant quand l'indice i croît. Ainsi, la somme peut être tronquée sans une grande perte (erreur $\leq 5\%$) dans la reconstitution des données initiales à partir de D , U et V . Par exemple les deux premières composantes peuvent capturer 90% de l'information contenue dans les données initiales. L'ACP est appliquée à des données de microarray dans [36, 19, 1, 46].

Les colonnes de la matrice U jouent le rôle des expériences et sont appelées "expériences propres" alors que les lignes de la matrice V jouent le rôle des gènes et sont appelées "gènes propres". Ici le mot "propre" est emprunté au nom de la décomposition (en valeurs propres) utilisée pour obtenir D , U et V . Il n'y a cependant pas de lien direct entre les composantes calculées avec les gènes et les expériences. Pour obtenir un lien direct, on utilise l'analyse de correspondance [25]. Notons aussi que l'outil de décomposition en valeurs propres utilisé par l'ACP recherche les combinaisons linéaires des variables initiales (gènes) qui maximisent la variance des données. Cela peut être insuffisant si la réduction de la dimension dans le sens du carré moyen n'est pas nécessaire. Il peut s'agir par exemple de trouver toute autre information non contenue dans la corrélation, pour cela on utilise la

méthode PP (Projection Pursuit). Cette dernière méthode, voir [14, 23, 33], permet de rechercher les projections intéressantes des données. Cela se fait en visitant toutes les projections pour retenir celle qui est intéressante. Précisons cependant qu'avec la méthode PP, on a pas de solution analytique comme c'est le cas avec l'ACP.

Au lieu de vouloir réduire la dimension des données, on peut aussi rechercher directement les liens qui existent entre elles. Cela est fait par les méthodes de classification. Dans un problème de classification on peut considérer deux situations : les classes ou groupes sont connus *a priori*, c'est-à-dire, nous disposons d'informations biologiques relatives à des gènes spécifiques et on voudrait répartir tous les gènes entre des groupes définis. Cela correspond à une *classification supervisée*. Pour la seconde situation, les groupes et le plus souvent leur nombre sont inconnus, le problème revient alors à déterminer les groupes avec les gènes qu'ils contiennent. Ceci correspond à une *classification non supervisée* ou regroupement automatique ou clustering.

3.2 Méthodes de classification supervisée

Les méthodes de classification supervisée permettent de classer les gènes ou les expériences dans des groupes de profils définis. Elles procèdent en deux étapes : apprentissage et reconnaissance. Pendant l'étape d'apprentissage, une partie des données est utilisée pour permettre à l'algorithme de reconnaître les caractéristiques de chaque classe puis, cette information est utilisée pour prédire la classe des autres données. Pour l'apprentissage, nous pouvons définir des profils idéaux puis calculer le coefficient de corrélation avec les gènes, voir [17] où une application à la discrimination de gènes de données tumorales est faite. La méthode SVM (Support Vector Machine) est utilisée dans [6] pour classer les gènes de données d'expression issues de microarray. La méthode des k plus proches voisins ou k-NN (k-Nearest Neighbor) peut être utilisée aussi.

3.3 Méthodes de regroupement automatique ou clustering

Les méthodes de clustering peuvent être divisées en deux grandes familles : les méthodes hiérarchiques et les méthodes de partitionnement. Soulignons que l'utilisation de certaines méthodes de regroupement sur des données conduit toujours à la formation de groupes même si naturellement les données n'en contiennent pas. Nous supposons par la suite que les données traitées contiennent naturellement des groupes, nous supposons aussi que le problème consiste à regrouper les gènes. Mais les méthodes présentées s'appliquent également au cas où l'on désire regrouper les expériences.

3.3.1 Méthodes hiérarchiques

Il y a deux types de méthodes hiérarchiques : ascendante et descendante. La méthode hiérarchique ascendante est la plus utilisée et procède comme suit. Initialement, chaque gène forme un groupe et une mesure d'aggrégation de deux groupes est choisie. Une nouvelle repartition est construite en réunissant les deux groupes les plus voisins au sens de la mesure choisie. Cette étape d'aggrégation est répétée jusqu'à ce que tous les gènes forment un seul groupe. En fonction de la façon dont la distance qui sépare deux groupes est calculée, on distingue plusieurs variétés d'algorithmes pour la méthode hiérarchique ascendante [18, 12, 22, 40]. La méthode hiérarchique descendante utilise l'opération inverse. C'est-à-dire, initialement tous les gènes forment un seul groupe. Un groupe est ensuite divisé en deux, puis l'opération de division est répétée jusqu'à ce que le nombre total des groupes soit égal au nombre des gènes.

Les résultats des méthodes hiérarchiques sont représentés dans des arbres appelés dendrogrammes où chaque branche donne un niveau de similarité pour les valeurs d'expression des gènes sur celle-ci. Pour plus de détails sur ces méthodes, voir [18, 12, 22, 40, 11]. Un avantage est que ces méthodes ne nécessitent pas de connaître le nombre de groupes dans les données.

Dans les méthodes hiérarchiques, une fois qu'un gène est placé dans un groupe, il reste dans celui-ci jusqu'à la fin de l'algorithme.

3.3.2 Méthodes de partitionnement

Les méthodes de partitionnement consistent à trouver la meilleure repartition de n gènes en K groupes de manière à ce qu'un critère, par exemple l'inertie totale des groupes, soit optimale. Ceci est un problème combinatoire bien posé. Une solution exhaustive consiste à tester toutes les combinaisons des n gènes en K groupes et à retenir celle qui optimise le critère choisi. Le nombre total des partitions possibles est $\approx \frac{K^n}{K!}$. Par exemple pour $n = 100$ et $K = 5$, on a $\sim 10^{67}$ partitions. Ce nombre est très grand, en effet avec un ordinateur pouvant faire 10^9 opérations (calcul du critère et comparaison) à la seconde, il faut $\approx 3 * 10^{50}$ années de travail! Ceci est irréaliste et connu sous la dénomination de problème NP -difficile en recherche opérationnelle. Dans la pratique, pour rechercher la partition solution, on utilise un algorithme itératif qui procède comme suit. À partir d'une solution acceptable et réalisable, l'améliorer par itérations successives, c'est-à-dire, changer la classe d'appartenance de certains gènes, jusqu'à obtenir une solution dans laquelle aucun gène ne change de classe d'une itération à la suivante. Dans les méthodes de partitionnement, on distingue les méthodes paramétriques et non paramétriques. Pour les méthodes paramétriques, on suppose connu *a priori* la distribution de probabilité des groupes puis on utilise l'approche Bayésienne pour affecter les gènes dans les groupes. Cela est fait par maximisation d'une fonction de vraisemblance [42, 46, 28]. La distribution de probabilité des groupes est généralement inconnue. C'est pour cela que l'on utilise souvent les méthodes non paramétriques. La méthode des cartes topologiques de Kohonen ou Self-Organizing Maps (SOM) est un algorithme neuronal non supervisé qui consiste à trouver les vecteurs représentant les gènes et réalisant dans le même temps un regroupement de l'espace d'entrée en un certain nombre de neurones (ou nœuds ou groupes) de dimension prédéfini. Pour des détails sur cette méthode, voir [24, 38, 41]. Une autre méthode très utilisée est l'algorithme K-Means.

L'algorithme K-Means : c'est une amélioration d'un algorithme proposé par Forgy [13]. Initialement les données sont réparties (aléatoirement ou à partir des résultats d'une méthode hiérarchique) en K groupes. Puis nous recherchons itérativement la répartition locale qui optimise le critère choisi. Pour cela alterner le calcul des centres de gravité des groupes et la répartition des gènes autour des centres de gravité jusqu'à stabilisation des calculs, c'est-à-dire, jusqu'à ce que les centres de gravité ne soient plus modifiés d'une itération à la suivante. Le critère choisi peut consister à maximiser les distances entre les K groupes ou à minimiser les distances entre gènes d'un même groupe. L'utilisation de cette méthode avec des données de microarray est faite dans [39].

L'algorithme Fuzzy C-Means : avec l'algorithme K-Means chaque gène appartient à seul un groupe. Au lieu de cela, nous pouvons attribuer à chaque gène un degré d'appartenance relativement aux K groupes. Ceci est réalisé avec la méthode FCM (Fuzzy C-Means) qui affecte à chaque gène une valeur, comprise entre 0 et 1, d'appartenance à chacun des K groupes. Ainsi si le degré d'appartenance d'un gène est proche de 1 pour une groupe, la chance que le dit gène appartienne à ce groupe est élevée. Inversement si ce degré est voisin de 0, il y a peu de chance que le gène appartienne au groupe indexé. Enfin, chaque gène est affecté au groupe avec lequel il possède le plus grand degré d'appartenance. On peut aussi affecter chaque gène au groupe dont le degré d'appartenance est supérieur à un seuil choisi. Des études ont montré que la méthode FCM est plus performante que la méthode K-Means [10]. L'utilisation de cette méthode avec des données de microarray est faite dans [10, 35, 9]. Le choix du paramètre qui contrôle le degré d'appartenance est discuté [9].

Sur le choix du nombre K de groupes. Le choix de K a fait l'objet d'une abondance littérature [31, 4, 16, 44]. Beaucoup de procédures proposées sont sensibles à une sous ou sur-estimation de

K . En plus il faut faire le regroupement pour différentes valeurs de K avant choisir la valeur de K satisfaisante. Pour pallier ce problème, une méthode de type hiérarchique descendante est proposée dans [43]. D'autres méthodes de regroupement sans connaissance *a priori* de K sont données dans [3, 37, 27].

4 Conclusion

Dans ce papier une présentation synthétique des méthodes statistiques utilisées pour analyser les données de microarray est faite. Le dépouillement de ce type de données fait intervenir les méthodes basées sur des tests d'hypothèses alors que les analyses avancées sont faites avec des méthodes de statistique exploratoire. L'analyse peut être poursuivie par une recherche dans des banques de données biologiques sur les gènes d'intérêt sélectionnés.

La multitude des méthodes utilisées pour analyser les données de puces à ADN s'explique d'une part, par la complexité du problème et d'autre part, par l'absence d'une démarche unifiée acceptée partout et donnant de bons résultats. Cela est également dû à la relative jeunesse de la technologie de microarray qui évolue encore. On assiste de plus en plus à l'apparition de puces contenant tous les gènes d'un génome ou des puces très spécialisées ne contenant que des gènes d'un intérêt particulier, e.g., ceux impliqués dans une maladie précise. Les outils d'analyse futurs vont être plus performants et faciles d'utilisation pour les non spécialistes de la statistique.

Références

- [1] O. Alter, P. O. Brown, and D. Bostein. Singular Value Decomposition for Genome-Wide Expression Data Processing and Modeling. *PNAS*, 97(18) :10101–10106, August 2000.
- [2] K. A. Baggerly, K. R. Coombes, K. R. Hess, D. N. Stivers, L. V. Abruzzo, and W. Zhang. Identifying Differentially Expressed Genes in cDNA Microarray Experiments. *Journal of Computational Biology*, 8(6) :639–659, 2001.
- [3] A. Ben-Dor, R. Shamir, and Z. Yakhini. Clustering Gene Expression Patterns. *Journal of Computation Biology*, 6(3/4) :281–297, 1999.
- [4] J. Bezdek. *Pattern Recognition With Fuzzy Objective Function Algorithms*. New York : Plenum Press, 1981.
- [5] M. Bittner, P. Meltzer, Y. Chen, Y. Jiang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon, Z. Yakhini, A. Ben-Dor, N. Sampas, E. Dougherty, E. Wang, F. Marincola, C. Gooden, J. Lueders, A. Glatfelter, P. Pollock, J. Carpten, E. Gillanders, D. Leja, K. Dietrich, C. Baudry, M. Berens, D. Alberts, V. Sondak, N. Hayward, and J. Trent. Molecular Classification of Cutaneous Malignant Melanoma by Gene Expression Profiling. *Nature*, 406 :536–540, 3 August 2000.
- [6] M. P. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. A. Jr, and D. Haussler. Knowledge-Based Analysis of Microarray Gene Expression Data by Using Support Vector Machines. *PNAS*, 97(1) :262–267, January 2000.
- [7] Y. Chen, E. R. Dougherty, and M. L. Bittner. Ratio-Based Decisions and the Quantitative Analysis of cDNA Microarray Images. *Journal of Biometrical Optics*, 2(4), October 1997.
- [8] E. Clark, T. Golub, E. Lander, and R. Hynes. Genomics Analysis of Metastasis Reveals an Essential Role for RhoC. *Nature*, 406 :532–535, 2000.
- [9] D. Dembélé and P. Kastner. Fuzzy C-Means Method for Clustering Microarrays Data. *Bioinformatics*, 19(8) :973–980, 2003.
- [10] E. R. Dougherty, J. Barrera, M. Brun, S. Kim, R. M. Cesar, Y. Chen, M. Bittner, and J. M. Trent. Inference From Clustering with Application to Gene-Expression Microarrays. *J of Computational Biology*, 9(1) :105–126, 2002.

- [11] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster Analysis and Display of Genome-Wide Expression Patterns. *PNAS*, 95 :14863–14868, December 1998.
- [12] B. S. Everitt. *Cluster Analysis*. Arnold, London, 3rd edition, 1993.
- [13] E. Forgy. Cluster Analysis for Multivariate Data : Efficiency versus Interpretability of Classification. In *Proc. of Biometric Society Meeting, Riverside, California Abstract in Biometrics, v21, n3, pp.768*, page 768, 1965.
- [14] J. H. Friedman. Exploratory Projection Pursuit. *Journal of the American Statistical Society*, 82(397) :249–266, March 1987.
- [15] G. Fuller, C. Rhee, K. Hess, L. Caskey, R. Wang, J. Bruner, W. Yung, and W. Zhang. Reactivation of Insulin-Like Growth Factor Binding Protein 2 Expression in Glioblastoma Multiform : A Revelation by Parallel Gene Expression Profiling. *Cancer Research*, 59 :4228–4232, 1999.
- [16] I. Gath and A. Geva. Unsupervised Optimal Fuzzy Clustering. *IEEE Pattern Analysis and Machine Intelligence*, 11(7) :773–781, July 1989.
- [17] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri, C. Bloomfield, and E. Lander. Molecular Classification of Cancer : Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, 286 :531–537, 1999.
- [18] J. Hartigan. *Clustering Algorithms*. Wiley, New York, 1975.
- [19] N. S. Holter, M. Mitra, A. Maritan, M. Cieplak, J. R. Banavar, and N. V. Fedoroff. Fundamental Patterns Underlying Gene Expression Profiles : Simplicity from Complexity. *PNAS*, 97(15) :8409–8414, July 2000.
- [20] D. C. Howell. *Méthodes Statistique en Science Humaine*. De Boeck Université Ed., 1999.
- [21] T. Ideker, V. Thorsson, A. F. Siegel, and L. E. Hood. Testing for Differentially-Expressed Genes by Maximum-Likelihood Analysis of Microarray Data. *Journal of Computational Biology*, 7(6) :805–817, 2000.
- [22] J. Jain and R. Dubes. *Algorithms for Clustering Data*. Prentice Hall, Englewood Cliff, New Jersey, 1988.
- [23] M. Jones and R. Sibson. What is Exploratory Projection Pursuit (with discussion). *Journal of the Royal Statistical Society, Serie A*, 150 :1–36, 1987.
- [24] T. Kohonen. *Self-Organizing Maps*. Springer, Berlin, 1997.
- [25] L. Lebart, A. Morineau, and M. Piron. *Statistique Exploratoire Multidimensionnelle*. Dunod, 3e Edition, 2000.
- [26] C. Li and W. H. Wong. Model-Based Analysis of Oligonucleotide Array : Expression Index Computation and Outlier Detection. *Proc Natl Acad Sci, USA*, 98(1) :31–36, 2001.
- [27] A. V. Lukashin and R. Fuchs. Analysis of Temporal Gene Expression Profiles : Clustering by Simulated Annealing and Determining the Optimal Number of Clusters. *Bioinformatics*, 17(5) :405–414, 2001.
- [28] G. J. McLachlan, R. W. Bean, and D. Peel. A Mixture Model-Based Approach to the Clustering of Microarray Expression Data. *Bioinformatics*, 18(3) :413–422, 2002.
- [29] Microarray. The Chipping Forecast. *Nature Genetics*, 21(supplement) :1–60, 1999.
- [30] Microarray. The Chipping Forecast II. *Nature Genetics*, 32(supplement) :461–552, 2002.
- [31] G. W. Milligan and M. C. Cooper. An Examination of Procedures for Determining the Number of Clusters in a Data Set. *Psychometrika*, 50(2) :159–179, June 1985.
- [32] M. Newton, C. Kendzioriski, C. Richmond, F. Blattner, and K. Tsui. On Differentially Variability of Expression Ratios : Improving Statistical Inference About Gene Expression Changes from Microarray Data. *Journal of Computational Biology*, 8(1) :37–52, 2001.

- [33] C. Posse. Projection Pursuit Exploratory Data Analysis. *Computational Statistics and Data Analysis*, 20 :669–687, 1995.
- [34] J. Quackenbush. Computational Analysis of Microarray Data. *Nature*, 2 :418–427, June 2001.
- [35] W. Raffelsberger, D. Dembélé, M. G. Neubauer, M. M. Gottardis, and H. Gronemeyer. Quality Indicators Increase the Reliability of Microarray Data. *Genomics*, 80(4) :385–394, October 2002.
- [36] S. Raychaudhuri, J. M. Stuart, and R. B. Altman. Principal Components Analysis to Summarize Microarray Experiments : Application to Sporulation Time Series. In *Pacific Symposium on Biocomputing, Honolulu, Hawaiï*, volume 5, pages 452–463, 2000.
- [37] R. Sharan and R. Shamir. CLICK : A Clustering Algorithm with Application to Gene Expression Analysis. In *AAAI - ISMB*, pages 307–316, 2000.
- [38] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub. Interpreting Patterns of Gene Expression with Self-Organizing Maps : Methods and Application to Hematopoietic Differentiation. *PNAS Genetics*, 96 :2907–2912, March 1999.
- [39] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. I. Cho, and G. M. Church. Systematic Determination of Genetic Network Architecture. *Nature Genetic*, 22 :281–285, July 1999.
- [40] S. Theodoridis and K. Kouthroumbas. *Pattern Recognition*. Academic Press, New York, 1999.
- [41] P. Törönen, M. Kolehmainen, G. Wong, and E. Castrén. Analysis of Gene Expression Data Using Self-Organizing Maps. *FEBS Letters*, 451 :142–146, 1999.
- [42] A. Webb. *Statistical Pattern Recognition*. John Wiley & Sons, Ltd, 2002.
- [43] N. Wicker, D. Dembélé, W. Raffelsberger, and O. Poch. Density of Points Clustering, Application to Transcriptomic Data Analysis. *Nucleic Acids Research*, 30(18) :3992–4000, 2002.
- [44] X. Xie and G. Beni. A Validity Measure for Fuzzy Clustering. *IEEE trans. PAMI*, 13(8) :841–847, August 1991.
- [45] K. Y. Yeung, C. Fraley, A. Murua, A. Raftery, and W. L. Ruzzo. Model-Based Clustering and Data Transformation for Genes Expression Data. *Bioinformatics*, 17(10) :977–987, 2001.
- [46] K. Y. Yeung and W. L. Ruzzo. Principal Component Analysis for Clustering Gene Expression Data. *Bioinformatics*, 17(9) :763–774, 2001.