THÈSE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ JOSEPH FOURIER

Discipline : Génie Biologique et Médical

Présentée et soutenue publiquement

par

MARTIN Olivier

Le 18 décembre 2002

APPROCHES STATISTIQUES POUR L'ANALYSE DE DONNÉES DE PUCES À ADN

Directeurs de thèse : CELEUX Gilles, TRILLING Laurent

M. ANTONIADIS Anestis	, Président
M. ROBIN Stéphane	, Rapporteur
M. RAFTERY Adrian	, Rapporteur
M. BARBRY Pascal	, Examinateur
M. TRILLING Laurent	, Directeur de thèse
M. CELEUX Gilles	, co-Directeur de thèse

Thèse préparée au sein du projet IS2 de l'Institut National de Recherche en Informatique et en Automatique (INRIA)

Remerciements

Cette thèse est le fruit de travaux menés au sein du projet IS2 de l'INRIA Rhône-Alpes. J'ai eu la chance d'y bénéficier d'un encadrement enrichissant et dynamique que j'ai longtemps cherché, et qui m'a permis de réaliser ce travail. Cette thèse est avant tout le reflet de nombreuses contributions, et je tiens à remercier ici les personnes qui ont participé à son élaboration.

Laurent Trilling, qui a accepté de diriger cette thèse, m'a laissé la plus grande liberté et toute sa confiance pour orienter et définir mes axes de recherche. Je lui suis extrêmement reconnaissant de son soutien et de l'appui qu'il m'a apporté à de nombreuses occasions.

J'exprime ma profonde gratitude à Gilles Celeux pour m'avoir accueilli et encadré au sein du projet IS2. Sa rigueur scientifique et son souci de perfectionnement m'ont fortement impressionné. Sa disponibilité et son intérêt pour mes travaux ont été une source importante de motivation et de persévérance dans les moments de doute. Il m'a transmis son souci de développer des méthodologies cohérentes et adaptées pour les problématiques biologiques abordées, et je suis persuadé que je tirerai encore durant de nombreuses années des bénéfices de son enseignement.

À Christian Lavergne, j'adresse mes sincères remerciements pour son aide indispensable ainsi que ses conseils qui m'ont permis d'orienter certains de mes travaux. Nos discussions m'ont beaucoup appris et le stage de biostatistique que j'ai encadré sous sa direction constitue une expérience précieuse.

Mes remerciements s'adressent également à Pascal Barbry. Les approches méthodologiques proposées sont directement issues de notre collaboration et de sa volonté de développer des approches adaptées pour l'analyse du transcriptome.

Concernant ce mémoire, je suis extrêmement flatté que Stéphane Robin et Adrian Raftery aient accepté d'en être les rapporteurs. Leurs remarques m'ont rassuré sur l'intérêt de ce travail et m'ont permis d'apporter plusieurs améliorations.

Et je tiens à remercier Anestis Antoniadis de m'avoir fait l'honneur de présider le jury de thèse.

Mais la thèse étant une expérience bien particulière, elle n'aurait jamais abouti sans le soutien et la compréhension des amis. Aussi, je dis un grand merci :

à Franck, Yann et Philippe, les "crevards" du bureau D117, pour la complicité et les bons moments que nous avons partagés,

à Guillaume pour m'avoir chaleureusement accueilli chez lui durant les derniers mois et à tous les membres d'IS2 pour leur bonne humeur,

à Claude pour son initiation à l'escalade, et pour tous ces moments formidables passés sur la paroi

à Florence pour m'avoir supporter durant son stage de biostatistique et pour son aide précieuse,

à Sylvain, Alexandre et Aurélien pour toutes les soirées avignonnaises que nous avons passées ensemble,

Je remercie du fond du coeur Stéphanie pour sa patience et l'amour qu'elle m'a témoigné durant la dernière année. J'adresse également un immense merci à ma soeur pour ses conseils et pour m'avoir soutenu dans les moments difficiles. Enfin, les valeurs que m'ont transmises mes parents et leurs encouragements sont les piliers de cette thèse : je ne pourrai jamais assez les remercier pour les sacrifices consentis durant toutes ces années d'étude.

À mes parents et ma soeur,

Table des matières

Table des figures

Liste des tableaux

Introduction

Du génome à la génomique fonctionnelle

La cellule, qui est l'unité élémentaire de tout organisme vivant, a pour destinée de croître et de se diviser pour donner naissance à des cellules filles. La croissance et la multiplication des cellules nécessitent de produire quantité de molécules. Parmi celles-ci, les protéines sont formées par l'association en longues chaînes de petites molécules appelées acides aminés. Le bon fonctionnement cellulaire est assuré par la présence d'enzymes qui déterminent quelles réactions chimiques doivent avoir lieu dans la cellule. L'intérêt des protéines s'est accru lorsque l'on a pu mettre en évidence la nature protéique des enzymes.

À chaque protéine correspond un nombre variable d'acides aminés arrangés en une séquence particulière. Ainsi, la cellule doit disposer de molécules pouvant stocker l'ordre des acides aminés des protéines. En fait, les chromosomes, qui portent les gènes contrôlant l'hérédité, contiennent également l'information nécessaire pour la fabrication des protéines. Ce n'est qu'après avoir établi la relation entre un gène et une protéine que la **génomique** a connu un essor important. Cette nouvelle démarche consiste à étudier complètement et systématiquement des génomes. Elle vise à connaître le nombre et la position des gènes sur les chromosomes, leur séquence et les produits cellulaires qui leur correspondent. Ainsi, de nombreux génomes d'organismes procaryotes et même des génomes eucaryotes ont été entièrement séquencés. Le séquençage a pour objectif de délimiter et de répertorier des fragments d'ADN. Mais le séquençage de l'ADN ne représente qu'une partie du travail nécessaire à la compréhension du génome. L'objectif ultime est de déterminer la fonction de ces fragments ; c'est la génomique fonctionnelle. La connaissance du code génétique peut renseigner en partie sur les fonctions de brins d'ADN, mais elle ne peut le faire précisément et systématiquement dans tous les cas. La génomique fonctionnelle et la “post-génomique” cherchent à déterminer quelles sont les gènes qui sont actifs à un moment donné pour permettre à un tissu cellulaire de se développer et de s'adapter à des conditions environnementales particulières.

Pendant de nombreuses années, les biologistes moléculaires ne pouvaient étudier simultanément qu'un nombre très faible de gènes. Grâce aux progrès biotechnologiques et au développement de la bioinformatique, il est maintenant possible d'accéder à de nombreuses informations sur l'ensemble des gènes d'un organisme. Ces techniques permettent de s'intéresser à plusieurs centaines ou plusieurs milliers de gènes et de protéines. Elles concernent trois domaines. D'une part, l'analyse du transcriptome permet d'étudier les différences et les variations du niveau d'expression des gènes. D'autre part, l'analyse du protéome permet d'identifier les protéines présentes dans une cellule. Enfin, la génomique structurale vise à déterminer la structure des protéines ; afin de prédire des interactions protéines-protéines. Les deux premiers domaines connaissent un essor important en raison des progrès technologiques et produisent de grandes quantités de données issues d'expériences biologiques. C'est l'analyse et l'interprétation de ces résultats qui permettront de mieux comprendre la fonction des gènes au sein d'un organisme. Le thème abordé dans cette thèse est l'étude du transcriptome.

Objectif du travail : étude du transcriptome

Connaissant l'ensemble ou une partie des gènes d'un organisme, différentes questions se posent :

- Quelle est la fonction d'un gène et quel est son rôle dans les processus cellulaires ?

- Comment un gène est-il régulé et comment les gènes et leurs produits interagissent ? Quels sont les réseaux d'interactions géniques ?

- Quelles sont les différences des niveaux d'expression pour différents types cellulaires et quels sont les changements d'expression pour différentes conditions ou différents traitements ?

La connaissance du niveau d'expression des gènes est une information primordiale pour espérer répondre à ces questions. Bien que l'ARNm ne soit pas le produit final des gènes, la transcription est la première étape dans la régulation génique. La relation entre les quantités d'ARNm et de protéines présentes dans la cellule n'est pas simple, mais l'absence d'ARNm implique nécessairement un niveau faible de protéines. Les niveaux d'expression des gènes constitue donc une information importante pour comprendre la régulation génique. Les puces à ADN (ou biopuces) sont des outils technologiques permettant de mesurer le niveau d'expression de plusieurs milliers de gènes simultanément. En mesurant le niveau d'expression d'un gène dans différents tissus au cours du développement cellulaire ou dans différentes conditions, on peut construire son *profil d'expression* qui donne une indication sur sa fonction dans le génome. On observe trois problématiques liées à l'étude de ces données.

- *La normalisation*
  Les données de puces à ADN sont le résultat d'une analyse d'image ainsi que d'un long protocole expérimental. En plus de la nature biologique des expériences, il existe de nombreux facteurs conduisant à une variabilité importante des mesures du niveau d'expression des gènes. Ces différents facteurs sont peu connus et rendent impossible une interprétation directe des résultats [97]. De nature différente, ces derniers introduisent

un biais dans l'estimation des niveaux d'expression. L'objectif de la normalisation est de corriger les observations afin de supprimer les effets de ces facteurs sur les mesures expérimentales.

- *L'identification de gènes exprimés de façon différentielle*
  Il s'agit ici d'identifier quels sont les gènes qui sont impliqués dans un processus cellulaire. De cette façon, il devrait être possible de déterminer les agents d'un réseau génétique ou de connaître les gènes dont l'expression est modifiée dans des tissus cancéreux par exemple.

- *L'étude de profils d'expression*
  À partir de différentes conditions, il est possible d'établir un profil d'expression de l'ensemble des gènes du génome. Par exemple, on peut étudier les variations d'expression des gènes suite à un stress cellulaire ou la réponse de l'organisme à un agoniste chimique. L'un des atouts des études de profils d'expression est qu'elles devraient permettre de proposer certaines caractéristiques fonctionnelles des gènes.

Dans cette thèse, nous visons à identifier les difficultés de l'étude des puces à ADN, et à proposer des solutions méthodologiques qui surmontent ces difficultés.

Plan de thèse

Le premier chapitre de cette thèse est consacré à des éléments de biologie. Il permet d'introduire les notions élémentaires de la biologie moléculaire. Bien que notre objectif soit l'étude du transcriptome, nous donnons quelques éléments sur la traduction, c'est-à-dire la phase suivant la transcription qui assure la synthèse des protéines. Nous souhaitons ainsi donner un large aperçu des principes biologiques. En fait, la fonction d'un gène dépend du rôle de la protéine au sein de l'activité cellulaire. De manière naturelle, l'étude de l'expression des protéines devrait être associée à celle des gènes. Une vue d'ensemble de ces mécanismes permet ainsi de mieux situer le cadre de nos travaux.
Dans ce même chapitre, nous donnons le principe de fonctionnement des biopuces. Quelles que soient les technologies considérées, il s'agit de mesurer le niveau de transcription de plusieurs milliers de gènes. Celles-ci reposent invariablement sur le principe fondamental de la biologie moléculaire : l'hybridation de deux brins complémentaires d'ADN simple brin. Nous détaillons plus particulièrement le cas de sondes déposées sur lames de verre et de cibles marquées par deux fluorophores : les cyanines Cy5 et Cy3. Ce choix tient à la nature des données expérimentales que nous analysons par la suite.

Le second chapitre présente les problèmes liés aux données brutes et l'étape de normalisation qu'ils impliquent. Pour cela, on applique en général une correction sur l'ensemble des valeurs qui consiste à se ramener à un a priori biologique [129]. Les trois chapitres suivants permettent d'aborder deux questions récurrentes posées par les biologistes : identifier des gènes exprimés différemment entre deux conditions expérimentales et étudier des profils d'expression de gènes (les profils d'expression donnent les variations d'expression des gènes dans différentes conditions par rapport à un état de référence). L'axe principal de cette thèse est de proposer des méthodes pour ces deux problèmes lorsque la variabilité des données d'expression est importante.

Dans un premier temps, nous décrivons des méthodes permettant d'identifier des gènes exprimés de façon différente entre deux conditions. Nous présentons les tests multiples pour ce problème ainsi que le contrôle de différents type d'erreur dans ce cadre. Suivant le modèle proposé par Newton *et al.* [91], nous envisageons une approche permettant d'identifier les gènes différentiellement en l'absence de mesures répétées. Les résultats obtenus sur des données expérimentales montrent que la grande variabilité des mesures nécessite de disposer de répétitions [78]. L'analyse de variance est alors une approche naturelle qui permet d'étudier la variabilité des données [69]. Elle offre la possibilité d'introduire différents facteurs pour la modélisation des niveaux d'expression des gènes. Un prolongement naturel de l'analyse de variance consiste à utiliser différents effets aléatoires dans la modélisation pour permettre une meilleure décomposition de la variance et représenter des structures de corrélation dans les données [126]. De tels modèles sont appelés des modèles mixtes [112]. Nous nous limitons au cas des modèles linéaires mixtes ou L2M.

Le quatrième chapitre est consacré à l'étude des profils d'expression [96]. D'un point de vue statistique, il s'agit généralement de résoudre des problèmes de discrimination ou de classification. Nous présentons un état de l'art sur ces méthodes en détaillant plus particulièrement la classification. Mais ces approches ne permettent pas de prendre en compte les problèmes de variabilité évoqués précédemment, et sont donc mal adaptées à notre domaine d'application. En utilisant l'information des mesures répétées, nous introduisons dans le cinquième chapitre une approche nouvelle pour la classification des profils d'expression qui permet de bien cerner les problèmes de variabilité.

Ainsi, le dernier chapitre aborde la classification de données répétées (ou répliquées) [86]. De part les remarques précédentes, les expériences de biopuces requièrent la mise en place de plans d'expériences de réplications des données en raison des problèmes de variabilité. Afin de prendre en compte les problèmes de variabilité dans la classification, il est nécessaire d'utiliser l'information des répétitions dans l'étude des profils. L'approche que nous proposons consiste à classer les unités statistiques (par exemple les gènes) avec l'ensemble des mesures répétées. Plus exactement, nous proposons un modèle de mélange pour des données répétées. Ce modèle prend en compte la variabilité des données par des modèles linéaires mixtes associés à chaque composant du mélange. Les simulations et l'application illustrent la capacité de notre modèle à prendre en compte efficacement la variabilité des données.

Notre démarche

Nous avons commencé par mettre clairement en évidence les problèmes de variabilité des données issues de biopuces. À partir de ces observations, notre objectif était de comprendre comment et dans quelle mesure les données de puces à ADN doivent être étudiées. Notre principale contribution tient à l'approche proposée pour la classification de données répétées. Cette méthodologie, fondée sur un modèle de mélange de modèles linéaires mixtes, est d'une portée générale dépassant le cadre de l'étude des profils d'expression. Mais, il est important de souligner que d'après les résultats obtenus pour la recherche de gènes exprimés différemment, les modèles mixtes sont bien adaptés pour la représentation des niveaux d'expression car ils permettent une bonne décomposition de la variance. Ainsi, les différents modèles de mélange de modèles mixtes que nous proposons s'avèrent bien adaptés à notre problème de classification de données répétées.

Chapitre 1

Transcriptome et puces à ADN

Le but de ce chapitre est de présenter les principes fondamentaux de la génétique ainsi que ceux des biopuces. L'idée n'est pas de faire un cours de biologie ou un exposé sur les techniques liées aux puces à ADN. L'intention de ce premier chapitre est plutôt de fournir quelques notions sur les deux sujets, d'abord pour comprendre le lien entre ces deux thèmes, ensuite pour introduire le vocabulaire qui va être utilisé et enfin pour présenter les problèmes liés au traitement des données de biopuces.

Dans un premier temps, nous donnons une introduction sur la biologie moléculaire et plus particulièrement sur la génétique. Nous décrivons les principes fondamentaux qui permettent aux cellules d'un organisme vivant de se diviser, de réagir à différentes conditions environnementales. Nous détaillons le dogme central de la biologie moléculaire qui repose sur la transmission de l'information génétique grâce aux étapes successives de transcription et de traduction. Sur ces aspects, des ouvrages plus complets peuvent être trouvés [21], [124].

Dans la deuxième et dernière partie, nous nous intéressons aux techniques qui permettent d'étudier le transcriptome, c'est-à-dire la quantité de gènes transcrits dans la cellule. S'il existe depuis longtemps des méthodes pour étudier les niveaux de transcription de quelques gènes, le développement des biotechnologies pour analyser l'ensemble des gènes d'un génome a permis une avancée extraordinaire dans le domaine de la biologie. Nous insistons plus particulièrement sur la technologie des puces à ADN avec marquage fluorescent [107], car c'est sur ce type de données que nous développons des approches statistiques pour l'étude du transcriptome.

1.1 Aspects biologiques

Depuis longtemps déjà, les agronomes et les biologistes ont remarqué que de nombreux caractères phénotypiques étaient observés lors de croisement de plantes ou d'animaux. Lorsqu'il a été admis qu'il y avait transmission d'une information chez les êtres vivants, il a fallu identifier sous quelle forme celle-ci était stockée ainsi que les mécanismes de l'hérédité. Il a été établi que pour transmettre cette information, appelée information génétique[1], il devait exister un matériel génétique (défini au paragraphe 1.1.3) possédant plusieurs propriétés.

Sachant que la cellule est l'unité élémentaire de tout organisme vivant et que la plupart des cellules sont amenées à se reproduire, celles-ci doivent maintenir intactes leurs caractéristiques fonctionnelles. Or, ces caractéristiques découlant justement de l'information génétique, il doit nécessairement exister des mécanismes permettant sa duplication et sa transmission lors de la reproduction cellulaire.

Aussi, comme les agents fonctionnels des organismes sont principalement les protéines, la cellule doit être capable de les synthétiser suivant ses besoins. Ainsi, lors de la duplication cellulaire, les cellules filles doivent hériter du plan d'élaboration de ces protéines.

Enfin, la survie et l'évolution des espèces montrent que cette information doit pouvoir être modifiée pour évoluer.

1.1.1 La cellule : unité élémentaire de tout organisme vivant

La plus petite partie de tout organisme vivant est la cellule. Elle est trop petite pour être observée à l'oeil nu mais suffisamment importante pour un microscope électronique. D'une façon générale, ce sont de très petites entités dont le diamètre est bien inférieur à 1mm. Chaque cellule est en fait un système complexe composé de plusieurs parties regroupées à l'intérieur d'une membrane appelée la membrane plasmique. Celle-ci constitue une barrière imperméable qui permet le passage de molécules nutritives et empêche le contenu de la cellule de se répandre à l'extérieur. L'intégrité de cette membrane est donc essentielle à la vie de la cellule. D'un point de vue cellulaire, on distingue des organismes unicellulaires (une seule cellule) et pluricellulaires. Par exemple, les bactéries sont des organismes unicellulaires capables de survivre et de se reproduire dans un environnement approprié.

Les composants de la cellule permettent de distinguer deux types d'organismes vivants : les procaryotes et les eucaryotes. Les bactéries sont des procaryotes (unicellulaires) tandis que l'homme, la souris ou le maïs sont des eucaryotes. Si cette distinction est importante, c'est que les mécanismes qui permettent aux organismes de vivre sont assez différents et généralement plus complexes chez les eucaryotes. Les procaryotes sont unicellulaires et plus petits que les eucaryotes. A l'intérieur des eucaryotes, on trouve un corps sphérique délimité par une membrane et entouré par le cytoplasme : c'est le noyau. Nous verrons que l'information génétique est contenu au sein du noyau chez les eucaryotes alors qu'elle est libre au sein des cellules procaryotes.

Une des caractéristiques essentielles de la plupart des cellules (eucaryotes ou procaryotes) est de croître et de se diviser pour donner naissance à des cellules filles. Pour remplir ces fonctions, la cellule dispose de quantité de molécules différentes. Leurs transformations sont l'un des fondements du bon fonctionnement de la cellule. On distingue suivant leur taille deux types de molécules : les petites molécules et les macromolécules. Avant de les décrire, citons les acides aminés pour les petites molécules, et les protéines et les acides nucléiques

[1] Considérons pour le moment que l'information génétique est l'information de caractères héréditaires. La notion de gène sera introduite plus tard.

qui forment la plus grande partie des macromolécules. Les macromolécules sont invariablement formées par l'association en longues chaînes de petites molécules (par exemple, les protéines sont composées d'acides aminés). Parmi les fonctions associées à ces molécules, les plus importantes permettent la dégradation de molécules nutritives en unités élémentaires facilement utilisables, le stockage de l'énergie provenant de cette dégradation, la synthèse de petites molécules pour la construction de macromolécules et l'assemblage de ces petites molécules.

Afin de bien comprendre la fonction et la place de ces molécules dans la vie cellulaire, nous présentons dans la suite certaines propriétés des protéines (macromolécules) qui correspondent comme nous venons de le préciser à l'association de petites molécules (les acides aminés). Pour finir cette présentation biologique, nous décrivons les macromolécules de type acides nucléiques pour expliquer comment l'information génétique présente dans l'ADN (acide désoxyribonucléique) est transcrite en ARN (acide ribonucléique) lors de la transcription, puis traduite en protéine lors de la traduction.

1.1.2 Les agents fonctionnels de la cellule : les protéines

Les protéines sont des polymères composés de petites molécules (monomères) qui sont associées en longues chaînes appelées acides aminés. L'intérêt lié aux protéines s'est accru lorsque l'on a reconnu le caractère protéique des enzymes dont on connaissait le rôle déterminant dans un grand nombre de réactions chimiques au sein de la cellule. Les enzymes sont des molécules dont les sous-unités, les acides aminés, sont liées par des liaisons peptidiques afin de former de longues chaînes linéaires, communément appelées chaînes polypeptidiques. Lorsque l'on a découvert qu'il existait 20 acides aminés, il a été émis l'hypothèse que chaque chaîne polypeptidique pouvait être caractérisée par une séquence unique d'acides aminés. Cette hypothèse fut confortée avec la découverte de la séquence des deux chaînes polypeptidiques de l'insuline. Le tableau 1.1 donne le nom des vingt acides aminés qui composent les chaînes polypeptidiques.

Un acide aminé est invariablement constitué d'une fonction amine NH_2, un groupement acide COOH (acide carboxylique), un atome d'hydrogène et une chaîne latérale R représentant un groupement organique. Le tout est lié comme le montre la figure 1.1 à un atome de carbone, appelé carbone α. Les propriétés d'un acide aminé, comme ses caractères hydrophobe ou acide, dépendent de la nature de sa chaîne latérale R. L'asymétrie entre les liaisons peptidiques implique une orientation sur la chaîne polypeptidique. Elle commence par le groupe amine (N terminal) vers le groupe carboxyl (C terminal).

Une fois la chaîne polypeptidique formée, plusieurs forces de nature chimique permettent à la protéine d'adopter une configuration particulière dans l'espace. Cette structure est à la fois très compacte et spécifique à la protéine en question. À quelques exceptions près, on peut considérer comme univoque la relation entre la chaîne d'acides aminés et la configuration spatiale d'une protéine. Cette structure permet l'existence de cavités dans lesquelles viennent se loger des substrats ou d'autres chaînes polypeptidiques selon un modèle clef-serrure qui définit ainsi le caractère fonctionnel de la protéine. Bien que la majorité des protéines soient

Acide aminé	Abréviation en trois lettres	Code à une lettre
Glycine	Gly	G
Alanine	Ala	A
Valine	Val	V
Isoleucine	Ile	I
Leucine	Leu	L
Sérine	Ser	S
Thréonine	Thr	T
Proline	Prol	P
Acide aspartique	Asp	D
Acide glutamique	Glu	E
Lysine	Lys	K
Arginine	Arg	R
Asparagine	Asn	N
Glutamine	Gln	Q
Cystéine	Cys	C
Méthionine	Met	M
Tryptophane	Trp	W
Phénylalanine	Phe	F
Tyrosine	Tyr	Y
Histidine	His	H

TAB. 1.1 – Les 20 acides aminés composant les protéines

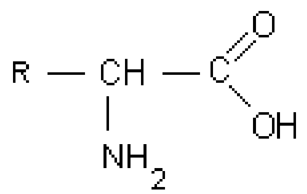


FIG. 1.1 – Formule générale des acides aminés

des enzymes, elles peuvent avoir une activité différente d'une fonction enzymatique. Il existe par exemple des protéines, telles que l'insuline ou les anticorps, dont l'activité se situe à l'extérieur de la cellule.

Les réactions chimiques sont déterminées par des enzymes spécifiques. Cependant, ces enzymes sont incapables d'intervenir dans la séquence des acides aminés qui composent les protéines. Si tel était le cas, à chaque protéine correspondrait une ou plusieurs enzymes capables de reconnaître la séquence des acides aminés. Ceci semble peu probable en raison du nombre important de protéines. Il doit donc exister d'autres molécules qui contiennent l'information permettant de sélectionner les bons acides aminés lors de l'élaboration des protéines. De plus, de telles molécules doivent aussi pouvoir être dupliquées, de sorte qu'après une division cellulaire, chaque cellule fille dispose d'une copie de cette information.

Lorsqu'il a été découvert que les chromosomes étaient le support de l'information génétique de la cellule et que chaque cellule fille en possédait le même nombre que ces cellules mères, les chromosomes ont semblé très similaires aux molécules hypothétiques donnant l'ordre des acides aminés d'une protéine.

1.1.3 L'ADN : le matériel génétique

C'est à partir d'expériences sur les pois que Mendel a conclu que beaucoup de leurs caractères (au sens de caractéristiques) étaient sous le contrôle de *marqueurs d'hérédité*. On considérera plus tard que le gène est une entité autonome, qui détermine les caractères tels que les avait définis Mendel. C'est aussi à partir de ces concepts que s'est faite la distinction entre phénotype et génotype, le génotype représentant la totalité de l'information génétique et le phénotype en étant son expression physique. La question de la substance des gènes fut abordée par des expériences sur des bactéries réalisées en 1928 par l'anglais Griffith. Lorsque l'on comprit que les chromosomes étaient le support de l'hérédité, l'ADN qui en est le composant principal devait contenir l'information génétique, autrement dit les gènes. Cependant, il n'était pas possible de concevoir la relation gène-protéine que nous allons décrire tant que la structure moléculaire du gène ,et donc de l'ADN, n'était pas connue.

Lorsque James Watson, Francis Crick et Maurice Wilking reçurent le prix Nobel en 1962 pour le modèle de la double hélice de l'ADN, une nouvelle voie de recherche, appelée biologie moléculaire, allait s'ouvrir.

La molécule d'ADN est formée de deux brins enroulés l'un autour de l'autre : on dit qu'elle est bicaténaire. Chacune des deux chaînes est un enchaînement de nucléotides qui sont de petites molécules composées d'un phosphate, d'un sucre (désoxyribose) et d'une base azotée. Il n'existe que quatre bases nucléotidiques : deux sont dites puriques (la guanine G et l'adénine A), les deux autres sont pyrimidiques (la cytosine C et la thymine T). Ce sont les quatre bases de l'alphabet de l'ADN. La composition chimique d'un nucléotide est présentée sur la figure 1.2 et le nom ainsi que le code des quatre nucléotides sont donnés dans le tableau 1.2. Les quatre nucléotides diffèrent par la base azotée qui les compose. La figure 1.3 présente les quatre bases azotées correspondant aux quatre nucléotides.

Les deux chaînes polynucléotides enlacées, formant la double hélice, constituent l'unité fondamentale de l'ADN. La configuration spatiale de cette structure est représentée sur la figure 1.4. Pour assurer la cohésion de cette structure, les deux brins sont reliés par des liaisons hydrogènes entre les bases azotées des nucléotides et selon une certaine complémentarité : la thymine se lie toujours à l'adénine et la cytosine se lie toujours à la guanine. Il est à noter que deux liaisons lient la thymine et l'adénine alors que trois sont nécessaires entre la cytosine et la guanine (figure 1.4). Malgré les forces relativement faibles de ces ponts hydrogènes, l'ADN

Nom	Code
Adénosine	A
Thymine	T
Guanine	G
Cytosine	C

TAB. 1.2 – Les quatre nucléotides composant l'ADN.

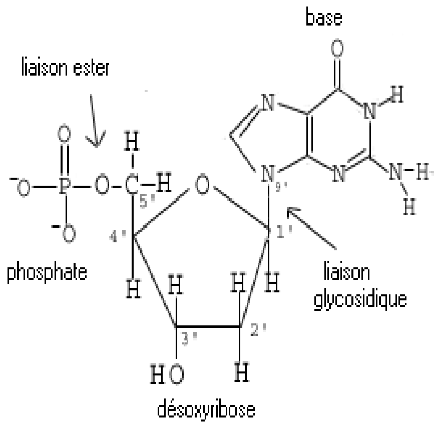


FIG. 1.2 – Un nucléotide. Dans cet exemple, la base est la guanine

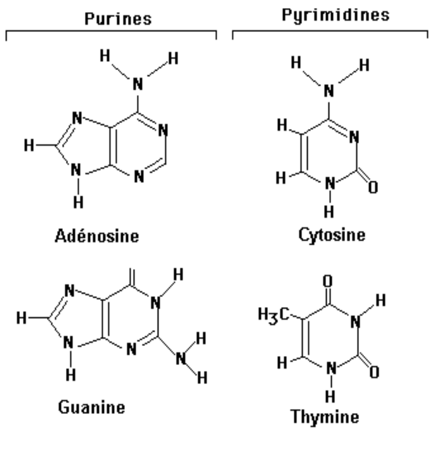


FIG. 1.3 – Les quatre bases azotées des nucléotides

contient tellement de bases que les deux brins ne se dissocient jamais spontanément. Lors d'une division cellulaire, la synthèse d'une nouvelle molécule d'ADN débute par la séparation des deux brins, puis construit une nouvelle molécule, nucléotide par nucléotide, en utilisant les règles d'appariement entre les bases (figure 1.5).

Sur un même brin, deux bases nucléotidiques sont liées entre le groupe phosphate d'un nucléotide et le sucre (désoxyribose) du nucléotide suivant. Cette liaison permet de définir une orientation sur chaque brin de l'ADN. Elle débute en $5'$ (phosphate) vers l'extrémité $3'$ (sucre). La représentation de l'ADN sur la figure 1.4 montre que l'orientation des deux brins est opposée.

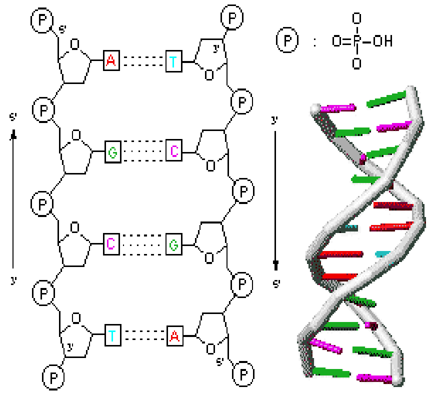


Fig. 1.4 – Molécule d'ADN

Ce sont donc les quatre lettres de l'alphabet de l'ADN (A, T, C, G), et elles seules qui doivent spécifier l'information génétique. La relation entre un gène et une protéine ne fut pas immédiate. Il fallut attendre de nombreux résultats sur la bactérie *Eschericia Coli (E. Coli)* avant de définir un gène comme une séquence de nucléotides adjacents qui spécifient la séquence des acides aminés des chaînes polypetidiques.

Cependant, on sait depuis longtemps que les protéines sont synthétisées presque exclusivement dans le cytoplasme. Chez les eucaryotes, l'ADN étant localisé dans le noyau, les protéines ne peuvent pas être synthétisées directement à partir de l'ADN. Il existe en effet un intermédiaire moléculaire de type ARN (acide ribonucléique) qui permet de transférer le code génétique en séquences polypeptidiques du noyau vers le cytoplasme. Ceci conduit à ce que l'on appelle habituellement le dogme central de la biologie moléculaire :

$$\text{ADN (gène)} \stackrel{transcription}{\longrightarrow} \text{ARN} \stackrel{traduction}{\longrightarrow} \text{Protéine.}$$

Nous détaillons dans la suite les deux étapes de transcription et de traduction.

1.1.4 La transcription

La phase de transcription constitue l'étape préliminaire pour la biosynthèse des protéines. Elle comprend l'ensemble des mécanismes par lequel un segment de la molécule d'ADN, cor-

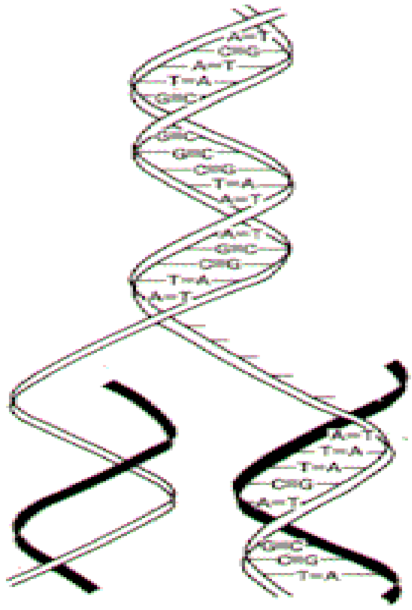


FIG. 1.5 – Duplication de l'ADN

respondant au gène, est copié en un brin d'ARN que l'on appelle l'ARN messager (ARNm). La transcription produit d'autres types d'ARN dont les principaux sont des ARNt (ARN de transfert) et des ARNr (ARN ribosomaux).

La transcription s'effectue sous l'action d'enzymes, les ARN polymérases qui permettent d'enchaîner les nucléotides, leur ordre étant spécifié par l'appariement de leurs bases à la séquence de désoxyribonucléique du brin d'ADN transcrit. La synthèse d'un ARNm s'effectue toujours dans le sens $5'$ à $3'$, de manière anti-parallèle et complémentaire à la portion d'ADN copiée (figure 1.6). Il y a respectivement correspondance entre les bases A, T, C, G de l'ADN et U (Uracile), A, G, C du brin d'ARNm. Dans l'ARN, la thymine est remplacée par l'uracile (figure 1.7).

La différence entre un brin d'ADN et d'ARNm tient aussi à la nature du sucre : il s'agit d'un ribose pour l'ARN et d'un désoxyribose (perte d'un atome d'oxygène) pour l'ADN (figure 1.8).

La transcription ne concerne qu'une portion de l'ADN, celle d'un gène. Il est évident qu'il existe des mécanismes définissant le début et la fin de la transcription, et qui permettent à l'ARN polymérase de se fixer et de transcrire la portion d'ADN. Ces mécanismes diffèrent entre les eucaryotes et les procaryotes. Chez les eucaryotes, la molécule d'ARNm synthétisée sort du noyau pour subir la phase d'épissage qui consiste à ne retenir que des sous-parties qui seront traduites en acides aminés pour former une protéine.

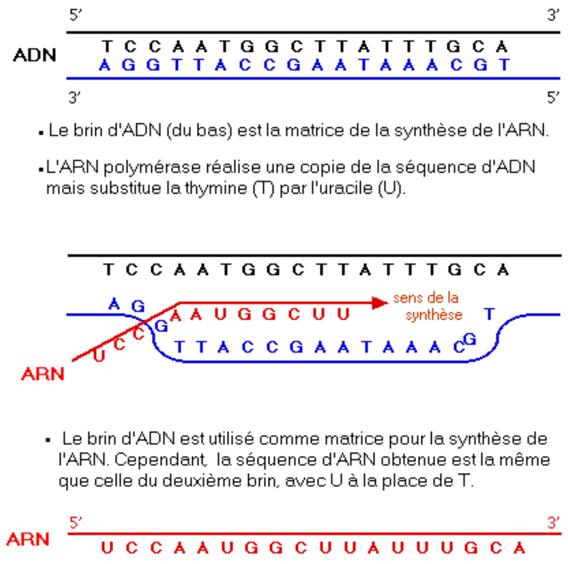


FIG. 1.6 – La phase de transcription

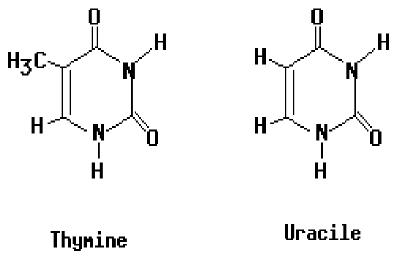


FIG. 1.7 – La thymine est remplacée par l'uracile dans les molécules d'ARN

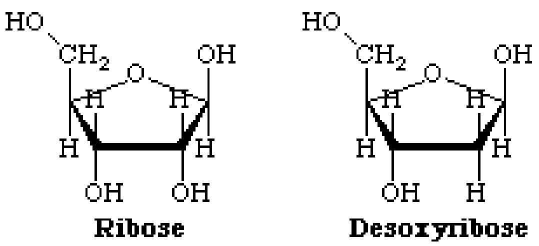


FIG. 1.8 – Un ribose et un désoxyribose

1.1.5 La traduction

Le mécanisme de la traduction d'un ARN messager en une chaîne polypeptidique est complexe et se déroule dans le cytoplasme. Ce sont les ribosomes, les ARN ribosomaux et les ARN de transferts qui constituent les éléments de la machinerie de la traduction qui convertit les séquences de nucléotides (U, C, A, G) des ARNm en séquences d'acides aminés dans les protéines. L'ARNm est le porteur du code, les ARN ribosoamaux sont des éléments de la machine à fabriquer des protéines et les ARN de transfert sont les clefs du code. Ces trois types d'ARN se trouvent dans la machine à décoder : le ribosome. Cependant, le nombre d'acides aminés étant de 20 et celui des nucléotides de 4, il est clair que c'est un groupe de nucléotides qui définit un acide aminé. Cette relation entre groupe de nucléotides et acides aminés, appelée code génétique, fut déchiffrée en 1966. Chacun des 20 acides aminés est codé par un groupement de trois nucléotides appelé codon. On remarque dans le tableau 1.3 que certains acides aminés sont codés par plus d'un codon : c'est la dégénérescence ou la redondance du code génétique.

1ére position	2ème position				3ème position
en $5'$	U	C	A	G	en $3'$
	Phe	Ser	Tyr	Cys	U
U	Phe	Ser	Tyr	Cys	C
	Leu	Ser	Stop	Stop	A
	Leu	Ser	Stop	Trp	G
	Leu	Pro	His	Arg	U
C	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
	Ile	Pro	Asn	Ser	U
A	Ile	Pro	Asn	Ser	C
	Ile	Pro	Lys	Arg	A
	Met	Pro	Lys	Arg	G
	Val	Pro	Asp	Gly	U
G	Val	Pro	Asp	Gly	C
	Val	Pro	Glu	Gly	A
	Val	Pro	Glu	Gly	G

TAB. 1.3 – Code génétique

1.1.6 Commentaires

En règle générale, il existe une ou plusieurs molécules d'ADN organisées en chromosome. Chez les eucaryotes, les chromosomes ont une structure particulière dans laquelle l'ADN est enroulé autour de protéines appelées les histones. Tous les organismes disposent d'un génome qui contient la plus grande partie de l'information héréditaire. À part de rares exceptions, la réplication de l'ADN à chaque division cellulaire implique que toutes les cellules d'un organisme contiennent le même génome.

L'ensemble des découvertes sur l'information portée par l'ADN a apporté une manière radicalement différente d'aborder la biologie. L'ADN est vu comme une substance dans laquelle on peut lire le passé (phylogénie et évolution du vivant), le présent (compréhension de phénomènes) et l'avenir (diagnostics et risques génétiques). Le génome étant le livre du vivant, il s'agit de le décrypter : identifier les séquences de l'ADN, localiser les gènes et prédire leurs fonctions.

La détermination des 4 lettres qui composent la molécule d'ADN est le séquençage. Le premier génome d'une bactérie fut séquencé en 1995. Le génome de la levure (*Saccharomyces cerevisiae*), eucaryote unicelluliare, a été obtenu en 1997, celui de la mouche (*Drosophila melanogaster*) en 2000 et celui d'une petite crucifère (*Arabidopsis thaliana*) en 2001. L'ensemble des séquences d'ADN (y compris les génomes complets) est stocké dans la base de données EMBL (EMBL-Bank) qui correspond en fait à la collaboration de trois bases de données : EMBL en Europe, GenBank aux États-Unis et la banque DDBJ au Japon. Ainsi, il est facile de comparer de nouvelles séquences à des séquences connues à l'aide de programmes dédiés (FASTA ou BLAST). Comme nous l'avons décrit, les gènes qui sont les porteurs de l'information génétique sont situés sur l'ADN mais leur localisation est un problème délicat, tout particulièrement chez les eucaryotes en raison de l'épissage. Cette étape, située juste aprés la transcription et précédant la traduction se déroule généralement à l'intérieur du noyau. Certaines parties, appelées introns, de l'ADN transcrit (pré ARN messager) sont supprimées alors que les sections restantes, appelées exons, sont assemblées pour former l'ARNm qui sera transcrit en protéine dans le cytoplasme. Il est à noter que dans certains cas le même pré-ARNm peut donner différents ARNm : c'est l'épissage alternatif. Les étapes de transcription et de traduction sont en réalité plus complexes que ce qui a été décrit précédemment, et il est difficile de donner une définition très précise d'un gène. De manière simplifiée, on considèrera un gène comme une région continue de l'ADN, à partir de laquelle la machinerie cellulaire est capable de synthétiser une ou plusieurs protéines.

Étant donné une séquence d'ADN, la localisation des gènes est un problème intéressant et difficile. Différents algorithmes, basés principalement sur des modèles de Markov à états cachés, ont été développés.

La génomique fonctionnelle peut se définir comme la recherche de la fonction et des interactions des gènes. La génomique comparative est probablement l'approche la plus directe qui s'est développée pour prédire la fonction des gènes. En comparant des génomes différents (génomes d'organismes de même espèce ou d'espèces différentes), il est alors possible de prédire la fonction de gènes, en se basant par exemple sur des similarités de séquences de gènes de fonction connue. Cependant, cette approche n'autorise pas une analyse systématique.

Aussi, l'étude fonctionnelle des gènes nécessite de s'intéresser directement aux produits des gènes : les ARNm et les protéines. Le transcriptome correspond à l'ensemble des gènes transcrits dans des conditions précises. Selon la cellule, mais aussi sa situation physiologique (stress, pathologie,...), un gène peut s'exprimer ou non, plus ou moins intensément (nombre d'ARNm transcrits plus ou moins important). L'étude du transcriptome participe clairement à la détermination de la fonction des gènes au sein d'une cellule. Le protéome représente toutes les protéines exprimées à un moment donné par le génome d'une cellule. Bien que différents facteurs comme la stabilité des ARN messagers, les modifications post-traductionnelles, le transport et la dégradation des protéines puissent l'influencer, une grande partie de la régulation de l'expression des gènes s'exerce au niveau de la transcription.

Parmi les outils biotechnologiques disponibles pour l'étude du transcriptome, les puces à ADN ou biopuces permettent de mesurer le niveau d'expression de plusieurs milliers de gènes simultanément, et sont en train de devenir un outil standard pour la recherche en biologie moléculaire [75]. Avant de présenter les problèmes et les objectifs liés à l'analyse de tels résultats, il est indispensable de préciser les principes de cette technologie.

1.2 Aspects techniques

Il existe différentes méthodes pour l'étude à grande échelle du transcriptome : le séquençage automatique [92], SAGE (Serial Analysis of Gene Expression) [122] ou le differential display et ses variantes [79]. Mais la simplicité relative de mise en oeuvre des puces à ADN et les perspectives envisagées ont permis un développement rapide de cette technologie. Les puces à ADN tirent parti des différents efforts de séquençage afin d'identifier les gènes exprimés dans un type cellulaire, à un moment particulier et sous certaines conditions. Une application courante cherche à comparer l'expression de gènes entre des cellules tumorales et saines, à étudier l'effet d'une drogue ou de différentes doses d'un traitement.

La fabrication des puces à ADN consiste à immobiliser de façon ordonnée sur un support solide des molécules d'ADN, appelées sondes, spécifiques des gènes que l'on souhaite étudier. L'ensemble de ces dépôts est parfois appelé la matrice des sondes (DNA array). Une fois fabriquées, les puces sont hybridées avec les cibles d'ADN marquées et extraites de cellules à analyser. Les puces se différencient par la nature du support, le nombre et le type de sondes utilisées, et le marquage des cibles. Le principe fondamental repose sur la réaction qui a fondé la biologie moléculaire : l'hybridation entre deux brins complémentaires d'ADN.

Suivant la densité des sondes déposées, on parle de macro-array (quelques dizaines ou centaines au centimètre carré) ou de micro-array (plusieurs milliers au centimètre carré). Deux approches sont possibles pour le greffage des sondes. Dans la première, les sondes sont directement synthétisées sur la puce, c'est la synthèse *in situ*. Dans ce cas, les sondes sont des polynucléotides relativement courts (entre 6 et 100 bases) que l'on appelle oligonucléotides et spécifiques des gènes étudiés. Ce procédé, largement développé par la société Affymétrix [81], reste néanmoins coûteux. La deuxième approche repose sur le transfert d'oligonucléotides ou d'ADN complémentaire (ADNc) préparés avant le greffage des sondes. L'ADNc correspond à de l'ADN simple brin synthétisé artificiellement et obtenu par une réaction de transcription inverse d'un ARNm. L'ADNc offre l'avantage d'être plus stable que la molécule d'ARNm, de pouvoir être stocké, copié et séquencé. L'utilisation de séquences d'ADNc nécessite de disposer d'un ensemble de sondes spécifiques des gènes que l'on veut étudier. Cet ensemble est obtenu par transcription inverse des ARN messagers d'une cellule et constitue alors une banque de clones d'ADNc. Une fois séquencé sous forme d'EST (Expressed Sequence Tag), ces clones sont identifiés. Ils sont ensuite amplifiés par PCR (polymerase chain reaction [22]) puis déposés de façon ordonnée sur une surface solide : une lame de verre ou une membrane de nylon.

Une fois la puce fabriquée, on peut étudier l'expression des gènes d'un échantillon biologique. Les cibles sont obtenues par transcription inverse des ARN messagers extraits de l'échantillon et par marquage simultané. Ce marquage est réalisé à l'aide de fluorophores spécifiques (cyanine5 et cyanine3) dans le cas de puce sur lame de verre. Lorsqu'il s'agit de puce sur membrane de nylon, le marquage est généralement radioactif. Dans les deux cas, ces solutions sont déposées sur la puce, de telle sorte que les cibles déposées en excès s'hybrident avec la séquence complémentaire sur la puce. Chacune des deux approches, marquage par radioactivité ou par fluorochrome, a ses propres avantages et inconvénients. Par exemple, la première est plus sensible et permet des détections d'un nombre d'ARN messagers plus faible que la seconde. En revanche, elle nécessite de manipuler des marqueurs radioactifs avec les contraintes que cela impose. Quant à la seconde, l'utilisation de deux fluorophores différents permet de déposer simultanément sur la puce deux échantillons différents, ce qui constitue un avantage indéniable pour des approches différentielles où l'on compare le niveau d'expression des gènes entre un tissu sain et tumoral, ou entre un organisme sauvage et mutant. Nous détaillons dans la suite les deux étapes de dépôt de sondes d'ADNc sur lame de verre et d'hybridation de cibles marquées à l'aide de fluorophores.

1.2.1 Dépôt des sondes

Avant de procéder à la fabrication de la puce, il est nécessaire d'établir un ensemble de sondes spécifiques d'un grand nombre de gènes. Cet ensemble est réalisé à partir de banques ou de bases de données telles que Genbank [12], dbEST [16] et UniGene [109]. L'utilisation des banques de données dans le cadre des puces à ADN est présentée dans [17]. Une fois les clones choisis, ils sont amplifiés par PCR puis purifiés et déposés de façon ordonnée au moyen d'un robot sur la lame de verre prétraitée (ce prétraitement permet l'accrochage des ADNc déposés). Le transfert des sondes est réalisé par un robot qui dépose de façon séquentielle un échantillon de sondes pour chacun des gènes. Afin d'accélérer ce processus, le robot possède différentes têtes de dépôt pour le greffage, en général 4, 8 ou 16.

1.2.2 Hybridation des cibles

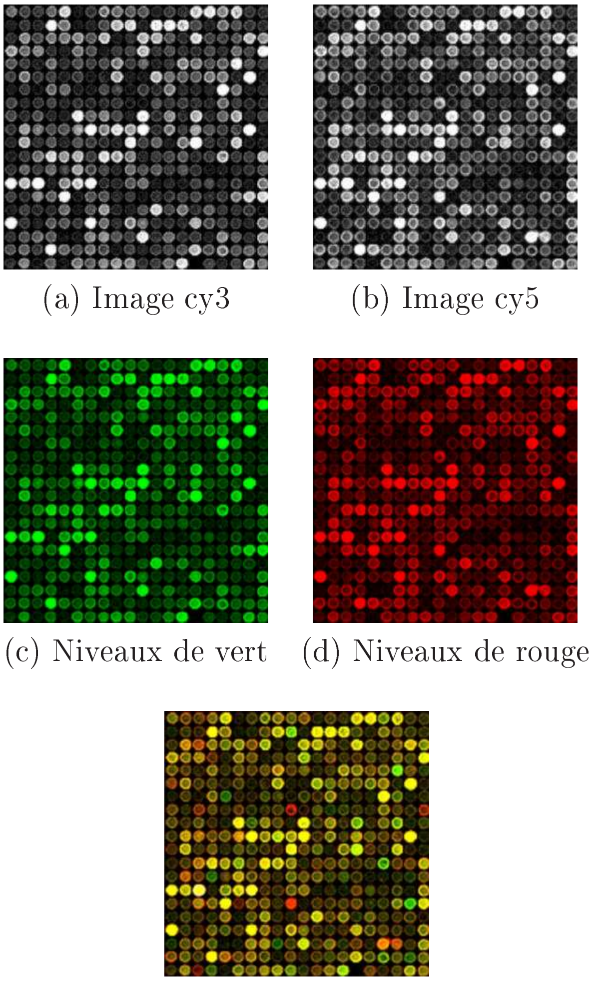
Ainsi fixées sur la lame, les sondes permettent de reconnaître certaines séquences d'ADN présentes dans un mélange qui serait déposé à la surface de la biopuce. Les cibles sont extraites de deux cultures cellulaires dont on veut comparer l'expression des gènes. Les ARN messagers sont extraits des cellules, puis amplifiés par PCR et transformés en ADN. Durant cette phase, l'ADN d'une culture est marqué avec un fluorochrome, la cyanine Cy3 tandis que l'ADN de l'autre culture est marqué avec le Cy5. Ces deux échantillons d'ADN marqués sont mélangés dans un tampon d'hybridation et déposés sur la puce, puis recouverts d'une lame de verre pour permettre une répartition homogène de l'échantillon. La phase d'hybridation se déroule dans une sorte d'incubateur (station fluidique) durant 16 heures environ à une température de 60 degrés. À cette température, un brin d'ADN marqué qui rencontre son complémentaire fixé sur la puce va s'apparier pour donner l'ADN double brin. Les ADN marqués sont alors hybridés avec des séquences spécifiques de gènes. Enfin, une étape de lavage permet de débarrasser la puce des cibles non hybridées.

1.2.3 Lecture des niveaux de fluorescence et analyse d'image

L'hybridation terminée, la quantité de cibles hybridées est mesurée en détectant la fluorescence émise par l'excitation des fluorophores. Pour réaliser cette détection, la configuration la plus classique utilise un laser pour exciter les fluorophores et un microscope confocal (ou scanner) couplé à un tube photo-multiplicateur (PMT) pour analyser les photons émis par les marqueurs. Le fonctionnement et la description des scanners à fluorescence sont détaillés dans *Microarrays : A Practical Approach* [106] et *Microarray biochip technology* [107]. Pour simplifier, le scanner génère deux images 16-bit (format TIFF généralement), une pour chaque fluorophore. Ces deux images en niveaux de gris représentent l'intensité de fluorescence lue par le scanner et donnent le niveau d'expression des gènes dans les deux conditions expérimentales. Une image en fausses couleurs sert souvent de représentation à ces deux images. Ces couleurs vont du vert, pour caractériser l'échantillon marqué en Cy3, au rouge pour l'échantillon marqué en Cy5. Le jaune indique que les cibles marquées en Cy5 et Cy3 se sont hybridées en proportion égale (figure 1.9).

Bien que les scanners disposent de plusieurs filtres pour ne retenir que le signal intéressant, il existe plusieurs sources de bruit qui affectent le signal obtenu. L'amplification des signaux de fluorescence en signal électrique, réalisé par le PMT, fait partie des sources de bruit dues au scanner. Le choix des lames, leur traitement de surface pour le greffage des sondes, le recouvrement des spectres d'émission et d'absorption des fluorophores font aussi partie des sources de bruit. Une image parfaite devrait refléter uniquement les intensités de fluorescence, alors qu'en réalité on obtient un signal qui correspond au signal de fluorescence voulu, auquel s'ajoutent différentes sources de variation.

Suivant les scanners, certains paramètres permettent d'ajuster la puissance du laser et l'amplification du PMT. Une forte puissance du laser génère un bruit important dans le signal obtenu. Un PMT mal réglé amplifie trop fortement les photons émis par les fluorophores, et génère également un bruit de fond important. En principe, il est souhaitable de choisir une puissance de laser forte plutôt que d'augmenter l'amplification du scanner. Le réglage du scanner est réalisé de sorte que la valeur des pixels soit en dessous du niveau de



(a) Image cy3 (b) Image cy5

(c) Niveaux de vert (d) Niveaux de rouge

(e) Image en fausses couleurs

FIG. 1.9 – Image en fausses couleurs d'une puce à ADN. À partir des deux images en noir et blanc obtenues pour chacun des deux échantillons (a) et (b), l'un marqué en cy3 (c) et l'autre au cy5 (d), on obtient une image en fausses couleurs (e).

saturation (valeurs inférieures à $2^{16} - 1 = 65535$).

L'analyse des deux images obtenues a pour but de mesurer pour chacun des spots la quantité de transcrits dans les échantillons d'ARN messagers. On estime également le bruit de fond de l'image et on évalue la qualité des spots. Cette étape d'analyse d'image se décompose en trois parties.

1. La localisation permet de préciser les coordonnées de chaque spot sur l'image. Si les techniques de dépôt sont fiables, il suffit en principe de positionner une grille respectant le plan de dépôt des sondes et les caractéristiques du robot ayant permis le dépôt des

sondes. En pratique, du fait du recouvrement de certains spots, certains ajustements doivent être apportés soit manuellement, soit à l'aide d'algorithmes et le plus souvent en combinant les deux. L'efficacité de cette localisation dépend fortement de la qualité des dépôts.

2. La segmentation permet de classer les pixels en tant que signal ou bruit de fond dans le voisinage de chacun des spots identifiés précédemment. Ce traitement doit être spécifique à chacun des spots car le bruit de fond est rarement uniforme. Il a pour origine différentes sources, comme par exemple de la fluorescence non spécifique telle que l'auto-fluorescence des lames. La forme, la taille et les irrégularités des spots sont des problèmes majeurs que doivent prendre en compte les algorithmes de segmentation. Yang *et al.* [128] mènent une étude comparative des méthodes de segmentation et décrivent différentes approches proposées par les logiciels. Le logiciel *QunatArray* [84] utilise la solution proposée par Chen *et al.* [26]. Une seconde méthode consiste à définir un cercle caractérisant le spot. Ce cercle peut être fixe comme dans le logiciel *ScanAlyse* [46] ou être estimé pour chaque spot *GenePix* [7]. Yang *et al.* [128] proposent une approche plus complexe [2], utilisant des algorithmes de morphologie mathématique pour prendre en compte les irrégularités des spots. Cette approche est intégrée dans le logiciel *Spot* [23].

3. Après l'étape de segmentation, un ensemble de valeurs est calculé pour caractériser le signal et le bruit de fond. La moyenne, la médiane et l'écart-type des pixels considérés comme signal et bruit de fond sont estimés. De plus, pour chaque spot, on retient aussi sa localisation sur la puce ainsi que la tête de dépôt du robot ayant permis de greffer les sondes correspondantes.

Nous avons longuement présenté les aspects biologiques et techniques liés aux puces à ADN. Bien que les biopuces offrent de nombreuses applications, nous nous intéressons dans la suite à l'étude des transcriptomes en utilisant cette technologie. Comme il a été décrit précédemment, les niveaux estimés intègrent de nombreuses sources de variations d'ordre biologique, physique, chimique et expérimental. Aussi, il est indispensable de mener une étape de normalisation qui doit permettre de supprimer les artefacts expérimentaux et les différences entre les deux fluorophores Cy5 et Cy3.

Chapitre 2

Représentation et normalisation des données

Il existe différentes sources de variation qui affectent les mesures des niveaux d'expression des gènes. Étant mal connues, ces variations rendent difficile l'analyse des données issues de l'analyse d'image, et donc la mise en relation entre les intensités mesurées et les niveaux d'expression des gènes. En effet, les deux molécules utilisées pour le marquage des ARN (les dyes [Cy5]dCTP et [Cy3]dCTP) possèdent des caractéristiques différentes quant à leur taux d'incorporation, leurs pics d'absorption et d'émission. L'objectif de la normalisation est de réaliser une transformation des intensités pour permettre l'étude des niveaux d'expression des gènes. Cette étape est primordiale car les études que nous menons dans la suite en dépendent fortement. Nous choisissons de définir la normalisation comme les méthodes permettant de se positionner dans un cadre d'analyse où certains a priori biologiques sont vérifiés. C'est pour cette raison que la normalisation doit précéder toute autre analyse. Un a priori courant consiste à admettre que pour une expérience donnée, un nombre relativement faible (moins de 10% ou 20%) de gènes ont une expression différente entre les deux conditions de l'étude. Si cette hypothèse est peu vraisemblable, il faut alors définir un ensemble de sondes (contrôles) dont on sait a priori que les niveaux d'expression sont identiques. Avant de présenter différentes normalisations, nous décrivons les représentations possibles qui permettent d'apprécier la qualité des données. Bien que simples, ces étapes descriptives sont très informatives et permettent d'orienter les analyses, d'améliorer les protocoles expérimentaux ou de définir des plans d'expériences. Nous insistons sur ce point pour sa simplicité de mise en œuvre et sa facilité d'interprétation.

Nous présentons d'abord la nature des données expérimentales ainsi que quelques méthodes pour les visualiser. Ensuite, nous abordons le problème de normalisation en décrivant différentes méthodes. Enfin, nous menons une étude comparative de ces approches sur une application.

2.1 Nature des données

Suivant le logiciel utilisé pour l'analyse d'image, plusieurs informations sont accessibles. Il est cependant difficile de tirer parti de toutes ces informations. Aussi, nous étudions uniquement les résultats donnant pour chaque spot : la moyenne ou la médiane des pixels du bruit de fond et du signal pour les deux fluorophores, et le numéro de la tête du robot ayant déposé les sondes ainsi que leur localisation sur la puce. Les principales informations que nous négligeons sont les distributions des pixels du bruit de fond et du signal.

Les notations suivantes seront utilisées durant le reste de la thèse. Les lettres R (red) et G (green) font référence aux dyes Cy5 et Cy3 respectivement. Pour un spot donné, nous notons Rf et Gf les moyennes (ou les médianes) des pixels du signal, et Rb et Gb les moyennes (ou les médianes) des pixels du bruit de fond. Bien souvent, le premier traitement vise à corriger le signal mesuré en fonction de bruit de fond lié à l'image. On définit R et G comme les valeurs obtenues par cette correction. Pour justifier cela, on admet qu'une part du signal émis par chaque spot correspond à du bruit de fond. En général, on estime simplement cette part par le bruit de fond qui a été mesuré autour du spot. Une correction simple consiste alors à poser $R = Rf - Rb$ et $G = Gf - Gb$. Cependant, cette correction du signal n'est pas toujours souhaitable. En effet, celle-ci sous-estime souvent la vraie valeur du signal et peut donner de mauvais résultats si le bruit de fond est élevé. Afin d'éviter cela, une approche plus rigoureuse consisterait à utiliser les caractéristiques de la distribution des pixels du signal et du bruit de fond comme le proposent Kooperberg *et al.* [72].

2.2 Études descriptives des données

Dans leurs premières applications, les puces à ADN ont permis d'étudier l'expression des gènes chez la levure *Saccharomyces cerevisiae* entre un stress cellulaire [37], [114] et un état de référence. Dans ces expériences, la référence est marquée par le fluorophore vert (Cy3) alors que le fluorophore rouge (Cy5) est utilisé pour l'échantillon stress. Ce choix d'affectation pour les marqueurs est encore largement utilisé. Une représentation simple consiste alors à placer les valeurs du Cy3 en abscisse et du Cy5 en ordonnée [104], [107]. Sur une telle représentation, les gènes induits par rapport à la condition de référence se situent au dessus de la droite $y = x$, et les gènes réprimés en dessous (figure 2.1(a)). Une seconde représentation, plus intéressante, consiste à utiliser une transformation logarithmique en base 2 (figure 2.1(c)). Il existe plusieurs raisons pour justifier cette transformation. La première est que la variation du logarithme des intensités est moins dépendante de la grandeur des intensités. Ensuite, elle permet d'obtenir des distributions avec moins de valeurs extrêmes et est la fonction de lien classique dans le cas de distributions à queue lourde. Le logarithme en base 2 est plus cohérent que la base naturelle ou la base 10 car les intensités des pixels sont comprises entre en 0 et $2^{16} - 1$.

Les représentations graphiques permettent de valider une expérience, peuvent guider des

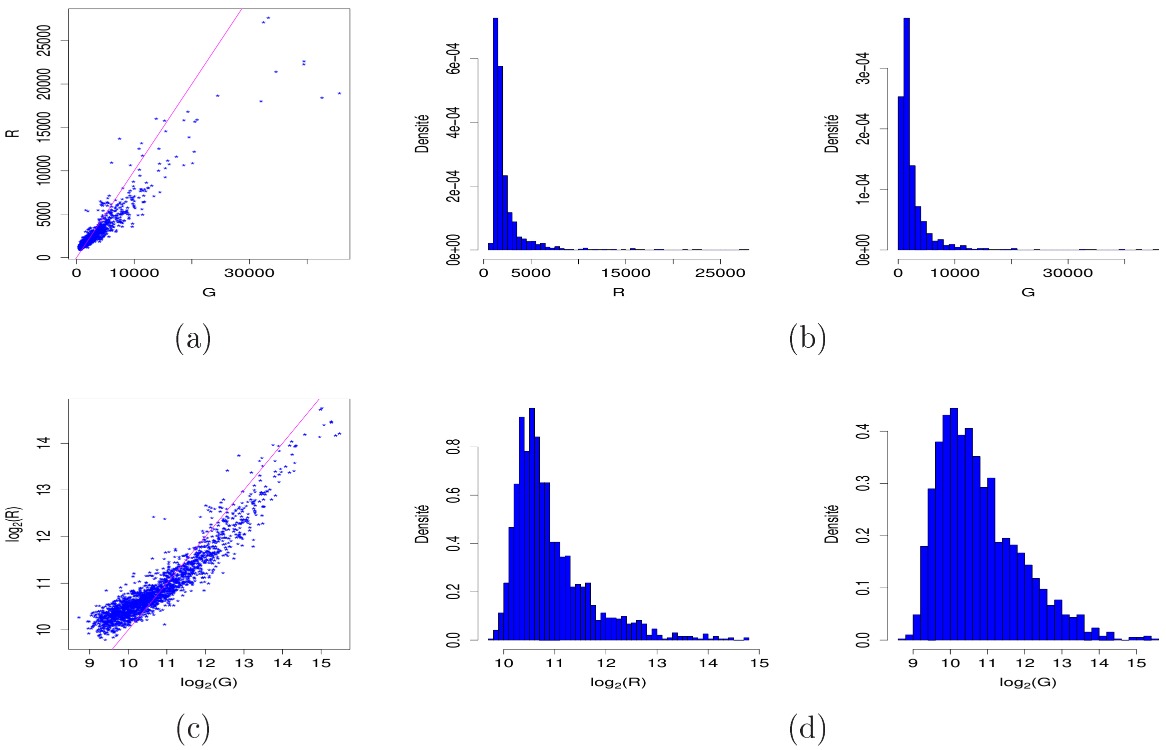
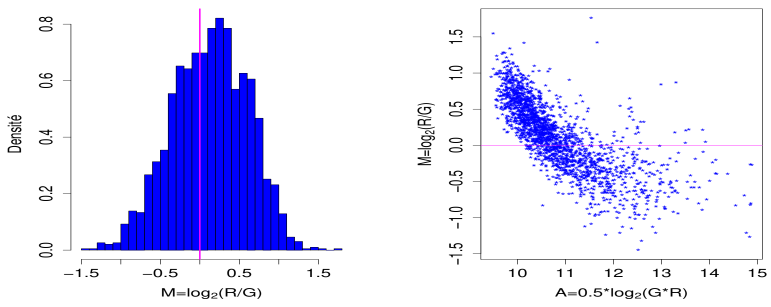


Fig. 2.1 – *Représentations de R (Cy5) et G (Cy3).* Une méthode de représentation simple consiste à placer G en abscisse et R en ordonnée (figure (a)). La droite $y = x$ en rouge sur les représentations (a) et (c) donne la droite théorique des gènes non différentiellement exprimés. Les figures (b) et (d) montrent que la transformation logarithmique permet de se rapprocher d'une distribution symétrique et offre une meilleure dispersion avec moins de valeurs extrêmes.

analyses et mettent en évidence des problèmes spécifiques. Dans la suite, nous illustrons certains de ces problèmes en nous appuyant sur différentes expériences.

Yang *et al.* [129] proposent une autre représentation en réalisant une rotation et un changement d'échelle sur les axes de la figure 2.1(c). En plaçant $A = 0.5 * \log_2(G * R)$ en abscisse et $M = \log_2(R/G)$ en ordonnée, on décrit les relations entre le logarithme des ratios des intensités et la moyenne des log-intensités en Cy3 et Cy5 (figure 2.2(b)). Sur cette représentation, l'utilisation de la base 2 pour la transformation logarithmique donne une induction (resp. répression) d'ordre 2 si $M = 1$ (resp. $M = -1$) et d'ordre 4 (resp. -4) si $M = 2$ (resp. $M = -2$). Si $M = 0$, il n'y a pas de différence d'expression entre les deux conditions.

Nous avons décrit jusqu'ici différentes méthodes de représentation des intensités en Cy5 (Rf) et Cy3 (Gf). Un étude du bruit de fond est aussi très utile pour juger de la qualité des données. Cette étude montre que, dans certains cas, la soustraction des bruits de fond Rb et



(a) (b)

FIG. 2.2 – *Histogramme des valeurs M et représentation de M en fonction de A.* Bien que l'histogramme des log-ratios (a) soit "presque" centrée autour de 0, on observe (b) que les log-ratios sont positifs pour des niveaux d'intensités faibles et négatifs pour des intensités fortes.

Gb aux valeurs Rf et Gf doit être évitée. Deux situations posent problèmes. Si les niveaux d'expression sont faibles, les valeurs corrigées R et G peuvent être alors très petites et les ratios M grands. Si les valeurs de bruit de fond sont importantes, on peut de nouveau obtenir des valeurs corrigées faibles. Dans ce cas, on risque aussi de surestimer la part du bruit de fond dans le signal. De manière générale, la transformation logarithmique étant sensible aux petites valeurs, l'étude des gènes faiblement exprimés est délicate.

La représentation du bruit de fond en fonction du signal donne un a priori intéressant sur la qualité des données. Une forme structurée de cette représentation reflète une répartition non uniforme du bruit de fond. Le graphique 2.3 permet de distinguer deux cas de figure.

1. Les représentations 2.3(a) donnent les intensités du bruit de fond Gb et Rb en fonction de Gf et Rf respectivement. Ces graphiques ne mettent pas de structure particulière en évidence, c'est-à-dire qu'il existe une faible corrélation entre le signal et le bruit de fond. La droite $y = x$ indique que tous les spots ont un bruit de fond inférieur au signal. Les représentations spatiales sur la puce du signal et du bruit de fond en (c) montrent une bonne répartition des valeurs. Ces observations permettent de conclure au bon déroulement expérimental ainsi qu'à la bonne qualité des résultats. Le bruit de fond étant plus faible que le signal, la soustraction du bruit de fond semble raisonnable.
2. La figure 2.3(b) met en évidence une relation particulière entre Gf et Gb. Comme le montre la représentation 2.3(d), ce problème apparaît lorsque l'on observe un bruit de fond élevé sur certaines zones de la puce. Sur l'exemple proposé, la représentation 2.3(d) fait apparaître des problèmes de bruit de fond en Cy3 alors que le bruit de fond en Cy5 est homogène sur la puce. De telles différences entre le Cy5 et le Cy3 interdisent une correction du signal basée sur la soustraction du bruit de fond.

Cette correction par le bruit de fond peut entraîner des modifications importantes sur les données. Sur les figures 2.4, nous donnons le classement des intensités avant et après soustraction du bruit de fond. La droite $y = x$ représente les valeurs pour lesquelles l'ordre

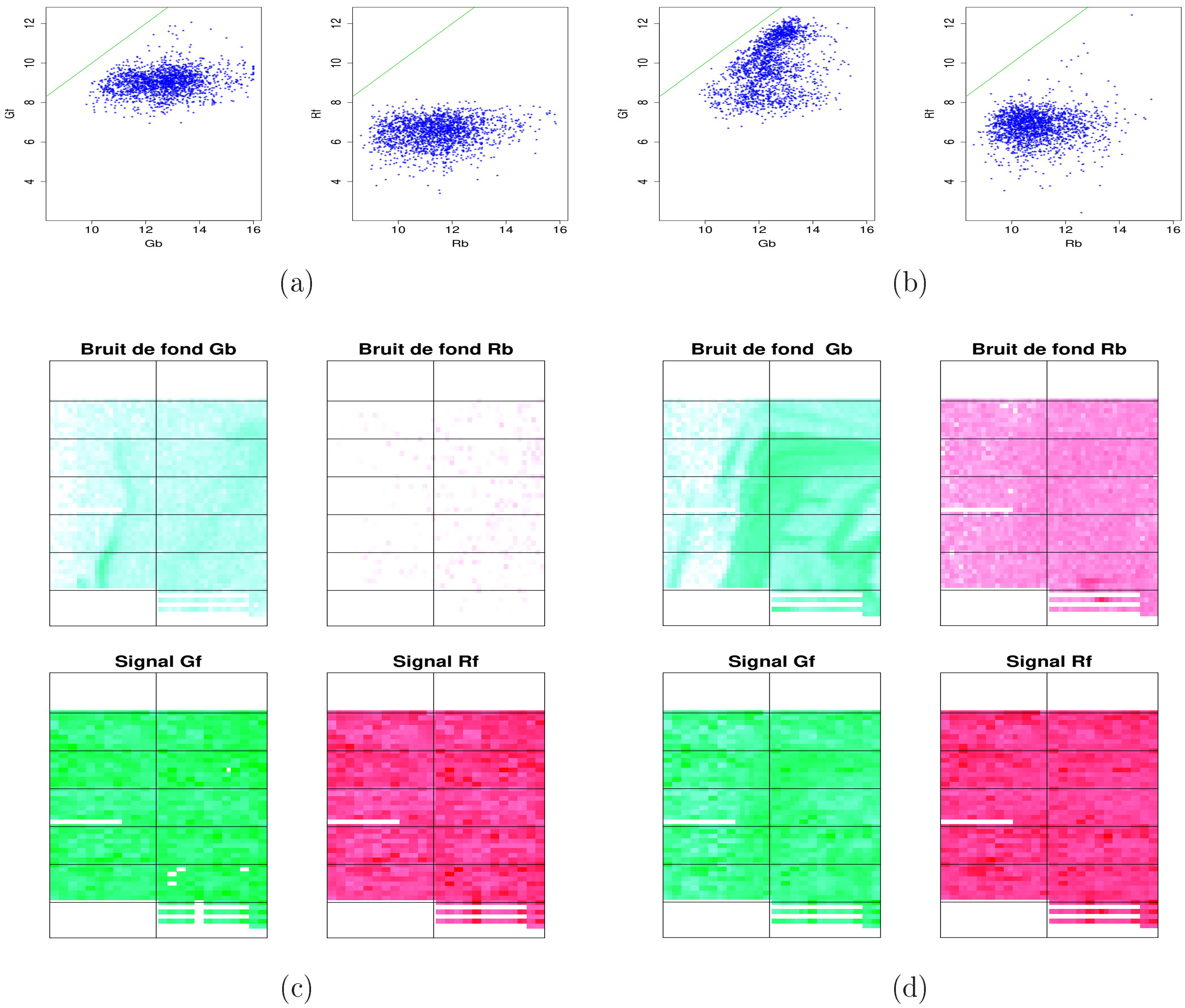
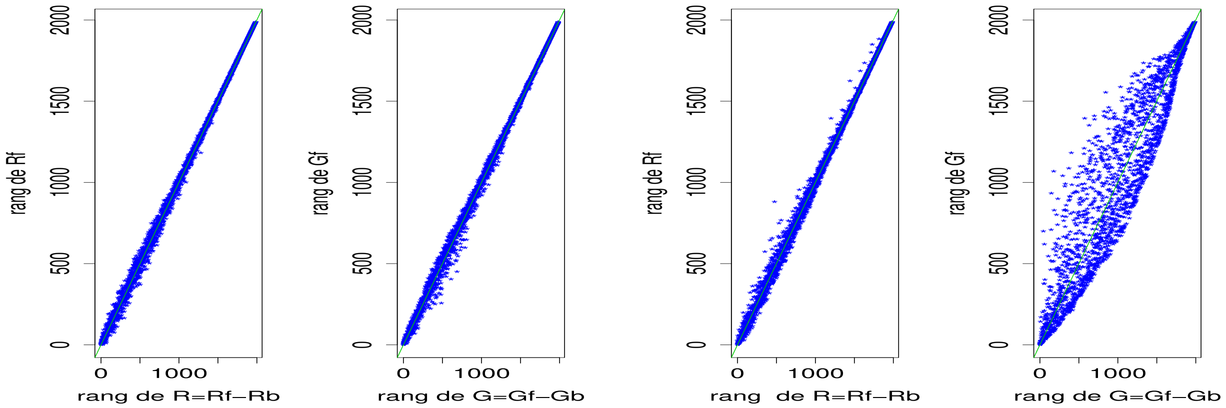


FIG. 2.3 – *Représentations du bruit de fond et du signal.* L'ensemble de ces graphiques soulignent divers problèmes sur le bruit de fond. À partir des deux expériences (a) et (b), nous illustrons les relations possibles entre le bruit de fond et le signal. Les représentations (c) et (d) donnent les répartitions spatiales du signal et du bruit de fond sur chacune des puces. Sur le graphique (a), on observe aucune relation particulière entre le bruit de fond et le signal pour le Cy3 et le Cy5 (corrélations de 0.18 et 0.14 respectivement). Ces faibles corrélations se traduisent par une bonne répartition spatiale des signaux (c). À l'opposé, les graphiques (b) et (d) mettent en évidence un problème de bruit de fond pour le Cy3 (corrélation de 0.49). On remarque aussi à partir des représentations (a) et (b) que les intensités en Cy5 sont en moyenne plus faibles que celles en Cy3.

est inchangé après soustraction du bruit de fond. Dans le cas où le bruit de fond est réparti de façon uniforme sur la puce (figure 2.3 (a) et (c)), le changement de classement des valeurs reste faible (2.4(a) et (c)). Les changements importants dans le cas du Cy3 ((b) et (d)) s'expliquent par les résultats des graphiques 2.3 (b) et (d). Enfin, on observe sur ces représentations que les fortes intensités sont moins sensibles que les faibles intensités.



(a) (b)

Fig. 2.4 – *Modification du classement des niveaux d'intensités avec une correction par le bruit de fond*

L'ensemble de ces illustrations montrent qu'il est possible de tirer de nombreuses conclusions en utilisant simplement des représentations graphiques. En raison des problèmes liés au bruit de fond, nous considérons par la suite que $R = Rf$ et $G = Gf$ dans la mesure où cette solution semble la plus prudente. Cependant, des travaux plus approfondis doivent être engagés afin d'envisager une solution plus rigoureuse.

Les représentations 2.1(c) et 2.2 soulèvent un autre problème. Sous l'hypothèse biologique que peu de gènes doivent varier entre les deux conditions expérimentales, l'histogramme des log-ratios M devrait être centré en 0. Or ces graphiques indiquent une relation particulière entre les intensités en Cy3 et Cy5 et ne valident pas notre hypothèse de départ. Ceci s'explique en partie par les différents réglages du scanner (puissance du laser, gain du PMT) ainsi que par les propriétés différentes des dyes. Comme nous l'avons précisé dans l'introduction, la normalisation a pour but de se ramener dans un cadre où certains a priori biologiques sont vérifiés.

Dans nos applications, les puces à ADN sont utilisées pour analyser les niveaux d'expression des gènes. En raison des problèmes évoqués précédemment, il est difficile à ce stade d'établir un lien entre les intensités mesurées lors de l'analyse d'image et le niveaux d'expression des gènes. C'est la normalisation, au sens où nous l'avons défini dans l'introduction, qui va le permettre.

2.3 Normalisation linéaire

Si l'on admet que peu de gènes varient entre les deux conditions expérimentales, l'ajustement le plus simple consiste à effectuer une translation des log-ratios $\log(R/G)$ des intensités de sorte que la moyenne ou la médiane soit égale 0.

Une seconde approche permet d'utiliser directement les valeurs R et G plutôt que les log-ratios M. Sous l'hypothèse que les intensités R et G sont linéairement liées, on pose le modèle de régression linéaire

$$R = aG + b + \epsilon$$

où ϵ est une loi normale centrée de variance σ^2. L'estimation des paramètres a et b permet de corriger les niveaux moyens en Cy5 en posant $\frac{R-\hat{b}}{\hat{a}}$ comme étant les mesures normalisées. Cependant, ce modèle ne permet de modéliser que des erreurs sur les valeurs de R. Sapir *et al.* [104] proposent alors un modèle de régression robuste et montrent l'intérêt de la transformation logarithmique pour ces modèles linéaires. La figure 2.5 donne une illustration de ces résultats et permet de comparer différentes transformations.

Bien que largement utilisées, ces solutions sont néanmoins d'un intérêt limité en raison de relations non linéaires entre les observations que nous considérons. Ramdas *et al.* [97] décrivent de tels problèmes dans leurs travaux. Ces problèmes doivent donc être pris en compte dans la normalisation. Pour cela, il est nécessaire d'intégrer différents facteurs expérimentaux.

2.4 Normalisation non linéaire

Yang *et al.* [129] proposent une approche non linéaire aux problèmes de normalisation et introduisent un cadre cohérent pour cela.

Une normalisation globale des données consiste à soustraire une constante c aux log-ratios M pour obtenir $M-c$ comme valeurs normalisées. Cette constante c est souvent choisie comme la médiane ou la moyenne des valeurs M. Si trop de gènes sont supposés varier, c doit être calculée à partir de contrôles. Kerr *et al.* [69] ainsi que Wolfinger *et al.* [126] proposent d'estimer c grâce à des modèles d'analyse de variance. Cependant, ces méthodes utilisent une correction globale pour l'ensemble des gènes et ne permettent de prendre en compte les problèmes de non linéarité. Yang *et al.* suggèrent ensuite d'introduire des variables telles que l'intensité des spots ou la localisation des gènes sur la puce. Comme le montre la figure 2.6, le log-ratio peut dépendre de l'intensité totale du spot. Une normalisation plus fine consiste alors à calculer c suivant la valeur de A. Les valeurs normalisées $M - c(A)$ sont obtenues en estimant la fonction c à l'aide d'une méthode de régression non linéaire. Ils proposent une méthode de régression locale [32], appelée *lowess* (*locally weighted scatterplot smoothing* [33], [34]). Pour l'intensité A_i, la correction $c(A_i)$ est obtenue par une méthode des moindres carrés pondérés à partir d'un ensemble de q points dans le voisinage de A_i. À partir du voisiange considéré, on estime les paramètres β_0 et β_1 qui minimisent

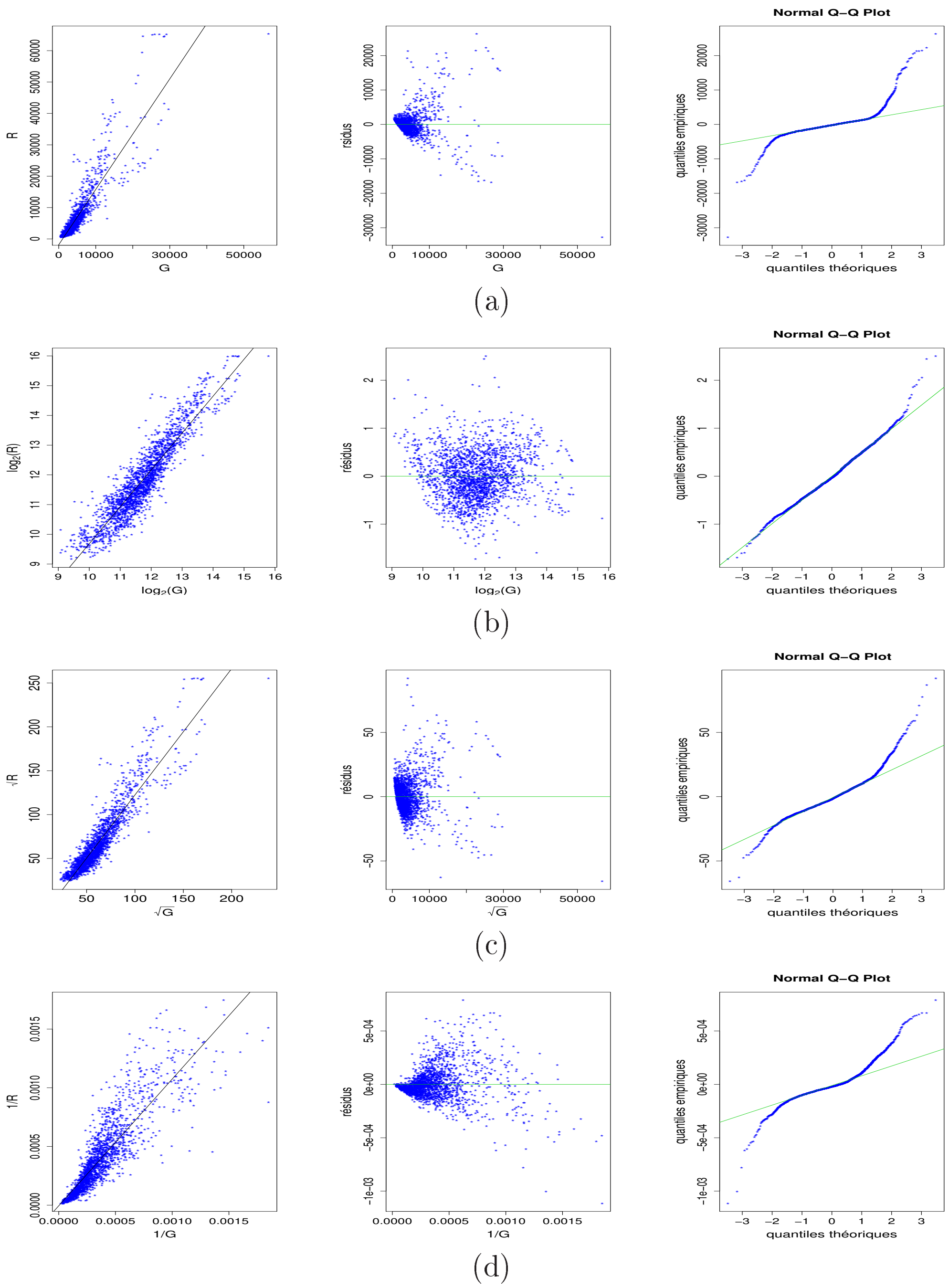


FIG. 2.5 – *Différentes fonctions de lien pour une normalisation linéaire entre R et G.* De toutes les transformations présentées ((a),(b),(c),(d)) la transformation logarithmique (b) est la seule pour laquelle les hypothèses de la régression linéaire sont acceptables. La régression linéaire sur les données brutes (a) montre que cette solution, encore trop souvent appliquée, est à proscrire.

$$\sum_{j=1}^{q} w_i(A_k)(M_k - \beta_0 - \beta_1 A_k)^2.$$

La fonction de pondération w_i affecte un poids faible aux points (A_k, M_k) pour lesquels la valeur A_k est éloignée de A_i et la valeur M_k est trop importante. Si on pose $w_i(A_k) = 1$ pour l'ensemble des points du voisinage, on retrouve alors le critère des moindres carrés habituel. Les points en dehors du voisinage considéré ont un poids nul. La fonction lowess est contrôlée par le paramètre f $(0 < f \leq 1)$ qui spécifie la proportion du nuage de points définissant le voisinage pour ajuster la droite des moindres carrés pondérés. Si f croît, on a tendance à lisser d'avantage car les voisinages choisis tendent alors à être tous identiques.

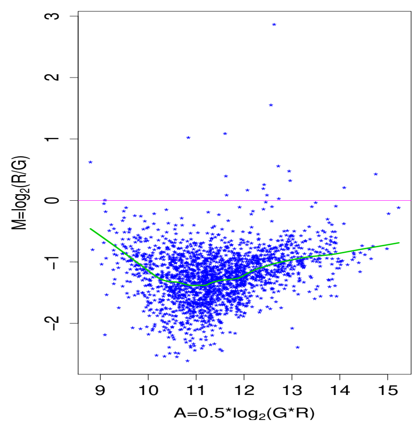


FIG. 2.6 – *Représentation* $M = \log(\frac{R}{G})$ *vs.* $A = 0.5 \times \log(R * G)$. Afin de prendre en compte la forme particulière du nuage, la correction effectuée $c(A)$ dépend de l'intensité du spot et est estimée par une méthode de régression localement linéaire. La valeur normalisée pour un gène k est $M_k - c(A_k)$ où $c(A_k)$ est la valeur estimée par la régression non linéaire.

D'autres solutions, dépendantes du niveau d'intensité, ont également été développées. Finkelstein *et al.* [49] décrivent une méthode itérative, basée sur plusieurs régressions linéaires, qui peut être vue comme une régression linéaire robuste. La fonction c est dans ce cas une fonction linéaire. Kepler *et al.* [67] utilisent une méthode de régression locale, différente de lowess, mais similaire à celle de Yang *et al.*

Une généralisation de cette normalisation consiste à l'appliquer pour chacune des têtes du robot permettant de déposer les sondes. Cela permet de prendre en compte des problèmes liés aux caractéristiques physiques des têtes du robot : différences entre les tailles des têtes, quantités de sondes prélevées et déposées, rinçage, etc. Ces caractéristiques font apparaître des variations systématiques entre les têtes. Afin de montrer que de tels effets ne sont pas négligeables, la figure 2.7(a) représente 15% des gènes dont le ratio est le plus grand. Cette représentation met en évidence des différences significatives entre les têtes de dépôt. Ces problèmes conduisent à considérer une normalisation différente pour l'ensemble des têtes du robot. La solution avancée s'écrit :

$$\log(\frac{R}{G}) - c_i(A) \tag{2.1}$$

où $c_i(A)$ est la correction apportée pour la $i^{\text{ième}}$ tête par une méthode de régression locale telle que lowess. Si i est la $i^{\text{ème}}$ tête de dépôt, le facteur de correction $c_i(A)$ est calculé à partir des sondes déposées par la $i^{\text{ème}}$ tête. Des résultats de cette approche sont présentés dans le paragraphe suivant.

2.5 Applications et comparaisons

La normalisation étant un problème difficile, la mise en place de plusieurs contrôles présente un grand nombre d'avantages. Cela permet d'apprécier la qualité des données ainsi que de valider la normalisation. Bien qu'il soit difficile de prendre en compte directement ces contrôles pour la normalisation, les graphiques définis précédemment permettent des conclusions d'ordre qualitatif. Nous illustrons à partir de résultats expérimentaux les différentes normalisations ainsi que l'intérêt des contrôles dans le protocole expérimental.

2.5.1 Nature des expériences

Le premier type de contrôles (contrôles I) facile à mettre en place correspond à des gènes dont l'expression ne doit pas varier entre les deux conditions. Le ratio d'expression de ces contrôles doit donc être égal à un. Pour cela, il suffit de choisir un ensemble de gènes, d'un organisme différent de celui étudié, et dont les séquences présentent le moins d'homologie possible avec les gènes étudiés. Pour pouvoir fixer a priori l'expression différentielle à un, ces contrôles sont déposés sur la puce avec des concentrations égales pour les deux fluorophores.

La mise en place de contrôles pour l'induction et la répression (contrôle II) reste tout aussi simple. Dans le mélange obtenu partir des ARNs totaux, il suffit d'ajouter des cibles marquées en Cy5 d'une part et des cibles marquées en Cy3 d'autre part. De tels contrôles ne présentent pas un intérêt à proprement dit pour la normalisation mais permettent d'avoir un œil critique sur la qualité des données.

En collaboration avec l'IPMC (Institut de Pharmacologie Cellulaire et Moléculaire, Nice), nous étudions le processus de réparation cellulaire de kératinocytes humaines blessés. L'objectif est d'analyser les changements transcriptionnels impliqués lors de la cicatrisation.

2.5.2 Résultats

Sur l'expérience que nous proposons d'analyser, les deux types de contrôles (I et II) sont utilisés. La soustraction du bruit de fond au signal n'étant pas une bonne approche, nous

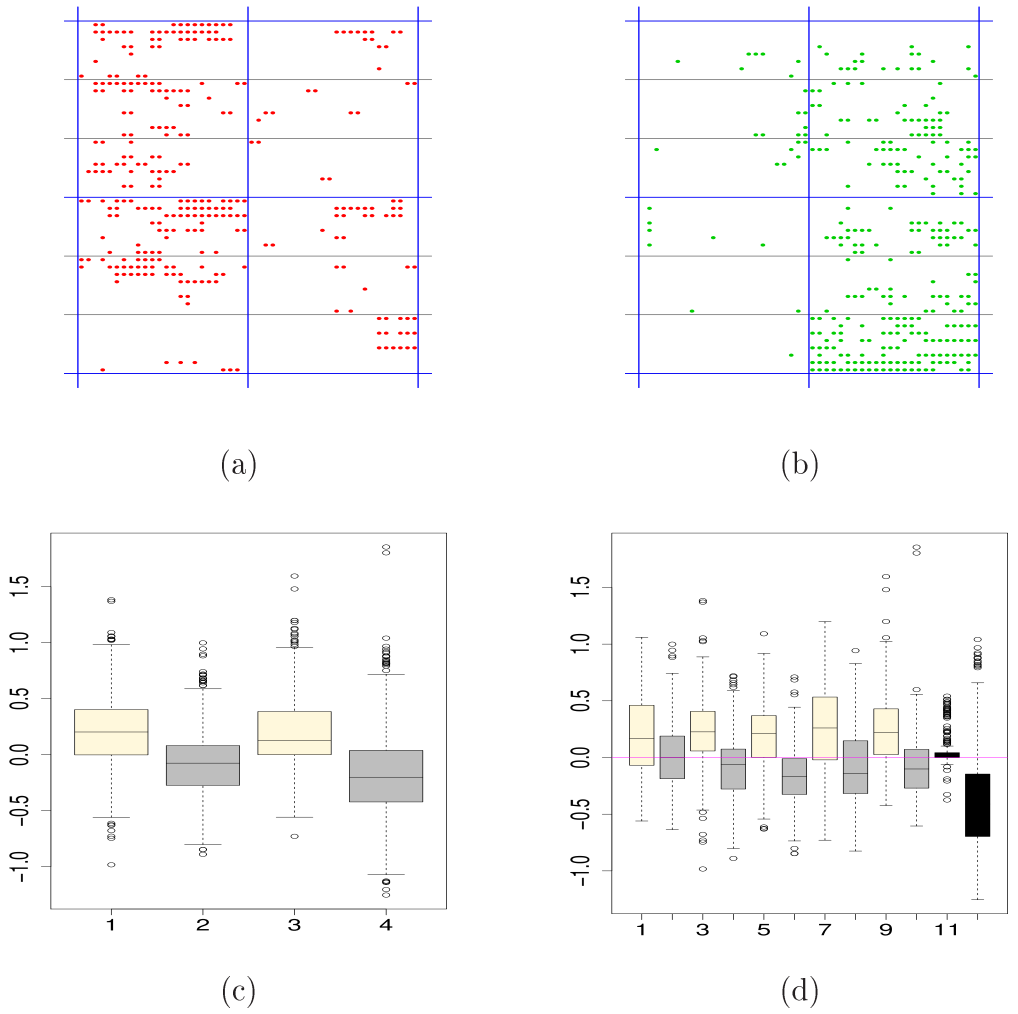


FIG. 2.7 – *Mise en évidence de la variabilité spatiale.* Les figures (a) et (b) donnent 15% des log-ratios les plus forts et les plus faibles respectivement. De façon évidente, les ratios forts sont plus nombreux sur la partie gauche de la puce que sur la partie droite. La représentation boxplot (c) montre effectivement que les têtes de dépôt 1 et 3 d'une part et 2 et 4 d'autre part sont caractéristiques des ratios forts et faibles. Le représentation (d) donne la représentation boxplot pour chacune des plaques à 192 puits. Ces plaques permettent de stocker les sondes avant leur transfert sur la puce et le robot effectue le prélèvement des sondes à partir des ces plaques. Cette représentation traduit ainsi le fait que les sondes provenant des plaques 1, 3, 5, 7 et 9 ont été déposées par les têtes 1 et 3 et les sondes provenant des plaques 2, 4, 6, 8, 10 par les têtes 2 et 4. Sur les figures (a) et (b), les têtes correspondent aux carrés bleus et les plaques à 192 puits aux rectangles noirs. Les sondes des plaques 11 et 12 sont utilisées comme contrôles expérimentaux.

choisissons de mener l'étape de normalisation sur les intensités Cy5 et Cy3, notées R et G respectivement.

2.5.2.1 Les données brutes

La représentation M *vs.* A est présentée sur la figure 2.8(a). La représentation 2.8(b) donne la répartition sur la puce des log-ratios M les plus forts. Les contrôles de type I sont représentés en jaune. En gris, il s'agit de contrôles pour l'induction et la répression (type II). La couleur noire est attribuée aux spots vides.

La première remarque que l'on peut faire est que les contrôles I se situent au centre du nuage, sensiblement en dessous de la droite $y = 0$ (2.8(a)). Ce problème est lié en particulier aux propriétés différentes des fluorophores utilisés. L'hypothèse selon laquelle peu de gènes sont différentiellement exprimés est confirmée par la position des contrôles de type I au centre du nuage. La normalisation va permettre de corriger le biais lié aux marqueurs.

En l'absence de sonde, la fluorescence émise est naturellement faible (valeurs de A petites pour les spots vides). Du fait qu'il ne devrait y avoir aucun signal en Cy5 et en Cy3, les intensités mesurées devrait être égale. Ceci se traduit effectivement par des valeurs de M proche de 0. Les spots vides peuvent être considérés comme des contrôles de type I pour des niveaux faibles d'expression.

Enfin, on constate une répartition particulière des contrôles de type II. Celle-ci s'explique par les problèmes de saturation. En effet, la valeur maximale des pixels est de $2^{16} - 1$. Si le signal en Cy3 est saturé, alors la valeur de M est donnée par la droite d'équation $M = 2A - 32$. À l'inverse, si le signal en Cy5 est saturé, alors la valeur de M est donnée par la droite $M = -2A + 32$. Si le signal est saturé en Cy5 et Cy3 alors $M = 8$. La forme alignée des contrôles de type II traduit donc soit des spots saturés en Cy5, soit des spots saturés en Cy3. On observe aussi que quelques contrôles de type I sont saturés en Cy3.

2.5.2.2 La normalisation

L'application de l'approche linéaire donne des résultats peu satisfaisants, comme le montre la figure 2.9(a), et peut conduire à de mauvaises interprétations. Si on utilise des normalisation globales telles que la moyenne ou la médiane, cela revient simplement à translater le nuage des points sur l'axe $M = 0$. Les résultats seraient donc similaires à ceux de la figure 2.9(a).

Le schéma 2.9(b) donne la distribution des valeurs de M suivant différentes normalisations. Bien que toutes les normalisations permettent de centrer approximativement les valeurs M sur 0, les queues de distribution sont nettement différentes entre une approche linéaire (ou basée sur la médiane) et une approche non linéaire. Or, l'étude des gènes différentiellement exprimés conduit à identifier les gènes dont le log-ratio des niveaux d'expression est fortement négatif ou positif. Les résultats d'une telle étude dépend donc directement du choix de la méthode de normalisation. Ce choix apparaît primordial pour la suite et doit se baser sur des contrôles expérimentaux.

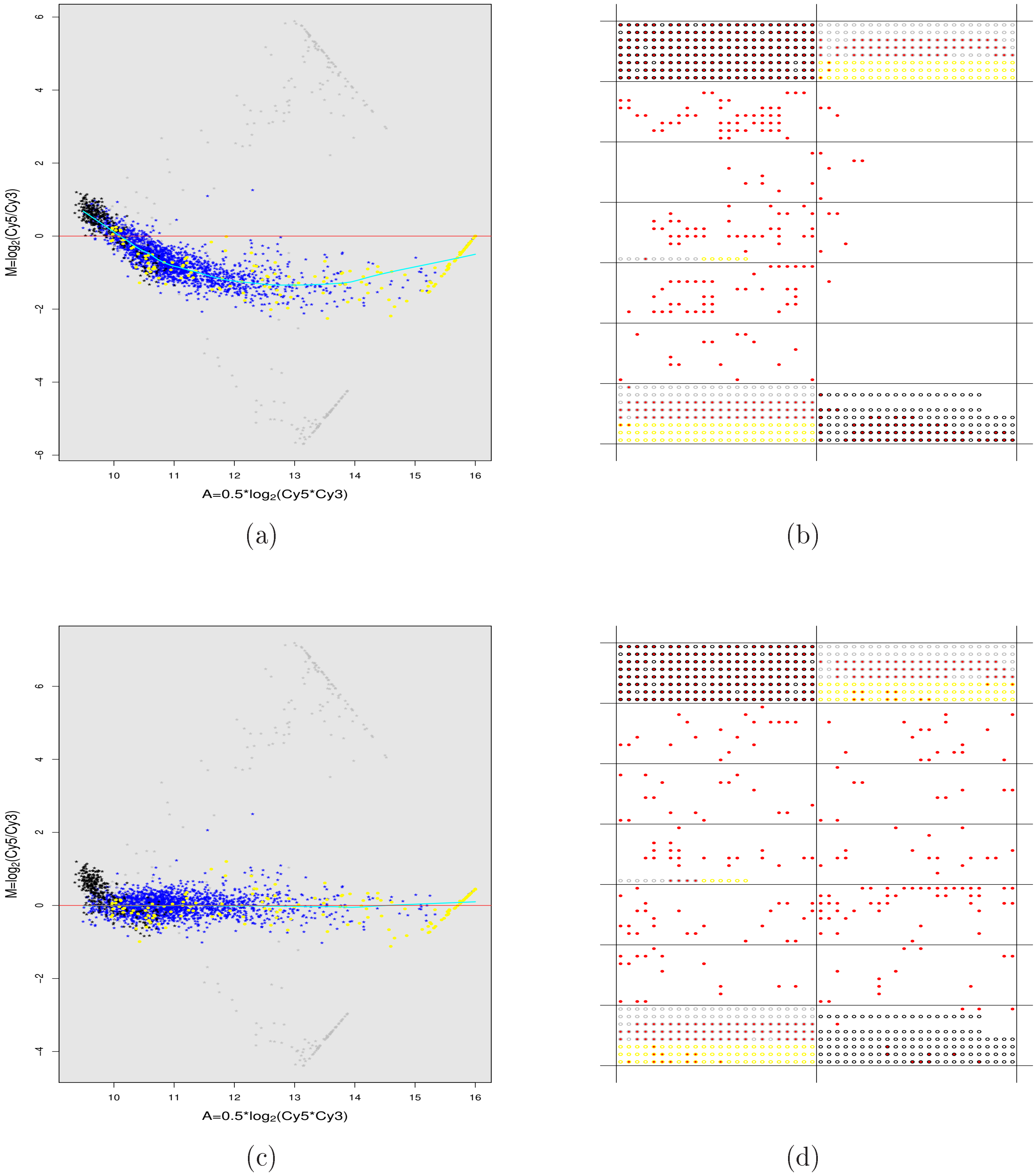
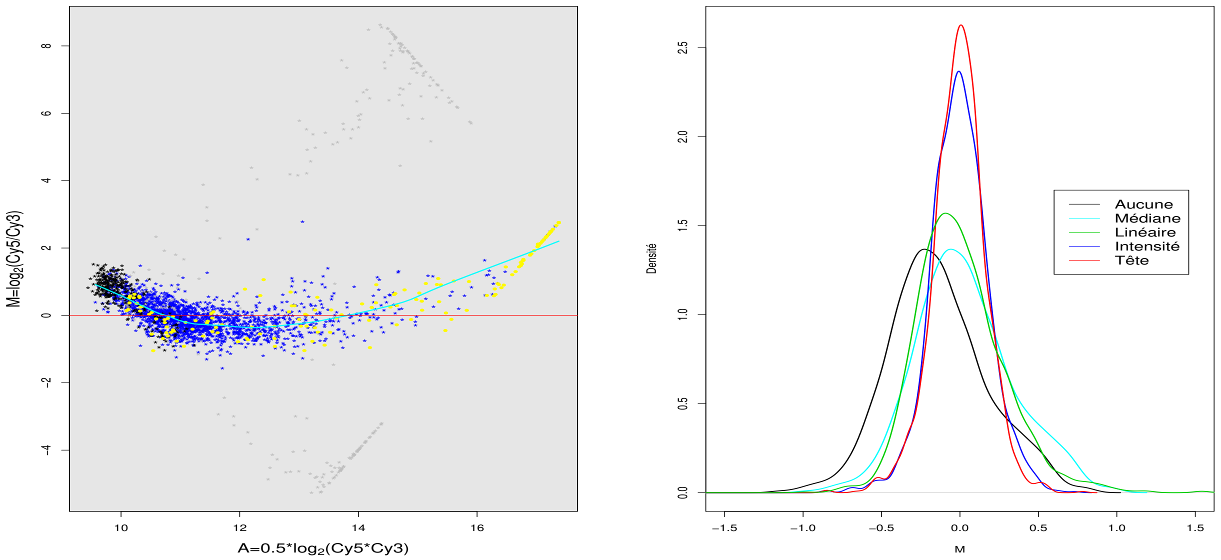


FIG. 2.8 – Résultats de la normalisation de type lowess



(a) (b)

FIG. 2.9 – Comparaison de méthodes de normalisation

Les figures 2.8(c) et (d) donnent le résultat d'une normalisation non linéaire, appliquée séparément pour chacune des têtes de dépôt. Contrairement à la normalisation linéaire 2.9(a), cette correction permet de positionner les contrôles de type I sur l'axe $M = 0$ et donne une meilleure répartition des ratios sur l'ensemble de la puce (figure 2.8(d)).

2.6 Commentaires

Ces résultats traduisent les difficultés importantes liées à la normalisation. Malheureusement, les solutions apportées sont peu satisfaisantes et la validation d'une méthode reste délicate. Dans ces conditions, des méthodes raisonnables consistent à définir autant que possible des ensembles de contrôles pour analyser tout d'abord les problèmes expérimentaux et ensuite pour s'assurer du bon sens de la normalisation.

En raison des différents problèmes que nous avons pu rencontrer lors de l'étape de normalisation, il est nécessaire de définir des plans de dépôt des sondes en introduisant un ensemble important de contrôles.

- Les contrôles pour l'induction, la répression et la non différence d'expression sont en général simples à mettre en place. Ils nous ont permis de remettre en question la normalisation linéaire sur nos applications. Nous recommandons leur mise en place systématique dans les protocoles expérimentaux.
- Sur les résultats présentés, les contrôles étaient déposés par une seule des têtes du

robot. Une meilleure répartition semble souhaitable.
- Enfin, la réalisation d'un ensemble de contrôles dont le rapport des concentrations en Cy5 et Cy3 serait connu entre les deux échantillons pourrait aussi être envisagée. L'utilisation de contrôles avec des gammes différentes d'intensité (meilleure répartition des valeurs A) permettrait d'éviter d'obtenir trop de spots saturés pour les contrôles de type I et II.

N'ayant pas étudié les problèmes de normalisation, nous choisissons une normalisation de type *lowess* pour chacune des têtes, afin de supprimer les problèmes liés à l'intensité du spot. Bien que arbitraire, ce choix nous paraît le moins mauvais. La normalisation reste un problème mal maîtrisé qui risque d'entacher largement les analyses ultérieures. En fait, on peut penser que les résultats d'une analyse de biopuces sont largement liés à la méthode de normalisation. De nombreux développements doivent donc être réalisés sur ces problèmes de normalisation.

Jusqu'à présent, nous avons décrit G et R comme des niveaux d'intensité en Cy3 et Cy5. Or, dans nos études les puces à ADN doivent permettre d'étudier les différences d'expression des gènes entre différentes conditions. Pour franchir ce pas, on admet en général que l'intensité en Cy5 ou Cy3 est proportionnelle au niveau d'expression du gène et par abus de langage, R et G sont les niveaux d'expression des gènes dans les conditions respectives. Cette convention n'a de sens que lorsque les données ont été normalisées. C'est pour cette raison que nous avons défini jusqu'à présent R et G comme des niveaux d'intensité. Dans la suite, nous parlerons donc des niveaux d'expression des gènes en faisant référence aux valeurs R et G normalisées.

Chapitre 3

Recherche de gènes différentiellement exprimés

Pour chaque gène déposé sur la puce[1], on obtient deux mesures normalisées de fluorescence, G (mesure en Cy3) et R (mesure en Cy5), représentant le niveau d'expression des gènes. La recherche de gènes différentiellement exprimés reste probablement la première question à laquelle le biologiste souhaite apporter une réponse (Callow *et al.* [25], Fridle *et al.* [52], Galitski *et al.* [54], Spellman *et al.* [114], Golub *et al.* [57]). Ce problème consiste à identifier les gènes pour lesquels on observe un changement d'expression entre les deux conditions expérimentales. Il peut s'agir alors soit d'une induction, soit d'une répression du gène par rapport à la condition de référence. Les premières expériences de puces à ADN n'ont cependant pas été réalisées dans ce sens, mais plutôt pour l'étude des profils d'expression. Cette question sera abordée dans le prochain chapitre. Mais la recherche de gènes différentiellement exprimés est importante pour l'interprétation des profils. De plus, c'est précisément les problèmes liés à cette problématique qui nous ont conduit à développer une méthode particulière pour l'étude des profils.

La principale difficulté de ce problème tient certainement à la variabilité des données d'expression ainsi qu'au manque de mesures pour un gène. Les premières approches étaient simplement basées sur la valeur des ratios R/G. Par exemple, Schena *et al.* [108, 109] considèrent comme différentiellement exprimés les gènes dont le ratio est inférieur à $1/5$ ou supérieure à 5. DeRisi *et al.* [37] proposent de fixer le seuil du log-ratio des expressions à $\pm\log(3)$. Le besoin de définir un seuil ou une approche rigoureuse traduisant une différence significative entre les niveaux d'expression a motivé de nombreux travaux. Nous nous efforçons dans la suite de présenter les principales approches pour ce problème. Il faut préciser que le développement de ces méthodes est fortement lié aux plans d'expériences ainsi qu'à la mise en place progressive de répétitions dans les protocoles expérimentaux. Afin de présenter ces méthodes, nous distinguons trois catégories. La première, basée sur des tests d'hypothèses, propose d'identifier les gènes différentiellement exprimés en comparant leurs niveaux d'ex-

[1]Nous utilisons la terminologie gène pour désigner tous les spots de la puce, bien qu'il n'y ait pas toujours correspondance entre un gène et un spot. Il peut aussi s'agir de séquence d'EST par exemple.

pression G et R. Ces approches fréquentistes nécessitent de disposer dans tous les cas d'un nombre plus ou moins important de répétitions. La seconde vise à caractériser les groupes de gènes qui sont différentiellement exprimés ou non. Pour cela, les niveaux d'expression sont représentés à l'aide de variables aléatoires. Enfin, on souhaite dans la dernière pouvoir identifier différentes sources de variabilité qui affectent les mesures.

3.1 Les tests d'hypothèses

Étant donné deux conditions 1 (Cy3) et 2 (Cy5), on note y_{gkr_i} le niveau d'expression pour la répétition r_i du gène g dans la condition k et N le nombre de gènes de l'étude. Pour chaque gène, on réalise n_1 répétitions dans la condition 1 et n_2 dans la condition 2. Le marquage fluorescent implique nécessairement $n_1 = n_2 = n$. On note aussi μ_{gk} le "vrai" niveau d'expression du gène g dans la condition k. L'hypothèse que le gène g n'est pas différentiellement exprimé dans les deux conditions est équivalente à poser $H_{0,g} : \mu_{g1} = \mu_{g2}$ alors qu'une différence d'expression s'écrit $H_{1,g} : \mu_{g1} \neq \mu_{g2}$. Afin d'identifier l'ensemble des gènes qui sont différentiellement exprimés, on réalise pour $g = 1, \dots, N$ les tests d'hypothèses avec $H_{0,g}$ l'hypothèse nulle et $H_{1,g}$ l'hypothèse alternative. Les gènes différentiellement exprimés sont alors ceux pour lesquels l'hypothèse $H_{0,g}$ est rejetée. Par convention, on note H_g l'hypothèse nulle $H_{0,g}$. Il s'agit de définir la procédure de test pour réaliser l'ensemble des tests $H_1, \dots, H_N$.

3.1.1 Le test de Student

Pour réaliser ces N tests, la statistique usuelle est

$$t_g = \frac{\bar{y}_{g2} - \bar{y}_{g1}}{\sigma_g} \tag{3.1}$$

où $\bar{y}_{g2}$ et $\bar{y}_{g1}$ sont les moyennes des niveaux d'expression dans les conditions 2 et 1 respectivement et σ_g est l'écart-type estimé de la différence des moyennes. Si on admet que les deux niveaux d'expression sont indépendants alors

$$\sigma_g^2 = \frac{s_{1g}^2}{n_1} + \frac{s_{2g}^2}{n_2} \tag{3.2}$$

avec $s_{1g}^2 = \frac{\sum_{r_i}(y_{g1r_i} - \bar{y}_{g1})^2}{n_1 - 1}$ et $s_{2g}^2 = \frac{\sum_{r_i}(y_{g2r_i} - \bar{y}_{g2})^2}{n_2 - 1}$. Dans la suite, nous notons T_g et t_g la variable aléatoire et la réalisation de la statistique de test. Il est important de noter que les réplications sont nécessaires pour une telle analyse et qu'en raison de coûts expérimentaux, leur nombre est parfois trop faible pour appliquer ces techniques. Ce point limite donc l'application de ces méthodes.

L'approche classique pour réaliser ces tests consiste à se placer sous l'hypothèse gaussienne et d'indépendance des variables Y_{g1} et Y_{g2} qui représentent le niveau d'expression des gènes g_1 et g_2. $\{y_{g1r_i}\}_{r_i\in\{1,\ldots,n_1\}}$ sont n_1 observations de la variables Y_{g1} et $\{y_{g2r_i}\}_{r_i\in\{1,\ldots,n_2\}}$ sont les n_2 réalisations de la variable Y_{g2}. La loi de T_g est, sous l'hypothèse $H_{0,g}$ une loi de Student à n_1+n_2-2 degrés de liberté et correspond au test traditionnel de comparaison de moyennes [105]. L'acceptation ou le rejet de l'hypothèse $H_{0,g}$ se base sur le calcul de la valeur critique (ou p-valeur) $P_{H_{0,g}}(|T_g| \geq |t_g|)$. Celle-ci correspond à la probabilité de l'événement $\{|T_g| \geq |t_g|\}$ sous l'hypothèse que $H_{0,g}$ soit vraie. Cette approche n'est pas sans inconvénient. Un premier problème évident concerne le faible nombre de répétions qui conduit souvent à de mauvaises estimations des variances. Baldi *et al.* [8] donnent un modèle bayésien pour pallier le manque de répétitions. La seconde difficulté est liée à l'utilisation répétée des tests et conduit aux problématiques des tests multiples. Nous détaillons dans la suite la cadre particulier d'une procédure de tests multiples.

3.1.2 Les tests de comparaisons multiples

L'utilisation des tests d'hypothèses pour la recherche de gènes différentiellement exprimés conduit à réaliser un test pour chaque gène. Pour cela, il faut alors s'intéresser aux méthodes dites de tests multiples [113, 125].

3.1.2.1 Les différentes erreurs dans les tests multiples

L'erreur de première espèce pour un test simple est appelée dans le cadre des tests multiples *comparison error rate* (CER). Elle correspond à la probabilité de rejeter une hypothèse nulle $H_i, i = 1, \ldots, N$ alors qu'elle est vraie. L'erreur CER pour le test H_i est définie par

$$CER_i = P(\text{rejeter } H_i | H_i \text{ est vraie}), \quad i = 1, \ldots, N.$$

Pour simplifier, nous pouvons supposer que toutes les erreurs CER_i sont égales à α. La probabilité de ne pas rejeter une hypothèse nulle H_i alors quelle est vraie vaut donc $1 - \alpha$. En supposant les N tests indépendants, la probabilité de ne pas rejeter l'ensemble des hypothèses H_i $(i = 1, \ldots, N)$ alors quelles sont vraies est $(1-\alpha)^N$. Par conséquent, la probabilité de rejeter au moins une hypothèse nulle alors qu'elles sont toutes vraies vaut $1-(1-\alpha)^N$ et se note EER (*experimentwise error rate*) ou FWE (*family wise error rate*). On écrit

$$FWE = P(\text{rejeter au moins une des hypothèses } H_1, \ldots, H_N | H_1, \ldots, H_N \text{ sont vraies}).$$

Si le nombre N de tests augmente, l'erreur EER aussi. Par exemple, si on suppose toutes les erreurs CER égales à $\alpha = 0.05$ et que l'on étudie 100 gènes, alors l'erreur EER vaut 0.994. Le nombre de faux positifs attendus est 100*0.05=5.

Dans le cas où les tests ne sont pas indépendants, le calcul de l'erreur EER dépend de la structure de corrélation des tests et il n'existe plus de formule simple. Il faut préciser que l'hypothèse selon laquelle toutes les hypothèses nulles sont vraies simultanément présente peut d'intérêt en pratique. L'erreur MEER (*maximum experimentwise error rate*) est la probabilité de rejeter à tort au moins une hypothèse H_i alors qu'elle est vraie, indépendamment du fait que d'autres hypothèses nulles soient vraies ou pas. Par conséquent, une procédure de test qui contrôle l'erreur MEER contrôle aussi l'erreur EER, mais la réciproque est fausse. Le contrôle de l'erreur MEER est une meilleure protection contre de mauvaises conclusions et conduit à une inférence statistique plus robuste. On parle parfois d'un contrôle fort de l'erreur FWE pour indiquer que l'on s'intéresse à l'erreur MEER et d'un contrôle faible de l'erreur FWE lorsque toutes les hypothèses nulles sont supposées vraies. Le dernier type d'erreur que nous considérons est l'erreur FDR (*false discovery rate*), introduite récemment par Benjamini *et al.* [11]. Cette erreur correspond à la proportion attendue d'hypothèses nulles rejetées à tort (c'est-à-dire rejeter l'hypothèse nulle alors qu'elle est vraie). Il s'agit donc, parmi les hypothèses nulles rejetées, de la proportion d'hypothèses qui est rejetée à tort. L'application des procédures de tests multiples pour l'analyse de données de biopuces concerne le contrôle des deux types d'erreur FWE et FDR. Dans le premier cas, citons par exemple les travaux de Dudoit *et al.* [40] ainsi que de Callow *et al.* [25]. L'étude de l'erreur FDR a suscité un nombre plus important de travaux [120, 74, 116, 115, 73].

3.1.2.2 L'erreur FWE ou MEER

Dans le cadre des tests simples, l'acceptation ou le rejet du test est basé sur le calcul de la valeur critique. La notion de p-valeur obtenue lors d'un test simple peut être généralisée dans le cadre des tests multiples. On définit pour cela les *p-valeurs ajustées*. Ces valeurs sont interprétables sur la même échelle que celles obtenues pour des tests simples et peuvent donc être comparées. Nous présentons quelques notions sur le calcul des *p-valeurs ajustées* suivant le type d'erreur envisagé.

Pour chacun des tests H_i, ($i \in \{1, \ldots, n\}$), on définit une statistique T_i. On note t_i la valeur observée de la statistique, p_i la p-valeur du test simple et $\tilde{p}_i$ la p-valeur ajustée. Pour simplifier les notations, nous supposons que les p-valeurs sont ordonnées avec $p_1 \leq p_2 \leq \ldots \leq p_G$. On note enfin α l'erreur MEER que l'on souhaite contrôler.

La méthode de Bonferroni [15] est probablement la plus connue et la plus utilisée pour le calcul des p-valeurs ajustées. Étant donné N tests, les p-valeurs ajustées sont données par $\tilde{p}_i = min(Np_i, 1)$. Cela équivaut à dire que la p-valeur ajustée est la p-valeur obtenue pour un test simple dont l'erreur de première espèce est fixée à α/N. Cette méthode est simple et applicable dans toutes les situations de tests multiples. Cependant, elle est souvent mal appropriée et tout particulièrement si le nombre de tests est grand. Cette correction devrait être utilisée lorsque le nombre de tests est faible (moins de 10) et lorsque les corrélations entre les statistiques de test sont faibles. D'autres approches plus intéressantes tiennent compte de l'ordre des p-valeurs et sont moins conservatrices. Une correction bien connue, inspirée de celle de Bonferroni, est la procédure de Holm [63] qui donne une méthode plus puissante [3]. Elle commence par comparer p_1 avec α/N. Si $p_1 \geq \alpha/N$, alors $H_1, \ldots, H_N$ sont acceptées. Sinon, p_2 est comparée à $\alpha/(N-1)$. Si $p_2 \geq \alpha/(N-1)$ alors $H_2, \ldots, H_N$ sont acceptées. Si H_2 est rejetée, la procédure continue.

Il existe d'autres modifications de la procédure de Bonferroni mais elle tendent toutes à être trop conservatrices et ne prennent pas en compte les structures de corrélation qui peuvent exister. Or, les niveaux d'expression de certains gènes, comme des gènes corégulés, peuvent être corrélés. L'application de tests multiples doit donc intégrer de telles configurations.

Westfall *et al.* [125] décrivent une procédure de rééchantillonage qui permet d'obtenir des p-valeurs ajustées en prenant en compte les corrélations entre les différentes statistiques de test. Pour des tests très corrélés, on peut montrer que cette procédure est nettement plus puissante que celles décrites précédemment. Callow *et al.* [25] l'appliquent pour l'étude de gènes différentiellement exprimés chez des souris transgéniques. Dudoit *et al.* [40] détaillent les aspects statistiques de cette application qui garantit donc que la probabilité d'obtenir au moins un faux positif est inférieure au seuil α choisi pour l'erreur FWE.

3.1.2.3 L'erreur FDR

Toutefois, les méthodes contrôlant l'erreur FWE peuvent être trop conservatrices pour des applications où peu de gènes sont différentiellement exprimés. Mais d'un autre coté, ne faire aucune correction peut entraîner un grand nombre de faux positifs. Un compromis raisonnable est de tolérer quelques faux positifs en contrôlant l'erreur FDR, c'est-à-dire la proportion de faux positifs parmi les gènes attendus comme différentiellement exprimés. De nombreux travaux se sont intéressés à cette erreur dans le cadre de l'analyse de données de biopuces.

Tusher *et al.* [120] décrivent une approche, SAM (*signifiance analysis of microarrays*), qui attribue à chaque gène un score à partir des mesures répétées (nous préciserons quel est ce score dans la section suivante). Leur approche consiste alors à choisir un seuil sur ce score pour identifier un ensemble de gènes différentiellement exprimés. Pour l'ensemble de ces gènes, cette procédure utilise des permutations sur les mesures répétées afin d'estimer le pourcentage de gènes sélectionnés par hasard comme différentiellement exprimés, c'est-à-dire l'erreur FDR. Ainsi, le seuil du score est choisi en fonction de l'erreur FDR que l'on tolère. Un aspect intéressant est que l'application de cette méthode ne nécessite pas un nombre trop important de répétitions. Dans leur application, Tusher *et al.* décrivent une expérience comportant simplement quatre répétitions par gène, obtenues à partir de quatre puces différentes. Ils comparent par ailleurs leur méthode à celle décrite par Westfall *et al.* et qui permet de contrôler l'erreur FWE. Cette dernière ne permettait d'identifier aucun gène différentiellement exprimés. Ceci semble indiquer que cette procédure est trop conservatrice lorsque peu de gènes sont différentiellement exprimés.

Korn *et al.* [74] donnent deux procédures intéressantes. Ils proposent de donner un intervalle de confiance sur le nombre de faux positifs obtenus ainsi que sur la proportion de faux positifs. La première procédure permet d'obtenir un nombre de faux positifs inférieur ou égal à une valeur fixée u avec un niveau de confiance donné. La seconde procédure permet de contrôler de manière similaire la proportion de faux positifs, qui est définie par le nombre d'hypothèses nulles rejetées à tort divisé par le nombre d'hypothèses rejetées. Cette proportion est notée FDP et son espérance est l'erreur FDR que nous venons de décrire.

Les travaux de Storey [116, 115] donnent des résultats plus complets sur l'étude théorique de l'erreur FDR ainsi que sur son application dans le cadre de l'analyse de données de puces à ADN.

3.1.2.4 Choix de la statistique de test

Dans leur approche, Tusher *et al.* et Efron *et al.* [45] utilisent une statistique de test légèrement modifiée par rapport à (3.1). Dans le cas où la somme des carrés sur les répétitions est faible, la variance estimée σ_g^2 (3.2) sera aussi faible, indépendamment du fait que la valeur $|\bar{y}_{g2} - \bar{y}_{g1}|$ soit faible ou grande. Cela implique alors des valeurs absolues grandes pour la statistique de test même s'il n'y a pas de différence d'expression. Ce problème s'observe plus particulièrement pour des niveaux faibles d'expression. Suivant les notations de la formule (3.2), on envisage alors le score suivant

$$t_g = \frac{\bar{y}_{g2} - \bar{y}_{g1}}{\sigma_g + a_0}. \tag{3.3}$$

où a_0 est une constante permettant de régulariser le score. D'autres travaux choisissent également cette modification. Efron *et al.* [44, 45] suggèrent de prendre pour a_0 le $9^{\text{ème}}$ décile de la distribution des écarts-types de tous les gènes (à noter toutefois qu'à la différence de Tusher *et al.*, Efron *et al.* n'utilisent pas de test d'hypothèses). Claverie [31] soulevait également ce problème, en précisant qu'il pouvait être d'autant plus important que le niveau d'expression des gènes étaient faibles. Pour tenir compte du niveau d'expression des gènes dans la recherche de gènes différentiellement exprimés, Ge *et al.* [1] donnent une approche intégrant la différence absolue des niveaux d'expression, le log-ratio des expressions ainsi que la variabilité des répétitions.

3.2 Approches probabilistes

Les premiers résultats de puces à ADN dont nous disposions ne permettaient pas l'utilisation des tests précédents en raison de l'absence de répétition. Dans une telle configuration, une solution envisagée est de modéliser, à l'aide de lois de probabilité, les niveaux d'expression des gènes. Chen *et al.* [26] sont certainement les premiers à proposer cette approche. Dans leurs travaux, ils utilisent le ratio des niveaux d'expression pour identifier les gènes différentiellement exprimés, et montrent comment, à partir d'une estimation de la loi de distribution des ratios, il est possible de proposer un test d'hypothèses pour résoudre ce problème.

L'une des limites de cette méthode est qu'elle ne s'intéresse qu'aux ratios, oubliant ainsi les niveaux d'expression. Newton *et al.* [91] décrivent une solution directement basée sur les niveaux d'expression. De cette manière, leur modèle traduit qu'il est d'autant plus difficile d'identifier un gène différentiellement exprimé que son niveau d'expression est faible. En nous appuyant sur les travaux de Newton *et al.*, nous développons un modèle qui permet de représenter différemment l'induction, la répression et l'absence de différence d'expression. L'un des objectifs est de comprendre s'il est utile de considérer ces trois cas de figure. Cette question n'ayant jamais été soulevée dans des travaux précédents, il nous semble intéressant d'y apporter des éléments de réponse.

Avant de détailler ces approches, nous donnons quelques notations. On considère que chaque spot de la puce correspond à un gène unique. Pour le spot k de la puce ($1 \leq k \leq N$), on note R_k et G_k les variables aléatoires représentant les niveaux d'expression pour le Cy5 et le Cy3 respectivement, μ_{R_k} et μ_{G_k} étant les vrais niveaux d'expression du gène k.

3.2.1 Le rapport des niveaux d'expression

Une approche intuitive pour identifier les gènes différentiellement exprimés est de s'intéresser aux rapports des niveaux d'expression des deux conditions expérimentales. Cette idée fut d'ailleurs la première méthode appliquée par les biologistes. En choisissant des seuils arbitraires, les gènes sont considérés comme induits ou réprimés si le ratio est supérieur ou inférieur à ce seuil. Chen *et al.* [26] introduisent un cadre cohérent en proposant une estimation de la densité des rapports.

On souhaite identifier les gènes différentiellement exprimés en réalisant un test

$$H_0 : \mu_{R_k} = \mu_{G_k}, \quad H_1 : \mu_{R_k} \neq \mu_{G_k}$$

en utilisant le ratio $T_k = R_k/G_k$. La démarche pour aborder ce problème est de calculer la loi de T_k sous l'hypothèse H_0. La solution proposée par Chen *et al.* repose sur un certains nombre d'hypothèses ainsi que sur certaines approximations numériques.

La première hypothèse est que le coefficient de variation est constant pour l'ensemble des gènes étudiés. Cela équivaut à supposer que

$$\sigma_{R_k} = c\mu_{R_k}, \quad \sigma_{G_k} = c\mu_{G_k}. \tag{3.4}$$

Sous l'hypothèse H_0, on a alors $\sigma_{R_k} = \sigma_{G_k} = c\mu_k$ avec $\mu_k = \mu_{R_k} = \mu_{G_k}$.

La seconde hypothèse tient aux lois de probabilité utilisées pour représenter les niveaux d'expression. On suppose que les variables R_k et G_k sont indépendantes et identiquement distribuées avec $R_k \sim \mathcal{N}(\mu_{R_k}, \sigma^2_{R_k})$ et $G_k \sim \mathcal{N}(\mu_{G_k}, \sigma^2_{G_k})$. En s'appuyant sur ces hypothèses, il est alors possible d'estimer la loi de T_k sous l'hypothèse H_0.

3.2.1.1 Loi de $T_k = R_k/G_k$ sous H_0

Pour calculer la loi de T_k, il est nécessaire de connaître la loi du ratio de deux variables aléatoires. Si X et Y sont deux variables aléatoires continues et indépendantes, la loi de $T = X/Y$ vaut

$$f_T(t) = \int_0^{\infty} y f_X(ty) f_Y(y)\, dy - \int_{-\infty}^{0} y f_X(ty) f_Y(y)\, dy. \tag{3.5}$$

En appliquant le résultat (3.5) sous les hypothèses précédentes, on obtient

$$f_{T_k}(t) = \int_0^{\infty} g f_{R_k}(tg) f_{G_k}(g)\, dg - \underbrace{\int_{-\infty}^{0} g f_{R_k}(tg) f_{G_k}(g)\, dg}_{\simeq 0}. \qquad (3.6)$$

Les niveaux d'expression étant positifs, Chen *et al.* admettent que le second terme de l'expression 3.6 est négligeable. La loi de T_k peut ainsi être approximée par

$$f_{T_k}(t) \simeq \int_0^{\infty} \frac{1}{\sigma_{R_k}\sqrt{2\pi}} \exp\left(-\frac{(tg-\mu_{R_k})^2}{2\sigma_{R_k}^2}\right) \times \frac{1}{\sigma_{G_k}\sqrt{2\pi}} \exp\left(-\frac{(g-\mu_{G_k})^2}{2\sigma_{G_k}^2}\right) dg. \qquad (3.7)$$

Sous l'hypothèse H_0, cette approximation donne alors

$$f_{T_k|H_0}(t) \simeq \frac{1}{2\pi c^2} \int_0^{\infty} \exp\left(-\frac{(tu-1)^2}{2c^2}\right) \times \exp\left(-\frac{(u-1)^2}{2c^2}\right) u\, du. \qquad (3.8)$$

Il n'existe pas de solution explicite de (3.8). Une seconde approximation consiste alors à intégrer sur $]-\infty,\infty[$. Enfin, par le changement de variable $u = g/\mu_k$, on obtient l'approximation

$$\begin{aligned} f_{T_k|H_0}(t) &\simeq \frac{1}{2\pi c^2} \int_{-\infty}^{\infty} \exp\left(-\frac{(tu-1)^2}{2c^2}\right) \times \exp\left(-\frac{(u-1)^2}{2c^2}\right) u\, du \\ &\simeq \frac{(1+t)\sqrt{1+t^2}}{c(1+t^2)^2\sqrt{2\pi}} \exp\left(-\frac{(t-1)^2}{2c^2(1+t^2)}\right). \end{aligned} \qquad (3.9)$$

Le résultat intéressant de cette approximation est que la loi de T_k sous l'hypothèse H_0 ne dépend ni de μ_k, ni de σ_k. Le seul paramètre de la densité $f_{T_k|H_0}(t)$ est le paramètre c. On note donc $f_{T_k|H_0}(t)$ par $f_T(t;c)$.

3.2.1.2 Estimation du paramètre du modèle

Suivant toutes ces hypothèses et ces approximations, le problème nécessite alors d'estimer le paramètre c de la loi $f_T(t;c)$. Cette estimation doit nécessairement être réalisée à partir de gènes dont on sait que l'expression n'est pas modifiée entre les deux conditions expérimentales. En effet, la loi $f_T(t;c)$ a été obtenue sous l'hypothèse H_0 qu'il n'y pas de différence d'expression. Admettant que peu de gènes sont différentiellement exprimés, nous estimerons c à partir de l'ensemble des données. Dans un cas différent, cette estimation doit être réalisée à partir de contrôles.

Supposant les observations indépendantes, la fonction de vraisemblance $\mathcal{L}$ pour le paramètre c s'écrit

$$\mathcal{L}(c) = \prod_{i=1}^{n} \frac{(1+t_i)\sqrt{1+t_i^2}}{c(1+t_i^2)^2\sqrt{2\pi}} \exp\left(-\frac{(t_i-1)^2}{2c^2(1+t_i^2)}\right) \tag{3.10}$$

où $t_1, \ldots, t_n$ sont les ratios pour l'ensemble des gènes considérés comme invariants.

Finalement, l'estimateur du maximum de vraisemblance est donné par

$$\hat{c} = \left[\frac{1}{n}\sum_{i=1}^{n} \frac{(t_i-1)^2}{(1+t_i^2)}\right]^{1/2} . \tag{3.11}$$

3.2.1.3 Gènes différentiellement exprimés

Ayant estimé la loi de la statistique T_k sous l'hypothèse H_0, et calculé l'ensemble des ratios $t_k = r_k/g_k$, on peut alors identifier les gènes différentiellement exprimés en réalisant le test d'hypothèses

$$H_0 : \mu_{R_k} = \mu_{G_k} \;\;,\;\; H_1 : \mu_{R_k} \neq \mu_{G_k}.$$

En notant α l'erreur de première espèce, on rejette H_0 si $t_k \geq q_{1-\alpha/2}$ ou si $t_k \leq q_{\alpha/2}$. Les valeurs $q_{1-\alpha/2}$ et $q_{\alpha/2}$ peuvent être calculées à partir de la loi $f_T(t;c)$. Chen *et al.* montrent qu'elles peuvent être estimées simplement à partir de l'estimation de c et suivant une fonction polynômiale du type

$$y = a_3c^3 + a_2c^2 + a_1c + a_0 \tag{3.12}$$

avec y représentant les valeurs inférieures ou supérieures de l'intervalle de confiance $[q_{\alpha/2}, q_{1-\alpha/2}]$ et c le coefficient de variation. Chen *et al.* donnent les valeurs (a_3, a_2, a_1, a_0) pour $q_{\alpha/2}$ et $q_{1-\alpha/2}$ suivant que $\alpha = 1\%$ ou $\alpha = 5\%$.

3.2.1.4 Commentaires

Bien que cette approche propose de représenter le niveau d'expression R_k et G_k par des lois de probabilité, la recherche de gènes différentiellement exprimés est en fait basée sur le ratio des expressions. Une remarque immédiate est donc que l'on ne prend pas en compte le niveau d'expression des gènes mais uniquement le rapport. Une meilleure modélisation consisterait donc à s'intéresser directement au niveau d'expression des gènes.

Par ailleurs, l'ensemble des résultats sont basés sur plusieurs approximations. Bien qu'elles semblent raisonnables, l'utilisation de lois définies sur $[0, \infty[$ pourrait permettre d'éviter ces approximations.

3.2.2 Proposition d'une modélisation des niveaux d'expression

Newton *et al.* [91] proposent une modélisation différente de celle de Chen *et al.* Dans un premier temps, ils montrent l'intérêt de la loi Gamma pour représenter les niveaux d'expression. Sous un certains nombre d'hypothèses, leur modélisation permet d'obtenir une estimation de la loi des ratios d'expression et conduit donc à une approche similaire à celle de Chen *et al.* Le second développement proposé consiste à représenter les gènes différentiellement exprimés et non différentiellement exprimés. Cette extension, basée sur un modèle bayésien hiérarchique, montre que l'on peut identifier les gènes différentiellement exprimés en prenant en compte leur niveau d'expression. Nous présentons une extension de ce modèle en montrant comment on peut caractériser différemment les gènes induits et réprimés. Nous verrons aussi que cette approche peut être vue comme un modèle de mélange.

3.2.2.1 Modèle proposé

Dans l'approche précédente, les niveaux d'expression sont représentés par une loi normale, et cela conduit à deux approximations liées au fait que le support de cette loi est $\mathbb{R}$. Comme le montre la figure 2.1, les hypothèses de normalité sur R et G semblent assez peu réalistes en raison de queues de distribution lourde. Pour résoudre en partie ces problèmes, Newton *et al.* proposent une modélisation utilisant la loi Gamma. Elle présente l'avantage d'être facile à manipuler, définie sur des valeurs positives et souple (proche d'une distribution gaussienne pour certaines valeurs des paramètres par exemple). On pose

$$\begin{aligned} R &\sim \mathcal{G}(a, \theta_R) \\ G &\sim \mathcal{G}(a, \theta_G) \end{aligned}$$

avec a le paramètre de forme commun à R et G, et θ_R et θ_G les paramètres d'échelle de R et G. La densité de la variable R s'écrit

$$f(r; a, \theta_R) = \frac{\theta_R^a}{\Gamma(a)} r^{a-1} \exp(-r\theta_R) \qquad \text{avec } r \geq 0. \tag{3.13}$$

À partir de ces hypothèses, il est possible de calculer la loi de $T = R/G$. En posant $\rho = \mu_R/\mu_G = \theta_G/\theta_R$, on obtient alors

$$f_T(t; a, \theta_G, \theta_R) = \frac{\Gamma(2a)}{\Gamma^2(a)} \times \frac{1}{\rho} \times \frac{(t/\rho)^{a-1}}{(1 + t/\rho)^{2a}} \qquad \text{avec } t \geq 0. \tag{3.14}$$

Si on suppose qu'il n'y pas de différence d'expression, c'est-à-dire $H_0 : \mu_R = \mu_G$, alors $\rho = 1$ et $\theta_R = \theta_G$ et

$$f_{T|H_0}(t; c) = \frac{\Gamma(2/c^2)}{\Gamma^2(1/c^2)} \frac{t^{(1/c^2)-1}}{(1+t)^{2/c^2}} \qquad \text{avec } t \geq 0, c = 1/\sqrt{a}. \tag{3.15}$$

Cette expression montre qu'il est possible, à ce stade, de suivre l'approche proposée par Chen *et al.* La loi du ratio sous l'hypothèse H_0 dépend alors uniquement du coefficient de variation $c = 1/\sqrt{a}$. La principale différence avec l'approche de Chen *et al.* est que l'on modélise les niveaux d'expression par des lois Gamma au lieu de lois normales.

Il y a un argument simple qui montre que les ratios d'expression seuls ne donnent pas toute l'information apportée par les deux mesures. Cette remarque est importante car en ne considérant que les ratios d'expression, on suppose alors que la distribution des ratios ne dépend pas du niveau d'expression. Pour montrer l'importance des niveaux d'expression, on peut s'intéresser à la distribution conditionnelle des ratios T sachant $S = R \times G$. La variable S est une combinaison des niveaux d'expression, et en supposant que $\theta_R = \theta_G = \theta$ ($\rho = 1$), on obtient la loi de S

$$f_S(s) = \frac{1}{(\Gamma(a))^2} \theta^{2\alpha} s^{\alpha-1} \int_0^\infty \frac{1}{r} \exp\left\{-\theta(r + \frac{s}{r})\right\} dr.$$

La densité jointe de (T, S) s'écrit

$$f_{(T,S)}(t, s) = \frac{1}{2t(\Gamma(a))^2} \theta^{2\alpha} s^{\alpha-1} \exp\left\{-\theta\sqrt{s}\left[\sqrt{t} + \frac{1}{\sqrt{t}}\right]\right\}.$$

Ceci conduit à la densité conditionnelle $T|S$ avec

$$f_{T|S=s}(t|s) = \frac{f_{(T,S)}(t, s)}{f_S(s)} \propto \frac{1}{t} \exp\left\{-\theta\sqrt{s}\left[\sqrt{t} + \frac{1}{\sqrt{t}}\right]\right\}.$$

La distribution de T présente ainsi une dispersion moins grande pour les grandes valeurs de S. Il est assez intuitif de penser que le ratio est plus variable lorsque les niveaux d'expression sont faibles. Le ratio de deux petites valeurs est plus sensible à de faibles variations du nominateur ou du dénominateur que le ratio de fortes valeurs.

À partir de ces remarques, il n'est donc pas rigoureux de rechercher les gènes différentiellement exprimés en se basant simplement sur le ratio des niveaux d'expression. Newton *et al.* suggèrent alors de plonger cette modélisation dans un cadre bayésien en considérant les paramètres θ_R et θ_G comme des variables aléatoires afin de prendre en compte les fluctuations sur les mesures. Dans la suite, nous montrons comment il est possible d'utiliser leur approche afin de modéliser l'induction, la répression et l'absence de différence d'expression. Le modèle que nous introduisons donne une représentation différente dans les trois cas. Cette distinction paraît a priori intéressante. En effet, un modèle qui représente différemment les gènes induits et réprimés semble plus naturel dans le cas où l'on ne dispose pas de répétition. La distinction de ces deux groupes est plus précise qu'une représentation des gènes différentiellement ou non différentiellement exprimés.

Cas des gènes non différentiellement exprimés

S'il n'y a pas de changement dans l'expression des gènes, alors les niveaux d'expression G et R suivent la même loi. Afin de prendre en compte différentes sources de variabilité, on considère que les paramètres d'échelle θ_G et θ_R sont des réalisations de la même loi exponentielle $\theta_R = \theta_G \sim \mathcal{E}(\nu) = \mathcal{G}(1,\nu)$.

En supposant que les variables G et R sont indépendantes et sous l'hypothèse qu'il n'y a pas de différence d'expression, la loi jointe du couple (G,R) s'écrit

$$\begin{aligned} f_0(g,r) &= \int_0^\infty f_{G,R}(g,r|\theta)\pi(\theta)d\theta = \int_0^\infty f_G(g|\theta)f_R(r|\theta)\pi(\theta)d\theta \\ &= \frac{\Gamma(2a+1)}{\Gamma^2(a)}\,\nu\,\frac{(gr)^{a-1}}{(g+r+\nu)^{2a+1}}. \end{aligned} \tag{3.16, 3.17}$$

Cas des gènes surexprimés

Une surexpression du gène k se traduit par le fait que $r_k > g_k$. Pour l'ensemble des gènes surexprimés (qui ne sont pas connus), on note D^+ la variable $R - G$. Nous proposons de modéliser les gènes surexprimés de la manière suivante

$$G \sim \Gamma(a,\theta_G) \tag{3.18}$$

$$D^+ \sim \Gamma(b,\theta_{D^+}) \tag{3.19}$$

On suppose que les paramètres d'échelle θ_G et θ_{D^+} sont deux lois indépendantes de distribution a priori $\mathcal{E}(\nu)$. Sous la condition d'indépendance entre G et D^+, la loi marginale du couple (G,R) s'écrit :

$$\begin{aligned} f^+(g,r) &= f^+_{G,R}(g,r) \\ &= f^+_{G,D^+}(g,r-g) \\ &= f_G(g) \times f_{D^+}(r-g) \\ &= \int_0^\infty f_G(g|\theta_G)\pi_{\theta_G}(\theta)d\theta \times \int_0^\infty f_{D^+}(r-g|\theta_{D^+})\pi_{\theta_{D^+}}(\theta)d\theta \\ &= (\nu)^2 \times a \times b \times \frac{g^{a-1}(r-g)^{b-1}}{(g+\nu)^{a+1}(r-g+\nu)^{b+1}}\mathbb{1}_{r>g}. \end{aligned} \tag{3.20}$$

Cas des gènes réprimés

De manière similaire au cas précédent, on pose D^- la variable $G - R$. La modélisation de la répression est représentée par :

$$\begin{matrix} R \sim \Gamma(a,\theta_R) \\ D^- \sim \Gamma(b,\theta_{D^-}) \end{matrix} \qquad \text{avec } \theta_R \text{ et } \theta_{D^-} \sim \mathcal{E}(\nu) \tag{3.21}$$

On a sous des hypothèses similaires au cas précédent

$$f^-(g,r) = (\nu)^2 \times a \times b \times \frac{r^{a-1}(g-r)^{b-1}}{(r+\nu)^{a+1}(g-r+\nu)^{b+1}} \mathbb{1}_{g>r}. \quad (3.22)$$

3.2.2.2 Estimation des paramètres du modèle

Dans une démarche bayésienne classique, on pourrait considérer ν comme un hyperparamètre et le fixer a priori. En se plaçant dans un cadre bayésien empirique analogue à celui de Newton *et al.*, nous préférons le considérer comme un paramètre fixe.

Un autre modèle bayésien hiérarchique consiste à poser les valeurs ν, ν^+ et ν^- comme paramètre sur les lois exponentielles. Par ailleurs, nous choisissons d'utiliser une loi exponentielle pour la loi a priori sur le paramètre d'échelle de la loi Gamma. Suivant Newton *et al.*, nous pourrions aussi adopter différentes loi Gamma telles que $\mathcal{G}(a_0,\nu)$, $\mathcal{G}(a_0,\nu^+)$ et $\mathcal{G}(a_0,\nu^-)$. Nous reviendrons sur ces extensions dans les études de simulation.

Le problème présenté par notre modèle est que l'on ne connaît pas l'identité des gènes dont l'expression est modifiée. Le calcul de la vraisemblance pour estimer les paramètres a, b, ν est donc difficile. La solution choisie est d'introduire une variable cachée Z telle que :

- Z vaut 0 si il n'y a pas de différence d'expression et $P(Z=0) = p_0$
- Z vaut 1 si le gène est induit et $P(Z=1) = p^+$
- Z vaut -1 si le gène est réprimé avec $P(Z=-1) = p^-$

On note $\eta = (a, b, \nu, p_0, p^+, p^-)$ le vecteur des paramètres du modèle, sous la condition $p_0 + p^+ + p^- = 1$. On peut alors, à l'aide de l'algorithme EM ([36],[89]) estimer le vecteur de paramètres η.

L'algorithme EM est un algorithme itératif qui permet de calculer le maximum de vraisemblance lorsque la structure des données associée au modèle est incomplète et rend alors le calcul de la vraisemblance difficile. Afin de déterminer les paramètres de notre modèle, on choisit de noter le couple (G, R) par la variable Y et $y \in \mathbb{R}^2$ une réalisation de Y. Les données manquantes, qui sont des réalisations de la variables Z, déterminent à quelle classe appartiennent les gènes. Les données complètes $x = (y, z)$ ne sont donc pas directement observées. Plus précisément, l'observation $y = y(x)$ est la projection d'une réalisation de l'espace χ (espace des données complètes) dans l'espace des observations. Une observation y détermine un sous-ensemble de χ noté $\chi(y)$ qui est l'image réciproque de la projection de χ sur y.

La densité de probabilité des données complètes est :

$$\begin{aligned} f_X(x|\eta) &= f(x|\eta) \\ &= (p_0 f_0(y))^{(1-z)(1+z)} \times (p^+ f^+(y))^{\frac{z(z+1)}{2}} \times (p^- f^-(y))^{\frac{z(z-1)}{2}}. \end{aligned}$$

La densité des données observées est

$$g(y|\eta) = \int_{\chi(y)} f(x|\eta) = p_0 f_0(r,g) + p^+ f^+(r,g) + p^- f^-(r,g).$$

Le principe de l'algorithme EM est de maximiser l'espérance de $\log f(x|\eta)$ sachant y et une valeur courante de η. L'algorithme EM se décompose en deux étapes. On note $\eta^{(i)}$ l'estimation du paramètre η à la $i^{\text{ième}}$ itération de l'algorithme.

- **Étape E : Calcul de l'espérance conditionnelle**
  La première étape de l'algorithme EM consiste à calculer l'espérance des données complètes sachant les données observées et η une valeur courante du paramètre.

$$\mathcal{Q}(\eta|\eta^{(i)}) = E(\log f(x|\eta)|y, \eta^{(i)})) = E(\ell(\eta)|y, \eta^{(i)}). \tag{3.23}$$

Si $x = (x_1, \dots, x_n)$ sont n réalisations indépendantes et identiquement distribuées selon la loi f_X, observées seulement à travers n réalisations $y = (y_1, \dots, y_n)$, la log-vraisemblance complétée s'écrit :

$$\begin{aligned}\ell(\eta) &= \sum_{k=1}^{n} \left[(1 - z_k^2)\log f_0(y_k) + \frac{z_k(z_k+1)}{2}\log f^+(y_k) + \frac{z_k(z_k-1)}{2}\log f^-(y_k)\right] \\ &+ \sum_{k=1}^{n} \left[(1 - z_k^2)\log p_0 + \frac{z_k(z_k+1)}{2}\log p^+ + \frac{z_k(z_k-1)}{2}\log p^-\right]\end{aligned}$$

avec $\eta = (a, b, \nu, p_0, p^+, p^-)$.
Le calcul de l'espérance des variables Z et Z^2 sachant les données observées donne

$$\begin{aligned}E(Z_k|Y = y_k) &= P(Z_k = 1|Y = y_k) - P(Z_k = -1|Y = y_k) \\ &= \frac{p^+ f^+(y_k) - p^- f^-(y_k)}{p^- f^-(y_k) + p_0 f_0(y_k) + p^+ f^+(y_k)} \\ &= \hat{z_k}\end{aligned} \tag{3.24}$$

$$\begin{aligned}E(Z_k^2|Y = y_k) &= P(Z_k = 1|Y = y_k) + P(Z_k = -1|Y = y_k) \\ &= \frac{p^+ f^+(y_k) + p^- f^-(y_k)}{p^- f^-(y_k) + p_0 f_0(y_k) + p^+ f^+(y_k)} \\ &= \widehat{z_k^2}.\end{aligned} \tag{3.25}$$

L'équation (3.23) s'écrit alors :

$$\begin{aligned}
\mathcal{Q}(\eta|\eta^{(i)}) &= \underbrace{\sum_{k=1}^{n}\left[(1-\widehat{z_k^2})\log f_0(y_k) + \frac{\widehat{z_k^2}+\widehat{z_k}}{2}\log f^+(y_k) + \frac{\widehat{z_k^2}-\widehat{z_k}}{2}\log f^-(y_k)\right]}_{A(a,b,\nu)} \\
&+ \underbrace{\sum_{k=1}^{n}\left[(1-\widehat{z_k^2})\log p_0 + \frac{\widehat{z_k^2}+\widehat{z_k}}{2}\log p^+ + \frac{\widehat{z_k^2}-\widehat{z_k}}{2}\log p^-\right]}_{B(p_0,p^+,p^-)} \\
&= A(a,b,\nu) + B(p_0,p^+,p^-). \qquad (3.26)
\end{aligned}$$

- **Étape M : maximisation de l'espérance conditionnelle**
  Lors de l'étape de maximisation, on calcule $\eta^{(i+1)}$ qui maximise $\mathcal{Q}(\eta|\eta^{(i)})$:

  $$\eta^{(i+1)} = \arg\max_{\eta} \mathcal{Q}(\eta|\eta^{(i)})$$

  Avec les notations prises ci-dessus, il vient :

  $$\begin{aligned}
(a,b,\nu)^{(i+1)} &= \arg\max_{a,b,\nu} A(a,b,\nu) \qquad (3.27)\\
(p^+)^{(i+1)} &= \frac{1}{n}\sum_{k=1}^{n}\frac{\widehat{z_k^2}+\widehat{z_k}}{2} \qquad (3.28)\\
(p^-)^{(i+1)} &= \frac{1}{n}\sum_{k=1}^{n}\frac{\widehat{z_k^2}-\widehat{z_k}}{2} \qquad (3.29)\\
(p_0)^{(i+1)} &= 1-(p^+)^{(i+1)}-(p^-)^{(i+1)}. \qquad (3.30)
\end{aligned}$$

Le point délicat est due à l'estimation des paramètres a, b et ν pour lesquels il n'existe pas de solution explicite pour maximiser $\mathcal{Q}(\eta|\eta^{(i)})$. Pour cela, la maximisation est réalisée par un algorithme d'optimisation.

L'algorithme EM consiste à choisir une valeur initiale $\eta^{(0)}$ pour η et à répéter les étapes E et M jusqu'à la convergence. En pratique, la convergence est admise lorsque $\|\eta^{(i)}-\eta^{(i-1)}\| < \epsilon$ pour ϵ et $\|.\|$ fixés.

Remarque :
Le calcul des proportions lors de l'étape M est équivalent à :

$$
\begin{aligned}
(p^+)^{(i+1)} &= \frac{1}{n}\sum_{k=1}^{n} P(Z_k = 1|y_k) && (3.31)\\
(p^-)^{(i+1)} &= \frac{1}{n}\sum_{k=1}^{n} P(Z_k = -1|y_k) && (3.32)\\
(p_0)^{(i+1)} &= \frac{1}{n}\sum_{k=1}^{n} P(Z_k = 0|y_k). && (3.33)
\end{aligned}
$$

Une fois l'ensemble des paramètres estimés, il est possible de calculer les probabilités a posteriori qu'un gène soit non différentiellement exprimés, induit ou réprimé. Ces trois probabilités sont données par les relations

$$
\begin{aligned}
P(Z_k = 0|y_k, \eta) &= \frac{p_0 f(y_k)}{p^- f^-(y_k) + p_0 f_0(y_k) + p^+ f^+(y_k)} && (3.34)\\
P(Z_k = 1|y_k, \eta) &= \frac{p^+ f^+(y_k)}{p^- f^-(y_k) + p_0 f_0(y_k) + p^+ f^+(y_k)} && (3.35)\\
P(Z_k = -1|y_k, \eta) &= \frac{p^- f^-(y_k)}{p^- f^-(y_k) + p_0 f_0(y_k) + p^+ f^+(y_k)}. && (3.36)
\end{aligned}
$$

À partir de ces expressions et de la définition des densités f^+ et f^-, il est immédiat que si $r_k > g_k$, alors $P(Z_k = -1|y_k, \eta) = 0$ et si $r_k < g_k$, alors $P(Z_k = 1|y_k, \eta) = 0$.

Pour identifier les gènes différentiellement exprimés, on peut appliquer, comme en classification, le principe du MAP (maximum a posteriori). Cela consiste à classer un gène dans la classe avec la plus forte probabilité a posteriori. De part la remarque précédente, un gène est alors considéré comme induit si $P(Z_k = 1|y_k, \eta) \geq 0.5$, réprimé si $P(Z_k = -1|y_k, \eta) \geq 0.5$ ou non différentiellement exprimé si $P(Z_k = 0|y_k, \eta) \geq 0.5$. Ne connaissant pas la vraie valeur de η pour le calcul de ces probabilités, on la remplace par son estimation $\widehat{\eta}$.

3.2.2.3 Résultats de simulations

Nous avons réalisé une simulation de Monte-Carlo afin de montrer que l'approche proposée permet d'estimer correctement l'ensemble des paramètres du modèle. Dans un souci de réalisme, nous avons supposé que peu de gènes étaient différentiellement exprimés. Nous choisissons donc de fixer à 8% le nombre de gènes différentiellement exprimés sur un total de 3000 gènes. On considère que 3% des gènes sont réprimés et que 5% sont induits. Les paramètres a, b, ν sont fixés arbitrairement à $a = 12$, $b = 20$, $\nu = 2$ et bien sûr on a $p^+ = 0.05$ et $p^- = 0.03$. Le tableau 3.1 donne les résultats pour 100 simulations du modèle ainsi paramétré.

paramètres	valeur	moyenne des estimations	écart-type des estimations
a	12	11.97	0.42
b	20	20.20	1.17
ν	2	2.02	0.01
p^+	0.05	0.0503	0.27×10^{-2}
p^-	0.03	0.0304	0.2×10^{-2}

TAB. 3.1 – Estimation des paramètres du modèle pour la recherche de gènes différentiellement exprimés

Nous avons également réalisé une simulation en supposant que les valeurs ν sont différentes pour chacun des trois cas : gènes induits, réprimés et non différentiellement exprimés. Les valeurs des paramètres sont $a = 12, b = 100, \nu = 10, \nu^+ = 2, \nu^- = 4, p^+ = 0.05$ et $p^- = 0.03$. Nous avons simulé 100 fois ce modèle pour 3000 gènes. Les résultats sont reportés dans le tableau 3.2.

paramètres	valeur	moyenne des estimations	écart-type des estimations
a	12	12.007	0.341
b	100	103.571	29.802
ν	10	10.007	0.381
ν^+	2	2.18	0.668
ν^-	4	4.200	1.213
p^+	0.05	0.0502	0.33×10^{-2}
p^-	0.03	0.0302	0.27×10^{-2}

TAB. 3.2 – Estimation des paramètres du modèle pour la recherche de gènes différentiellement exprimés

Les résultats de ces simulations sont satisfaisants. Les paramètres a et ν sont très bien estimés. En revanche, on constate un biais et une incertitude importante pour l'estimation des paramètres b, ν^+, ν^- en raison du faible nombre de points représentant les gènes différentiellement exprimés. Ce problème n'apparaît pas lorsque $\nu = \nu^- = \nu^+$. Le modèle le plus paramétré doit donc être appliqué avec prudence sur des données réelles si peu de gènes sont différentiellement exprimés. Du fait qu'il représente différemment les trois classes de gènes, il peut toutefois être plus réaliste. Nous comparerons ces deux modèles dans la partie application, afin de mieux comprendre les différences entre ces deux modèles.

Comme le propose Newton *et al.*, on peut également envisager d'utiliser des lois Gamma pour les distributions a priori. Les résultats obtenus ont montré qu'il est alors impossible d'estimer les paramètres. Ce problème est lié au fait que nous ne disposons pas d'équations explicites pour l'étape M de l'algorihme EM. L'algorithme d'optimisation est dans ce cas trop instable, et ne permet d'envisager une telle paramétrisation sur des données réelles.

3.2.3 Lien et comparaison avec les modèles de mélange

Nous montrons ici que le modèle proposé est un cas particulier de modèles de mélange. Dans le cadre d'une application, nous comparons notre approche à d'autres propositions, dont celle de Newton *et al.*

3.2.3.1 Le modèle proposé vu comme un modèle de mélange

Dans le paragraphe précédent, le problème de gènes différentiellement exprimés nécessitait l'introduction d'un test d'hypothèses. Ici, la solution proposée peut être vue comme un modèle de mélange. Dans le modèle que nous avons présenté, on distingue trois composants : les gènes non différentiellement exprimés, les gènes induits et les gènes réprimés. Étant donné une observation $y = (g, r)$ et suivant les notations introduites précédemment, la densité en ce point s'écrit

$$f(y) = p_0 f_0(y) + p^+ f^+(y) + p^- f^-(y).$$

Sapir *et al.* [104] décrivent une approche beaucoup plus simple. Le principe consiste à étudier les résidus obtenus lors de la régression entre les niveaux d'expression r et g. Afin d'analyser de façon symétrique les deux fluorophores, on choisit d'utiliser les résidus orthogonaux et non les résidus des moindres carrés. La distribution de ces résidus orthogonaux e_i $(1 \leq i \leq N)$ est représentée par le mélange

$$f(e_i) = (1-p) f_0(e_i) + p f_D(e_i) \tag{3.37}$$

où

- f_D est la distribution associée au composant caractérisant les gènes non différentiellement exprimés. f_D est la densité d'une loi normale $\mathcal{N}(0, \sigma^2)$
- f_0 est la distribution associée au composant caractérisant les gènes différentiellement exprimés. f_0 est la densité d'une loi uniforme sur $[a, b]$.
- p est la proportion de gènes différentiellement exprimés.

L'ensemble des paramètres (σ^2, a, b, p) qui définit le modèle est estimé par l'algorithme EM. Cependant Sapir *et al.* ne justifient pas le choix de leur modèle. L'idée de modéliser différemment les résidus obtenus par la régression semble intéressante. Cependant, le choix des distributions n'est pas justifié et semble discutable. Dans le cas où peu de gènes sont différentiellement exprimés, la droite de régression devrait être sensiblement la droite $r = g$, qui représente les gènes non différentiellement exprimés. Les gènes différentiellement exprimés correspondent alors aux gènes les plus éloignés de cette droite, c'est-à-dire avec des résidus importants. On justifie donc difficilement l'hypothèse d'une répartition uniforme des résidus pour les gènes différentiellement exprimés.

Lee *et al.* [78] donnent également une solution basée sur un modèle de mélange. Dans leur approche, les log-ratios des expressions sont représentés par un mélange de deux loi normales.

Suivant le modèle (3.37), les lois f_0 et f_D sont des lois gaussiennes dont les paramètres sont estimés par l'algorithme EM. Tout comme dans le cas précédent, le choix de cette modélisation n'est pas justifié. Ce modèle paraît tout aussi discutable. En effet, les log-ratios sont négatifs si les gènes sont réprimés et positifs si les gènes sont induits. Intuitivement, la distribution des gènes différentiellement exprimés paraît donc bimodale. On comprend difficilement la justification d'une loi normale (donc unimodale) pour représenter les gènes différentiellement exprimés. En revanche, la représentation des gènes non différentiellement exprimés par une distribution symétrique des log-ratios et centrée 0 est compréhensible.

3.2.3.2 Application

Le thème biologique abordé dans cette application est l'étude des microtubules chez la levure *Saccharomyces cerevisiae*. Les microtubules jouent un rôle fondamental dans les régulations du contrôle et du déroulement de la division cellulaire. La tubuline est la protéine constitutive majeure de ces structures. Sans détailler les questions biologiques abordées, il s'agit de comparer deux souches différentes de levure : une souche mutante ayant subi une délétion d'un gène intervenant dans la fabrication de la tubuline et une souche sauvage. Nous étudions l'ensemble des gènes connus de la levure afin d'identifier ceux dont l'activité est modifiée par cette mutation. Parmi les 6400 spots de la puce (les spots étant les gènes ou des ORFs *open reading frame*), on cherche ceux qui sont différentiellement exprimés : la souche mutante est marquée avec le Cy5 et la souche sauvage avec le Cy3. Dans cette analyse, nous disposons seulement de contrôles pour l'absence de différence d'expression. Comme dans le cas de la normalisation, ces contrôles sont identifiés par la couleur jaune sur la figure 3.3.

Dans un premier temps, nous nous intéressons aux différences obtenues entre les deux approches proposées. Dans les deux cas, le choix des classes est réalisé suivant le principe du maximum a posteriori. La figure 3.1 donne les courbes de niveau obtenues suivant ce principe pour les deux modèles. Les résultats pour les deux approches sont relativement différents. On observe que le niveau moyen d'expression du gène influe sur l'identification des gènes différentiellement exprimés. Dans les deux modèles, on considère plus difficilement qu'un gène i est différentiellement exprimés si son niveau d'expression moyen A_i est faible. En choisissant des valeurs différentes pour l'induction et la répression ($\nu^+ \neq \nu^-$), on sélectionne 28 gènes supplémentaires représentés par des ronds pleins (traits en pointillé). Les ronds vides sont les gènes identifiés lorsque l'on suppose $\nu^+ = \nu-$ (traits pleins). La principale différence entre ces deux modèles est que l'on choisit d'avantage de gènes différentiellement exprimés lorsque le niveau d'expression moyen est faible (A faible). L'objectif de ces méthodes étant d'éviter des faux positifs liés à des niveaux d'expression faible, nous préférons baser nos résultats sur le modèle le moins paramétré en supposant que $\nu^+ = \nu^-$. Dans une démarche expérimentale, il conviendrait dans un premier temps de s'intéresser aux gènes sélectionnés suivant ce modèle.

Nous souhaitons dans un deuxième temps comparer le modèle proposé par Newton *et al.* et le modèle que nous venons de sélectionner. En modélisant différemment les gènes induits et

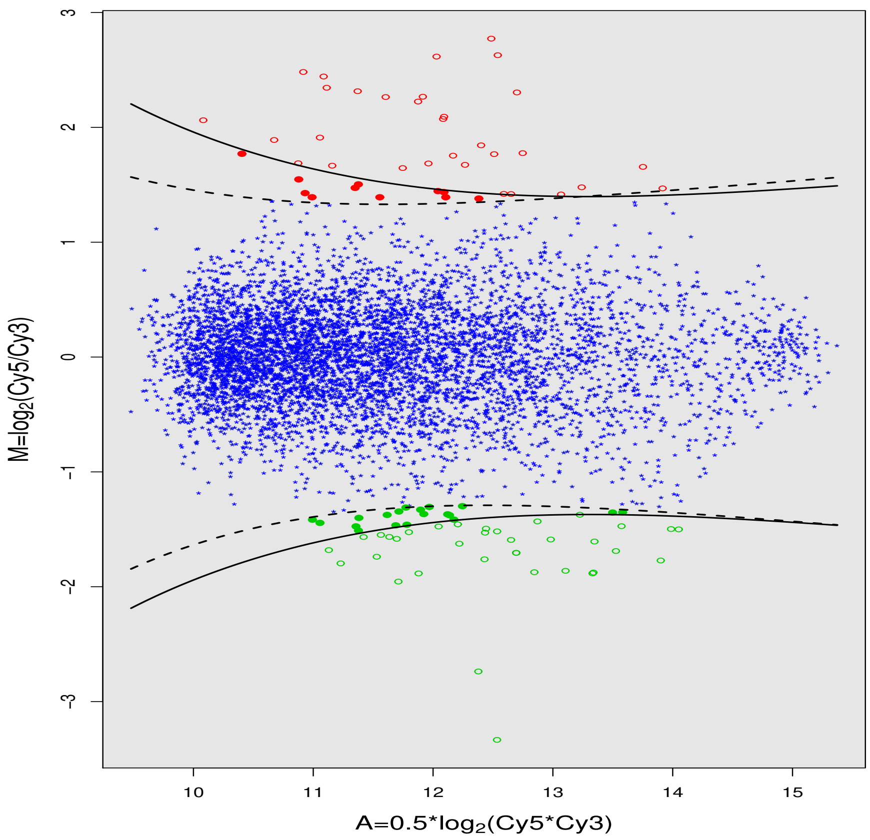


FIG. 3.1 – *Modélisation de l'induction, la répression et l'absence de différence d'expression.*

réprimés, on sélectionne 6 gènes supplémentaires représentés par des ronds pleins. Ce résultat est tout à fait cohérent même si les différences ne concernent qu'un faible nombre de gènes. Plus le niveau d'expression d'un gène est fort (A grand) et plus on peut avoir confiance sur la valeur du ratio M. À ratio égal, la probabilité qu'un gène i soit différentiellement exprimé est donc plus importante si le niveau moyen d'expression A_i est grand. Le modèle proposé par Newton *et al.* ne montre pas une telle propriété.

À l'aide de la figure 3.3, nous comparons notre approche avec les modèles proposés par Sapir *et al.* [104] et Chen *et al.* [26].

L'approche de Sapir *et al.* est également basée sur un modèle de mélange. Pour comparer les différents résulats, nous utilisons donc a nouveau le principe du MAP. Pour cette approche, les gènes sont induits si le ratio d'expression est supèrieur à 2.72 ($M = 1.44$) et réprimés si le ratio est infèrieur à -2.72 ($M = -1.44$). Par rapport à l'approche que nous proposons, on sélectionne ici plus de gènes faiblement exprimés ; ce qui peut conduire à un nombre de faux positifs plus important.

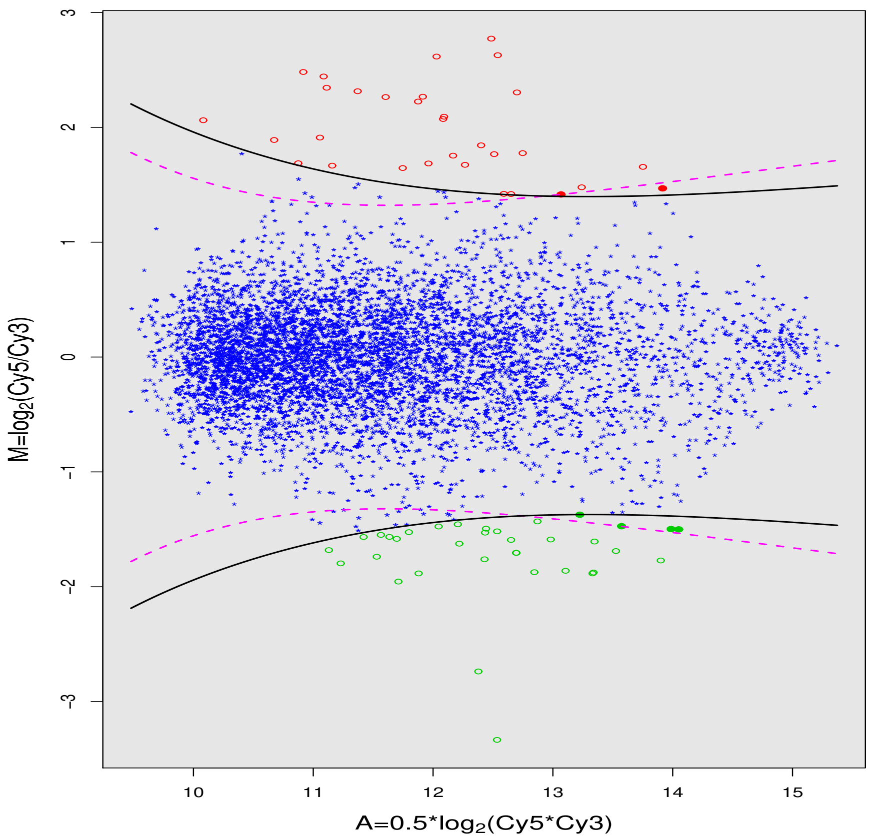


FIG. 3.2 – *Comparaison des 2 modèles proposés pour la recherche de gènes différentiellement exptimés.*

L'identification des gènes différentiellement exprimés selon l'approche de Chen *et al.* est basée sur l'estimation de la densité des ratios d'expression ainsi que sur un test d'hypothèses. Nous donnons les résultats pour deux valeurs du risque de première espèce : $\alpha = 5\%$ et $\alpha = 1\%$. Dans le premier cas, cela semble conduire a de nombreux positifs. Pour $\alpha = 1\%$, si les gènes réprimés sont sensiblement les mêmes que ceux de notre approche, les gènes induits sont plus nombreux. Ces différences sont plus importantes que dans le cas du modèle de Sapir *et al.* On peut penser qu'elles sont liées aux problèmes de tests multiples. Une approche plus rigoureuse pour l'application du test proposé par Chen *et al.* consisterait à introduire une correction de type Bonferroni par exemple.

3.2.3.3 Discussion sur le choix de modèles

Ces études montrent que les analyses dépendent fortement du choix du modèle. Ce choix peut bien entendu être basé sur des tests d'hypothèses ou sur des critères d'infor-

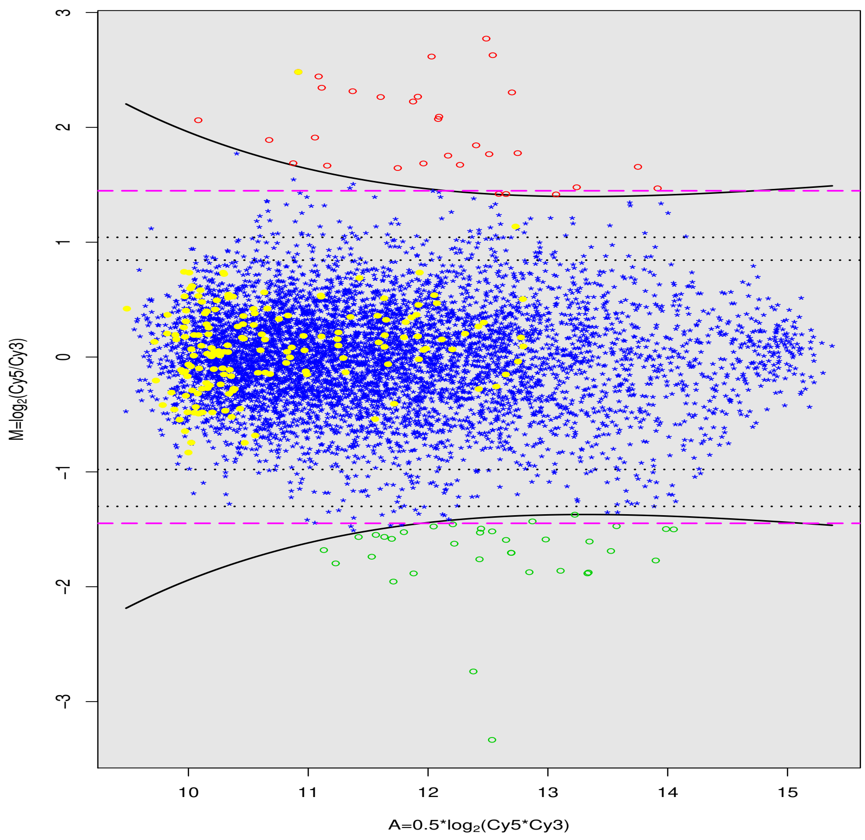


FIG. 3.3 – *Gènes différentiellement exprimés selon différents modèles*

mation. Néanmoins, le choix d'un modèle peut également dépendre de ses propritétés de représentation.

De notre point de vue, un modèle devrait satisfaire deux principes. Premièrement, il doit permettre une représentation des phénomènes biologiques que l'on cherche à mettre en évidence. Deuxièmement, il doit rester cohérent avec les connaissances que l'on a priori du problème. Le modèle que nous avons proposé répond clairement à ces deux points.

- Notre approche donne une représentation différente des gènes induits et réprimés. Ces deux classes de gènes étant bien différentes d'un point de vue biologique, cette distinction dans la modélisation semble naturelle.
- Pour la recherche de gènes différentiellement exprimés, l'intérêt majeur de l'approche proposée par Newton *et al.* est d'être plus sélective pour des gènes faiblement exprimés (figure 3.2). Newton *et al.* justifient cela de la manière suivante :
  "As we expected, we are less confident about naive fold change at low signal intensity".

Ainsi, leur modèle cherche à refléter une variabilité importante pour des intensités faibles. Mais l'inconvénient de leur approche est qu'elle s'avère également sélective pour les gènes fortement exprimés (figure 3.2). Cette caractéristique est contraire à l'intuition. En effet, plus un gène est exprimé plus la variabilité des mesures devrait être faible, et c'est sur ce point que nous avons axé notre modélisation. Dans notre approche, nous avons été attentifs à pallier cet inconvénient. Cela nous a conduit à distinguer les gènes réprimés et induits.

Pour ces deux raisons, les résultats obtenus nous paraissent plus cohérents que ceux du modèle de Newton *et al.* On peut ajouter que notre méthode conduit à être plus sélectif que Newton pour des gènes faiblement exprimés, ce qui dans une démarche expérimentale semble nécessaire. Néanmoins, il est très délicat de comparer les modèles présentés à partir de simulations et de déterminer ainsi celui qui permet le mieux d'identifier les gènes différentiellement exprimés. De même, Newton *et al.* admettent qu'il est difficile d'évaluer de tels modèles :
“ The simulation highlights difficulties with infering differential expression, but we point out that specific error rates may depend on the application”.
Notre objectif était de voir s'il était possible de pousser plus loin la modélisation afin de prendre en compte des caractéristiques biologiques supplémentaires. Une solution souhaitable pour évaluer les performances d'un modèle serait d'analyser des données où les niveaux d'expression et les gènes différentiellement exprimés seraient connus a priori. Pour cette raison, nous pensons que la mise en place de contrôles et de plans expérimentaux apporteront un élément de réponse important pour la validation des modèles.

3.2.3.4 Problèmes de variabilité

Les approches présentées permettent d'identifier les gènes différentiellement lorsque que l'on ne dispose pas de répétition, et ont été adoptées pour nos premières études. Cependant, il est impossible de valider les résultats biologiques sans réaliser de nouvelles expériences. Pour cela, nous nous intéressons donc aux problèmes de variabilité. Nous avons appliqué la méthode proposée sur un premier jeu de données (Réplication 1). Afin de valider ces résultats, nous avons ensuite analysé deux autres jeux de données correspondant à la même expérience biologique (Réplications 2 et 3). Les tableaux 3.3 donnent les tableaux de contingence croisant les résulats pour deux réplications ainsi qu'un tableau croisant les résulats pour trois réplications.

Le croisement de ces résultats met en évidence un problème de variabilité important entre les différentes expériences. Les comparaisons deux à deux montrent par exemple que seulement un quart des gènes différentiellement exprimés pour la puce 2 le sont pour la puce 3, et réciproquement. On constate qu'en comparant les puces 2 et 1, ce rapport est d'environ un tiers. Si on compare les trois puces ensemble, on trouve seulement 15 gènes différentiellement exprimés pour les 3 réplications. Cela semble faible au regard des 106, 85 et 89 gènes identifiés sur les répétitions 1, 2 et 3.

	Réplication 2		
Rép. 1	NE=6133	DE=85	Total
NE=6112	6063	49	6112
DE=106	70	**36**	106
Total	6133	85	6218

	Réplication 3		
Rép. 1	NE=6129	DE=89	Total
NE=6112	6045	67	6112
DE=106	84	**22**	106
Total	6129	89	6218

	Réplication 3		
Rép. 2.	NE=6129	DE=89	Total
NE=6133	6065	68	6133
DE=85	64	**21**	85
Total	6129	89	6218

	Réplication 3				
	NE=6129		DE=89		
	Réplication 2		Réplication 2		
Réplication 1	NE=6133	DE=85	NE=6133	DE=85	Total
NE	6002	43	61	6	6112
DE	63	21	7	**15**	106
Total	6065	64	68	21	6218

NE : Non différentiellement exprimé
DE : Différentiellement exprimé (induit ou réprimé)

TAB. 3.3 – *Étude des gènes différentiellement sur trois puces.* Croisements des résultats pour trois expériences similaires.

Ces problèmes ont déjà été signalés par Lee *et al.* [78] et méritent plusieurs remarques. D'abord, il semble peu raisonnable de mener une étude sans réaliser de répétition. L'absence de répétition risque en effet de conduire à sélectionner de nombreux faux positifs ou faux négatifs. Ensuite, la mise en place de mesures répétées nécessite de développer des outils permettant de les analyser. Enfin, différents facteurs doivent être à la source de cette variabilité importante. Il s'agit donc de pouvoir les identifier et les analyser. En conclusion, les approches que nous avons présentées sont en pratique d'un intérêt limité car elles concernent des expériences où l'on ne dispose pas de répétition.

3.2.4 Approches existantes pour les répétitions

Dans cette section, nous donnons quelques méthodes qui peremettent de prendre en compte les répétitions. Nous distinguons les méthodes basées sur les modèles de mélange et les méthodes donnant une représentation des problèmes de variabilité.

3.2.4.1 Les approches mélanges pour les répétitions

Comme nous l'avons vu, une expérience ne comportant qu'une seule mesure par gène ne permet pas une analyse fiable. Des répétitions doivent donc être réalisées. Les travaux proposant d'aborder la recherche de gène différentiellement exprimés sous l'angle des mélanges peuvent être classés en deux catégories.

La première permet l'étude de données issues d'une puce ne contenant qu'un spot par gène et dépend d'hypothèses paramétriques fortes. Les études que nous avons menées s'inscrivent dans ce cadre.

La seconde catégorie d'approche utilise des données répétées (plusieurs puces ou plusieurs spots pour un même gène sur la même puce). Les précédents modèles de mélange se basent sur des hypothèses paramétriques. Efron *et al.* [44, 45, 43] ainsi que Pan *et al.* [93, 95] montrent que l'on peut s'affranchir totalement ou en partie de ces hypothèses paramétriques dans le cas de mesures répétées.

Bien que ces deux approches soient appliquées sur des données obtenues à partir de puces Affymetrix, elles restent suffisamment générales pour être présentées dans notre cadre d'application. Elles visent toutes les deux à définir un score pour chaque gène à partir des répétitions disponibles. On note W ce score et $\{W_i\}$ l'ensemble des scores obtenus pour les N gènes. Dans le cas de Efron *et al.*, ce score est identique à celui proposé par Tusher *et al.* [120], et que nous avons présenté lors des tests d'hypothèses (3.3). Quant à celui proposé par Pan *et al.*, il est un peu plus complexe. Mais dans les deux cas, on représente la distribution f de ces scores suivant le même modèle de mélange

$$f = p_0 f_0 + p_D f_D \tag{3.38}$$

où

- f_0 est la densité pour W que l'expression d'un gène ne soit pas modifiée ($Z = 0$)
- f_D est la densité pour W que l'expression d'un gène soit modifiée ($Z = 1$)
- p_D est la probabilité qu'un gène soit différentiellement exprimé
- $p_0 = 1 - p_D$ est la probabilité qu'un gène ne soit pas différentiellement exprimé.

Suivant la règle de Bayes, on peut calculer les probabilités conditionnelles qu'un gène soit différentiellement exprimés ou non. Elles s'écrivent

$$\begin{aligned} P(Z_i = 1|W_i) &= 1 - p_0 \frac{f_0(W_i)}{f(W_i)} \\ P(Z_i = 0|W_i) &= p_0 \frac{f_0(W_i)}{f(W_i)} \end{aligned} \tag{3.39}$$

Sans hypothèse paramétrique, le modèle (3.38) n'est pas identifiable. Un second type de score, noté z_i, est alors calculé à partir des répétitions pour estimer la loi f_0. Pour cela, Pan *et al.* approximent f_0 en utilisant un modèle de mélange gaussien. La justification tient au fait que toute distribution peut être estimée par un mèlange de distributions gaussiennes. Les paramètres du mélange sont estimés via l'algorithme EM et le nombre de composants est choisi grâce aux critères BIC ou AIC. La solution de Efron *et al.* est différente dans la mesure où aucune hypothèse paramétrique sur la distribution de f_0 n'est faite. Leur méthode utilise un modèle de régression logistique. Pan *et al.* [93] donnent une comparaison de ces deux appproches ansi que de celle proposée par Tusher *et al.* Efron *et al.* [43] montrent également que leur modèle est proche d'une méthode fréquentiste dans laquelle on souhaite contrôler l'erreur FDR. Enfin, Pan *et al.* [94] expliquent comment, à partir de leur modèle, le nombre de répétitions nécessaire peut être choisi.

3.2.4.2 Modélisation de l'erreur de mesure

D'autres travaux proposent d'étudier les gènes différentiellement exprimés en introduisant directement les problèmes de variabilité dans la modélisation des niveaux d'expression.

Ideker *et al.* [64] présentent un modèle pour l'erreur de mesure en lui associant un test pour identifier les gènes différentiellement exprimés. On note r_{kj} et g_{kj} les niveaux d'expression pour le gène k $(1 \leq k \leq N)$ et pour la $j^{\text{ième}}$ répétition. Le modèle proposé s'écrit

$$r_{kj} = \mu_{R_k} + \mu_{R_k}\epsilon_{R_{kj}} + \delta_{r_{kj}} \tag{3.40}$$

$$g_{kj} = \mu_{G_k} + \mu_{G_k}\epsilon_{G_{kj}} + \delta_{g_{kj}} \tag{3.41}$$

où

- (μ_{R_k}, μ_{G_k}) sont les vrais niveaux d'expression du gène k
- Les facteurs multiplicatifs $(\epsilon_{r_{kj}}, \epsilon_{g_{kj}})$ sont des réalisations d'une loi normale bivariée $\mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{\epsilon_R} & \rho_\epsilon \\ \rho_\epsilon & \sigma^2_{\epsilon_G} \end{pmatrix}\right)$
- Les termes d'erreur $(\delta_{r_{kj}}, \delta_{g_{kj}})$ sont des réalisations d'une loi normale bivariée $\mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{\delta_R} & \rho_\delta \\ \rho_\delta & \sigma^2_{\delta_G} \end{pmatrix}\right)$.

Le modèle ainsi défini comporte $2N + 6$ paramètres qui peuvent être estimés par maximum de vraisemblance. En raison du nombre important de paramètres, Ideker *et al.* donnent une procédure en plusieurs étapes pour l'estimation. À partir de ce modèle, ils introduisent alors un test de vraisemblance pour déterminer les gènes différentiellement exprimés. Le nombre considérable de paramètres pose cependant un problème difficile quant à la qualité des estimations.

Dans leurs travaux, Rocke *et al.* [102] construisent un modèle pour évaluer l'erreur de mesure en fonction du niveau d'expression. Dans cette approche, la variabilité du niveau d'expression est représentée par deux composantes. Pour des niveaux faibles d'expression, la variabilité est approximativement constante ; alors que pour des niveaux plus importants, le coefficient de variation est constant. Ces résultats montrent que le choix des gènes différentiellement exprimés doit dépendre du niveau d'expression. Ils remettent ainsi en cause l'hypothèse de Chen *et al.* où le coefficient de variation est constant pour tous les gènes.

3.2.5 Conclusions

De nombreuses approches basées sur des modèles probabilistes ont été proposées dans le but de modéliser des propriétés de la distribution de l'expression des gènes [26, 44, 64, 91, 120]. Nous pouvons dégager quelques remarques. La variance de l'expression des gènes est liée au niveau d'expression des gènes [91, 64, 102] et l'identification des gènes différentiellement exprimés doit en tenir compte. En disposant de répétitions, certains travaux ont montré que l'on pouvait s'affranchir d'hypothèses paramétriques sur la distribution des niveaux d'expression [44, 45, 95, 120], souvent mal justifiées.

De manière évidente, la difficulté principale de ces analyses concerne les problèmes de reproductibilité des mesures. Ces problèmes impliquent la nécessité de réaliser des répétitions afin de valider les conclusions biologiques et les modèles statistiques. Le modèle que nous avons introduit est donc d'un intérêt limité car il ne permet pas d'analyser des mesures répétées. Pour répondre à ce problème, nous avons présenté quelques travaux qui permettent d'intégrer les répétitions dans l'analyse des gènes différentiellement exprimés.

Cependant, il existe différents facteurs responsables de la variabilité. Dans la suite, nous cherchons à représenter les niveaux d'expression en introduisant différents facteurs qui permettent d'expliquer ces problèmes. Parmi ces facteurs, nous devons différencier les facteurs expérimentaux et les facteurs d'intérêt biologique qui vont nous permettre d'identifier les gènes différentiellement exprimés.

3.3 Analyse de variance

Bien que les puces à ADN soient en train de devenir un outil fondamental en biologie moléculaire, les résultats précédents montrent que l'interprétation des résultats est difficile. Les problèmes de variabilité sont encore mal connus et limitent fortement les applications possibles. Ils doivent donc être considérés avec la plus grande attention. Pour cela, on a recours classiquement aux plans d'expériences, aux mesures répétées et à la mise en place de contrôles. De nombreux développements doivent être réalisés sur ces trois points. Dans ce paragraphe, nous nous intéressons aux méthodes permettant d'introduire les répétitions dans l'analyse une fois l'expérience réalisée. Nous n'aborderons pas le problème du choix du plan d'expériences. Les réplications peuvent être introduites à différents niveaux de l'analyse : les gènes sont spottés à plusieurs reprises sur la puce, plusieurs puces sont réalisées avec les mêmes échantillons d'ARN ou différents échantillons sont utilisés pour plusieurs puces. D'ailleurs, on considère le dernier cas comme des mesures répliquées alors que les deux premiers correspondent à des mesures répétées. Enfin, ces trois cas donnent a priori des mesures avec une variabilité croissante.

Dans cette partie, nous ne considérons que le cas des mesures répétées (répétitions intra-puce et inter-puces) obtenues par un marquage fluorescent. Chaque résultat obtenu pour un spot de la puce est une combinaison spécifique des facteurs suivants : la puce, le fluorophore, la condition expérimentale et le gène. On note y_{ijkgr} le logarithme de l'expression du gène G_g pour la puce ou array A_i. La condition expérimentale ou variété V_k est marquée avec le fluorophore ou dye D_j. L'indice r correspond à la r$^{\text{ième}}$ répétition du gène g sur la plaque i dans la condition k, marquée avec le fluorophore j. La variété V_k fait référence à la condition k de notre étude avec $k \in \{1, 2\}$ pour la comparaison de deux échantillons biologiques.

Une méthode naturelle pour prendre en considération tous ces facteurs est l'analyse de variance. Elle est sans doute l'approche la plus appliquée pour l'analyse des données et a été largement développée par Kerr *et al.* [69]. Notre travail se base en partie sur leurs résultats. Comme nous le précisons par la suite, l'analyse de variance reste tout de même délicate pour l'étude de données de biopuces. Aussi, nous commençons par décrire le cadre d'application de ces méthodes.

3.3.1 Principes

L'analyse de variance permet d'apprécier l'effet de variables qualitatives (les facteurs) sur une variable numérique (le niveau d'expression des gènes). De part la nature des données expérimentales, on considère les 4 facteurs A, D, V, G. Ces facteurs sont considérés comme étant des facteurs à effets fixes, c'est-à-dire avec un nombre fixe et fini de niveaux. Ils en possèdent I, J, K, N respectivement. On suppose le plan d'expériences équilibré : le nombre de répétitions est identique pour tous les gènes et égal à R.

3.3.1.1 Les facteurs

On souhaite représenter le niveau d'expression des gènes suivant ces différents facteurs. Il reste toutefois à décrire le facteur traduisant la variabilité biologique. Il traduit comment le niveau d'expression d'un gène g varie suivant une condition k. Ce facteur correspond à l'interaction des deux facteurs principaux V et G : on le note VG_{kg}. Le niveau d'expression du gène se décompose ainsi

$$y_{ijkgr} = \mu + A_i + D_j + V_k + G_g + VG_{kg} + \epsilon_{ijkgr} \tag{3.42}$$

où :

- y_{ijkgr} est le logarithme des niveaux d'expression,
- μ est l'effet global,
- A_i est l'effet de la $i$ème puce (facteur A à I niveaux),
- D_j est le $j$ième niveau de l'effet fluorophore D à $J = 2$ niveaux,
- V_k correspond au $k$ième niveau de l'effet variété V à $K = 2$ niveaux,

- G_g traduit l'effet du $g$ième gène pour le facteur G à N niveaux,
- VG_{kg} est l'interaction entre la $k$ième variété et le $g$ième gène,
- ϵ_{ijkgr} représente les erreurs de mesures indépendantes et que l'on considère de loi normale centrée et de variance σ^2,
- les contraintes du modèle sont données par $\sum_i A_i = \sum_j D_j = \sum_k V_k = \sum_g G_g = \sum_k VG_{kg} = \sum_g VG_{kg} = 0$.

Le terme μ représente le signal moyen sur tous les facteurs. Le terme A_i prend en compte les différences entre les puces moyennées sur tous les gènes, variétés et fluorophores. Cette variabilité est due par exemple à la conception de la puce. Le terme D_j représente les différences entre les signaux moyens de chaque fluorophore. Cette différence est généralement liée au fait que le Cy5 est plus brillant que le Cy3. Le terme V_k prend en compte le traitement des expériences ; cela peut représenter différents types d'organismes, de tissus, organes, traitements. Les variations s'expliquent par des concentrations variables d'ARNm dans chaque échantillon marqué, ou à des niveaux de transcription différents. Le terme G_g fait référence à la moyenne des effets de chaque gène déposé sur les puces.

Les facteurs principaux A, D, V, G permettent ainsi de prendre en compte différentes sources de variabilité. Le terme intéressant est l'interaction VG_{kg} car il traduit les variations dues à la combinaison spécifique de la variété k et du gène g. Ce terme d'interaction est présent dans tous les modèles d'analyse de variance que nous développerons car il représente l'information biologique que l'on souhaite étudier. Ainsi, la recherche de gènes différentiellement exprimés entre deux conditions V_1 et V_2 consiste à identifier les gènes g_0 pour lesquels $VG_{1g_0} - VG_{2g_0} = 2 * VG_{1g_0}$ est significativement différent de 0. Avant cela, nous souhaitons développer ces modèles d'analyse de variance. Si l'écriture semble aisée dans un premier temps, certaines difficultés doivent être prises en compte.

3.3.1.2 Le principe du *dye-swap* ou flip-flop

Dans nos applications, les données disponibles sont issues de plans d'expériences basés sur le dye-swap. Nous concentrons donc notre travail sur l'étude de données obtenues à partir de ces plans. Kerr *et al.* [69, 70] donnent plusieurs résultats sur les problèmes de choix de plans d'expériences. N'ayant pas eu l'occasion de travailler sur la mise en place de protocoles expérimentaux, nous n'aborderons pas ce sujet.

Les plans basés sur le *dye-swap* sont largement utilisés car ils permettent de prendre en compte la variabilité due aux marqueurs. Kerr *et al.* [69] ont proposé dans un premier temps d'appliquer directement l'analyse de variance pour des données brutes sans réaliser de normalisation, en considérant les termes A, D et V comme des facteurs de normalisation. Toutefois, en examinant les résidus estimés par ce type de modèle, Kerr *et al.* [68] ont constaté que certains effets non linéaires n'autorisent pas une approche aussi directe et proposent donc une étape de normalisation pour pouvoir appliquer un modèle d'analyse de variance. Dans nos analyses, les données ont été normalisées suivant l'approche proposée par Yang *et al.* ([129], [40]) en raison des problèmes de non linéarité qui peuvent survenir dans les mesures [97]. Ces problèmes ont déjà été signalés au second chapitre.

Afin de limiter le biais lié à l'utilisation des marqueurs fluorescents (différence d'incorporation, de fluorescence), on a souvent recours aux plans basés sur le *dye-swap* ou *flip-flop*. Le principe consiste à utiliser deux puces de la manière suivante : la première est hybridée avec l'échantillon traitement (2) marqué avec le Cy5 (Dye 2) et avec l'échantillon référence (1) marqué avec le Cy3 (Dye 1) ; la seconde puce est hybridée avec l'échantillon traitement (2) marqué avec le Cy3 et avec l'échantillon (1) marqué avec le Cy5. Cela revient donc à échanger les marqueurs par rapport aux échantillons sur les deux puces. On représente ce plan par le tableau 3.4.

	Dye 1 : Cy3	Dye 2 : Cy5
Array 1	Condition référence	Condition test
Array 2	Condition test	Condition référence

TAB. 3.4 – Principe du *dye-swap*

Tous les modèles que nous considérons seront basés sur ce plan. Les variantes concernent le nombre de répétitions intra-puce ainsi que le nombre de *dye-swap* réalisés.

3.3.1.3 Les modèles d'analyse de variance

L'ensemble des facteurs possibles et des interactions de tout ordre sont : A, D, V, G, AD, AV, AG, DV, DG, VG, ADV, ADG, AVG, DVG et $ADVG$. Tous ces facteurs[2] ne sont pas toujours estimables. Selon le plan d'expériences envisagé, certains de ces facteurs seront partiellement ou totalement confondus avec d'autres.

Le nombre de spots sur une puce est NR. Nous faisons l'hypothèse que les plans considérés sont équilibrés, et donc tous les gènes ont le même nombre R de répétitions. On note I le nombre de puces utilisées ; il est de 2 si un *swap* est réalisé, de 4 si deux *swaps* sont utilisés et ainsi de suite.

Le modèle (3.42) est le plus simple que l'on puisse proposer dans la mesure où seuls les facteurs principaux et l'interaction VG d'ordre 2 sont considérés. Cependant, des modèles plus compliqués, faisant intervenir d'autres interactions, peuvent être envisagés.

Avant d'aborder ces différents modèles, nous détaillons les propriétés des plans basés sur le *dye-swap*. Étant donné un modèle d'analyse de variance, nous lui associons l'écriture

[2]Nous utilisons le terme facteur au sens large : facteur pouvant désigner un facteur propre ou une interaction (traitement)

canonique suivante

$$Y = X\theta + \epsilon \tag{3.43}$$

où

- Y variable aléatoire à expliquer
- X matrice d'expériences connue et associée au plan d'expériences
- θ vecteur des différents facteurs pris en compte dans le modèle
- ϵ variable aléatoire normale des erreurs.

Afin d'illustrer cette écriture pour le modèle (3.42), nous considérons le cas particulier où l'on dispose de $I = 2$ puces avec $G = 3$ gènes. On dispose bien évidemment de $J = 2$ dyes et de $K = 2$ variétés ou conditions expérimentales. On note 1_R et 0_R le vecteur unité et le vecteur nul de dimension R. Les expressions de X et θ sont

$$X = \begin{pmatrix} 1_R & 1_R & 1_R & 1_R & 1_R & 0_R & 1_R & 0_R \\ 1_R & 1_R & 1_R & 1_R & 0_R & 1_R & 0_R & 1_R \\ 1_R & 1_R & 1_R & 1_R & -1_R & -1_R & -1_R & -1_R \\ 1_R & 1_R & -1_R & -1_R & 1_R & 0_R & -1_R & 0_R \\ 1_R & 1_R & -1_R & -1_R & 0_R & 1_R & 0_R & -1_R \\ 1_R & 1_R & -1_R & -1_R & -1_R & -1_R & 1_R & 1_R \\ 1_R & -1_R & 1_R & -1_R & 1_R & 0_R & -1_R & 0_R \\ 1_R & -1_R & 1_R & -1_R & 0_R & 1_R & 0_R & -1_R \\ 1_R & -1_R & 1_R & -1_R & -1_R & -1_R & 1_R & 1_R \\ 1_R & -1_R & -1_R & 1_R & 1_R & 0_R & 1_R & 0_R \\ 1_R & -1_R & -1_R & 1_R & 0_R & 1_R & 0_R & 1_R \\ 1_R & -1_R & -1_R & 1_R & -1_R & -1_R & -1_R & -1_R \end{pmatrix} \text{ et } \theta = \begin{pmatrix} \mu \\ A_1 \\ D_1 \\ V_1 \\ G_1 \\ G_2 \\ VG_{11} \\ VG_{12} \end{pmatrix}.$$

Dans ce modèle, θ est composé de 8 paramètres libres. Les contraintes du modèle d'analyse de variance donnent les relations suivantes : $A_1 + A_2 = 0, D_1 + D_2 = 0, V_1 + V_2 = 0, G_1 + G_2 + G_3 = 0$. On obtient également des relations similaires pour les termes d'interactions VG_{kg}.

3.3.2 Effets confondus

Dans les définitions suivantes, A et B sont deux facteurs du modèle (3.43) avec n_A et n_B niveaux respectivement. On note X_A et X_B les matrices d'incidence correspondant aux facteurs A et B (X_A et X_B sont un sous-ensemble de colonnes de la matrice X). Par exemple, dans l'exemple précédent, X_D correspond à la troisième colonne de la matrice X. De même, X_G est la matrice formée des cinquième et sixième colonnes.

Définition 3.1 *Facteurs A et B confondus*
A et B sont confondus si et seulement si les espaces vectoriels engendrés par X_A et X_B sont les mêmes.

Définition 3.2 *Facteur A confondu dans le facteur B*
A est confondu dans B si et seulement si l'espace vectoriel engendré par X_A est inclus dans celui engendré par X_B.

Définition 3.3 *Facteurs A et B partiellement confondus*
Soient r_A et r_B le rang des matrices X_A et X_B, et r le rang de la matrice $[X_A|X_B]$.
A et B sont partiellement confondus si et seulement si $max(r_A, r_B) < r < r_A + r_B$.

Définition 3.4 *Facteurs A et B orthogonaux*
Deux facteurs A et B sont orthogonaux si et seulement si les sous-espaces vectoriels engendrés par X_A et X_B sont orthogonaux.

Remarque : Si deux facteurs sont orthogonaux, leurs estimations sont indépendantes et le fait d'ajouter ou d'enlever un facteur dans le modèle ne modifie pas les estimations obtenues pour l'autre.

3.3.2.1 Plan d'expériences *dye-swap* simple

Si $I = 2$, on en déduit que deux des facteurs A, D et V déterminent le troisième. Ainsi, A est confondu avec DV, D avec AV et V avec AD. Par exemple, si le dye et la variété sont connus, alors la puce est également connue et réciproquement.
Le tableau 3.5 présente le plan qui répartit les 16 facteurs expérimentaux en 8 paires de facteurs confondus. Les facteurs qui ne sont pas confondus sont orthogonaux aux autres facteurs.

μ	$\sim$	ADV
A	$\sim$	DV
D	$\sim$	AV
V	$\sim$	AD
G	$\sim$	$ADVG$
VG	$\sim$	ADG
AG	$\sim$	DVG
DG	$\sim$	AVG

Tab. 3.5 – Facteurs confondus : *dye-swap* simple

Il faut noter que l'estimation de VG correspond en réalité à l'estimation de deux facteurs confondus : VG et ADG. L'estimation de VG n'a donc de sens qu'en supposant que l'interaction d'ordre 3 ADG est nulle. Le même type de remarque peut être fait pour les facteurs A, D, V et G. Mais dans le cas de VG, cette hypothèse est importante car elle concerne le facteur d'intérêt biologique.

Ainsi, un modèle prenant en compte les $2^4 = 16$ facteurs s'écrit

$$y_{ijkgr} = \mu + A_i + D_j + V_k + G_g + AG_{ig} + DG_{jg} + VG_{kg} + \epsilon_{ijkgr}. \tag{3.44}$$

Dans le cas où $I = 2$, c'est-à-dire le cas simple de deux puces, les caractéristiques de l'analyse de variance du modèle (3.44) sont données dans le tableau 3.6. Le nombre de degrés de liberté pour l'erreur dépend du nombre de répétitions et il est nécessairement non nul. D'après ce tableau, ceci amène à considérer les plans sans répétition ($R = 1$) et avec répétitions ($R > 1$).

Source	Degré de liberté
Array A	1
Dye D	1
Variété V	1
Gène G	$N-1$
Variété*Gène VG	$N-1$
Array*Gène AG	$N-1$
Dye*Gène DG	$N-1$
Erreur ϵ	$4N(R-1)$
Nombre total de degré de liberté	$4RN-1$

TAB. 3.6 – Tableau d'analyse de variance : *dye-swap* simple.

- $R = 1$
  On se place dans le cas $R = 1$ d'un seul spot par gène. Le modèle (3.44) est alors saturé car 4N paramètres doivent être estimés avec 4N observations : le terme d'erreur a 0 degré de liberté. Aussi, bien que le terme DG puisse être significatif, Kerr et *al.* [69] suggèrent de le considérer comme l'erreur. Le modèle choisi est alors

$$y_{ijkg} = \mu + A_i + D_j + V_k + G_g + AG_{ig} + VG_{kg} + \epsilon_{ijkg}. \tag{3.45}$$

- $R > 1$
  L'ensemble des facteurs du modèle (3.44) peut être estimé, l'erreur ayant un degré de liberté non nul.

Ainsi, l'estimation des facteurs VG est la même pour les trois modèles (3.42), (3.44) et (3.45). Ces modèles diffèrent bien entendu sur les valeurs des résidus estimés et sur la variance estimée $\widehat{\sigma^2}$. Bien que l'estimation de VG_{kg} ne dépende pas du modèle, sa précision est liée à l'estimation de σ^2. Le choix du modèle est donc un problème important car il détermine la qualité de l'estimation de VG_{kg}.

3.3.2.2 Plan d'expériences *dye-swap* multiple

Pour augmenter la précision et l'efficacité du plan d'expériences par *dye-swap*, on peut augmenter le nombre de puces et disposer de $I = 4, 6, 8$ puces ou plus. Ce type de plan est encore rarement utilisé mais devrait se développer avec la baisse des coûts expérimentaux. La structure de ce plan est plus complexe que dans le cas $I = 2$. En effet, les facteurs qui étaient précédemment confondus (tableau 3.5) ne le sont que partiellement ici. Dans tous les cas ($R = 1$ et $R > 1$), les modèles (3.42), (3.44) ou (3.45) peuvent être choisis pour l'analyse et il n'est pas nécessaire de considérer DG comme non significatif.

Cependant, dans ce modèle les facteurs principaux ne saturent pas le plan d'expériences recouvert par les plaques, les marqueurs et les variétés, c'est-à-dire que certaines interactions des facteurs principaux ne sont pas prises en compte par le modèle le plus complet (3.44). En effet, le facteur V a un degré de liberté et est partiellement confondu dans AD qui a $I-1$ degrés de liberté. Pour s'en convaincre, il suffit de voir que si le numéro de la puce et du dye sont connus, alors la variété est également connu. Cependant, la réciproque étant fausse, les facteurs sont partiellement confondus.

En remplaçant V par AD, on obtient alors le modèle suivant

$$y_{ijkgr} = \mu + A_i + D_j + AD_{ij} + G_g + AG_{ig} + DG_{jg} + VG_{kg} + \epsilon_{ijkgr}. \tag{3.46}$$

Ce modèle sature le plan d'expériences de telle manière que toutes les combinaisons entre les plaques, les marqueurs et les variétés sont prises en compte directement ou indirectement (V est confondu dans AD). On montre aussi que AV et AD sont partiellement confondus. On pourrait donc proposer de façon équivalente au modèle (3.46) les modèles

$$y_{ijkgr} = \mu + A_i + D_j + V_k + AD_{ij} + G_g + VG_{kg} + AG_{ig} + DG_{kg} + \epsilon_{ijkgr} \tag{3.47}$$
$$y_{ijkgr} = \mu + A_i + D_j + V_k + AV_{ik} + G_g + VG_{kg} + AG_{ig} + DG_{kg} + \epsilon_{ijkgr} \tag{3.48}$$

Dans les deux modèles (3.47) et (3.48), le nombre de degrés de liberté de V est 1, celui de AD et de AV est $I - 2$.

Les modèles (3.46), (3.47) et (3.48) étant équivalents, nous donnons le tableau d'analyse de variance (tableau 3.7) pour I puces et R réplications suivant l'écriture du modèle (3.46). Deux modèles sont dits équivalents si l'espace engendré par les deux matrices d'incidence X, correspondant aux deux modèles, définissent le même espace vectoriel.

3.3.3 Recherche de gènes différentiellement exprimés

Quelque soit le modèle choisi, la recherche de gènes différentiellement exprimés vise à identifier les gènes g tels que $dVG_g = VG_{2g} - VG_{1g}$ est significativement différent de 0. Pour cela, il est donc nécessaire d'estimer la valeur dVG_g.

Source	Degré de liberté
Array A	$I-1$
Dye D	1
Array*Dye	$I-1$
Gène G	$N-1$
Array*Gène	$(I-1)(N-1)$
Variété*Gène	$N-1$
Dye*Gène	$N-1$
Erreur	$(2R-1)NI-2N-I+2$
Nombre total de degrés de liberté	$2RNI-1$

TAB. 3.7 – Tableau d'analyse de variance : *dye-swap* multiple.

3.3.3.1 Estimation des paramètres d'intérêt biologique

Ayant posé la contrainte $VG_{1g}+VG_{2g}=0$ dans les modèles, l'estimation de dVG_g se ramène à l'estimation du paramètre VG_{1g} (ou de VG_{2g}) et $d\widehat{VG}_g = 2*\widehat{VG_{2g}} = -2*\widehat{VG_{1g}}$.

De manière générale, l'estimation des paramètres d'un modèle linéaire se résout en utilisant la matrice d'expériences X qui dépend directement du modèle et du plan d'expériences. Il suffit alors d'inverser la matrice X^tX de dimension (p,p) pour estimer l'ensemble des p paramètres. Cependant, cette approche ne peut pas être appliquée dans notre cas en raison du fait que p est de l'ordre de plusieurs milliers, et qu'il est difficile de calculer $(X^tX)^{-1}$. L'estimateur des moindre carrés correspond à la valeur du paramètre qui minimise la somme des carrés des erreurs (SCE). Sous l'hypothèse de normalité des erreurs, cet estimateur est également l'estimateur du maximum de vraisemblance. En considérant la somme des carrés des erreurs associée au modèle (3.47), on obtient alors la relation :

$$\text{SCE} = \sum_{ijkgr}(y_{ijkgr}-\mu-A_i-D_j-V_k-AD_{ij}-G_g-VG_{kg}-AG_{ig}-DG_{jg})^2 \qquad (3.49)$$

La somme se fait sur les indices i, j, k, g et r qui apparaissent dans le plan d'expériences équilibré. Les estimateurs sont obtenus en prenant les dérivées partielles de la SCE par rapport aux paramètres.

Le plan d'expériences étant équilibré, le calcul des dérivés partielles par rapport à l'interaction d'intérêt VG_{kg} permet d'obtenir

$$\frac{\text{SCE}}{\partial VG_{kg}} = -2*\sum_{ijr}(y_{ijkgr}-\mu-A_i-D_j-V_k-G_g-AD_{ij}-AG_{ig}-VG_{kg}-DG_{jg}). \quad (3.50)$$

Les indices k et g étant fixés, la somme est réalisée sur tous les indices (i,j,r) tels que

i, j, k, g, r représentent un ensemble d'indices pour le plan d'expériences.

On résoud l'équation $\frac{\partial SCE}{\partial VG_{kg}} = 0$ sous les contraintes de l'analyse de variance

$$\begin{aligned}
\sum_i A_i &= \sum_j D_j = 0 && (3.51)\\
\sum_i AD_{ij} &= \sum_j AD_{ij} = 0 && (3.52)\\
\sum_k VG_{kg} &= \sum_g VG_{kg} = 0 && (3.53)\\
\sum_i AG_{ig} &= \sum_g AG_{ig} = 0 && (3.54)\\
\sum_j DG_{jg} &= \sum_g DG_{jg} = 0. && (3.55)
\end{aligned}$$

On obtient alors la relation

$$\sum_{ijr} (y_{ijkgr} - \mu - V_k - G_g - VG_{kg}) = 0. \quad (3.56)$$

L'estimation des moindres carrés du paramètre VG_{kg} est donc donnée par

$$\begin{aligned}
\widehat{VG}_{kg} &= \bar{y}_{..kg.} - \widehat{V}_k - \widehat{G}_g - \widehat{\mu} && (3.57)\\
\widehat{VG}_{kg} &= \bar{y}_{..kg.} - \bar{y}_{..k..} - \bar{y}_{...g.} + \bar{y}_{.....}. && (3.58)
\end{aligned}$$

À l'aide des estimations de V et G, on obtient

$$d\widehat{VG}_g = \widehat{VG}_{2g} - \widehat{VG}_{1g} = \bar{y}_{..2g.} - \bar{y}_{..1g.} - \bar{y}_{..2..} + \bar{y}_{..1..}. \quad (3.59)$$

Afin d'identifier les gènes différentiellement exprimés, une approche simple consiste à estimer la loi de dVG_g sous l'hypothèse que le gène g n'est pas différentiellement exprimé. Ceci permet de réaliser le test $\mathrm{H}_0 : dVG_g = 0$ contre $\mathrm{H}_1 : dVG_g \neq 0$, l'hypothèse H_0 traduisant le fait qu'il n'y pas de différence d'expression entre les deux conditions expérimentales, et l'hypothèse H_1 donnant une différence d'expression.

On note $n = 2IRG$ le nombre total d'observations pour I puces. La valeur G est le nombre de gènes de l'étude pour lesquels les mesures sont répliquées R fois sur chaque puce. Étant donné un modèle d'analyse de variance, on note p le nombre de paramètres libres à estimer. L'écriture canonique associée à un modèle linéaire que nous avons introduit en (3.43) s'écrit

$$Y = X\theta + \epsilon \tag{3.60}$$

avec

- Y le vecteur aléatoire à expliquer et dont y_{ijkgr} est une réalisation. On note $\mathbf{y} \in I\!R^n$ le vecteur des observations.
- X est la matrice d'incidence connue telle que $X = \begin{bmatrix} \dot{X} \\ \vdots \\ \dot{X} \end{bmatrix}$, avec X de taille (n, p).

  La matrice $\dot{X}$ de taille $2RG$ est la matrice d'expériences correspondant à un *dye-swap* (c'est-à-dire 2 puces). Elle est de taille $(4RG, p)$.
- θ est le vecteur de taille p des paramètres du modèle.
- ϵ le vecteur des erreurs de taille n, réalisations indépendantes d'une loi $\mathcal{N}(0, \sigma^2)$.

Sous l'hypothèse gaussienne des résidus, l'estimateur du maximum de vraisemblance $\widehat{\theta}$ est défini par $\widehat{\theta} = (X'X)^{-1}X'Y$. De plus, $\widehat{\theta} \sim \mathcal{N}_p(\theta, \sigma^2(X'X)^{-1})$ est un estimateur sans biais de variance minimale de θ. On estime la variance de $\widehat{\theta}$ par

$$\widehat{\text{var}}(\widehat{\theta}) = \widehat{\sigma}^2(X'X)^{-1} \tag{3.61}$$

avec $\widehat{\sigma}^2 = \frac{1}{n-p}(\mathbf{y} - X\widehat{\theta})'(\mathbf{y} - X\widehat{\theta})$. Nous obtenons donc le résultat

$$\forall k, g \quad \widehat{\text{var}}(\widehat{VG}_{kg}) = \frac{\widehat{\sigma}^2}{2}\frac{G-1}{IRG}. \tag{3.62}$$

Avec la relation $\widehat{dVG}_g = 2\widehat{VG}_{2g}$, la variance de l'estimateur de dVG_g est donnée par

$$\forall g \quad \widehat{\text{var}}(\widehat{dVG}_g) = \widehat{\sigma}^2\frac{2(G-1)}{IRG}. \tag{3.63}$$

3.3.3.2 Tests d'hypothèses

Un gène est alors considéré comme différentiellement exprimé avec un niveau de confiance $1 - \alpha$ si $|\widehat{dVG}_g| \geq q_{1-\alpha/2}$ où $q_{1-\alpha/2}$ est le quantile d'ordre $1 - \frac{\alpha}{2}$ d'une loi normale centrée et de variance $\widehat{\sigma}^2 \times \frac{2(G-1)}{IRG}$.

Ainsi, le choix des gènes differntiellement exprimés dépend de l'estimation de σ et donc du modèle choisi pour l'analyse de variance. Un problème immédiat pour identifier ces gènes concerne donc le choix du modèle pour représenter le niveau d'expression des gènes.

Remarques :
Quels que soient les modèles ou les hypothèses, les approches précédentes consistent à réaliser N tests d'hypothèses, un pour chaque gène. On retrouve alors les problèmes liés aux tests multiples.

Dans les applications proposées, nous appliquons une correction de type Bonferroni. Cette correction n'est pas très appropriée dans la mesure où elle conduit à de nombreux faux négatifs (gènes classés à tort comme non différentiellement exprimés). Dans un premier temps, nous la choisissons pour sa simplicité de mise en oeuvre.

3.3.4 Choix du modèle d'analyse de variance

Nous avons présenté différents modèles pour représenter le logarithme des ratios d'expression. Afin d'identifier les gènes différentiellement exprimés, il convient de choisir un modèle. Pour résoudre ce problème, on propose d'utiliser des critères, souvent appelés critères d'information, tels que les critères AIC (*An Information Criterion* ou *Akaike Information Criterion*; [4, 5]) et BIC (*Bayesian Information Criterion*; [66, 111]).

La vraisemblance est une mesure de la qualité d'ajustement du modèle aux données. Mais le défaut de la vraisemblance est que plus le modèle est complexe, plus il peut s'adapter aux données et plus grande est la vraisemblance des paramètres de ce modèle. Les deux critères AIC et BIC proposent une pénalisation de la vraisemblance suivant la complexité du modèle. Ils cherchent à réaliser un compromis entre la complexité du modèle et la qualité d'ajustement des données au modèle.

Nous rappelons donc les principes de construction de ces deux critères, que nous utiliserons à nouveau dans le cadre de choix de modèles pour la classification.

3.3.4.1 Critère AIC

Le critère AIC, proposé par Akaike [4, 5], est basé sur la divergence de Kullback-Leibler (KL). En nous appuyant sur l'ouvrage de Burnham [24], nous rappelons les principaux aspects de ce critère. On note f la loi de probabilité dont sont issues nos observations et $g_\theta(x) = g(x;\theta)$ une famille paramétrique où le paramètre $\theta \in I\!\!R^d$. On cherche θ_0 tel que g_{θ_0} soit le plus proche de f au sens de la divergence de Kullback-Leibler. Pour simplifier cette présentation, on suppose que f fait partie de la famille de modèles considérés. Cette divergence entre deux distributions de probabilité f et g_θ est définie par

$$\begin{aligned} KL(f, g_\theta) &= \int f(x) \log \frac{f(x)}{g(x;\theta)} dx \\ &= E_f[\log(f(x))] - E_f[\log(g_\theta(x))]. \end{aligned}$$

Le premier terme de l'expression est une constante qui dépend uniquement de la vraie distribution. Le calcul de $E_f[\log(g(x;\theta))$ donne donc la valeur de $KL(f, g_\theta)$ à une constante près. On définit θ_0 comme la valeur du paramètre qui minimise la divergence de Kullback-Leibler, c'est-à-dire :

$$\theta_0 = \arg\min_\theta KL(f, g_\theta). \tag{3.64}$$

Cependant, cette minimisation nécessite de connaître la "vraie" loi f qui n'est pas connue. La valeur de θ étant également inconnue, on peut l'estimer à partir des observations. La variabilité liée à cette estimation doit alors être prise en compte. Si y est un ensemble d'observations, on note $\widehat{\theta}(y)$ l'estimateur du maximum de vraisemblance. La distinction entre les cas où θ est connu et θ est inconnu modifie l'approche de sélection de modèles. La divergence de Kullback-Leibler entre f et $g_{\widehat{\theta}(y)}$ vaut alors

$$KL(f, g_{\widehat{\theta}(y)}) = \int f(x) \log \frac{f(x)}{g(x; \widehat{\theta}(y))} dx.$$

Cette divergence est une variable aléatoire qui dépend des données observées y. On cherche alors à minimiser l'espérance de cette divergence :

$$\min \left[E_y\{KL(f, g_{\widehat{\theta}(y)})\} \right] = \min \left\{ \int f(x) \log(f(x)) dx - E_y \int f(x) \log g(x; \widehat{\theta}(y)) dx \right\}.$$

Ce qui est équivalent à

$$\max \left[Ey\, Ex\, \log g(x; \widehat{\theta}(y)) \right].$$

Une série d'approximations, basées sur des développements de Taylor et sur l'hypothèse que f appartient à la famille de modèles considérés, permet d'obtenir le critère AIC. Pour cette famille, le critère AIC pour un modèle M avec q paramètres libres, et dont $\ell(\widehat{\theta}(y)$ est la log-vraisemblance maximale est donné par l'expression

$$\mathrm{AIC}(M) = -2\ell(\widehat{\theta}(y)) + 2q. \tag{3.65}$$

Le modèle sélectionné est alors le modèle qui minimise le critère AIC. Intuitivement, le premier terme de l'équation (3.65) favorise un modèle pour lequel la vraisemblance maximale est forte et le second terme pénalise un modèle avec trop de paramètres.

3.3.4.2 Critère BIC

Ayant défini un ensemble de modèles probabilistes, les outils de la théorie bayésienne sont disponibles pour la sélection de modèles. Le critère BIC consiste à retenir le modèle le plus probable a posteriori. La probabilité a posteriori du modèle M_k, sachant les observations y, est donnée par

$$P(M_k|\mathbf{y}) = \frac{f(\mathbf{y}|M_k)P(M_k)}{\sum_M f(\mathbf{y}|M)P(M)}. \tag{3.66}$$

Dans cette expression, $P(M_k)$ est la probabilité a priori du modèle M_k. Le terme $f(\mathbf{y}|M_k)$ est appelé la vraisemblance marginale ou intégrée. Elle est définie par

$$f(\mathbf{y}|M_k) = \int f(\mathbf{y}|\theta_k, M_k)\pi(\theta_k|M_k) d\theta_k \tag{3.67}$$

où θ_k est le paramètre du modèle M_k, $f(\mathbf{y}|M_k, \theta_k)$ est la vraisemblance associée et $\pi(\theta_k|M_k)$ est la distribution a priori du paramètre θ_k. Parmi l'ensemble des modèles considérés, on choisit un modèle M_0 de référence. Pour comparer les modèles M_k et M_0, on s'intéresse aux rapports des probablités a posteriori (3.66) sous l'hypothèse que les modèles sont équiprobabales a priori. Si Δ est l'ensemble des modèles, on obtient

$$P(M_j) = \frac{1}{\text{card}(\Delta)} \qquad \forall M_j \in \Delta. \tag{3.68}$$

Le rapport des probabilités a posteriori est également nommé facteur de Bayes du modèle M_k contre le modèle M_0

$$B_{M_k M_0} = \frac{P(M_k|\mathbf{y})}{P(M_0|\mathbf{y})} = \frac{f(\mathbf{y}|M_k)}{f(\mathbf{y}|M_0)}. \tag{3.69}$$

La probabilité a posteriori du modèle M_k peut s'écrire

$$P(M_k|\mathbf{y}) = \frac{B_{M_k M_0} P(M_k)}{\sum_L B_{M_k M_L} P(M_L)}. \tag{3.70}$$

D'où suivant l'hypothèse (3.68)

$$\begin{aligned} P(M_k|\mathbf{y}) &\propto B_{M_k M_O} && (3.71)\\ &\propto f(\mathbf{y}|M_k) && (3.72) \end{aligned}$$

Comparer les probabilités a poteriori des différents modèles est donc équivalent à comparer les facteurs $B_{M_k M_0}$ ou les vraisemblances intégrées $f(\mathbf{y}|M_k)$. Le calcul de la vraisemblance intégrée nécessite de fixer les densités a priori $\pi(\theta_k|M_k)$ (voir 3.67). Le calcul du facteur de Bayes est réalisé de manière approchée en utilisant un critère (critère de Schwarz)

$$S = \ell(\widehat{\theta}_{M_k}) - \ell(\widehat{\theta}_{M_0}) - \frac{1}{2}(q_k - q_0) \log(n) \tag{3.73}$$

où $\ell(\widehat{\theta}_{M_k})$ et $\ell(\widehat{\theta}_{M_0})$ sont les maximum de vraisemblance des modèles M_k et M_0, q_k et q_0 le nombre de paramètres libres et n est le nombre d'observations. Sous certaines hypothèses de régularité, la quantité S vérifie

$$\frac{S - \log(B_{M_k M_0})}{\log(B_{M_k M_0})} \overset{n\to\infty}{\longrightarrow} 0 \tag{3.74}$$

La quantité S peut être vue comme une approximation du logarithme du facteur de Bayes. Pour comparer deux modèles M_k et $M_k^{'}$, on calcule le facteur de Bayes entre chacun des ces modèles et le modèle de référence M_0, puis on retient celui avec le plus grand facteur. Cette procédure est indépendante du modèle M_0 et revient à choisir le modèle qui minimise le critère BIC donné par

$$\text{BIC}(M_k) = -2\ell(\widehat{\theta}_{M_k}) + q_k \ln(n). \tag{3.75}$$

On retrouve dans la formule (3.75) la même interprétation que pour AIC. Cependant la pénalisation du nombre de paramètres est beaucoup plus forte car elle vaut $q_k \ln(n)$, avec n le nombre d'observations.

3.3.5 Hypothèse de non normalité des résidus

Une fois le modèle choisi, si l'hypothèse de normalité sur les résidus semble peu raisonnable, Kerr *et al.* [69] propose d'utiliser la technique du bootstrap [42] afin d'estimer la loi de $d\widehat{VG}_g$.

Le principe étant identique pour tous les modèles, nous choisissons le modèle (3.42). Après avoir estimé les paramètres A, D, V, G, VG en minimisant la somme des carrés des erreurs, on calcule l'ensemble des résidus associés au modèle par $\widehat{\epsilon}_{ijkgr} = y_{ijgr} - \widehat{y}_{ijkgr}$ avec $\widehat{y}_{ijkgr} = \widehat{\mu} + \widehat{A}_i + \widehat{D}_j + \widehat{V}_k + \widehat{G}_g + \widehat{VG}_{kg}$. Le principe consiste à simuler des données $y^{\star}_{ijkgr}$ à partir des paramètres et des résidus estimés avec

$$y^{\star}_{ijkgr} = \widehat{\mu} + \widehat{A}_i + \widehat{D}_j + \widehat{V}_k + \widehat{G}_g + \widehat{VG}_{kg} + \epsilon^{\star}_{ijkgr}. \tag{3.76}$$

Les valeurs $\epsilon^{\star}_{ijkgr}$ correspondent à des tirages aléatoires de la loi $\sqrt{\frac{n}{n-p}}\widehat{F}$ où $\widehat{F}$ est la distribution empirique des résidus. Le terme $\sqrt{\frac{n}{n-p}}$ permet d'obtenir une distribution avec la même variance que la distribution des erreurs et un estimateur consistant de la variance de $\widehat{dVG}$ [127]. Pour chaque échantillon simulé, on calcule alors $d\widehat{VG}^{\star}_g = 2\widehat{VG}^{\star}_{1g}$ pour $g = 1, \ldots, N$ d'après la formule (3.57). L'intervalle de confiance au niveau $1 - \alpha$ est obtenu à partir des quantiles d'ordre $\frac{\alpha}{2}$ et $1 - \frac{\alpha}{2}$ de la distribution empirique des $d\widehat{VG}_g^{\star}$.

L'approche bootstrap permet alors de valider les gènes sélectionnés de façon plus robuste.

3.3.6 Applications 1

On dispose de deux expériences comportant 994 gènes déposés en deux exemplaires sur chaque lame. Pour chaque expérience, un seul *dye-swap* est réalisé. On appelle data1 et data2 les données issues des deux expériences. Nous présentons tout d'abord les 6 modèles dont les termes d'interaction ont un intérêt dans cette étude. Le choix d'un modèle pour sélectionner les gènes d'intérêt se fera à l'aide de la représentation des résidus et du calcul des critères AIC et BIC.

3.3.6.1 Les modèles

Nous proposons de considérer les 6 modèles suivants.

$$\begin{array}{lcl}
(\mathrm{M1}) : y_{ijkgr} & = & \mu + A_i + D_j + V_k + G_g + VG_{kg} + \epsilon_{ijkgr} \\
(\mathrm{M2}) : y_{ijkgr} & = & \mu + A_i + D_j + V_k + G_g + VG_{kg} + DG_{jg} + \epsilon_{ijkgr} \\
(\mathrm{M3}) : y_{ijkgr} & = & \mu + A_i + D_j + V_k + G_g + VG_{kg} + AG_{ig} + \epsilon_{ijkgr} \\
(\mathrm{M4}) : y_{ijkgr} & = & \mu + A_i + D_j + V_k + G_g + VG_{kg} + AG_{igr} + \epsilon_{ijkgr} \\
(\mathrm{M5}) : y_{ijkgr} & = & \mu + A_i + D_j + V_k + G_g + VG_{kg} + AG_{ig} + DG_{jg} + \epsilon_{ijkgr} \\
(\mathrm{M6}) : y_{ijkgr} & = & \mu + A_i + D_j + V_k + G_g + VG_{kg} + AG_{igr} + DG_{jg} + \epsilon_{ijkgr}
\end{array}$$

Les modèles ($M3$, $M5$) et ($M4$, $M6$) diffèrent par la prise en compte de l'indice r de répétition dans le terme AG_{igr}. Kerr et *al.* [69] ont décrit ce facteur comme l'effet spot.

3.3.6.2 Représentation des résidus et critères AIC/BIC

La représentation des résidus en fonction des valeurs estimées par le modèle permet de contrôler la non corrélation des résidus et de juger rapidement si l'hypothèse de normalité est acceptable.

Si on considère les données data1, la représentation des résidus est donnée sur les figures (3.4) pour les 6 modèles envisagés. On observe deux cas.

- Les modèles $M1$, $M2$, $M5$ et $M6$ mettent en évidence un aspect structuré des résidus. En raison de l'hypothèse de normalité des résidus, ces modèles donnent une mauvaise représentation des niveaux d'expression.
- Seuls les résultats des modèles $M3$ et $M4$ nous confortent dans le choix de l'hypothèse gaussienne.

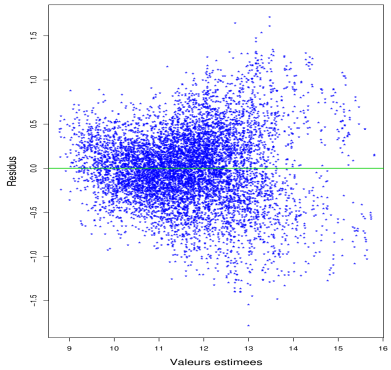
Le tableau 3.8 donne les valeurs du kurtosis (coefficient d'aplatissement) et du skewness (coefficient d'asymétrie) des résidus pour les 6 modèles. Bien que pour les 6 modèles, ces critères soient assez proches de ceux d'une loi normale, seuls les modèles $M3$ et $M4$ peuvent être envisagés en raison des remarques précédentes.

Data1	$M1$	$M2$	$M3$	$M4$	$M5$	$M6$	Loi normale
Kurtosis	3.49	3.57	4.22	4.10	5.5	5.86	3
Skewness	-5.10^{-4}	-10^{-4}	-6.10^{-3}	3.10^{-16}	-2^{-15}	10^{-15}	0

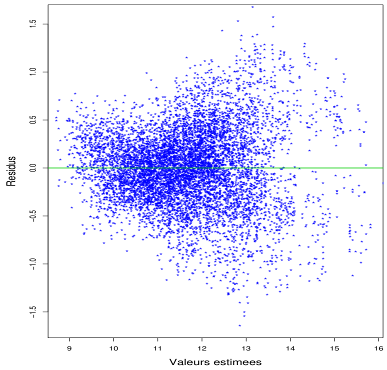
TAB. 3.8 – *Kurtosis et skewness (données Data1)*

Cette première approche par la représentation des résidus permet de retenir deux modèles intéressants (les modèles $M3$ et $M4$) qui diffèrent par le paramètre r de répétition. Le modèle $M4$ (Kerr et *al.* [69]) nécessite d'estimer un nombre plus important de paramètres en comparaison du modèle $M3$. Suivant le principe de parcimonie, nous souhaitons connaître le modèle le plus approprié pour la modélisation des niveaux d'expression.

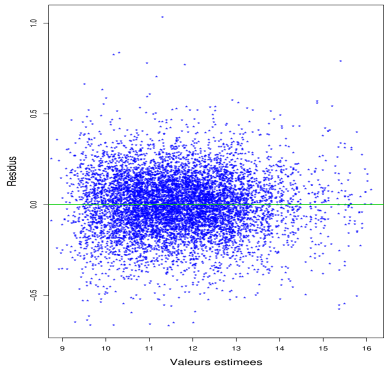
Nous avons calculé les critères AIC et BIC pour chacun des 6 cas analysés ci-dessus pour les données data1. Le tableau 3.9 présente les valeurs obtenues pour les critères AIC et BIC en fonction des différents modèles selon leur niveau de complexité croissant. On constate que plus le modèle est complexe, plus le critère AIC décroît, ce qui ne permet pas de choisir un modèle. Cela confirme la tendance du critère AIC à choisir des modèles trop complexes. Le critère BIC indique le modèle $M3$ comme le plus approprié.



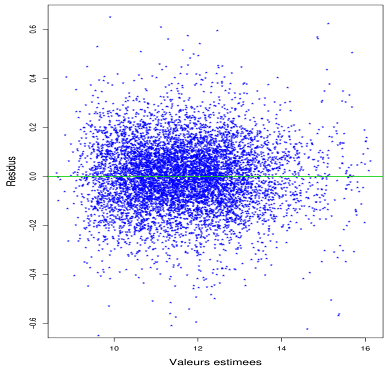
(a) Modèle $M1$



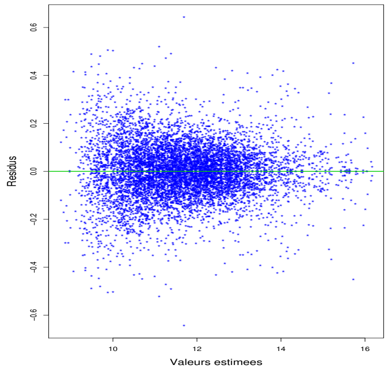
(b) Modèle $M2$



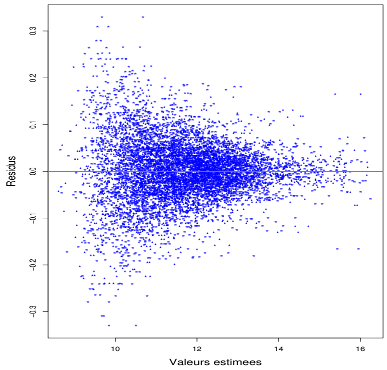
(c) Modèle $M3$



(d) Modèle $M4$



(e) Modèle $M5$



(f) Modèle $M6$

FIG. 3.4 – *Répartition des résidus suivant différents modèles d'analyse de variance*

Ayant choisi le modèle $M3$ pour représenter le niveau d'expression des gènes, nous pouvons réaliser l'étude des gènes différemment exprimés. Le graphique 3.5 correspond aux résultats sur la recherche de gènes différemment exprimés. L'erreur CER est fixée à $\frac{1\%}{N}$. Nous

Modèles	$M1$	$M2$	$M3$	$M4$	$M5$	$M6$
AIC	7168	8493	-5766	-6851	-9636	**-16439**
BIC	21060	29318	**15058**	20906	18120	18250

TAB. 3.9 – *Critères AIC et BIC pour différents modèles d'analyse de variance (données Data1)*

représentons les estimations de dVG_g à l'aide d'une représentation Q-Qplot (figure 3.5). Cette représentation traduit bien que l'hypothèse gaussienne est acceptable. L'approche bootstrap donne par ailleurs exactement les mêmes résultats pour les gènes différemment exprimés.

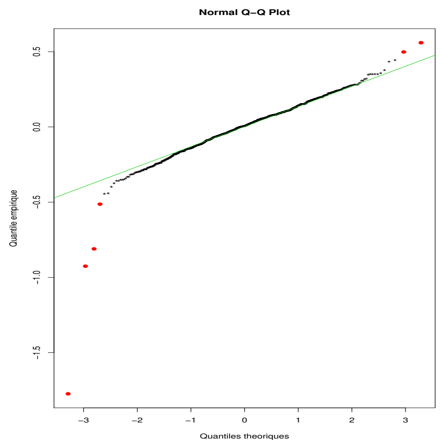


FIG. 3.5 – *Gènes différentiellement exprimés (données Data1)*

3.3.7 Application 2

Dans cette deuxième application, nous souhaitons mettre en évidence que dans certains cas l'analyse peut être plus difficile, traduisant les limites des modèles d'analyse de variance pour représenter les niveaux d'expression et les problèmes de variabilité. Nous continuons à utiliser l'ensemble des 6 modèles définis précédemment.

Sur le second jeu de données data2, seuls les modèles $M3$ et $M4$ donnent une représentation intéressante des résidus. Ce sont les seuls modèles pour lesquels l'homosédasticité semble respectée.

Le tableau 3.10 donne les valeurs du kurtosis et du skewness des résidus pour les 6 modèles.

Ces résultats traduisent que seul le modèle $M4$ permet de représenter le niveau d'expression des gènes avec nos hypothèses. C'est en effet le seul des modèles envisagés pour lequel les hypothèses de normalité semblent respectées. De ce fait, il est difficile de choisir un

Data4	$M1$	$M2$	$M3$	$M4$	$M5$	$M6$	Loi normale
Kurtosis	10.92	13.07	36.15	4.25	94.87	5.95	3
Skewness	0.61	0.72	0.14	6.10^{-17}	-10^{-14}	10^{-15}	0

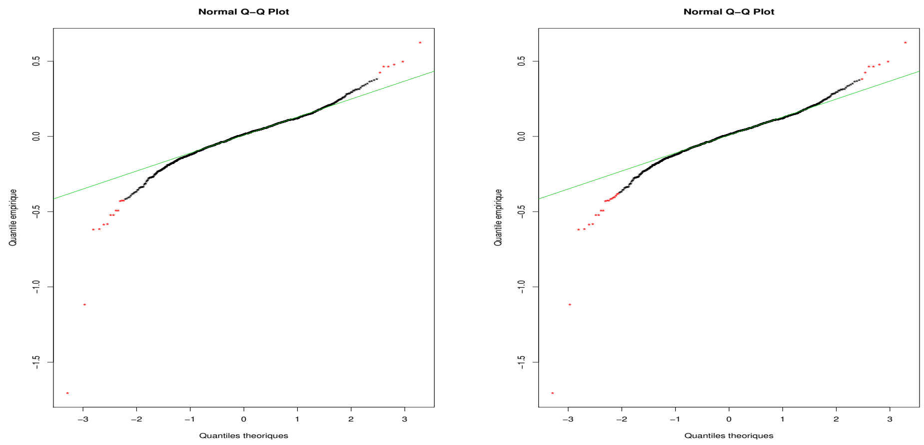
TAB. 3.10 – *Kurtosis et skewness (données Data2)*

modèle suivant les critères AIC et BIC si les hypothèses de normalité ne sont pas vérifiées. Le tableau 3.11 montre qu'en basant notre seul choix sur le critère BIC, nous aurions alors choisi le modèle $M3$.

Modèles	$M1$	$M2$	$M3$	$M4$	$M5$	$M6$
AIC	3268	4195	-4052	-6979	-6157	-1588
BIC	17161	25020	16771	20777	21599	18800

TAB. 3.11 – *Critères AIC et BIC selon les différents modèles (données Data2)*

Nous choisissons le modèle $M4$ pour l'étude de gènes différemment exprimés. Même si pour ce modèle l'hypothèse de normalité reste acceptable, une approche bootstrap peut être envisagée. Sur la figure 3.6, nous comparons les résulats pour le modèle $M4$ entre une approche bootstrap et une approche basée sur l'hypothèse gaussienne des résidus.



(a) (b)

FIG. 3.6 – *Gènes différentiellement exprimés suivant l'hypothèse gaussienne (a) et une approche bootstrap (b) (données Data2)*

L'approche bootstrap donne des résultats sensiblement différents. Sous l'hypothèse gaussienne, le nombre de gènes induits est de 6 et le nombre de gènes réprimés de 14. Avec une approche bootstrap, ces valeurs sont plus grandes et valent 7 et 20 respectivement.

3.3.8 Commentaires et perspectives

Les analyses menées sur ces deux expériences mettent en évidence la difficulté du choix d'un modèle pour identifier des gènes différentiellement exprimés et certaines limites quant à l'utilisation des modèles d'analyse de variance.

Dans un premier temps, nous avons décrit plusieurs modèles pour la représentation des niveaux d'expression. L'utilisation de l'analyse de variance dans le cadre des données d'expression mérite donc une attention particulière pour le choix du modèle. De plus, si le cadre habituel de l'hypothèse de normalité des résidus n'est pas vérifiée, l'approche proposée par Kerr *et al.*, basée sur le bootstrap est une alternative intéressante.

La nature des modèles sélectionnés entre ces deux expériences est différente. Un point intéressant concerne le nombre de paramètres des modèles $M3$ et $M4$. Le dernier a un nombre nettement plus important de paramètres que le premier. Pour le modèle $M3$, le nombre de degrés de liberté du terme AG est $(I-1)*(G-1)$, alors que pour le modèle $M4$ il en comporte $R(I-1)*(G-1)$. Kerr *et al.* introduisent l'indice de répétition dans le terme AG et décrivent l'effet AG_{igr} comme un effet spot. Ils proposent d'expliquer cet effet par des variations de concentrations de sondes déposées sur la puce. Dans l'étude des données data1, le modèle $M3$ s'ajuste bien aux données expérimentales et il n'est pas nécessaire d'introduire l'indice de répétition dans le modèle. Nous pensons que le nombre important de paramètres du modèle $M4$ pour l'étude des données Data2 traduit d'avantage une mauvaise modélisation des problèmes de variabilité qu'un effet spot.

On peut envisager une décomposition plus précise des structures de variance et une prise en compte d'éventuelles corrélations dans les données. Nous suggérons de nous intéresser aux modèles mixtes qui permettent de prendre en compte non seulement des effets fixes (comme dans les modèles précédents), mais aussi des effets aléatoires.

3.4 Les modèles mixtes

En raison de la nature de certains effets, il semble raisonnable de distinguer deux types de facteurs : les facteurs à effets fixes et ceux à effets aléatoires. Les facteurs à effets fixes ont un nombre fini de niveaux et les données se répartissent sur ces différents niveaux. Les facteurs à effets aléatoires ont un nombre potentiellement infini de niveaux et les données (en nombre fini) fournissent un échantillon de ces niveaux. De tels modèles sont appelés modèles mixtes (Searle *et al.* [112]) et ont été appliqués pour la normalisation des données et la recherche de gènes différentiellement exprimés par Wolfinger *et al.* [126]. En introduisant différentes composantes de la variance, ils sont ainsi plus précis. Nous nous intéressons au cas des modèles linéaires mixtes que l'on note L2M.

Les approches que nous avons décrites précédemment considéraient que tous les effets étaient fixes. Certains facteurs, comme le facteur gène (G) ou le facteur array (A), pourraient être représentés par des effets aléatoires. Les modèles mixtes semblent donc bien appropriés. L'objectif de cette présentation sur les modèles mixtes est double. D'abord, nous souhaitons comprendre comment les différents effets peuvent intervenir dans la modélisation des niveaux d'expression. Ensuite, cela nous permet d'introduire les notions fondamentales qui seront utilisées pour la classification de profils d'expression. Ce lien avec la classification traduit ainsi les connections importantes entre les problèmes de classification, de recherche de gènes différentiellement exprimés et de représentation des niveaux d'expression.

Dans une première partie, nous rappelons les principes des modèles mixtes en décrivant plus particulièrement les méthodes d'estimation des paramètres de ces modèles. Cette présentation s'appuie sur les travaux de Trottier [119] et de Searle *et al.* [112]. En deuxième partie, nous essaierons de voir comment ces modèles peuvent s'appliquer aux données de biopuces.

3.4.1 Modélisation avec effets aléatoires

Comme nous l'avons décrit, le problème de variabilité est fondamental pour l'étude des données de biopuces. Les modèles mixtes permettent de mieux comprendre ces problèmes. Nous présentons d'abord le cas des modèles linéaires mixtes. Nous donnons ensuite quelques résultats sur l'application de ces modèles pour l'étude de données de biopuces.

3.4.1.1 Les modèles linéaires mixtes ou L2M

L'écriture canonique associée à un modèle linéaire mixte est

$$Y = \underbrace{X\beta}_{\text{partie effets fixes}} + \underbrace{U\xi}_{\text{partie effets aléatoires}} + \epsilon \tag{3.77}$$

où

- Y : vecteur aléatoire à expliquer, de taille N, et dont $y = (y_1, \ldots, y_N)'$ est une réalisation.
- β : vecteur des paramètres inconnus des effets fixes, de taille p, et X, de dimension (N, p), sa matrice d'incidence supposée fixe et connue.
- ξ : vecteur des effets aléatoires de taille p. Si L est le nombre d'effets aléatoires considérés, ce vecteur se décompose en $\xi = (\xi_1', \ldots, \xi_L')'$. Chaque composant ξ_j est un vecteur aléatoire de dimension q_j. Il est constitué des q_j réalisations du $j^{\text{ème}}$ effet aléatoire, observées au sein des données ($\sum_{j=1}^{L} q_j = q$). La matrice d'incidence U, de taille (N, q), est connue.
- ϵ : vecteur aléatoire des erreurs de taille N indépendant de ξ et $\epsilon \sim \mathcal{N}_N(0, \sigma_0^2 \Lambda_0)$. On note $S = \sigma_0^2 \Lambda_0$.

On suppose de plus que $\forall j \in \{1, \ldots, L\}, \xi_j \sim \mathcal{N}_{q_j}(0, \sigma_j^2 B_j)$. Par ailleurs, les effets aléatoires ξ_i et ξ_j sont indépendants $\forall (i, j) \in \{1, \ldots, L\}^2$. De ce fait, $\xi \sim \mathcal{N}_N(0, B)$ où B est une matrice diagonale par blocs. La matrice d'incidence U des effets aléatoires est connue et formée des différentes matrices d'incidence U_j de chacun des effets aléatoires.

Suivant ces hypothèses, on note

$$\begin{aligned} \mathrm{var}(Y) = \Gamma &= S + UBU^{'} & (3.78)\\ &= \sigma^2\Lambda_0 + \sum_{j=1}^{L} \sigma_j^2 U_j B_j U_j^{'}. & (3.79) \end{aligned}$$

Avec $\forall j \in \{1, \dots, L\}, \Lambda_j = U_j B_j U_j^{'}$, cela donne

$$\Gamma = \sum_{j=0}^{L} \sigma_j^2 \Lambda_j. \quad (3.80)$$

L'expression de Γ montre que la variance est scindée en plusieurs composantes, que l'on appelle les composantes de la variance. De part les hypothèses de normalité sur les effets aléatoires, il s'en déduit les propriétés suivantes

- $Y|\xi \sim \mathcal{N}_N(X\beta + U\xi, S)$
- $Y \sim \mathcal{N}_N(X\beta, \Gamma)$
- $\begin{pmatrix} Y \\ \xi \end{pmatrix} \sim \mathcal{N}_{N+q}\left(\begin{pmatrix} X\beta \\ 0 \end{pmatrix}, \begin{pmatrix} \Lambda_0 + UBU^{'} & UB \\ BU^{'} & B \end{pmatrix}\right)$

Dans un modèle mixte de la forme (3.77), le vecteur des effets fixes β ainsi que le vecteur des $L+1$ paramètres de variance sont inconnus. Nous présentons dans la suite une méthode d'estimation de ces paramètres.

3.4.1.2 Estimation dans les L2M : algorithme EM pour ML

Il existe différentes approches pour l'estimation des paramètres d'un modèle linéaire mixte. L'ouvrage de Searle *et al.* [112] donne une présentation détaillée de ces méthodes ainsi que leurs propriétés. Ici, nous limitons notre présentation à l'application de l'algorithme EM pour la maximisation de la vraisemblance (ML ou maximum likelihood).

Suivant les notations de Trottier [119], $\sigma^2 = (\sigma_0^2, \sigma_1^2, \dots, \sigma_L^2)^{'}$ est le vecteur des composantes de la variance. Le vecteur des observations est noté y. La vraisemblance associée à notre modèle s'écrit

$$\mathcal{L}(\beta, \sigma^2; y) = \frac{1}{(2\pi)^{N/2}|\Gamma|^{1/2}} \exp -\frac{1}{2}(y - X\beta)'\Gamma^{-1}(y - X\beta). \quad (3.81)$$

La fonction de log-vraisemblance est donnée par

$$\ell(\beta, \sigma^2; y) = -\frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln(|\Gamma|) - \frac{1}{2}(y - X\beta)^{'}\Gamma^{-1}(y - X\beta). \quad (3.82)$$

À partir de l'équation (3.82), on envisage d'estimer les paramètres qui maximisent la vraisemblance. Diverses alternatives peuvent être proposées [112]. Nous présentons l'application de l'algorithme EM [36] pour l'estimation des paramètres du maximum de vraisemblance dans le cas particulier des L2M.

Les données observées (ou incomplètes suivant la terminologie EM) sont constituées du vecteur y. Le vecteur x des données complètes est défini par $x = (y^{'}, \xi^{'})^{'}$ avec ξ le vecteur des effets aléatoires $\xi^{'} = (\xi_1^{'}, \ldots, \xi_L^{'})^{'}$. La fonction de vraisemblance associée aux données complètes s'écrit

$$\ell(\beta, \sigma^2; x) = -\frac{1}{2}(\sum_{j=0}^{L} q_j) \ln(2\pi) - \frac{1}{2}\sum_{j=0}^{L}(q_j \ln(\sigma_j^2) + \ln(|B_j|)) - \frac{1}{2}\sum_{j=0}^{L} \frac{\xi_j^{'}\Lambda_j^{-1}\xi_j}{\sigma_j^2} \tag{3.83}$$

avec $\xi_0 = \epsilon = y - X\beta - \sum_{j=1}^{L} U_j\xi_j, q_0 = N, \Lambda_0 = B_0, U_0 = I_N$.

Si l'on observait une réalisation du vecteur aléatoire ξ, les estimateurs du maximum de vraisemblance associée aux données complètes s'obtiendraient par

$$\begin{aligned}
\hat{\sigma}_j^2 &= \frac{\xi^{'}\Lambda_j^{-1}\xi_j}{q_j} \quad \forall j \in \{0, \ldots, L\} \\
(X^{\prime}\Lambda_0^{-1}X)\hat{\beta} &= X^{'}\Lambda_0^{-1}(y - \sum_{j=1}^{L} U_j\xi_j).
\end{aligned}$$

Or, comme on ne dispose que des observations y, l'algorithme EM consiste à remplacer les statistiques exhaustives $\xi_j^{'}\Lambda_j^{-1}\xi_j$ et $y - \sum_{j=1}^{L} U_j\xi_j$ par leurs espérances conditionnelles sachant y et les valeurs courantes des paramètres. Pour cela, on utilise les résultats suivants sur le conditionnement des variables aléatoires normales. $\forall j \in \{0, \ldots, L\}$

$$\begin{aligned}
E(\xi_j|y) &= \sigma_j^2\Lambda_j U_j^{'}\Gamma^{-1}(y - X\beta) \\
E(\xi_j^{'}\Lambda_j^{-1}\xi_j|y) &= \sigma_j^4(y - X\beta)^{'}\Gamma^{-1}\Lambda_j\Gamma^{-1}(y - X\beta) + \mathrm{tr}(\sigma_j^2 I_{q_j} - \sigma_j^4 U_j^{'}\Lambda^{-1}U_j B_j) \\
&= \sigma_j^4(y - X\beta)^{'}\Gamma^{-1}\Lambda_j\Gamma^{-1}(y - X\beta) + q_j\sigma_j^2\mathrm{tr}(\sigma_j^2 I_{q_j} - \sigma_j^4 U_j^{'}\Lambda^{-1}U_j B_j)
\end{aligned}$$

$$\begin{aligned}
E(Y - \sum_{j=1}^{L} U_j\xi_j|y) &= y - \sum_{j=1}^{K} \sigma_j^2\Lambda_j\Gamma^{-1}(y - X\beta) \\
&= y - (\Gamma - \sigma_0^2\Lambda_0)\Gamma^{-1}(y - X\beta) \\
&= X\beta + \sigma_0^2\Lambda_0\Gamma^{-1}(y - X\beta).
\end{aligned}$$

L'application de l'algorithme EM pour l'estimation des paramètres du maximum de vraisemblance conduit au schéma itératif suivant. À partir des estimations obtenues à l'itération m, les nouvelles valeurs des paramètres à l'étape $m + 1$ sont données par

$$\begin{aligned}
q_j \sigma_j^{2(m+1)} &= \sigma_j^{4(m)}(y - X\beta^{(m)})' \Gamma^{-1(m)} \Lambda^{-1(m)} (y - X\beta^{(m)}) + q_j \sigma_j^{2(m)} - \sigma_j^{4(m)} \mathrm{tr}(\Gamma^{-1(m)} \Lambda_j) \\
X\beta^{(m+1)} &= X(X'\Lambda_0^{-1}X)^{-1} X' \Lambda_0^{-1} (X\beta^{(m)} + \sigma_0^{2(m)} \Lambda_0 \Gamma^{-1(m)} (y - X\beta^{(m)})).
\end{aligned}$$

Nous venons de présenter une méthode d'estimation des paramètres d'un modèle linéaire mixte construite sur la maximisation de la vraisemblance via l'algorithme EM. L'estimation du maximum de vraisemblance restreint (REML) se focalise d'avantage sur l'estimation des composantes de la variance. Pour cela, on fait disparaître momentanément les effets fixes pour ne maximiser que la partie de la vraisemblance concernant les composantes et indépendante des effets fixes. Dans ce cadre, la méthodologie EM peut également être appliquée. L'ouvrage de Searle *et al.* [112] détaille cette approche ainsi que d'autres.

3.4.2 Application à l'étude des niveaux d'expression

Les modèles d'analyse de variance que nous avons exposés donnent une représentation des niveaux d'expression pour l'ensemble des mesures obtenues. Cette solution globale paraît néanmoins trop directe. Une approche plus fine mais plus fastidieuse consiste à réaliser une analyse de variance par gène. Cette extension de l'analyse de variance à une analyse gène par gène est envisagée par Dudoit *et al.* [40] en conclusion de leur travaux. Elle conduit alors à estimer un effet *dye* D et un terme d'erreur différents pour tous les gènes, ce qui donne une meilleure représentation des niveaux d'expression. Wolfinger *et al.* [126] et Chu *et al.* [30] envisage cette solution tout en proposant des modèles plus complexes que l'analyse de variance. Dans leur approche, Wolfinger *et al.* considèrent les effets A et AV comme des effets aléatoires et introduisent ainsi pour chaque gène des corrélations entre les observations relatives à une même puce. Les gènes différentiellement exprimés sont identifiés grâce à un test du rapport de vraisemblance sur les effets fixes VG_{kg}. Néanmoins, cette approche requiert un nombre suffisant de répétitions pour estimer l'ensemble des paramètres des modèles mixtes et n'a pu être envisagée dans nos études. Avec quatre observations par gène, un modèle comportant l'effet moyen μ et les trois facteurs principaux A, D, V est déjà saturé.

Un nombre trop faible de répétition impose donc une analyse globale. L'utilisation des modèles d'analyse de variance telle que nous l'avons décrite implique un nombre très important de paramètres à estimer. Cela peut conduire à des valeurs extrêmes pour l'estimation de certains effets. Une approche permettant d'éviter ce problème consisterait à considérer les facteurs gène G et VG comme des effets aléatoires. Kerr *et al.* [69] donne une remarque similaire dans leur commentaire. Une telle approche complique néanmoins l'identification de gènes différentiellement exprimés. En effet, il n'est alors plus possible de réaliser un test d'hypothèses comme nous l'avons décrit précédemment. Une axe de recherche serait d'exploiter le principe du BLUP *Best Linear Unbiased Predictor* [101] qui permet de prédire la réalisation des effets aléatoires VG_{kg}.

3.5 Conclusion

Ce chapitre relativement long traduit les nombreux développements qui ont été proposés pour le problème de recherche de gènes différemment exprimés. Le problème de variabilité lié à l'étude des données de biopuces est un problème fondamental pour l'analyse des résultats. Pour cette raison, plusieurs travaux doivent être menés sur ce point. Deux axes de recherche semblent intéressants.

- Afin de mener une étude gène par gène, un plan d'expériences comportant suffisamment de répétitions doit être envisagé pour une estimation correcte de l'ensemble des paramètres. Les modèles mixtes qui semblent bien appropriés peuvent alors être étudiés.
- Le second point concerne l'identification de gènes différentiellement exprimés. À partir des modèles précédents, il faut alors montrer comment les erreurs FDR ou FWE peuvent être contrôler dans le processus de sélection des gènes.

En conclusion, le problème de variabilité est un problème fondamental dans l'étude des données de biopuces. Un travail important doit encore être fait sur ce problème si l'on souhaite utiliser et développer les techniques de puces à ADN pour l'étude de l'expression des gènes.

Chapitre 4

Étude des profils d'expression

Les possibilités des puces à ADN ne se limitent pas à la comparaison des niveaux de transcription entre deux échantillons. Afin de déterminer des caractéristiques fonctionnelles des gènes, des expériences plus complexes doivent être réalisées : étude de la différenciation cellulaire, réponse transcriptionnelle à différents facteurs, comparaison de différents tissus, effets de différentes drogues. Bien que les problématiques biologiques soient très différentes, les méthodes liées à l'analyse de ces données restent assez proches. On peut les diviser en deux catégories.

- Définir une règle de discrimination entre différents types cellulaires, entre différents tissus (trouver par exemple une règle permettant de discriminer des cellules tumorales de cellule saines).
- Identifier différents types cellulaires ayant des caractéristiques communes (classer par exemple des tissus cellulaires).

Les mêmes problèmes peuvent être posés lorsque l'étude porte directement sur des gènes. On souhaite par exemple définir une règle de décision qui permet d'affecter un gène, dont la fonction est inconnue, dans une classe dont certaines caractéristiques fonctionnelles sont connues. Ou encore, étant donné un ensemble de mesures sur plusieurs milliers de gènes, on veut définir des classes de gènes coexprimés au cours du temps.

Golub *et al.* [57] définissent ces deux problématiques par *class prediction* et *class discovery*. D'un point de vue statistique, il s'agit de problèmes bien connus : la discrimination (classification supervisée) et la classification (classification non supervisée).

Dans ce chapitre, nous rappelons les méthodes les plus utilisées pour l'étude de profils d'expression, en développant plus particulièrement celles de la classification. Nous introduisons pour cela certaines notations. Les données expérimentales sont représentées par une matrice X de taille (N, p). L'ensemble des N unités statistiques (gènes ou tissus) est décrit par p variables explicatives $X^1, \ldots, X^p$. Chaque individu $\mathbf{x}_i$ se défini par p mesures (les p variables descriptives) et $\mathbf{x}_i \in I\!R^p$.

4.1 L'analyse discriminante

Dans le cadre de l'analyse discriminante, les unités statistiques sont caractérisées par une variable qualitative Z qui définit K groupes connus a priori. L'objectif de l'analyse discriminante est d'expliquer et de prédire pour un nouvel individu son appartenance à l'un des groupes. Cette prédiction s'effectue grâce aux variables explicatives.

Il existe de nombreuses méthodes d'analyse discriminante et il serait trop long de dresser un panorama. Parmi ces techniques, les plus utilisées sont l'analyse discriminante linéaire et quadratique [105, 87], la méthode des k plus proches voisins [62], les arbres de décisions [18] et les réseaux de neurones [99]. D'autres approches, plus sophistiquées, ont également été développées : la modélisation par mélange [60] ou les *support vector machines* (SVM) [121].

Dudoit *et al.* [39] mènent une étude très complète sur différentes méthodes (discrimination linéaire et quadratique, arbre de décision, boosting [62]) pour trois jeux de données différents [6, 57, 103]. Une conclusion intéressante de leurs travaux est que des classifieurs simples, comme la méthode du plus proche voisin, donne des résultats satisfaisants en comparaison à des méthodes plus complexes.

Ben-Dor *et al.* [10] présentent également une analyse comparative de différentes approches. Le problème abordé vise à prédire si un tissu est sain ou cancéreux. Dans ces travaux, les unités statistiques sont des tissus et les variables descriptives des gènes dont on mesure le niveau de transcription. Comme il semble évident que certaines variables (gènes) n'apportent aucune information pour la règle de décision, il est utile d'identifier les variables les plus discriminantes. Dudoit *et al.* soulevaient ce problème dans leurs conclusions. Ben-Dor *et al.* définissent un score pour chaque gène qui permet ainsi de sélectionner un sous-ensemble de gènes les plus discriminants.

Dans les deux travaux précédents, l'unité statistique correspond à un tissu cellulaire et les variables descriptives à des gènes. Le problème de discrimination peut également se poser en considérant les gènes comme unités statistiques. Brown *et al.* [19] décrivent une méthode pour la classification fonctionnelle, en se basant sur l'expression des gènes dans différentes conditions. Au préalable, ils déterminent un ensemble de 6 classes fonctionnelles définies à l'aide d'une base de données (MIPS Yeast Genome Database, MYGD). La technique des SVM est ensuite utilisée pour prédire une classe aux gènes de fonction inconnue. Cette discrimination est réalisée à l'aide de différentes expériences de puces à ADN sur la levure *Saccharomyces cerevisiae* : étude du shift diuaxic [37], de la division cellulaire [114], de la sporulation [29]. Ce sont d'ailleurs les mêmes données que Eisen *et al.* [48] analysent par une méthode de classification.

Les méthodes de classification (non supervisées) semblent plus utilisées que l'analyse discriminante. Cependant, l'une des applications rapides des puces à ADN pourrait être liée au diagnostic médical qui équivaut à prédire l'appartenance d'un individu à une ou plusieurs classes connues a priori. L'approche décrite par Brown *et al.* [19] montrent l'intérêt de l'analyse discriminante pour aborder la classification fonctionnelle. Bien que peu développée en raison de l'influence des travaux de Eisen *et al.* [48] sur l'étude des profils d'expression, l'analyse discriminante présente des perspectives intéressantes pour l'étude fonctionnelle des gènes.

4.2 La classification

De même qu'en analyse discriminante, l'unité statistique peut tout aussi bien être un tissu ou un gène dans les problèmes de classification. Dans la plupart des applications, c'est cependant le gène qui est considéré comme l'unité à classer. Les variables descriptives définissent le niveau d'expression des gènes pour différents temps (cinétique), dans différentes conditions, pour différents traitements ou pour différentes concentrations d'une drogue. Les études qui ont été le plus développées sont sans doute celles basées sur une cinétique d'expression.

L'application des méthodes de classification a pour objectif d'identifier des gènes coexprimés, qui pourraient alors être corégulés. Cette démarche vise ainsi à regrouper des gènes partageant des caractéristiques fonctionnelles communes. Ainsi, Tavazoie *et al.* [118] identifient des classes dont les gènes présentent des similarités de fonction importantes. Chu *et al.* [29] obtiennent des résultats similaires à partir de profils d'expression obtenus durant la sporulation de la levure.

Un second domaine d'application de la classification concerne la recherche de motifs dans les régions promotrices des gènes classés identiquement. Par exemple, Tavazoie *et al.* [118] identifient 18 motifs significatifs dans les régions promotrices de gènes de même classe. Ces résultats semblent indiquer qu'il est possible d'inférer des mécanismes de corégulation à partir des gènes co-exprimés (la classification est également largement développée pour l'analyse de tissus tumoraux [57, 6]).

Lors d'une étude sur des profils d'expression, chaque gène est caractérisé par un vecteur $x_i \in \mathbb{R}^p$, donnant les valeurs pour les p variables descriptives. Ce vecteur x_i est appelé le profil d'expression du gène. De nombreuses méthodes de classification ont été utilisées pour étudier les données de puces à ADN. Celles-ci se répartissent en deux groupes : les méthodes de partitionnement et les méthodes hiérarchiques. Quackenbush [96] donnent une très bonne synthèse de ces méthodes dans le cadre des données de puces à ADN. Nous nous intéressons aux approches les plus courantes : la classification hiérarchique, les Kmeans et les cartes de Kohonen ou *self-organizing map* (SOM). Une dernière approche, plus robuste et plus intéressante, correspond aux modèles de mélange.

4.2.1 La classification hiérarchique

Suite aux travaux de Eisen *et al.* [48], la classification hiérarchique s'est largement développée en raison de sa simplicité. On distingue deux types de classification hiérarchique.

- La classification ascendante : on part de la partition la plus fine où toutes les classes se composent d'un seul élément, pour arriver par agrégations successives des classes à la

partition la plus grossière comportant une seule classe où sont réunies toutes les unités statistiques.
- La classification descendante ou segmentation : on part de la partition triviale pour arriver par divisions successives à la partition grossière.

La classification ascendante (CAH) est la plus utilisée car plus robuste que la classification descendante. Dans ce cas, on définit des indices δ d'agglomération (appelés improprement distances) entre deux classes A et B. Le poids des parties A et B, somme des poids de leurs éléments, sont notés p_A et p_B. Leurs barycentres sont notés g_A et g_B. La distance, le plus souvent euclidienne, entre deux individus est notée d. Les différents indices sont les suivants.

- $\delta(A, B) = \inf(d_{ij}, i \in A, j \in B)$ (saut minimum ou simple lien ou *single-linkage*)
- $\delta(A, B) = \max(d_{ij}, i \in A, j \in B)$ (saut maximum ou *complete-linkage*)
- $\delta(A, B) = moy(d_{ij}, i \in A, j \in B)$ (saut moyen ou UPGMA ou *average-linkage*)
- $\delta(A, B) = d(g_A, g_B)$ (distance des barycentres)
- $\delta(A, B) = \frac{p_A p_B}{p_A + p_B} d^2(g_A, g_B)$ (critère de Ward [123])

L'algorithme CAH est le suivant. D'abord, on calcule la matrice des indices d'agrégation entre les groupes. Puis, on recherche les deux classes les plus similaires au sens de l'indice d'agrégation. Ces deux classes sont alors groupées pour former une nouvelle classe. Enfin, les distances entre cette nouvelle classe et toutes les autres sont calculées. Il n'est pas nécessaire de recalculer toutes les distances car seulement celles relatives à la nouvelle classe changent. Ces étapes sont répétées jusqu'à ce que l'on obtienne une seule classe, la classe contenant l'ensemble tout entier. Le résultat de algorithme se représente par un arbre hiérarchique (ou dendrogramme) qui donne les valeurs des indices d'agrégation à chaque agglomération de deux classes.

L'un des problèmes de la CAH est le choix de l'indice d'agrégation et des différences importantes peuvent apparaître suivant le critère choisi. Goldstein *et al.* [56] décrivent les difficultés liées à l'application de ces méthodes pour la classification de différents tissus en raison du choix de ce critère. Ils montrent ainsi que les résultats sont très dépendants de ce choix pour la classification. De plus, il n'existe pas de critère pour le choix du nombre classe et il est difficile de comparer différents résultats d'une classification hiérarchique. Les problèmes de variabilité décrits dans les précédents chapitres ne peuvent pas être pris en compte par ces méthodes. Pour toutes ces raisons, ces méthodes ne sont pas adaptées à la classification de profils d'expression et sont trop souvent utilisées à tort.

4.2.2 La méthode des Kmeans

Les méthodes de partitionnement, dont les plus connues sont certainement celle des Kmeans et des centres mobiles, sont une bonne alternative aux méthodes hiérarchiques. Pour k fixé, on recherche à partitionner notre ensemble d'unités statistiques en k groupes.

La méthode des centres mobiles se déroule en 4 étapes. (i) On commence par définir au hasard, ou selon des considérations extérieurs à la méthode, k points dans l'espace des individus, appelés noyaux. Ces noyaux sont généralement k individus. (ii) À chaque noyau, on associe l'ensemble des points les plus proches. On obtient ainsi une partition en k classes. (iii) On remplace alors les k noyaux précédents par les centres de gravité des classes. (iv) On itère le processus jusqu'à ce que l'algorithme converge vers un optimum local. Cet algorithme revient à minimiser la somme des distances au carré des unités x_i , $i \in \{1, \dots, N\}$ au centre de gravité $g_{c(i)}$ le plus proche. En choisissant pour $d(.,.)$ la distance euclidienne, ce critère s'écrit

$$E_K = \sum_{i=1}^{N} d^2(x_i - g_{c(i)}) \tag{4.1}$$

Dans cet algorithme, on attend que tous les individus aient été affectés à une classe pour recalculer les noyaux. L'algorithme des Kmeans est légèrement différent de celui des centres mobiles puisque les centres sont recalculés après l'affectation de chaque individu (cet algorithme a été utilisé par Tavazoie *et al.* [118]).

Pour ces deux algorithmes, il s'agit d'optimiser un critère de type inertie. Pour un ensemble de points donnés, on note I_{totale} l'inertie totale du nuage de points dans l'espace $\mathbb{R}^p$ et $I_1, \dots, I_K$ les inerties associées à chacune des classes. L'inertie est la moyenne des carrés des distances au centre de gravité. On a la relation suivante

$$I_{\text{totale}} = I_{\text{intra}} + I_{\text{inter}}$$

où $I_{\text{intra}} = \sum_{k=1}^{K} I_k$ est l'inertie intraclasse et $I_{\text{inter}} = \sum_{k=1}^{K} p_k d^2(g, g_k)$ est l'inertie interclasse avec p_k le poids de la classe k (le poids d'une classe est égale à la somme des poids des éléments qui la constituent).

Les deux algorithmes reviennent à minimiser l'inertie intraclasse, ce qui équivaut donc à maximiser l'inertie interclasse. D'un point de vue géométrique, il s'agit d'obtenir des classes compactes et homogènes (minimiser l'inertie intraclasse) tout en étant les plus distinctes possibles (maximiser l'inertie interclasse). Un des inconvénients est lié au fait que l'on peut obtenir des résultats différents suivant l'initialisation choisie. Un défaut majeur de cet algorithme est qu'il ne peut identifier que des structures simples entre les classes. Une surdispersion ou une sous-dispersion des groupes peuvent par exemple conduire à de mauvais résultats de cette méthode.

4.2.3 Algorithme de réseaux de neurones

La quantification vectorielle consiste à représenter une distribution (finie ou non) dans un espace $\mathbb{R}^p$ de dimension p à l'aide d'un sous-ensemble fini de K éléments (les *prototypes* ou *vecteurs-codes*) dans cet espace. L'algorithme des Kmeans est connu comme l'algorithme de Lloyd [80] dans ce domaine. Le processus de classification est appelé codage et les représentants des classes (les centres de gravité pour les Kmeans) sont les prototypes.

Les *self-organizing maps* ou cartes toplogiques de Kohonen [71] ont été appliquées par Tamayo *et al.* [117]. L'algorithme des SOM peut être vu comme une version contrainte des Kmeans, où les prototypes sont contraints à appartenir à un espace de dimension réduite (généralement un plan). Le codage consiste à remplacer chaque point x_i de la distribution par le numéro d'un vecteur code.

Par exemple, on quadrille un rectangle dans cet espace de faible dimension. On associe à K prototypes $g_k \in I\!R^p$ un couple d'entiers l_k pour chacun des K noeuds de la grille, $l_k \in \mathcal{Q}_1 \times \mathcal{Q}_2$, avec $\mathcal{Q}_1 = \{1, \ldots, q_1\}$, $\mathcal{Q}_2 = \{1, \ldots, q_2\}$ et $K = q_1 q_2$. La contrainte consiste à chercher des positions des prototypes dans l'espace des observations telles que les prototypes voisins dans l'espace des observations soient aussi voisins dans l'espace de faible dimension. Cette contrainte s'exprime dans la règle de mise à jour des paramètres de position des prototypes g_k. Les données x_i sont prises en compte séquentiellement. On identifie le prototype g_j le plus proche de x_i dans l'espace $I\!R^p$ des observations, et tous les voisins g_k de g_j sont mis à jour. Cette mise à jour s'effectue par exemple suivant la relation

$$g_k \leftarrow g_k - \eta\, G(l_k, l_j)(g_k - x_i)$$

où η correspond au pas de l'algorithme et $G(l_k, l_j)$ est une fonction de voisinage des prototypes k et j dans l'espace de faible dimension. La fonction G donne un poids plus important aux prototypes g_k avec des indices l_k proches de ceux de l_j.

Cet algorithme nécessite cependant de fixer certains paramètres : la forme de la grille dans l'espace de faible dimension, le nombre de classes K, la fonction G et le paramètre α. Cette approche peut être vue comme une projection non linéaire de l'espace original vers l'espace de faible dimension. Les cartes de Kohonen sont pour cette raison plus des méthodes de représentation (telle que l'analyse en composantes principales ou ACP [105]) que des méthodes de classification. Nous lui préférons l'algorithme des Kmeans plus simple et plus approprié aux problèmes de classification.

4.2.4 Méthodes combinées

De nombreuses combinaisons de méthodes sont possibles. Une combinaison bien connue consiste à réaliser dans un premier temps une classification hiérarchique en utilisant le critère de Ward [123]. La classification obtenue est alors améliorée dans un deuxième temps en utilisant l'algorithme des K-means. La combinaison de ces deux méthodes est très logique pour la raison suivante.

La classification hiérarchique selon le critère de Ward est certainement une méthode de référence [105]. Le critère de Ward est défini comme la perte d'inertie interclasse résultant du regroupement de deux classes. En d'autres termes, il permet d'associer les classes qui donnent la plus grande inertie interclasse après regroupement (ou qui assure la plus faible inertie intraclasse). Ce critère donne donc, après chaque regroupement, des classes les plus compactes et les plus distinctes possibles. Le résultat de cette classification est donc de bon aloi pour initialiser l'algorithme des K-means (c'est une solution raisonnable pour espérer atteindre le minimum global de l'inertie intraclasse).

Dans certains cas, il est intéressant de combiner cette approche à d'autres méthodes d'analyse. Une classification sur les coordonnées de l'ACP permet par exemple de donner une représentation très informative des classes sur les différents plans de l'ACP. Lorsque la dimension est trop importante, l'ACP permet également de réduire la dimension de l'espace des observations et rend la classification plus stable. Curieusement, ces combinaisons n'ont jamais été proposées. Dans bien des domaines, elles s'avèrent pourtant meilleures que la plupart des approches, souvent trop directes, qui ont été proposées [77].

4.2.5 Les modèles de mélange

Le dernier type d'approche s'appuie sur une hypothèse de mélange où chaque classe $\mathcal{C}_k$ est représentée par une loi normale multivariée. Cette distribution est caractérisée par son espérance (μ_k) et sa variance (la matrice de variance Σ_k). La densité ϕ_k du composant $\mathcal{C}_k$ est une loi normale dans un espace euclidien $I\!R^p$ et s'écrit

$$\phi_k(x;\theta_k) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}}\exp(-\frac{1}{2}(x-\mu_k)'\Sigma_k^{-1}(x-\mu_k)) \tag{4.2}$$

avec $\theta_k = (\mu_k, \Sigma_k)$ les paramètres du composant $\mathcal{C}_k$. La densité f d'un mélange de lois normales est donnée par

$$f(x;\theta) = \sum_{k=1}^{K} p_k\phi_k(x;\theta_k). \tag{4.3}$$

où $p_k > 0$, $k = 1,\ldots,K$, sont les proportions du mélange et $\sum_{k=1}^{K} p_k = 1$. Cette densité est paramétrée par

$$\theta = ((p_k)_{k=1,\ldots,K-1}, (\theta_k)_{k=1,\ldots,K}). \tag{4.4}$$

Les proportions p_k des différents composants s'interprètent comme les probabilités des composants $\mathcal{C}_k$, $p_k = P(Z = k)$, la variable aléatoire cachée Z identifiant le composant associé à chaque unité statistique. L'estimation du paramètre θ du mélange consiste à inférer une valeur de θ à partir des observations $x_i \in I\!R^p$, qui sont des réalisations indépendantes de la loi de mélange $f(x;\theta)$. La classification est réalisée suivant le principe du maximum a posteriori. Un individu x_i est affecté au composant $k_0^{(i)}$ le plus probable a posteriori, ce qui s'écrit

$$k_0^{(i)} = \arg\max_{k=1,\ldots,K} P(Z = k|x_i). \tag{4.5}$$

L'expression (4.5) faisant intervenir la valeur θ, inconnue, du paramètre du mélange, on la remplace par l'estimation $\hat{\theta}$ obtenue (en général) par la méthode du maximum de vraisemblance.

Les intérêts d'une approche mélange sont nombreux. Tout d'abord, l'approche probabiliste fournit un cadre formel cohérent pour les problèmes de classification. Par exemple, les problèmes de choix du nombre de classes deviennent des problèmes de choix de modèles

[13, 50]. Ceci présente un avantage important sur les précédentes approches de classification pour lesquelles il n'existe pas de méthode rigoureusement établie pour choisir le nombre de classes. De plus, suivant les contraintes sur les valeurs des proportions, sur les matrices Σ_k, on dispose de nombreux modèles de mélange pour la classification. La classification des données d'expression sous une approche mélange a été proposée à plusieurs reprises [130, 55, 88]. Dans le prochain chapitre, nous utilisons un modèle de mélange pour la classification de données répétées.

4.2.6 Autres approches

D'autres approches, plus spécifiques que les précédentes ont été proposées. Hastie *et al.* [61] s'intéressent à constituer des classes pour lesquelles les unités statistiques ont une grande variance suivant les variables descriptives, ce qui traduit des changements importants entre les différentes conditions. Pour cela, les classes sont formées à partir des premières composantes de l'ACP. Lazzeroni *et al.* [76] décrivent un algorithme dans lequel un gène peut appartenir à plusieurs classes ou n'être affecté à aucune d'entre elles.

Dans certains cas, il peut être intéressant de mener une classification simultanée sur les lignes et les colonnes. L'approche la plus utilisée consiste dans ce cas à réaliser deux classifications hiérarchiques : la première sur les lignes puis la seconde sur les colonnes. Pour ce problème, Cheng *et al.* [27] introduisent le concept de *bicluster*.

4.2.7 Conclusion

Toutes ces approches présentent le grand défaut de ne pas prendre en compte les problèmes de variabilité des mesures que nous avons signalés lors de la recherche de gènes différentiellement exprimés. Ainsi, elle n'utilisent pas l'information des mesures répétées. Dans le chapitre suivant, nous présentons une nouvelle approche pour l'étude des profils d'expression qui propose une solution à ces problèmes.

Chapitre 5

Classification de données répétées

En raison des problèmes de variabilité, de plus en plus de chercheurs se rendent compte qu'il est indispensable de réaliser différentes mesures pour un même gène pour l'identification de gènes différentiellement exprimés [78]. Nous avons vu lors du troisième chapitre le risque lié à une analyse sans répétition. De manière évidente, ces problèmes concernent également l'étude des profils d'expression. Cependant, les méthodes de classification actuelles ne tiennent pas compte de ces difficultés. Pour ces raisons, leur interprétation est hasardeuse. Il est donc nécessaire de proposer des méthodes de classification qui tirent parti des répétitions réalisées.

5.1 Introduction

Les modèles de mélange gaussien donnent des solutions bien adaptées pour le problème de classification. Cependant, ces techniques ne permettent pas de prendre en compte la variabilité issue de variables aléatoires. Deux attitudes sont généralement adoptées devant un tel problème.

- Négliger le problème en utilisant une seule mesure pour chaque unité statistique.
- Restreindre le problème de variabilité à un effet moyen obtenu en moyennant R mesures répétées et indépendantes pour chaque unité statistique.

Ces deux attitudes ne sont pas satisfaisantes. La première peut conduire à des modèles donnant une très mauvaise représentation des observations. La seconde est meilleure mais suppose que la variabilité ne dépend pas de covariable ou des unités statistiques, ce qui peut s'avérer irréaliste.

Nous voulons nous attaquer au problème crucial de la classification sur données répétées (ou répliquées) en tirant parti du cadre des mélanges de distributions qui nous ouvre de nombreuses solutions de modélisation et fournit des outils solides pour évaluer les performances des classifieurs qu'ils induisent. Aussi, nous nous attachons à proposer différents modèles de mélange permettant de prendre en compte la variabilité issue d'effets aléatoires.

Ces modèles ont pour objectif de bien cerner les sources de variabilité en analysant l'information des mesures répétées par des modèles linéaires mixtes construits conditionnellement aux composants du mélange.

Nous supposons que pour I unités statistiques indépendantes (gènes ou tissus), on dispose de R mesures répétées à T différents temps ou pour T conditions différentes. On note alors y_{itr} la $r^{\text{ième}}$ mesure au temps t pour l'unité statistique i (les valeurs y_{itr} correspondent généralement aux log-ratios des niveaux d'expression $\log(R/G)$).

5.2 Modèle de mélange pour la classification de données répétées

Nous proposons d'aborder la classification de données d'expression dans un cadre probabiliste. Nous cherchons dans un premier temps à modéliser la mesure y_{itr} afin de prendre en compte les répétitions. Nous abordons ensuite le problème de la classification de données répétées sous une hypothèse de mélange gaussien.

5.2.1 Modélisation de l'expression différentielle

En considérant le cas le plus simple, c'est-à-dire celui où nous ne disposons pas de répétition ($R = 1$), une approche pour modéliser la mesure y_{it} de l'unité statistique i au temps t peut s'écrire :

$$(\mathrm{E}_0) : y_{it} = \beta_t + \epsilon_{it} \tag{5.1}$$

où

- β_t est la "vraie" valeur de la mesure au temps t
- $\epsilon_{it} \sim \mathcal{N}(0, \sigma^2)$ correspond à une erreur de mesure pour le gène i au temps t.

Une extension simple du modèle (5.1) permettant de prendre en compte des répétitions consiste à représenter la mesure y_{itr} par

$$(\mathrm{E}_0) : y_{itr} = \beta_t + \epsilon_{itr} \tag{5.2}$$

où

- β_t est la "vraie" valeur de la mesure au temps t
- $\epsilon_{itr} \sim \mathcal{N}(0, \sigma^2)$ correspond à une erreur de mesure pour la r^e répétition du gène i au temps t.

Pour les deux modèles (5.1) et (5.2), nous pouvons adopter l'écriture canonique d'un modèle linéaire :

$$Y = X\beta + \epsilon \tag{5.3}$$

où

- Y : vecteur aléatoire à expliquer, de taille $N = ITR$, et dont $\mathbf{y} = (\mathbf{y}_1^{'}, \ldots, \mathbf{y}_I^{'})^{'}$ est une réalisation avec $\mathbf{y}_i = (y_{i11}, \ldots, y_{i1R}, \ldots, y_{iT1}, \ldots, y_{iTR})^{'}$ l'ensemble des observations pour l'unité i.
- ϵ : vecteur aléatoire des erreurs de taille N et $\epsilon \sim \mathcal{N}(0, \sigma^2 \mathrm{Id}_N)$.
- β : vecteur des paramètres inconnus, de taille T, et X, de dimension (N, T). En notant 1_R et 0_R les vecteurs $(1, \ldots, 1)^{'}$ et $(0, \ldots, 0)'$ de taille R, la matrice d'incidence X s'écrit :

$$X_{(N,T)} = \begin{bmatrix} \dot{X} \\ \vdots \\ \dot{X} \end{bmatrix} \quad \text{en notant} \quad \dot{X}_{(RT,T)} = \begin{bmatrix} 1_R & 0_R & \cdots & 0_R \\ 0_R & \ddots & \cdots & \vdots \\ \vdots & \cdots & \ddots & 0_R \\ 0_R & \cdots & 0_R & 1_R \end{bmatrix}.$$

Cette écriture traduit le fait que $Y \sim \mathcal{N}(X\beta, \sigma^2 Id_N)$ et suppose donc que toutes les unités statistiques ont la même mesure théorique β. Grâce au modèle de mélange, nous pouvons différencier les unités statistiques suivant leurs mesures.

5.2.2 Modèle de mélange

Sous cette représentation, nous abordons le problème de la classification sous une hypothèse de mélange gaussien. Chacun des composants $\mathcal{C}_k$ du mélange permet de décrire chacune des K classes à identifier. On associe à chaque composant une variable aléatoire $Y^k \sim \mathcal{N}_{N_k}(X^k\beta_k, \sigma_k^2 Id_{N_k})$ de dimension $N_k = I_k TR$ où I_k est le nombre d'individus (gènes ou tissus) dans la classe k et $\beta_k \in I\!R^T$ caractérise le profil moyen de la classe k ($\beta_k = (\beta_{kt}, t = 1, \ldots, T)$). La matrice X^k est obtenue en superposant I_k matrices $\dot{X}$. On note $\mathbf{y}^k = (\mathbf{y}_1^{'}, \ldots, \mathbf{y}_{N_k}^{'})^{'}$ une réalisation de Y^k.
Conditionnellement au composant $\mathcal{C}_k$ ($1 \leq k \leq K$), la densité pour un individu i s'écrit

$$g_k(\mathbf{y}_i|\theta_k) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y}_i - \dot{X}\beta_k)^{'}\Sigma_k^{-1}(\mathbf{y}_i - \dot{X}\beta_k)\right) \tag{5.4}$$

avec $d = RT$, $\theta_k = (\beta_k, \Sigma_k)$, et Σ_k la matrice de variance-covariance (matrice symétrique définie positive de taille (d, d)). Sous l'hypothèse $\Sigma_k = \sigma_k^2 Id_{TR}$, on a

$$g_k(\mathbf{y}_i|\theta_k) = \frac{1}{(2\pi)^{d/2}(\sigma_k^2)^{d/2}} \exp\left(-\frac{1}{2\sigma_k^2}(\mathbf{y}_i - \dot{X}\beta_k)^{'}(\mathbf{y}_i - \dot{X}\beta_k)\right). \tag{5.5}$$

La densité pour un individu i du mélange s'exprime par

$$g(\mathbf{y}_i|\theta, \mathbf{p}) = \sum_{k=1}^{K} p_k g_k(\mathbf{y}_i|\theta_k) \tag{5.6}$$

où $\theta = (\theta_1, \ldots, \theta_K)$ est le vecteur des paramètres des composants et $\mathbf{p} = (p_1, \ldots, p_K)$ est le vecteur des proportions du mélange. Supposant que les observations sont des réalisations indépendantes, on obtient la vraisemblance du paramètre $(\theta, \mathbf{p})$

$$g(\mathbf{y}|\theta, \mathbf{p}) = \prod_{i=1}^{I} \sum_{k=1}^{K} p_k g_k(\mathbf{y}_i|\theta_k). \tag{5.7}$$

Afin de réaliser une classification des données, il s'agit alors d'estimer le paramètre θ et le vecteur $\mathbf{p}$ des proportions du mélange. Cette estimation peut être réalisée par maximisation de la log-vraisemblance en utilisant l'algorithme EM (Espérance-Maximisation) [89] qui tire parti de la structure incomplète des données. Les données manquantes sont ici les vecteurs $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})$ indicateurs d'appartenance à l'un des composants, $z_{ik} = 1$ si $i \in \mathcal{C}_k$ et 0 sinon. Dans la section suivante, nous détaillons l'algorithme EM pour le modèle précédent.

5.2.3 Algorithme EM pour un mélange de modèles linéaires E_0

Ayant observé les données $\mathbf{y}$ issues de la loi de mélange, nous souhaitons estimer l'ensemble des paramètres $\mathbf{p} = (p_k, k = 1, \ldots, K)$ et $\theta = (\theta_1, \ldots, \theta_K)$. Le vecteur $\mathbf{x} = (\mathbf{y}, \mathbf{z})$ est le vecteur des données complètes avec $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_I)$. La densité des données complètes s'écrit

$$f(\mathbf{x}|\theta, \mathbf{p}) \;=\; \prod_{i=1}^{I} f(\mathbf{y}_i, \mathbf{z}_i|\mathbf{p}, \theta) = \prod_{i=1}^{I}\prod_{k=1}^{K} \left(p_k g_k(\mathbf{y}_i|\theta_k)\right)^{z_{ik}}. \tag{5.8}$$

La fonction de log-vraisemblance associée aux données complètes $\mathbf{x} = (\mathbf{y}, \mathbf{z})$ pour le vecteur de paramètres $(\mathbf{p}, \theta)$ est

$$\begin{aligned} l(\mathbf{p}, \theta|\mathbf{x}) &= \sum_{i=1}^{I}\sum_{k=1}^{K} z_{ik} \ln(p_k) + \sum_{i=1}^{I}\sum_{k=1}^{K} z_{ik} \ln(g_k(\mathbf{y_i}|\theta_k)) \\ &= \sum_{i=1}^{I}\sum_{k=1}^{K} z_{ik} \ln(p_k) + \sum_{i=1}^{I}\sum_{k=1}^{K} z_{ik} h(\theta_k|\mathbf{y_i}) \end{aligned} \tag{5.9}$$

où

$$h(\theta_k|\mathbf{y_i}) = -\frac{1}{2}\ln(2\pi) - \frac{TR}{2}\ln(\sigma_k^2) - \frac{1}{2}\frac{(\mathbf{y}_i - \dot{X}\beta_k)'(\mathbf{y}_i - \dot{X}\beta_k)}{\sigma_k^2}. \tag{5.10}$$

Chaque itération de l'algorithme EM est constituée de deux étapes. À l'itération $[q]$, partant de la valeur $\theta^{[q]}$ et $\mathbf{p}^{[q]}$ des paramètres, elle se décompose en :

- *étape E* : calcul de l'espérance conditionnelle $\mathbb{E}(l(\mathbf{p}, \theta|\mathbf{x})|\mathbf{y}, \theta^{[q]}, \mathbf{p}^{[q]})$.
- *étape M* : calcul des nouvelles valeurs $\theta^{[q+1]}$ et $\mathbf{p}^{[q+1]}$ des paramètres qui maximisent l'espérance conditionnelle obtenue à l'étape précédente.

Les équations de l'algorithme EM pour notre modèle E_0 donnent les relations suivantes.

5.2.3.1 Étape E

Cette étape consiste donc à calculer une espérance conditionnelle.

$$\begin{aligned}\mathcal{Q}(\theta, p|\theta^{[q]}, \mathbf{p}^{[q]}) &= \mathbb{E}\left(l(\mathbf{p}, \theta|\mathbf{y}, z)|\mathbf{y}, \theta^{[q]}, \mathbf{p}^{[q]}\right) \\ &= \sum_{i=1}^{I}\sum_{k=1}^{K} t_i^{[q]}(k)\ln(p_k) + \sum_{i=1}^{I}\sum_{k=1}^{K} t_i^{[q]}(k)\mathbb{E}[h(\theta_k|\mathbf{y}_i)|\mathbf{y}, \theta^{[q]}] \end{aligned} \quad (5.11)$$

avec

$$t_i^{[q]}(k) = P(i \in \mathcal{C}_k|\mathbf{y}_i, \theta^{[q]}, \mathbf{p}^{[q]}) = \frac{p_k^{[q]} g_k(\mathbf{y}_i|\theta_k^{[q]})}{\sum_{l=1}^{K} p_l^{[q]} g_l(\mathbf{y}_i|\theta_l^{[q]})}. \quad (5.12)$$

Il faut remarquer qu'en pratique le calcul effectif de $\mathcal{Q}$ n'est pas nécessaire. À cette étape, il est juste nécessaire d'actualiser les probabilités $t_i(k)$.

5.2.3.2 Étape M

Les valeurs qui maximisent l'expression (5.11) sont données de façon explicite pour chacun des $k = 1, \ldots, K$ composants par les formules suivantes :

$$p_k^{[q+1]} = \sum_{i=1}^{I} \frac{t_i^{[q]}(k)}{I} \quad (5.13)$$

$$\sigma_k^{2[q+1]} = \frac{1}{TR\sum_{i=1}^{I} t_i^{[q]}(k)} \sum_{i=1}^{I} t_i^{[q]}(k) \left(\mathbf{y}_i - \dot{X}\beta_k^{[q]}\right)' \left(\mathbf{y}_i - \dot{X}\beta_k^{[q]}\right) \quad (5.14)$$

$$\beta_k^{[q+1]} = \frac{1}{\sum_{i=1}^{I} t_i^{[q]}(k)} \sum_{i=1}^{I} t_i^{[q]}(k)(\dot{X}'\dot{X})^{-1}\dot{X}'\mathbf{y}_i. \quad (5.15)$$

Cette approche permet de prendre en compte les répétitions des unités statistiques, chaque gène étant décrit par le vecteur de ses répétitions. Cependant, elle ne nous renseigne pas sur les relations entre les répétitions d'un gène. Nous proposons donc d'aborder ce problème en introduisant des effets aléatoires dans la modélisation du niveau d'expression.

Avant cela, notons que l'absence de répétition ($R = 1$) implique $\dot{X} = Id_T$. Les équations sont alors celles obtenues pour un mélange gaussien dont les matrices de variance sont proportionnelles à l'identité avec des déterminants différents [90]. Par ailleurs, si on suppose $\sigma_k^2 = \sigma^2$, $\forall k = 1, \ldots, K$, il s'agit alors du modèle sphérique où les matrices de variance sont égales et proportionnelles à l'identité. Dans le cas $R = 1$, de nombreux modèles peuvent être choisis pour la classification suivant les hypothèses posées sur la matrice de variance-covariance Σ_k [9]. Les travaux de Yeund *et al.* [130] et Ghosh *et al.* [55] s'appuient sur ces différents modèles de classification.

5.2.4 Lien avec les Kmeans

Bien que compliqué en apparence, le modèle défini par (5.5) et (5.6) consiste à réaliser une moyenne sur les répétitions. Pour le voir, notons $\tilde{\mathbf{y}}_i = (\bar{y}_{i1.}, \bar{y}_{i2.}, \dots, \bar{y}_{iT.})'$ avec $\bar{y}_{ij.} = \frac{\sum_{r=1}^{R} y_{ijr}}{R}$. Suivant ces notations, $\tilde{\mathbf{y}}_i$ correspond aux données de l'unité statistique i pour laquelle on réalise la moyenne sur les répétitions aux différents temps. On obtient alors $\tilde{\mathbf{y}}_i = \frac{\dot{X}'\mathbf{y}_i}{R}$. On définit la densité $\tilde{g}_k$ par

$$\tilde{g}_k(\tilde{\mathbf{y}}_i|\tilde{\theta}_k) = \frac{1}{(2\pi)^{\tilde{d}/2}|\tilde{\Sigma}_k|^{1/2}} \exp\left(-\frac{1}{2}(\tilde{\mathbf{y}}_i - \dot{X}\beta_k)'\tilde{\Sigma}_k^{-1}(\tilde{\mathbf{y}}_i - \dot{X}\beta_k)\right) \tag{5.16}$$

avec $\tilde{d} = T$, $\tilde{\theta}_k = (\beta_k, \tilde{\Sigma}_k)$, β_k caractérisant le profil moyen de la classe k ($\beta_k \in I\!R^T$) et $\tilde{\Sigma}_k$ la matrice de variance-covariance (matrice symétrique définie positive de taille $(\tilde{d}, \tilde{d})$). Sous l'hypothèse $\tilde{\Sigma}_k = \sigma_k^2 Id_T$, on a

$$\tilde{g}_k(\tilde{\mathbf{y}}_i|\tilde{\theta}_k) = \frac{1}{(2\pi)^{\tilde{d}/2}(\sigma_k^2)^{\tilde{d}/2}} \exp\left(-\frac{1}{2\sigma_k^2}(\tilde{\mathbf{y}}_i - \dot{X}\beta_k)'(\tilde{\mathbf{y}}_i - \dot{X}\beta_k)\right). \tag{5.17}$$

En conséquence, on montre facilement que l'étape E décrite précédemment pour le calcul des probabilités a posteriori revient à remplacer la densité g et le vecteur $\mathbf{y}_i$ par la densité $\tilde{g}$ et le vecteur $\tilde{y}$. De même, nous pouvons écrire pour chacun des $k = 1, \dots, K$ composants l'étape M décrite précédemment de la manière suivante

$$p_k^{[q+1]} = \sum_{i=1}^{I} \frac{t_i^{[q]}(k)}{I} \tag{5.18}$$

$$\sigma_k^{2[q+1]} = \frac{1}{T\sum_{i=1}^{I} t_i^{[q]}(k)} \sum_{i=1}^{I} t_i^{[q]}(k) \left(\tilde{\mathbf{y}}_i - \beta_k^{[q]}\right)' \left(\tilde{\mathbf{y}}_i - \beta_k^{[q]}\right) \tag{5.19}$$

$$\beta_k^{[q+1]} = \frac{1}{\sum_{i=1}^{I} t_i^{[q]}(k)} \sum_{i=1}^{I} t_i^{[q]}(k)\tilde{\mathbf{y}}_i. \tag{5.20}$$

Ainsi, le modèle que nous avons présenté est équivalent à un modèle de mélange gaussien dont les matrices de variance sont proportionnelles à l'identité. Les unités statistiques sont les individus pour lesquels on réalise la moyenne sur les répétitions à chaque temps.

Nous avons supposé que les paramètres de variance étaient différents suivant les composants du mélange. Nous obtenons de manière évidente le même résultat si ces paramètres sont supposés égaux. Il faut également noter que la procédure des Kmeans est très proche d'un mélange gaussien dans lequel les composants ont tous la même structure de variance $\sigma^2 Id$. En effet, il s'agit dans les deux cas de minimiser le critère d'inertie intra-classe.

Ainsi, l'approche précédente reste équivalente à réaliser une moyenne sur les répétitions. Nous allons voir dans la suite comment il est possible d'améliorer ces modèles afin de mieux caractériser les problèmes de variabilité pour la classification de données répétées.

5.3 Les modèles linéaires mixtes

Les modèles linéaires mixtes, notés L2M, permettent d'enrichir les modèles linéaires classiques ne contenant que des effets fixes. Nous avons déjà montré dans le chapitre sur la recherche de gènes différentiellement exprimés l'intérêt de ces modèles pour la modélisation des niveaux d'expression. Pour le problème de classification, nous proposons les modèles linéaires mixtes suivants afin de prendre en compte un effet aléatoire "individu" sur la mesure y_{itr}

$$\begin{aligned}
(\mathrm{E}_1) : y_{itr} &= \beta_t + \xi_i + \epsilon_{itr} && (5.21)\\
(\mathrm{E}_2) : y_{itr} &= \beta_t + \xi_{it} + \epsilon_{itr} && (5.22)\\
(\mathrm{E}_3) : y_{itr} &= \beta_t + (\xi_i + \xi_{it}) + \epsilon_{itr} && (5.23)
\end{aligned}$$

où
- β_t représente l'effet fixe du temps
- $\xi_i \sim \mathcal{N}(0, \omega^2)$ est l'effet aléatoire du gène i sur la mesure
- $\xi_{it} \sim \mathcal{N}(0, \tau^2)$ est l'effet aléatoire du gène i au temps t
- $\epsilon_{itr} \sim \mathcal{N}(0, \sigma^2)$ est l'erreur de mesure
- les différents effets aléatoires et erreurs de mesure sont mutuellement indépendants.

Dans un souci de clarté, nous développons dans un premier temps notre approche en nous basant sur le modèle E_2 qui nous semble le plus à même de prendre en compte les répétitions dans de nombreuses situations. L'écriture canonique associée que nous utiliserons par la suite est :

$$Y = \underbrace{X\beta}_{\text{partie effets fixes}} + \underbrace{U\xi}_{\text{partie effets aléatoires}} + \epsilon \qquad (5.24)$$

où
- Y : vecteur aléatoire à expliquer, de taille $N = ITR$, et dont $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_I)$ est une réalisation.
- ξ : vecteur des effets aléatoires de taille IT et $\xi \sim \mathcal{N}_{IT}(0, \tau^2 \mathrm{Id}_{IT})$. La matrice d'incidence $U_{(N,IT)}$ s'écrit : $U_{(N,IT)} = \begin{bmatrix} 1_R & 0_R & \cdots & 0_R \\ 0_R & \ddots & \cdots & \vdots \\ \vdots & \cdots & \ddots & 0_R \\ 0_R & \cdots & 0_R & 1_R \end{bmatrix}$ avec 1_R et 0_R les vecteurs $(1, \ldots, 1)$ et $(0, \ldots, 0)$ de taille R.
- ϵ : vecteur aléatoire des erreurs de taille N indépendant de ξ et $\epsilon \sim \mathcal{N}_N(0, \sigma^2 \mathrm{Id}_N)$.
- β : vecteur des paramètres inconnus des effets fixes, de taille T, et X, de dimension (N, T), sa matrice d'incidence définie précédemment en (5.3).

Les trois modélisations envisagées vont permettre d'introduire des structures de covariance particulières entre les observations. Avant de choisir un modèle, il est important de voir quelles sont les dépendances impliquées.

5.3.1 Les structures de covariance

En notant $\delta_k^{k'} = 1$ si $k = k'$ et 0 sinon, nous explicitons les covariances liées à chacun des trois modèles.

5.3.1.1 Modèle $\mathbf{E}_1$

Une donnée y_{itr} est représentée par le modèle linéaire $y_{itr} = \beta_t + \xi_i + \epsilon_{itr}$. La covariance entre deux observations y_{itr} et $y_{i'\,t'\,r'}$ est

$$\mathrm{cov}(y_{itr}, y_{i'\,t'\,r'}) = \omega^2 \delta_i^{i'} + \sigma^2 \delta_i^{i'} \delta_t^{t'} \delta_r^{r'}. \tag{5.25}$$

Ainsi, pour une unité statistique donnée la covariance entre deux mesures (indépendamment du temps et de la répétition) vaut ω^2. La variance d'une observation est $\omega^2 + \sigma^2$.

5.3.1.2 Modèle $\mathbf{E}_2$

Une donnée y_{itr} est représentée par le modèle linéaire $y_{itr} = \beta_t + \xi_{it} + \epsilon_{itr}$. La covariance entre deux observations y_{itr} et $y_{i'\,t'\,r'}$ est

$$\mathrm{cov}(y_{itr}, y_{i'\,t'\,r'}) = \tau^2 \delta_i^{i'} \delta_t^{t'} + \sigma^2 \delta_i^{i'} \delta_t^{t'} \delta_r^{r'}. \tag{5.26}$$

Dans ce modèle, seule la covariance entre les répétitions d'une unité statistique pour un temps fixé n'est pas nulle et vaut τ^2. Pour une observation donnée, la variance vaut $\tau^2 + \sigma^2$.

5.3.1.3 Modèle $\mathbf{E}_3$

Une donnée y_{itr} est représentée par le modèle linéaire $y_{itr} = \beta_t + \xi_i + \xi_{it} + \epsilon_{itr}$. La covariance entre deux observations y_{itr} et $y_{i'\,t'\,r'}$ est

$$\mathrm{cov}(y_{itr}, y_{i'\,t'\,r'}) = \omega^2 \delta_i^{i'} + \tau^2 \delta_i^{i'} \delta_t^{t'} + \sigma^2 \delta_i^{i'} \delta_t^{t'} \delta_r^{r'}. \tag{5.27}$$

Ce modèle est le plus complet. En effet, entre deux temps différents, la covariance entre deux mesures pour une même unité statistique vaut ω^2. À temps fixé, la covariance entre les répétitions est de $\omega^2 + \tau^2$ et la variance d'une mesure est $\omega^2 + \tau^2 + \sigma^2$.

5.3.1.4 Remarques

Ces structures de covariance étant relativement proches, plusieurs points doivent être précisés :

– Le choix initial du modèle E_2 tient au fait qu'il traduit le plus simplement possible les corrélations qui peuvent apparaître sur des données expérimentales. Le modèle E_3 semble intéressant car il permet de modéliser des corrélations entre les temps et les répétitions. Cependant, notre présentation se concentre sur le modèle E_2 car son étude est plus simple.

– Ces trois modèles permettent d'introduire des relations de covariance entre les répétitions ou les différents temps. Pour les trois modèles, on considère que les unités statistiques sont indépendantes. Ceci implique que les covariances sont nulles lorsque les observations ne correspondent pas une même unité statistique.

Dans tous les cas, il s'agit de modèles à structure de données incomplètes où les données manquantes sont les effets aléatoires ξ_i. L'ensemble des paramètres de ce type de modèle peut être estimé par l'algorithme EM [89], [112] qui utilise explicitement la structure incomplète des données. Dans l'optique d'utiliser l'algorithme EM pour les mélanges de modèles linéaires mixtes, rappelons que la fonction de log-vraisemblance associée aux données complètes $(\mathbf{y}, \xi)$ s'écrit dans le cas du modèle E_2 :

$$\begin{aligned} l(\beta, \tau^2, \sigma^2 | \mathbf{y}, \xi) \;&=\; -\frac{1}{2}(N + IT)\ln(2\pi) - \frac{1}{2}N\ln(\sigma^2) - \frac{1}{2}IT\ln(\tau^2) \\ &\quad -\frac{1}{2}\frac{(\mathbf{y} - X\beta - U\xi)'(\mathbf{y} - X\beta - U\xi)}{\sigma^2} - \frac{1}{2}\frac{\xi'\xi}{\tau^2}. \end{aligned} \tag{5.28}$$

5.4 Mélange de modèles linéaires mixtes

Nous proposons d'étudier les mélanges de modèles mixtes afin d'analyser les données d'expression. Pour cela, nous commençons par introduire formellement le modèle et nous détaillons ensuite comment ses paramètres peuvent être estimés à l'aide de l'algorithme EM. Enfin, nous donnons quelques résultats de simulations. Une application sera donnée dans la prochaine section.

5.4.1 Modèles de mélange de L2M

Les unités statistiques se répartissent en K composants et les unités statistiques qui appartiennent au composant $\mathcal{C}_k$ définissent un vecteur aléatoire Y_k de dimension $N_k = I_k TR$ où I_k est le nombre d'individus de $\mathcal{C}_k$. On note $\mathbf{y}^k$ le vecteur des observations des éléments du composant $\mathcal{C}_k$. L'hypothèse sous-jacente est que toutes les observations d'une même unité statistique sont issues du même composant. On suppose que le vecteur des observations $\mathbf{y}^k$ du composant $\mathcal{C}_k$ est une réalisation de la variable aléatoire

$$Y^k = X^k\beta_k + U^k\xi^k + \epsilon^k. \tag{5.29}$$

où β_k est un effet fixe, ξ^k est un effet aléatoire et, X^k et U^k sont des matrices d'incidence.

Pour clarifier cette présentation, nous détaillons le cas du modèle E_2. Les vecteurs $\mathbf{y}^k, k = 1, \ldots, K$ vérifient l'équation

$$y^k_{itr} = \beta^k_t + \xi^k_{it} + \epsilon^k_{itr}. \tag{5.30}$$

Ceci est équivalent à dire que $\mathbf{y}^k$ est une réalisation de la variable aléatoire Y^k où :

$$Y^k \sim \mathcal{N}_{N_k}(X^k\beta_k, \tau^2_k U^k(U^k)' + \sigma^2 \mathrm{Id}_{N_k}) \tag{5.31}$$

avec
- $X^k_{(N_k,T)}$ de même structure que X
- β_k vecteur de dimension $(T, 1)$ spécifique au k^e composant et $\beta_k = (\beta^k_t)_{t=1,\ldots,T}$
- $U^k_{(N_k,I_kT)}$ de même structure que U
- τ^2_k variance spécifique au k^e composant
- σ^2 le paramètre de variance sur les erreurs de mesure supposé indépendant des composants du mélange.

De plus, les variables aléatoires ξ^k et ϵ^k sont supposées indépendantes. La variable ξ^k est définie par $\xi^k = ((\xi^k_1)', \ldots, (\xi^k_I)')'$ avec $\xi^k_i = (\xi^k_{it}, t = 1, \ldots, T)$.

Ainsi, la densité pour un individu $\mathbf{y}_i$ du mélange s'écrit

$$g(\mathbf{y}_i|\theta, \mathbf{p}) = \sum_{k=1}^{K} p_k g_k(\mathbf{y}_i|\theta_k) \tag{5.32}$$

où $\mathbf{p} = (p_1, \ldots, p_K)$ et $p_k \geq 0, k = 1, \ldots, K$ sont les proportions du mélange verifiant $\sum_{k=1}^{K} p_k = 1$, $g_k(.|\theta_k)$ est la densité d'une loi gaussienne de paramètre $\theta_k = (\beta_k, \Gamma_k)$ avec $\dot{X}\beta_k$ l'espérance et Γ_k la matrice de variance-covariance. Ainsi, connaissant le composant dont est issue une observation, sa distribution conditionnelle est une loi gaussienne d'espérance $\dot{X}\beta_k$ et de variance Γ_k.

Remarque :
On a ici considéré un modèle de mélange où les paramètres β_k et τ^2_k dépendent du composant $\mathcal{C}_k$ alors que σ^2 est identique pour tous les composants. On peut envisager différents modèles de mélange emboîtés. Dans le cas du modèle (E_2), les modèles emboités que nous choisissons d'étudier sont :
- M_1 : $(\beta_k, k = 1, \ldots, K; \sigma^2, \tau^2)$
- M_2 : $(\beta_k, \tau^2_k, k = 1, \ldots, K; \sigma^2)$
- M_3 : $(\beta_k, \tau^2_k, \sigma^2_k, k = 1, \ldots, K)$.

Le mélange M_1 suppose que les composants se différencient seulement par les effets fixes. Pour le mélange M_2, les composants se distinguent à la fois par l'effet fixe et par l'effet aléatoire. Enfin, le mélange M_3 est une généralisation du modèle M_2 pour lequel le paramètre σ^2_k de l'erreur de mesure dépend du composant $\mathcal{C}_k$.

Pour identifier un mélange de L2M, nous considérons l'approche du maximum de vraisemblance et nous utilisons l'algorithme EM qui tire parti de la structure incomplète des données. Les données manquantes sont de deux types :

(i) les variables indicatrices $(\mathbf{z}_i, i = 1, \ldots, I)$ d'appartenance des unités statistiques à l'un des composants du mélange, $(z_{ik} = 1$ si $i \in \mathcal{C}_k$ et 0 sinon)

(ii) les effets aléatoires $\xi_i^k = (\xi_{it}^k, t = 1, \ldots, T), i = 1, \ldots, I$ pour chaque composant $\mathcal{C}_k, k = 1, \ldots, K$.

5.4.2 Algorithme EM pour le mélange de L2M

Nous exposons la méthodologie pour le modèle de mélange M_2 de modèles mixtes E_2 car il nous apparaît le plus intéressant. On note alors $\mathbf{p} = (p_1, \ldots, p_K)$ les proportions du mélange, $\theta_k = (\beta_k, \tau_k, \sigma)$ les paramètres du modèle linéaire mixte associé au composant $\mathcal{C}_k$ et $\theta = (\theta_1, \ldots, \theta_K)$.

Le logarithme de la vraisemblance du modèle M_2 associé aux données complètes s'écrit :

$$l(\theta, p|\mathbf{y}, \mathbf{z}, \xi) = \sum_{i=1}^{I} \sum_{k=1}^{K} z_{ik} \ln\left(p_k f_k(\mathbf{y_i}, \xi_i^k|\theta_k)\right) \tag{5.33}$$

où $\xi_i^k = (\xi_{i1}^k, \ldots, \xi_{iT}^k)$ est le vecteur de taille T des effets aléatoires pour l'unité statistique i pour le composant $\mathcal{C}_k$ et $\mathbf{y_i}$ de taille TR représente les valeurs prises par l'unité statistique i. Sachant que l'unité statistique i est issue du composant $\mathcal{C}_k$, y_i correspond à une observation de la loi $\mathcal{N}(\dot{X}\beta_k, \Gamma_k)$ avec $\Gamma_k = \tau_k^2 \dot{U}\dot{U}' + \sigma^2 Id_{TR}$ et $\dot{U}_{(TR,T)} = \begin{bmatrix} 1_R & 0_R & \cdots & 0_R \\ 0_R & \ddots & \cdots & \vdots \\ \vdots & \cdots & \ddots & 0_R \\ 0_R & \cdots & 0_R & 1_R \end{bmatrix}$ la matrice d'incidence associée à $\mathbf{y}_i$.

La densité f_k des données complètes conditionnellement à la classe $\mathcal{C}_k$ est une loi normale de moyenne $\mu_k = \begin{bmatrix} \dot{X}\beta_k \\ 0_T \end{bmatrix}$ et de matrice de variance $\Sigma_k = \begin{bmatrix} \Gamma_k & \tau_k^2 \dot{U} \\ \tau_k^2 \dot{U}' & \tau_k^2 Id_T \end{bmatrix}$.

La log-vraisemblance des données complètes s'écrit

$$\begin{aligned} l(\mathbf{p}, \theta|\mathbf{y}, \xi, \mathbf{z}) &= \sum_{i=1}^{I} \sum_{k=1}^{K} z_{ik} \ln(p_k) + \sum_{i=1}^{I} \sum_{k=1}^{K} z_{ik} \ln(f_k(\mathbf{y_i}, \xi_i^k|\theta_k)) \\ &= \sum_{i=1}^{I} \sum_{k=1}^{K} z_{ik} \ln(p_k) + \sum_{i=1}^{I} \sum_{k=1}^{K} z_{ik} h(\theta_k|\mathbf{y_i}, \xi_i^k)) \end{aligned} \tag{5.34}$$

où

$$
\begin{aligned}
h(\theta_k|\mathbf{y_i},\xi_i^k) &= -\frac{1}{2}(TR+T)\ln(2\pi)-\frac{1}{2}TR\ln(\sigma^2)-\frac{1}{2}T\ln(\tau_k^2) \\
&\quad -\frac{1}{2}\frac{(\mathbf{y_i}-\dot{X}\beta_k-\dot{U}\xi_i^k)'(\mathbf{y_i}-\dot{X}\beta_k-\dot{U}\xi_i^k)}{\sigma^2}-\frac{1}{2}\frac{(\xi_i^k)'(\xi_i^k)}{\tau_k^2}. \qquad (5.35)
\end{aligned}
$$

5.4.2.1 Étape E

L'espérance conditionnelle de la vraisemblance complète du modèle de mélange [90] s'écrit ($[q]$ désignant l'indice d'itération) :

$$
\begin{aligned}
\mathcal{Q}(\theta,p|\theta^{[q]},p^{[q]}) &= \mathbb{E}\left(l(p,\theta|\mathbf{y},z,\xi)|\mathbf{y},\theta^{[q]},p^{[q]}\right) \\
&= \sum_{i=1}^{I}\sum_{k=1}^{K}t_i^{[q]}(k)\ln(p_k)+\sum_{i=1}^{I}\sum_{k=1}^{K}t_i^{[q]}(k)\mathbb{E}[h(\theta_k|\mathbf{y_i},\xi_i^k)|\mathbf{y},\theta^{[q]}] \qquad (5.36)
\end{aligned}
$$

avec

$$
t_i^{[q]}(k)=P(i\in\mathcal{C}_k|\mathbf{y}_i,\theta^{[q]},p^{[q]})=\frac{p_k^{[q]}f_k(\mathbf{y_i}|\theta_k^{[q]})}{\sum_{l=1}^{K}p_l^{[q]}f_l(\mathbf{y_i}|\theta_l^{[q]})}. \qquad (5.37)
$$

la probabilité conditionnelle que l'unité i soit issue du composant $\mathcal{C}_k$.

Du fait que $(\xi_i^k)'(\xi_i^k)$ et $(\mathbf{y}_i-\dot{U}\xi_i^k)$ sont des statistiques exhaustives pour le modèle complété [112], il n'est pas nécessaire de calculer explicitement $\mathbb{E}\left(l_\theta(\theta_k;\mathbf{y_i},\xi_i^k)|\mathbf{y},\theta^{[q]},\mathbf{p}^{[q]}\right)$. Afin de maximiser $\mathcal{Q}(\theta,\mathbf{p}|\theta^{[q]},\mathbf{p}^{[q]})$, nous avons simplement besoin de calculer les espérances conditionnelles des statistiques exhaustives $(\xi_i^k)'(\xi_i^k)$ et $(\mathbf{y}_i-\dot{U}\xi-i^k)$, sachant les données observées $\mathbf{y}_i$ et une valeur courante des paramètres $\theta^{[q]}$ et $\mathbf{p}^{[q]}$.

Sachant que $\mathbf{y}_i\in\mathcal{C}_k$, ces espérances conditionnelles s'écrivent (cf. Trottier [119] (p. 49))

$$
\begin{aligned}
\mathbb{E}(\xi_i^{k'}\xi_i^k|\mathbf{y_i},\theta^{[q]}) &= \tau_k^4(\mathbf{y_i}-\dot{X}\beta_k)'\Gamma_k^{-1}\dot{U}\dot{U}'\Gamma_k^{-1}(\mathbf{y_i}-\dot{X}\beta_k) \\
&\quad +R\tau_k^2-\tau_k^4\mathrm{tr}(\Gamma_k^{-1}\dot{U}\dot{U}') \qquad (5.38) \\
\mathbb{E}(\mathbf{y_i}-\dot{U}\xi_i^k|\mathbf{y_i},\theta^{[q]}) &= \dot{X}\beta_k+\sigma^2\Gamma_k^{-1}(\mathbf{y}-\dot{X}\beta_k). \qquad (5.39)
\end{aligned}
$$

5.4.2.2 Étape M

Cette étape consiste à trouver les paramètres qui maximisent $\mathcal{Q}(\theta,p|\theta^{[q]},p^{[q]})$. En dérivant l'expression (5.36) suivant chacun des paramètres du modèle, cette maximisation conduit à

$$
p_k^{[q+1]}=\sum_{i=1}^{I}\frac{t_i^{[q]}(k)}{I},\ \text{for },k=1,\ldots,K, \qquad (5.40)
$$

et à résoudre les équations pour les paramètres β_k,τ_k^2, $k=1,\ldots,K$ et le paramètre σ^2

$$
\sum_{i=1}^{I}t_i^{[q]}(k)\frac{\partial\mathbb{E}[h(\theta_k|\mathbf{y_i},\xi_i^k)|\mathbf{y},\theta^{[q]}]}{\partial\beta_k}=0, \qquad (5.41)
$$

$$\sum_{i=1}^{I} t_i^{[q]}(k) \frac{\partial \mathbb{E}[h(\theta_k | \mathbf{y_i}, \xi_i^k) | \mathbf{y}, \theta^{[q]}]}{\partial \tau_k^2} = 0, \tag{5.42}$$

$$\sum_{i=1}^{I} \sum_{k=1}^{K} t_i^{[q]}(k) \frac{\partial \mathbb{E}[h(\theta_k | \mathbf{y_i}, \xi_i^k) | \mathbf{y}, \theta^{[q]}]}{\partial \sigma^2} = 0. \tag{5.43}$$

En utilisant les espérances conditionnelles définies en (5.38) et (5.39), cela conduit aux expressions explicites pour $k = 1, \ldots, K$

$$\begin{aligned}
\sigma^{2[q+1]} &= \frac{1}{N} \sum_{i=1}^{I} \sum_{l=1}^{K} t_i^{[q]}(l) \Big[\sigma^{4[q]} (\mathbf{y_i} - \dot{X}\beta_l^{[q]})' \Gamma_l^{-2[q]} (\mathbf{y_i} - \dot{X}\beta_l^{[q]}) \\
&\quad + RT\sigma^{2[q]} - \sigma^{4[q]} \mathrm{tr}(\Gamma_l^{-1[q]}) \Big] && (5.44) \\
\tau_k^{2[q+1]} &= \frac{1}{T \sum_{i=1}^{I} t_i^{[q]}(k)} \sum_{i=1}^{I} t_i^{[q]}(k) \Big[\tau_k^{4[q]} (\mathbf{y_i} - \dot{X}\beta_k^{[q]})' \; \Gamma_k^{-1[q]} \dot{U}\dot{U}' \Gamma_k^{-1[q]} (\mathbf{y_i} - \dot{X}\beta_k^{[q]}) \\
&\quad + T\tau_k^{2[q]} - \tau_k^{4[q]} \mathrm{tr}(\Gamma_k^{-1[q]} \dot{U}\dot{U}') \Big] && (5.45) \\
\beta_k^{[q+1]} &= \frac{1}{\sum_{i=1}^{I} t_i^{[q]}(k)} \sum_{i=1}^{I} t_i^{[q]}(k) \Big[\sigma^{2[q]} (\dot{X}'\dot{X})^{-1} \dot{X}' \Gamma_k^{-1[q]} (\mathbf{y_i} - \dot{X}\beta_k^{[q]}) + \beta_k^{[q]} \Big]. && (5.46)
\end{aligned}$$

5.4.3 Remarques

Avant de présenter des résultats de simulations et d'une application, l'introduction de ces modèles mérite quelques remarques.

D'abord, il est clair que l'on peut étendre ces modèles de mélange de L2M à tout autre mélange de modèles linéaires mixtes vérifiant l'équation (5.24). Dans notre présentation, nous nous sommes limité au cas où l'effet aléatoire porte sur les unités statistiques car la classification porte précisément sur celles-ci. Si un tel effet aléatoire existe, il est indispensable de le prendre en considération durant le processus de classification. Suivant le type d'application, il peut s'agir du gène, d'un tissu ou d'un patient. Nous ne précisons pas la nature de la mesure réalisée pour les unités statistiques. Dans la partie application, l'unité statistique considérée est le gène. Dans leurs travaux, Kerr *et al.* [69] considère l'effet gène comme fixe, tout en admettant que l'on pourrait également le considérer comme un effet aléatoire. Les mesures que nous considérerons pour la classification de profils d'expression correspondent à un score issu d'une probabilité a posteriori. L'approche proposée est donc très générale et pourrait trouver d'autres applications pour l'analyse de données de biopuces.

Concernant l'estimation des paramètres des mélanges de L2M, nous avons présenté l'algorithme EM afin de maximiser la vraisemblance. Une alternative au maximum de vraisemblance pour l'estimation des paramètres d'un modèle L2M est la méthode du maximum de vraisemblance restreint (REML). Cette dernière permet de se focaliser sur les composantes de la variance, en faisant disparaître momentanément les effets fixes. Le principe consiste à projeter le modèle sur l'orthogonal du sous espace vectoriel engendré par les colonnes de la matrice X des effets fixes (cf. 5.24). L'estimation du maximum de vraisemblance restreint n'est autre que l'estimation du maximum de vraisemblance sur l'espace projeté. Dans un contexte de mélange, nous n'envisageons pas l'estimation des paramètres suivant la méthode du REML. Nous pensons qu'une telle approche pourrait conduire à des complications techniques sans fournir le maximum de vraisemblance pour les effets fixes $\beta_k, k = 1, \ldots, K$ [112]. Ceci est un inconvénient majeur car nous utilisons ces paramètres dans le calcul des critères BIC et ICL. Sans une bonne estimation de ces paramètres, il est donc difficile de comparer correctement les modèles entre eux.

Par ailleurs, nous ne présentons que trois modèles de mélange. D'autres modèles pourraient également être envisagés :

1. $(\beta, \tau^2, \sigma_k^2), k = 1, \ldots, K$
2. $(\beta, \tau_k^2, \sigma^2), k = 1, \ldots, K$
3. $(\beta, \tau_k^2, \sigma_k^2), k = 1, \ldots, K$
4. $(\beta_k, \tau^2, \sigma_k^2), k = 1, \ldots, K$
5. $(\beta_k, \tau^2, \sigma^2), k = 1, \ldots, K$
6. $(\beta_k, \tau_k^2, \sigma^2), k = 1, \ldots, K$
7. $(\beta_k, \tau_k^2, \sigma_k^2), k = 1, \ldots, K$

D'un point de vue classification, les trois premiers ne présentent pas d'intérêt a priori car les effets fixes qui caractérisent les centres des classes sont égaux. Cependant, ils offrent différentes décompositions de la variance qui peuvent s'avérer pertinentes. Les trois derniers sont ceux que nous avons présentés et que nous illustrerons lors de l'application. Quant au quatrième modèle, bien que nous ne l'ayons pas proposé dans notre application, il pourrait être envisagé.

5.4.4 Simulations

Nous présentons des résultats issus de simulations de Monte-Carlo pour montrer d'une part que l'algorithme proposé permet d'estimer correctement les valeurs des paramètres des modèles simulés et pour illustrer d'autre part l'intérêt de prendre en compte les répétitions dans la classification.

Ces simulations sont proposées sur deux modèles (A) et (B) de type ($E_2.M_2$). Dans les deux cas, I est fixé à 200, le nombre de répétions à 4 et le nombre de classes à 3. Les centres des classes sont caractérisés par $\beta_1 = (0, 0, 2)'$, $\beta_2 = (-1, 0, -1)'$ et $\beta_3 = (1, 2, 0)'$ de sorte qu'ils soient à peu près équidistants. D'autre part, nous avons choisi pour les variances des effets aléatoires des valeurs différentes pour chacune des 3 classes $\tau_1^2 = 0.2, \tau_2^2 = 0.5$ et $\tau_3^2 = 1$. Ainsi, entre les modèles (A) et (B) seule la variance de l'erreur de mesure est différente avec $\sigma_{(A)}^2 = 2$ et $\sigma_{(B)}^2 = 3$.

Nous avons réalisé 100 simulations des modèles (A) et (B). Dans le tableau 5.1, nous donnons les moyennes et les variances des estimations des paramètres. Dans le tableau 5.2, nous donnons la moyenne des taux de bon classement en utilisant la méthode du *maximum a posteriori* (MAP) qui consiste à affecter le gène i au composant $\mathcal{C}_k$ le plus probable a posteriori, c'est-à-dire $k = \arg\max_{1 \leq l \leq K} t_i(l)$.

Les résultats des modèles (A) et (B) montrent que l'on estime en moyenne raisonnablement bien les valeurs des paramètres des modèles simulés. On retrouve bien que la précision des estimations dépend de la variance des effets aléatoires : plus cette variance est importante, plus l'écart-type des estimations est grand pour l'ensemble des paramètres. De même, le pourcentage moyen de bon classement après estimation des paramètres (tableau 5.2) se détériore lorsque la variance des effets aléatoires augmente. On remarque aussi que si la variance σ^2 augmente (modèle (B)), la qualité des estimations est moins bonne (écart-type plus important) bien que la moyenne des estimations reste très correcte. Enfin, en comparant les résultats du tableau 5.2 pour les modèles (A) et (B), on constate que le taux de bon classement baisse lorsque σ^2 augmente. Ces remarques traduisent simplement le bon comportement de la procédure d'estimation du modèle suivant la nature des variations prises en considération.

De plus, pour illustrer l'intérêt de prendre en compte les répétitions, nous réalisons à nouveau 100 simulations d'un modèle (A') identique à (A) excepté pour le nombre de répétitions fixé à $R = 2$. Les résultats du modèle (A') (tableau 5.2) montrent clairement comment le taux moyen de bon classement baisse en l'absence de répétition. Bien qu'il s'agisse ici de données simulées, la prise en compte de répétitions semble être une source d'information importante pour la validation des classes.

5.5 Application à l'analyse de données d'expression

Après une blessure, une régénération rapide de l'épithélium est nécessaire pour maintenir la fonction de barrière. Le processus complexe de cicatrisation implique plusieurs étapes telles que la migration des cellules situées au bord de la blessure et leur prolifération. Ce processus de réparation cellulaire peut être altéré par différents facteurs tels que des radicaux libres, des protéases, et des cytokines inflammatoires qui prolongent l'inflammation et retardent la cicatrisation.

La technologie des puces à ADN a été utilisée pour étudier la réparation cellulaire dans le cas d'une blessure d'un épithélium. Cette étude importante concerne plusieurs types d'expériences et un grand nombre de mesures. Les résultats que nous présentons concernent l'étude de kératinocytes humains blessés et simultanément traités à la blessure avec une drogue telle que SB203580, un inhibiteur spécifique de l'activité de la MAP kinase p38 afin de tester les effets sur la réparation cellulaire. Cette application qui porte sur un ensemble de 999 gènes (sondes plus exactement) vise l'analyse des profils d'expression d'une cinétique à $T = 4$ temps (30 min, 90 min, 180 min et 360 min après avoir blessé les cellules). À chaque instant t de la cinétique, l'expérience réalisée consiste à marquer l'échantillon test (blessé et

TAB. 5.1 – Résultats de l'estimation des paramètres pour 100 simulations des modèles (A) et (B).

	Paramètre	Modèle	Composant 1	Composant 2	Composant 3
(1)	effets fixes $t = 1$		$\beta_1^1 = 0$	$\beta_1^2 = -1$	$\beta_1^3 = 1$
(2)		(A) : $\sigma^2 = 2$	0.029 (0.124)	-1.007 (0.125)	1.071 (0.274)
(2)		(B) : $\sigma^2 = 3$	0.019 (0.141)	-1.044 (0.156)	1.035 (0.420)
(1)	effets fixes $t = 2$		$\beta_2^1 = 0$	$\beta_2^2 = 0$	$\beta_2^3 = 2$
(2)		(A) : $\sigma^2 = 2$	-0.008 (0.131)	0.004 (0.131)	2.061 (0.373)
(2)		(B) : $\sigma^2 = 3$	-0.009 (0.163)	-0.014 (0.163)	2.030 (0.401)
(1)	effets fixes $t = 3$		$\beta_3^1 = 2$	$\beta_3^2 = -1$	$\beta_3^3 = 0$
(2)		(A) : $\sigma^2 = 2$	1.994 (0.153)	-0.992 (0.123)	0.008 (0.256)
(2)		(B) : $\sigma^2 = 3$	1.970 (0.208)	-0.999 (0.165)	-0.037 (0.330)
(1)	proportions		$p_1 = 0.3$	$p_2 = 0.5$	$p_3 = 0.2$
(2)		(A) : $\sigma^2 = 2$	0.301 (0.030)	0.501 (0.041)	0.197 (0.045)
(2)		(B) : $\sigma^2 = 3$	0.308 (0.043)	0.487 (0.060)	0.204 (0.069)
(1)	effets aléatoires		$\tau_1^2 = 0.2$	$\tau_2^2 = 0.5$	$\tau_3^2 = 1$
(2)		(A) : $\sigma^2 = 2$	0.211 (0.092)	0.484 (0.117)	0.901 (0.270)
(2)		(B) : $\sigma^2 = 3$	0.216 (0.134)	0.449 (0.153)	0.866 (0.303)
	erreur de mesure				
(2)		(A) : $\sigma^2 = 2$		2.005 (0.069)	
(2)		(B) : $\sigma^2 = 3$		2.995 (0.095)	

(1) Valeurs des paramètres des simulations
(2) Moyenne et (écart-type) des estimations

TAB. 5.2 – Moyennes des taux (%) de bon classement pour 100 simulations des modèles (A), (B) et (A').

Simulation	Composant 1	Composant 2	Composant 3
modèle (A)	91.70	93.62	76.35
modèle (B)	88.43	88.84	71.12
modèle (A')	85.38	86.18	65.45

(A) Moyennes des taux de bon classement avec $R = 4$ répétitions et $\sigma^2 = 2$.
(B) Moyennes des taux de bon classement avec $R = 4$ répétitions et $\sigma^2 = 3$.
(A') Moyennes des taux de bon classement avec $R = 2$ répétitions et $\sigma^2 = 2$.

traité) avec le fluorophore Cy5 et l'échantillon de référence (non blessé) avec le Cy3. De plus, lors de chaque expérience, des duplications sont réalisées de sorte à obtenir deux mesures pour un même gène. Ces duplications sont obtenues en déposant les mêmes sondes sur deux spots différents de la lame. Pour chaque gène, nous disposons ainsi de $R = 2$ couples (Cy5, Cy3) de mesures. Avant de présenter les résultats, nous détaillons le traitement préalable réalisé.

5.5.1 Analyses préliminaires

Avant de rechercher à classer les profils d'expression, deux étapes ont été successivement menées.

1. *Normalisation des données*
   Afin de supprimer certaines sources de variabilité liées à des problèmes techniques ou biologiques, les données sont normalisées suivant les travaux exposés par Yang *et al.* [129]. Cette normalisation que nous avons présentée au chapitre II, consiste à apporter une correction dépendant du niveau d'intensité du spot et des têtes du robot utilisées pour déposer les sondes.
2. *Recherche de gènes différentiellement exprimés*
   Nous choisissons de mener l'étude des profils d'expression sur les gènes que nous considérons comme différentiellement exprimés. Pour cela, nous avons choisi l'approche que nous avons proposée dans le chapitre III [85] et qui est basée sur les travaux de Newton *et al.* [91]. Plus précisément, nous pouvons obtenir pour chacun des gènes un score dans l'intervalle [-1,1]. Un score négatif signifie que le gène est réprimé alors qu'un score positif indique qu'il est induit. De cette manière, nous obtenons une mesure facilement interprétable et normalisée qui permet d'éviter l'utilisation des ratios entre les niveaux d'expression. En effet, la standardisation de ces ratios entraîne souvent de grandes valeurs pour de faibles variations au cours du temps. Nous avons déjà vu un problème similaire lors de l'utilisation de la statistique de Student pour la recherche de gènes différentiellement exprimés. Parmi les 999 gènes de l'étude, nous en avons ainsi sélectionné $I = 58$.

5.5.2 Résultats de la classification

L'un des avantages de la classification basée sur des modèles probabilistes est qu'il est possible de comparer différentes solutions. Cela permet de confronter différents modèles de classification ainsi que le nombre de classes. Un des critères les plus utilisés pour sélectionner un modèle de mélange est le critère BIC que nous avons déjà introduit. Un second critère, dans le contexte de la classification, est le critère ICL (Integrated Completed Likelihood) [14] qui est une approximation de type BIC pour la vraisemblance intégrée associée aux données complètes.

Pour nos données, nous comparons les performances de 77 modèles de mélange. Les neuf modèles $E_i.M_j$ pour $i = 1, 2, 3$ et $j = 1, 2, 3$ sont considérés. De plus, nous envisageons également de réaliser la moyenne sur les répétitions (modèle E_0). On peut considérer les paramètres de variance des classes égaux (modèle $E_0.M_1$) ou différents (modèle $E_0.M_3$). Pour ces 11 modèles, le nombre de classe varie de $K = 2$ à $K = 8$. Le tableau 5.3 donne les valeurs des critères ICL et BIC pour le modèle $E_2.M_3$. Ce modèle se caractérise par un effet fixe, un effet aléatoire et une erreur de mesure différents pour chaque classe. Pour tous les autres modèles, BIC et ICL donnent des valeurs supérieures à 240. Ainsi, les deux critères sélectionnent fortement le modèle $E_2.M_3$. À partir de ce tableau, il apparaît que les deux critères préfèrent un modèle à $K = 4$ composants et que le cas où $K = 5$ semble intéressant. Les estimations des paramètres pour $K = 4$ sont données dans le tableau 5.4. On constate que la troisième classe est très petite et se compose de deux gènes. Nous avons alors réalisé une nouvelle classification avec les 77 modèles en les retirant de l'analyse. Ces deux gènes sont des contrôles expérimentaux utilisés par les biologistes et, de ce fait la classe 3 peut-être considérée comme une classe de gènes atypiques. Sans ces contrôles, les critères ICL et BIC proposent le modèle $E_2.M_3$ avec trois composants identiques aux trois précédents. Nous concluons donc à une classification stable et robuste de nos données vis-à-vis des gènes atypiques et proposons de poursuivre notre interprétation.

	2 classes	3 classes	4 classes	5 classes	6 classes	7 classes	8 classes
BIC	251.45	214.19	189.41	192.90	206.70	201.50	242.83
ICL	243.82	208.15	183.13	187.42	202.53	199.12	241.64

TAB. 5.3 – Critères BIC et ICL pour le modèle $E_2.M_3$

	Paramètre β	Proportion des classes	variance de l'effet aléatoire	variance erreur de mesure
Classe 1	$\beta_1 = (0.073, 0.321, 0.416, 0.347)$	$p_1 = 0.367$	$\tau_1^2 = 0.078$	$\sigma_1^2 = 0.018$
Classe 2	$\beta_2 = (-0.057, -0.19, -0.313, -0.517)$	$p_2 = 0.195$	$\tau_2^2 = 0.066$	$\sigma_2^2 = 0.004$
Classe 3	$\beta_3 = (0.763, 0.928, 0.997, 0.955)$	$p_3 = 0.034$	$\tau_3^2 = 0.01$	$\sigma_3^2 = 0.000$
Classe 4	$\beta_4 = (-0.122, -0.023, -0.156, -0.075)$	$p_4 = 0.403$	$\tau_4^2 = 0.053$	$\sigma_4^2 = 0.063$

TAB. 5.4 – Estimation des paramètres de la classification

Sur la figure 5.1, les quatre classes obtenues pour le modèle $E_2.M_3$ par l'opérateur du MAP sont représentées suivant le principe décrit par Eisen *et al.* [48]. Les gènes induits figurent en rouge alors que les gènes réprimés sont en vert. De plus, dans chaque classe, les deux mesures du gène sont représentées.

Avec les résultats du tableau 5.4, cette représentation invite à deux types de remarques. Les premières concernent l'interprétation biologique et les secondes la qualité des mesures au sein d'une classe ainsi que la similarité des gènes d'une classe.

– Les classes 1 et 2 sont en opposition aux troisième et quatrième temps. La classe 1 est caractéristique de gènes induits alors que les gènes de la seconde classe sont de plus en

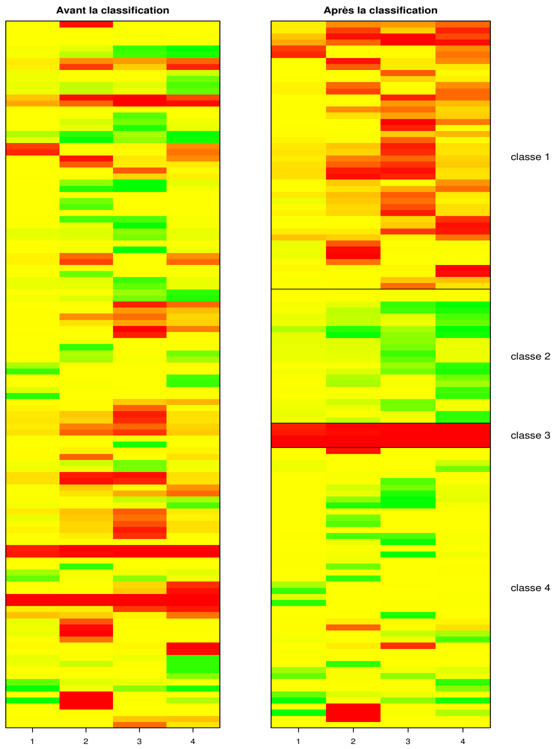


FIG. 5.1 – Classification des profils d'expression.

plus réprimés.
La classe 3 contient donc seulement deux gènes fortement induits qui sont des contrôles introduits par le biologiste. Cette classe ne présente aucun intérêt biologique, mais il est plutôt agréable d'avoir isolé cette classe de gènes atypiques.
La classe 4 contient des gènes dont l'expression ne semble pas différente entre les deux conditions au cours du temps.

– Les classes 1 et 2 ont la plus grande variance sur l'effet aléatoire. Cela suggère que pour ces classes l'effet aléatoire est le plus significatif. En effet, la représentation de type Eisen montre qu'il existe des différences entre les membres des classes. Mais, la variance de l'erreur de mesure est faible. Nous concluons donc à une bonne reproductibilité des mesures pour les gènes de ces classes. Ainsi, la variabilité des mesures au sein de ces classes s'expliquent en partie par des différences biologiques entre les profils. Ces différences se traduisent par un effet gène important au sein des classes.
La variance de l'erreur de mesure est la plus forte pour la classe 4. Ceci traduit une mauvaise reproductibilité des mesures. Il est donc difficile d'interpréter facilement cette classe d'un point de vue biologique.

Pour terminer cette analyse, il est intéressant de voir les résultats obtenus lorsque la variabilité est restreinte à un effet moyen en négligeant de possibles effets aléatoires. Le tableau 5.5 donnent les critères ICL et BIC pour le modèle que nous avons appelé $E_0.M_3$. Il est intéressant de noter que les valeurs des deux critères diminuent lorsque K croît. Cela montre que ce modèle n'est pas adapté pour ce jeu de données et qu'il peut conduire à une classification en un trop grand nombre de classes, car il ne tient pas compte des problèmes de variabilité liés aux gènes.

	2 classes	3 classes	4 classes	5 classes	6 classes	7 classes	8 classes
BIC	415.11	401.59	349.01	328.59	355.95	317.14	285.12
ICL	414.38	400.65	348.14	326.45	353.01	315.37	284.22

TAB. 5.5 – Critère BIC et ICL pour le modèle $E_0.M_3$

5.5.3 Simulations basées sur les paramètres de la classification

Afin de valider les résultats obtenus sur les données expérimentales, nous avons simulé des échantillons de 58 gènes avec 2 répétitions et 4 temps à partir des paramètres estimés sur les données expérimentales. Nous étudions la moyenne et l'écart-type des estimations des paramètres pour 100 simulations d'un tel modèle (dit modèle *expérimental*). Nous voulons nous assurer que la classification obtenue sur les données expérimentales met bien en évidence une structure particulière des données.

Du fait que la classe 3 de l'étude expérimentale ne contient que deux gènes et que sa supression ne modifiait pas les résultats, les simulations de Monte Carlo sont réalisées pour les trois autres classes. Le tableau 5.6 donne les valeurs des paramètres pour 100 simulations ainsi que les moyennes et variances des estimations des paramètres. Ces résultats montrent une bonne estimation des paramètres et rassurent sur la fiabilité du modèle estimé à partir des 58 gènes.

Le tableau de confusion 5.7 donne la moyenne des taux de bon classement sur les 100 simulations précédentes. Les résultats montrent que l'on retrouve bien les classes dont sont issues les données simulées. On remarque également que les gènes mal classés des classes 1 et 2 se retrouvent très majoritairement dans la classe 3. Ceci s'explique par le fait que c'est la classe pour laquelle la variance de l'erreur de mesure est la plus grande. C'est également la classe 3 qui a le taux de bon classement le moins bon (91.25 %). Ces simulations valident l'approche proposée et montre l'intérêt d'introduire l'information issue des répétitions dans la classification.

5.6 Commentaires et perspectives

Nous avons proposé d'aborder la classification de données répétées grâce à des modèles de mélange de modèles mixtes. Dans un problème de classification, cette approche nous a

TAB. 5.6 – Résultats des estimations des paramètres pour 100 simulations du modèle expérimental.

	Paramètre	Composant 1	Composant 2	Composant 3
(1)	effets fixes $t = 1$	$\beta_1^1 = 0.1$	$\beta_1^2 = -0.1$	$\beta_1^3 = -0.1$
(2)		0.103 (0.067)	-0.116 (0.083)	-0.094 (0.066)
(1)	effets fixes $t = 2$	$\beta_2^1 = 0.3$	$\beta_2^2 = -0.2$	$\beta_2^3 = 0$
(2)		0.288 (0.075)	-0.213 (0.100)	0.000 (0.058)
(1)	effets fixes $t = 3$	$\beta_3^1 = 0.4$	$\beta_3^2 = -0.3$	$\beta_3^3 = -0.1$
(2)		0.412 (0.073)	-0.294 (0.093)	-0.092 (0.071)
(1)	effets fixes $t = 4$	$\beta_4^1 = 0.5$	$\beta_4^2 = -0.5$	$\beta_4^3 = -0.1$
(2)		0.512 (0.073)	-0.500 (0.093)	-0.105 (0.077)
(1)	proportions	$p_1 = 0.4$	$p_2 = 0.2$	$p_3 = 0.4$
(2)		0.387 (0.045)	0.209 (0.020)	0.403 (0.051)
(1)	effets aléatoires	$\tau_1^2 = 0.08$	$\tau_2^2 = 0.07$	$\tau_3^2 = 0.05$
(2)		0.072 (0.012)	0.062 (0.015)	0.047 (0.013)
(1)	erreurs de mesure	$\sigma_1^2 = 0.02$	$\sigma_2^2 = 0.004$	$\sigma_3^2 = 0.06$
(2)		0.019 (0.003)	0.005 (0.009)	0.059 (0.011)

(1) Valeurs des paramètres des simulations
(2) Moyenne et (écart-type) des estimations

TAB. 5.7 – Tableau de confusion pour 100 simulations du modèle expérimental.

		Classes estimées			
		Classe 1	Classe 2	Classe 3	Total
Classes simulées	Classe 1	**92.60**	0.22	7.17	100
	Classe 2	0.32	**95.34**	4.33	100
	Classe 3	4.68	4.06	**91.25**	100

permis d'obtenir des résultats plus pertinents que d'autres méthodes car elle tire parti de la méthodologie des L2M. Dans de nombreuses applications où les problèmes de variabilité sont importants, il peut être très judicieux de distinguer les unités statistiques suivant leur variabilité.

En ce qui concerne les biopuces, ces modèles pourraient avoir plusieurs applications et devenir une méthode de référence pour l'étude de profils d'expression lorsque la variabilité est importante. De plus, l'utilisation du BLUP (Best Linear Unbiased Predictor) [112] pour la prédiction des effets aléatoires pourrait permettre d'obtenir des informations plus précises sur les gènes au sein d'une classe et suggérer de nouvelles méthodes de représentation des classes.

Comme nous l'avons décrit, les modèles de mélange de L2M peuvent conduire à, de nombreux modèles. C'est la raison pour laquelle il est important de proposer une bonne méthode de sélection des modèles. Dans ces travaux de thèse, cette sélection s'effectue grâce aux critères BIC et ICL. Nos premières analyses montrent des résultats encourageants pour ces deux critères. Il faudra poursuivre l'étude de leur capacité à fournir des mélanges de L2M pertinents.

Conclusion générale

Bilan

Nous avons considéré les questions les plus fréquentes liées à l'analyse de données d'expression issues de puces à ADN. La normalisation, la recherche de gènes différentiellement exprimés ainsi que l'étude des profils d'expression sont les problèmes les plus importants concernant l'analyse des données d'expression. Notre objectif était de présenter et de développer des solutions pour la recherche de gènes différentiellement exprimés et la classification de profils d'expression.

La variabilité des données expérimentales est actuellement l'un des inconvénients majeurs des techniques liées aux biopuces. Dans ce travail, nous avons naturellement axé nos méthodes sur la recherche de solutions tenant compte de cet aspect crucial.

Les premières données expérimentales dont nous disposions étaient issues d'une seule expérience. Dans le chapitre III, nous nous sommes alors intéressé aux méthodes permettant d'identifier des gènes différentiellement exprimés (entre deux conditions expérimentales) en l'absence de répétition. C'est en comparant les résultats issus de trois expériences que nous avons observé le manque de reproductibilité lié aux données. La mise en place de plans d'expériences pour l'analyse des données est donc apparue comme une étape importante dans le protocole expérimental et nous a conduit à étudier d'autres méthodes. Parmi ces méthodes, l'analyse de variance semblait bien adaptée. Nous avons illustré en détail comment celle-ci peut s'appliquer sur les données des puces à ADN. Cependant, en raison de la nature de certains facteurs, il semblait intéressant de prolonger ces modèles par les modèles mixtes et plus particulièrement les modèles linéaires mixtes (L2M). Sur ce point, nous avons mis en évidence les bonnes propriétés de ces modèles pour notre problème, ouvrant ainsi de larges perspectives. L'introduction d'effets aléatoires pour la modélisation des niveaux d'expression nous a permis de mieux comprendre certaines caractéristiques de ces données. Suite à cela, nous avons envisagé l'analyse des profils d'expression.

Nous avons présenté au quatrième chapitre les méthodes les plus utilisées pour l'étude des profils d'expression. Cependant, celles-ci ne permettent pas de prendre en compte les problèmes de variabilité décrits précédemment. L'objectif du cinquième chapitre était de voir comment il est possible de réaliser proprement une classification sur des données répétées. Les modèles probabilistes donnant un cadre intéressant et cohérent à la classification, nous avons choisi les modèles de mélange pour ce problème. La solution proposée tire parti des précédentes conclusions et modélise chaque unité statistique par un modèle linéaire mixte. Pour le problème de classification, on envisage alors différents modèles de mélange de modèles mixtes. Les 9 modèles proposés se différencient par le modèle linéaire mixte représentant les unités statistiques ainsi que par la nature du mélange. Dans le cas où l'effet aléatoire n'est pas pris en compte, cela revient à considérer un mélange gaussien sur les moyennes des répétitions avec des matrices de variance-covariance proportionnelles à l'identité. Enfin, le choix du modèle de mélange est réalisé à l'aide des critères d'information BIC et ICL.

Dans l'étude que nous avons présentée, nous observons que ces critères ne permettent pas de choisir un modèle si le problème de classification est simplifié en moyennant les répétitions. En revanche, ces deux critères indiquent clairement que parmi les 9 modèles de mélanges de L2M qui proposent une description plus fine des données, il en un est qui donne une bonne représentation des données. Celui-ci permet une meilleure interprétation de la dispersion de nos classes : elle se caractérise soit par des problèmes de mesure, soit par un effet individu au sein des classes.

Perspectives

Dans la continuité de ce travail, nous voyons plusieurs perspectives sur l'étude des données d'expression. Nous distinguons deux aspects essentiels.

- D'un point de vue statistique, de nombreux axes de recherche peuvent être envisagés. Si nous avons décrit l'intérêt des L2M et des mélanges de L2M pour la classification de données répétées, un prolongement naturel de ce travail serait de tirer parti du cadre des modèles mixtes. Par exemple, la prédiction des effets aléatoires (BLUP) donne des perspectives intéressantes pour différencier les individus statistiques (par exemple les gènes) au sein d'un classe. On pourrait également rechercher à appliquer de tels modèles pour l'étude des gènes différentiellement exprimés.

- Dans certaines analyses de données de puces à ADN, l'analyse discriminante est plus appropriée que la classification. Nous avons signalé quelques unes de ces situations au quatrième chapitre. Les mélanges de modèles mixtes peuvent être naturellement adaptés à l'analyse discriminante sur données répétées. En raison des bons résultats obtenus en classification, cette solution mérite sans aucun doute d'être étudiée.

- L'étude du protéome connaît également un essor important. Des problèmes similaires à l'étude de profils d'expression de gènes se posent pour l'étude de profils de protéines [59]. Le niveau d'expression des protéines au sein de la cellule est une information importante pour comprendre la structure des réseaux géniques qui en contrôlent l'expression. Nous pensons que l'étude fonctionnelle des gènes doit tirer parti des informations issues du protéome et du transcriptome [58]. Un problème intéressant serait alors de comprendre les liens qui existent entre les variations d'expression des gènes et des protéines.

En conclusion, l'analyse de données transcriptome et protéome constitue un nouveau domaine d'étude pour la biologie et promet des résultats intéressants en génomique fonctionnelle. La combinaison de ces informations devrait permettre de progresser sur l'inférence de réseaux génétiques que proposent D'haeseleer [38] ou Friedman *et al.* [53]. Une difficulté essentielle qui limitent ces techniques tient aux problèmes de variabilité important. Ceux-ci sont liés d'une part au fait qu'il s'agit d'étudier des phénomènes biologiques et d'autre part à des problèmes technologiques. Pour ces raisons, l'analyse de ces données nécessite l'étude des facteurs de variabilité afin de voir dans quelle mesure nous pourrons en retirer des résultats intéressants d'un point de vue biologique.

Bibliographie

[1] *PIDEX : A statistical approach for screening differentially expressed genes using microarray analysis*, 2002. Poster, soumis à publication.

[2] A. Adams and L. Bishop. Seeded region growing. *IEEE Transactions on pattern analysis and machine learning*, 16 :641–647, 1994.

[3] M. Aickin and H. Gensler. Adjusting for multiple testing when reporting research results : the Bonferroni vs. Holm methods. *American Journal of Public Health*, 86 :726–728, 1996.

[4] H. Akaike. Information theory as an extension of the maximum likelihood principle. In B. Petrov and F. Csaki, editors, *Second International Symposium on Information Theory*, pages 267–281. Budapest, Akademiai Kiado, 1973.

[5] H. Akaike. A new look at the statistical identification model. *IEEE Transactions on Automatic Control*, 19 :716–723, 1974.

[6] A. Alizadeh, M. Eisen, R. Davis, M. C., I. Lossos, A. Rosenwald, J. Boldrick, H. Sabet, T. Tran, X. Yu, J. Powell, L. Yang, G. Marti, T. Moore, J. Hudson Jr, L. Lu, D. Lewis, R. Tibshirani, G. Sherlock, W. Chan, T. Greiner, D. Weisenburger, J. Armitage, R. Warnke, R. Levy, W. Wilson, M. Grever, J. Byrd, D. Botstein, P. Brown, and L. Staudt. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, 403 :503–511, 2000.

[7] I. Axon Instruments. Genepix 4000a users's guide, 1999.

[8] P. Baldi and A. Long. A Bayesian framework for the analysis of microarray expression data : regularized t-test and statistical inferences of gene changes. *Bioinformatics*, 17 :509–519, 2001.

[9] J. Banfield and A. Raftery. Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49 :802–821, 1993.

[10] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini. Tissue classification with gene expression profiles. *Journal of Computational Biology*, 7, 2000.

[11] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate : a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 85 :289–300, 1995.

[12] D. Benson, M. Boguski, D. Lipman, and J. Ostell. Genbank. *Nucleic Acids Research*, 25 :1–6, 1997.

[13] C. Biernacki. *Choix de modèles en classification*. PhD thesis, Université de Technologie de Compiègne, 1997.

[14] C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE, Trans. on PAMI*, 22 :719–725, 2000.

[15] J. Bland and D. Altman. Multiple significance tests : the Bonferroni method. *British Medical Journal*, 310 :170, 1995.

[16] M. Boguski, T. Lowe, and C. Toltohev. dbest-database for "expressed sequence tags". *Nature Genetics*, 4 :332–333, 1993.

[17] D. Bowtell. Options available - from start to finish - for obtaining expression data by microarray. *Nature Genetics supplement*, 21 :25–32, 1999.

[18] L. Breiman, J. Friedman, R. Ohlsen, and C. Stone. *Classification and Regression Trees*. Wadsworth, 1984.

[19] M. Brown, W. Grundy, D. Lin, C. Sugnet, M. Ares, and D. Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences*, 97 :262–267, 2000.

[20] P. Brown and D. Botstein. Exploring the new world of the genome with DNA microarrays. *Nature Genetics Supplement*, 21 :33–37, 1999.

[21] T. Brown. *Genomes*. BIOS Scientific Publishers Ltd, 1999.

[22] T. Brown. *Genomes*, chapter 2, page 20. BIOS Scientific Publishers, 1999.

[23] M. Buckley. The spot user's guide. http ://www.cmis.csiro.au/IAP/Spot/spotmanual, 2000. CSIRO mathematical and information Sciences.

[24] K. Burnham and D. Anderson. *Model selection and inference : a practical information-theoretic approach*. Springer-Verlag, New York, 1998.

[25] M. Callow, S. Dudoit, E. Gong, T. Speed, and E. Rubin. Microarray experiment profiling identifies genes with altered expression in hdl-deficient mice. *Genome Research*, 10 :2022–2029, 2000.

[26] Y. Chen, E. Dougherty, and M. Bitnner. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Journal of Biomedical Optics*, 2(4) :364–374, 1997.

[27] Y. Cheng and G. Church. Biclustering of expression data. In *ISMB 2000*, 2000.

[28] V. Cheung, M. Morley, F. Aguilar, M. A., R. Kucherlapati, and G. Childs. Making and reading microarrays. *Nature Genetics Supplement*, 21 :15–19, 1999.

[29] S. Chu, J. DeRisi, M. Eisen, J. Mulholland, D. Botstein, P. Brown, and I. Herskowitz. The transcriptional program of sporulation in budding yeast. *Science*, 282 :699–705, 1998.

[30] T. Chu, B. Weir, and R. Wolfinger. A systematic statistical linear modeling approach to oligonucleotide array experiments. *Mathematical Biosciences*, pages 35–51, 2002.

[31] J. Claverie. Computational methods for the identification of differential and coordinated gene expression. *Human Molecular Genetics*, 8(10) :1821–1832, 1999.

[32] W. Cleveland and C. Loader. *Smoothing by local regression : principles and methods.* Statistical Theory and Computational Aspects of Smoothing, Springer, New York, 1996.

[33] W. S. Cleveland. Lowess : a program for smoothing scatterplots by robust locally weighted regression. *The American Statistician*, 35 :54, 1981.

[34] W. S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368) :829–836, 79.

[35] J. Daudin, S. Ghachem, M. Descender, S. Robin, A. Hénault, A. Sekowska, and A. Danchin. Comparison of statistical methods for differential analysis of macroarray data, 2001. soumis.

[36] A. Dempster, N. Laird, and D. Rubin. Maximum likelhiood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B 39*, pages 1–38, 1977.

[37] J. DeRisi, V. Iyer, and P. Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278 :680–686, 1997.

[38] P. D'haeseleer. *Reconstructing gene networks from large scale gene expression data.* PhD thesis, University of New Mexico, 2000.

[39] S. Dudoit, J. Fridlyand, and T. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97(457) :77–87, 2002.

[40] S. Dudoit, Y. Yang, T. Speed, and M. Callow. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, 12(1) :111–139, 2002.

[41] D. Duggan, M. Bitnner, Y. Chen, P. Meltzer, and J. Trent. Expression profiling using cDNA microarrays. *Nature Genetics Supplement*, 21 :10–14, 1999.

[42] B. Efron and R. Tibshirani. Bootstrap methods for standard errors, confidence intervals, and others measures of statistical accuracy. *Statistical Science*, 1(1) :54–77, 1986.

[43] B. Efron and R. Tibshirani. Empirical bayes methods and false discovery rates for microarrays. *Genetic Epidemiology*, 23(1) :70–86, 2002.

[44] B. Efron, R. Tibshirani, V. Goss, and G. Chu. Microarrays and their use in a comparative experiment. Technical report, Department of Statistics, Stanford University, 2000.

[45] B. Efron, R. Tibshirani, J. Storey, and V. Tusher. Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96 :1151–1160, 2001.

[46] M. Eisen. Scanalyze, 1999. http ://rana.Stanford.EDU/software.

[47] M. Eisen and P. Brown. DNA arrays for analysis of gene expression. *Methods in enzymology*, 1999.

[48] M. Eisen, P. Spellman, P. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95 :14863–14868, 1998.

[49] D. Finkelstein, J. Gollub, R. Ewing, F. Sterky, S. Somerville, and J. Cherry. Iterative linear regression by sector. In S. Lin, K. Johnson, and K. Academic, editors, *Methods of Microarray Data Analysis. Paper from CAMDA 2000*, pages 57–68, 2001.

[50] C. Fraley and A. E. Raftery. How many clusters ? Which clustering method ? Answers via model-based cluster analysis. *Computer Journal*, 41 :578–588, 1998.

[51] C. Fraley and A. E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97 :611–631, 2002.

[52] C. Fridle, T. Koga, E. Rubin, and J. Bristow. Expression profiling reveals distinct sets of genes altered during induction and regression of cardiac hypertrophy. *Proceedings of the National Academy of Sciences*, 97 :6745–6750, 2000.

[53] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, 7 :601–620, 2000.

[54] T. Galitski, A. Saldanha, C. Styles, E. Lander, and G. Fink. Ploidy regulation of gene expression. *Science*, 285 :251–254, 1999.

[55] D. Ghosh and A. Chinnaiyan. Mixture modelling of gene expression data from microarray experiments. *Bioinformatics*, 18(2) :275–286, 2002.

[56] D. Goldstein, D. Ghosh, and E. Conlon. Statistical issues in the clustering of gene expression data. *Statistica Sinica*, 2002.

[57] T. Golub, D. Slonim, T. Tamayo, C. Huard, M. Gaasenbeek, H. Mesirov, J.P. and-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.

[58] S. Gygi, Y. Rocho, B. Franza, and Aebersold. Correlation between protein and mRNA in yeast. *Molecular and Cellular Biology*, 19 :1720–1730, 1999.

[59] R. Harris, A. Yang, R. Stein, K. Lucy, L. Brusten, A. Hearth, R. Parekh, M. Wterfield, M. O'Hare, M. Neville, M. Page, and M. Zvelebil. Cluster analysis of an extensive human breast cancer cell line protein expression map database. *Proteomics*, 2 :212–213, 2002.

[60] T. Hastie and R. Tibshirani. Discriminant analysis by Gaussian mixtures. *Journal of the Royal Statistics Society*, 58 :155–176, 1996.

[61] T. Hastie, R. Tibshirani, M. Eisen, A. Alizadeh, R. Levy, L. Staudt, W. Chan, D. Botstein, and P. Brown. Gene shaving as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biology*, 1 :1–21, 2000.

[62] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer-Verlag, New York, 2001.

[63] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6 :65–70, 1979.

[64] T. Ideker, V. Thorsson, A. Siegel, and L. Hood. Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. *Journal of Computational Biology*, 7(6) :805–817, 2000.

[65] R. Ihaka and R. Gentelman. R : A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5 :299–314, 1996.

[66] R. Kass and A. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430) :773–794, 1995.

[67] T. Kepler, L. Crosby, and K. Morgan. Normalization and analysis of DNA microarray data by self-consistency and local regression. Santa Fe institute Working Paper, Santa Fe, New Mexico, 2000.

[68] K. Kerr, C. Afshari, L. Bennett, P. Bushel, J. Martinez, N. Walker, and G. Chuchill. Statistical analysis of a gene expression microarray experiment with replication. *Statistica Sinica*, 12, 2002.

[69] K. Kerr, M. Martin, and G. Churchill. Analysis of variance for gene expression microarray data. *Journal of Computational Biology*, 7 :819–837, 2000.

[70] M. Kerr and G. Churchill. Experimental design for gene expression microarrays. *Biostatistics*, 2 :183–201, 2001.

[71] T. Kohonen. *Self-Organizinp maps*. Springer, New York, 1997.

[72] C. Kooperberg, T. Fazzio, J. Delrow, and T. Tsukiyama. Improved background correction for spotted DNA microarrays. *Journal of Computational Biology*, 9(1) :55–66, 2002.

[73] C. Kooperberg, S. Sipione, M. LeBlanc, A. Strand, E. Cattaneo, and J. Olson. Evaluating test-statistics to select interesting genes in microarray experiments. *Human Molecular Genetics*, 2002. à paraître.

[74] E. Korn, J. Troendle, L. McShane, and R. Simon. Controlling the number of false discoveries : Application to high-dimensional genomic data. Technical report, National Cancer Institute, 2001.

[75] E. Lander. Array of hope. *Nature Genetics Supplement*, 21 :3–4, 1999.

[76] L. Lazzeroni and A. Owen. Plaid models for gene expression data. *Statistica Sinica*, 12 :61–86, 2002.

[77] L. Lebart, M. Morineau, and M. Piron. *Statistique exploratoire multidimensionnelle*. Dunod, Paris, 2000.

[78] M. Lee, F. Kuo, G. Whitmore, and J. Sklar. Importance of replication in microarray gene expression studies : Statistical methods and evidence from repetitive cDNA

hybridizations. *Proceedings of the National Academy of Sciences*, 97(18) :9834–9839, 2000.

[79] P. Liang, L. Averboukhl, and A. Pardee. Distribution and cloning of eucaryotes mRNAs by means of differential display : refinements and optimization. *Nucleic Acids Research*, 21 :3269–3275, 1993.

[80] Y. Linde, A. Buzzo, and R. Gray. An algorithm for vector quantizer design. *IEEE Transaction on Communication*, pages 84–95, 1980.

[81] R. Lipshutz, S. Fodor, T. Gingeras, and D. Lockhart. High density synthetic oligonucleotide microarray. *Nature Genetics Supplement*, 21 :20–24, 1999.

[82] I. Lönnstedt and T. Speed. Replicated microarray data. *Statistica Sinica*, 2001. à paraitre.

[83] D. Lockhart, H. Dong, Byrne.M.C., M. Follettie, M. Gallo, and M. Chee. Expression monitoring by hybridation to high-density oligo-nucleotide arrays. *Nature Biotechnol*, 14 :1675–1680, 1996.

[84] G. Lumonics. Quantarray analysis software, operator's manual, 1999.

[85] O. Martin. Puces à ADN et analyse de l'expression des gènes. In *XXXIII^{e} Journées de Statistique*, pages 566–570, 2001.

[86] O. Martin, G. Celeux, and C. Lavergne. Classification de données répétées issues de puces à ADN. Application à l'analyse de profils d'expression. In *JOBIM*, 2002.

[87] G. McLachlan. *Discriminant analysis and statistical pattern recognition.* Wiley, New York, 1992.

[88] G. McLachlan, R. Bean, and D. Peel. A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, 18(3) :413–422, 2002.

[89] G. McLachlan and T. Krishnan. *The* EM *algorithm and extensions.* Wiley, 1997.

[90] G. McLachlan and D. Peel. *Finite mixture models.* Wiley, 2000.

[91] M. Newton, C. Kendziorski, C. Richmond, F. Blattner, and K. Tsu. On differential variability of expression ratios : improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology*, 8(1) :37–52, 2001.

[92] K. Okubo, N. Hori, and R. Matoba. Large scale cDNA sequencing analysis of quantitative and qualitative aspects of gene expression. *Nature Genetics*, 2 :173–179, 1992.

[93] W. Pan. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiment. *Bioinformatics*, 18 :536–545, 2002.

[94] W. Pan and C. Lin, J. C Le. How many replicates of arrays are required to detect gene expression changes in microarray experiments ? A mixture model approach. *Genome Biology*, 2002.

[95] W. Pan and C. Lin, J. C Le. Model-based cluster analysis of microarray gene expression data. *Genome Biology*, 3(2), 2002.

[96] J. Quackenbush. Computational analysis of microarray data. *Nature Review Genetics*, 2 :418–427, 2001.

[97] L. Ramdas, K. Coombes, K. Baggerly, L. Abruzzo, W. Highsmith, T. Krogmann, S. Hamilton, and W. Zhang. Sources of nonlinearity in cDNA microarray expression measurements. *Genome Biology*, 2001.

[98] C. Richmond, J. Glasner, R. Mau, H. Jin, and B. F.R. Genome-wide expression profiling in *Escherichia coli*. *Nucleic Acids Research*, 27(19) :3821–3835, 1999.

[99] B. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge, 1996.

[100] C. Robert. *L'analyse statistique bayésienne*. Economica, 1992.

[101] G. Robinson. That BLUP is a good thing : the estimation of random effects. *Statistical Science*, 6 :15–51, 1991.

[102] D. Rocke and B. Durbin. A model for measurement error for gene expression arrays. *Journal of Computational Biology*, 6(8) :557–569, 2001.

[103] D. Ross, U. Scherf, M. Eisen, C. Perou, C. Rees, P. Spellman, V. Iyer, S. Jeffrey, M. Van de Rijn, M. Waltham, A. Pergamenschikov, J. Lee, D. Lashkari, D. Shalon, T. Myers, J. Weinstein, D. Botstein, and P. Brown. Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics*, 24 :227–234, 2000.

[104] M. Sapir and G. Churchill. Estimating the posterior probability of differential gene expression from microarray data, 2000. Poster. The jackson Laboratory, Bar Harbor, ME.

[105] G. Saporta. *Probabilités, Analyse des données et Statistique*. Paris technip, 1996.

[106] M. Schena, editor. DNA *microarrays : a practical approach*. Oxford University Press, 1999.

[107] M. Schena, editor. *Microarray biochip technology*. Eaton, 2000.

[108] M. Schena, D. Shalon, R. Davis, and P. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270 :460–470, 1995.

[109] M. Schena, D. Shalon, R. Heller, A. Chai, P. Brown, and R. Davis. Parallel human genome analysis : Microarray-based expression monitoring of 1000 genes. *Proceedings of the National Academy of Sciences*, 93 :10614–10619, 1996.

[110] G. Schuler. A gene map of the human genome. *Science*, 274 :540–546, 1996.

[111] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6 :461–464, 1978.

[112] S. Searle, G. Casella, and C. Mc Culloch. *Varience Components*. John Wiley and Sons, 1992.

[113] J. Shaffer. Multiple hypothesis testing. *Annual Review of Psychology*, 46 :561–584, 1995.

[114] P. Spelleman, G. Sherlock, M. Zhang, V. Iyer, K. Anders, M. Eisen, P. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces by microarray hybridization. *Molecular Biology of the Cell*, 9(12) :3273–3297, 1998.

[115] J. Storey. *False Discovery Rates : Theory and Applications to* DNA *Microarrays*. PhD thesis, Department of Statistics, Stanford University, 2002.

[116] J. Storey and R. Tibshirani. Estimating false discovery rates under dependence, with applications to DNA microarrays. Technical report, Department of Statistics, Stanford University, 2001. Submitted to Journal of the American Statistical Association.

[117] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. Lander, and T. Golub. Interpreting patterns of gene expresion with self-organizing maps : methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences*, 96 :2907–2912, 1999.

[118] S. Tavazoie, J. Hughes, M. Campbell, R. Cho, and G. Chirch. Systematic determination of genetic network architecture. *Nature Genetics*, 22 :281–285, 1999.

[119] C. Trottier. *Estimation dans les modèles linéaires généralisés à effets aléatoires*. PhD thesis, Institut National de Polytechnique de Grenoble, 1998.

[120] V. Tusher, R. Tibshirani, and T. Hastie. Signifiance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98 :5116–5121, 2001.

[121] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag,New York, 1996.

[122] V. Velculescu, L. Zhang, B. Vogelstein, and K. Kinzler. Serial analysis of gene expression. *Science*, 270 :467–470, 1995.

[123] J. Ward. Hierarchical groupings to optimize an objective function. *Journal of the American Statistical Association*, 58 :234–244, 1963.

[124] J. D. Watson, M. Gilman, J. Witkowski, and M. Zoller. *ADN Recombinant*. De Boeck-Wesmael, Bruxelles, 2nd edition, 1994.

[125] P. Westfall and S. Young. *Resampling-based multiple testing : examples and methods for p-values adjustment*. Wiley series in probability and mathematical statics. Wiley, 1993.

[126] R. Wolfinger, G. Gibson, E. Wolfinger, L. Bennett, H. Hamadeh, , P. Bushel, C. Afshari, and R. Paules. Assessing gene signifiance from cDNA microarray expression data via mixed models. *Journal of Computational Biology*, 8 :625–637, 2001.

[127] C. Wu. Jacknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics*, 14(4) :1261–1295, 1986.

[128] Y. Yang, M. J. Buckley, S. Dudoit, and T. P. Speed. Comparison of methods for image analysis on cDNA microarray data. *Journal of Computational and Graphical Statistics*, 11(1), 2002.

[129] Y. Yang, S. Dudoit, P. Luu, and T. Speed. Normalization for cDNA microarray data. In M. Bittner, Y. Chen, A. Dorsel, and E. Dougherty, editors, *Microarrays : Optical Technologies and Informatics, Proceedings of SPIE*, volume 4266, 2001.

[130] K. Yeung, C. Fraley, A. Murua, A. Raftery, and W. Ruzzo. Model-based clustering and data transformations for gene expression data. *Bioinformatics*, 17 :977–987, 2001.

TITRE : Approches statistiques pour l'analyse de données de puces à ADN

Résumé :
Les puces à ADN permettent d'étudier simultanément le niveau d'expression de plusieurs milliers de gènes. Ce travail est consacré à l'étude et au développement de méthodes statistiques pour l'analyse de données de puces à ADN. Une fois les données normalisées, on distingue deux problématiques : la recherche de gènes différentiellement exprimés entre deux conditions expérimentales et l'étude de profils d'expression. En raison de la variabilité expérimentale importante, des méthodes statistiques appropriées doivent être mises en place.
Concernant l'identification de gènes différentiellement exprimés, nous montrons la nécessité de disposer de répétitions pour une analyse robuste. L'analyse de variance apparaît dans un premier temps une approche naturelle permettant de prendre en compte différents facteurs de variabilité. Une méthode plus complexe consisterait à introduire différentes composantes de la variance, en représentant les niveaux d'expression des gènes par des modèles linéaires mixtes.
La classification de profils d'expression est abordée avec pour objectif de prendre en compte les problèmes de variabilité. La classification de données répétées permet de séparer pour chaque classe la variabilité due aux erreurs de mesure des effets aléatoires liés à des covariables. Notre approche est basée sur des modèles de mélange de modèles linéaires mixtes. Les unités statistiques d'une classe peuvent être caractérisées par différents modèles mixtes et plusieurs modèles de mélange sont envisagés. Choisissant un cadre probabiliste pour le problème de classification, la question du choix de modèle et du nombre de classes est résolue à l'aide de critères d'information.

Mots-clés : Puce à ADN, Analyse des données d'expression, Analyse de variance, Classification, Modèle mixte, Modèle de mélange, Mélange de modèles mixtes.

TITLE : Statistical methods for the analysis of cDNA microarray data

Abstract :
cDNA microarrays are a method for measuring simultaneously the expression levels of thousands of genes. In this work, we consider statistical methods for gene expression data analysis. After the normalization step, two problems can be considered : first the determination of a statistically significant change in gene expression between two experimental conditions and second the analysis of gene expression profiles. Data variability is one of the main difficulties that led us to develop specific methods.
To detect differentially expressed genes, we describe the need for experimental replication to robustify the analysis. First, analysis of variance is investigated to determine which factors of variability are relevant. An ongoing work considers more complex approach that introduces different variance components. In this model, gene expression levels are represented by linear mixed models.
In a next step, in order to account for data variability in gene expression profiles clustering, we propose to properly exploit the data repetitions. Repeated data clustering is an approach to isolate error measurements from covariable random effects. Our approach is based on mixed model mixture. For each class, statistic units can be characterized by different mixed models and several mixture models are considered. Within a probabilistic framework, the problem of selecting a particular mixed model mixture is solved minimizing information criteria.

Keywords : DNA chip, Gene expression data analysis, Variance analysis, Mixed model, Unsupervised classification, Mixture model, Mixture of mixed models.