REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Mohamed Khider – BISKRA
Faculté des Sciences Exactes, des Sciences de la Nature et de la Vie
Département d'informatique

N° d'ordre : SIOD19/M2/2018

Mémoire

Présenté pour obtenir le diplôme de master académique en

Informatique

Parcours : SIOD

Analyse de données à haut débit issues de puces à ADN

Par :

MOUNIB MERIAM

Soutenu le 05/06/2017, devant le jury composé de :

ZERARKA Med Faouzi	MCB	Président
MERABET Rabia	MAA	Rapporteur
MOUSSAOUI Manel	MAA	Examinateur

Dédicace

A mes très chers parents,

Qui ont toujours été là pour moi, et qui m'ont donné un magnifique modèle de labeur et de persévérance. Ma mère, la plus belle créature que Dieu a créée sur terre, la lanterne qui éclaire mon chemin. Mon père, le signe d'amour et ma raison de vivre. J'espère qu'ils trouveront dans ce travail toute ma reconnaissance et tout mon amour.

A ma chère grand-mère, la source de tendresse, de patience et de générosité « Naziha ».

A mes chers frères et sœurs.

A mes meilleurs amis.

A tout ma famille.

Je dédie ce mémoire

Remerciement

« (Et lorsque votre Seigneur proclama : "Si vous êtes reconnaissants, très certainement J'augmenterai [Mes bienfaits] pour vous) »

[Coran S14.V7]

Avant tout, je remercie Dieu le très haut qui m'a donné le courage et la volonté de réaliser ce modeste travail.

« (CELUI QUI NE REMERCIE PAS LES GENS, NE REMERCIE PAS ALLAH.) »

[Authentique Hadith].

Un grand merci à Dr CHERIET Abdelhakim pour ses conseils toujours pleins Cie bon sens. Je remercie également CHALA Abdelouahed pour son dévouement et sa très grande gentillesse.

Je tiens à remercier tous mes amis, plus spécifiquement, Fatema kadir et seif eddine Brahimi. Leur amitié est un cadeau précieux, auquel je tiens énormément.

Je remercie très chaleureusement mes frères et sœurs pour tout ce qu'il a fait pour moi. Ses douceurs et ses humours ont favorisé mon acclimatation à ce nouvel environnement et m'ont apporté l'équilibre indispensable à la réussite de ces 2 années de maîtrise.

Je ne pouvais terminer ces remerciements sans penser à mes très **chers parents**. Merci pour avoir cru en moi, pour m'avoir inlassablement encouragé et pour avoir toujours été à mes côtés. Merci pour avoir réalisé mon rêve d'enfant en me donnant l'opportunité de poursuivre mes études jusqu' à ce point.

Résumé

Dans le cadre de données d'expression génétique, nous nous intéressons aux méthodes qui permettent d'identifier les gènes significativement différentiellement exprimés entre deux situations biologiques.

Nous allons comparer une méthode classique d'analyse par tests d'hypothèses à des méthodes d'analyse différentielle par régression régularisée. La difficulté de ce genre de jeu de données est la profusion de variables (les gènes) pour assez peu d'individus (les profils d'expression). La stratégie usuelle consiste à mettre en œuvre autant de tests qu'il y a de variables et de considérer que les variables principales sont celles qui ont la « meilleure » p-value. Une stratégie alternative pourrait consister à choisir de classer les variables non plus en fonction de leur significativité (pour un test), mais plutôt de le classer suivant leur poids dans le modèle régularisé obtenu. L'approche qui nous allons utiliser pour identifier les gènes différentiellement exprimés sont dits '*filter'* par la méthode *ebayes*. Le cadre ressemble à celui de l'apprentissage supervisé, car on dispose de profils d'expression géniques pour si possible l'ensemble du génome d'un organisme, chaque puce appartenant à une classe (situation biologique particulière).

L'implémentation des méthodes évoquées dans ce mémoire a été effectuée sous R.

Table des matières

Table des Figures

Liste Des Tables

Liste des Abréviations

ADN	Acide désoxyribonucléique
ADNc	ADN complémentaire
ARN	Acide ribonucléique
ARNc	Acide ribonucléique complémentaire
ARNm	Acide ribonucléique messager
ARNt	Acide ribonucléique de transfert
ARNr	Acide ribonucléique ribosomal
GEO	Gene Expression Omnibus
MIAME	*Minimum Information about a Microarray Experiment*)
PM	Perfect Match
MM	MisMatch
MAS5	Multi Array Suite 5.0
RMA	Robust Multi Array Average
GCRMA	Gene Chip RMA
AvgDiff	Average differential
FC	*Fold Change*, Facteur de changement
FARMS	Factor Analysis for Robust Microarray Summarization
SAM	Significance Analysis of Microarray
Bdf/BG	Bruits de fond/Background
PCR	Polymérisation Chain Réaction
CIBEX	Center for Information Biology gene Expression database
DDBJ	DNA Data Bank of Japan

Introduction Générale

La bio-informatique moderne est née de la convergence de deux aspects de la recherche en biologie : le stockage des séquences moléculaires sur ordinateurs sous la forme de bases données et l'application d'algorithmes mathématiques pour l'alignement des séquences d'acides nucléiques et protéiques. Discipline hybride en constante évolution, la bio-informatique et ses domaines d'applications se précisent.

Les puces d'ADN constituent une nouvelle technologie à haut débit d'analyse d'un grand nombre de gènes de façon simultanée. Cette technologie permet d'étudier en une seule expérimentation le transcriptome,

La technologie des microarray demeure à ce jour un outil important pour la mesure de l'expression génique. Au-delà de la technologie elle-même, l'analyse des données provenant des microarray constitue un problème statistique complexe, ce qui explique la myriade de méthodes proposées pour le prétraitement et en particulier, l'analyse de l'expression différentielle. Toutefois, l'absence de données de calibration ou de méthodologie de comparaison appropriée a empêché l'émergence d'un consensus quant aux méthodes d'analyse optimales. En conséquence, la décision de l'analyste de choisir telle méthode plutôt qu'une autre se fera la plupart du temps de façon subjective, en se basant par exemple sur la facilite d'utilisation, l'accès au logiciel ou la popularité.

Ce mémoire présente une approche nouvelle à l'évaluation de différentes méthodes pour l'analyse de l'expression différentielle sur des données de puces à ADN. Cette approche repose sur l'hypothèse que la probabilité de Co-expression différentielle de gènes associes dans la littérature (Co régulation, même fonction, etc.) est supérieure à celle de gènes choisis au hasard. Ainsi, les résultats produits par la meilleure méthode d'analyse devraient, par conséquent, le mieux refléter les associations tirées de la littérature (signatures moléculaires).

Chapitre 1 :
Introduction au Puce à ADN

1. Introduction

Le terme bio-informatique est apparu pour la première fois dans une publication de Paulien Hogeweg et Ben Hesper, en référence à l'étude des processus d'information dans les systèmes biotique.

La bio-informatique est constituée par l'ensemble des concepts et des techniques nécessaires à l'interprétation informatique de l'information biologique. C'est un champ de recherche à la croisée de plusieurs disciplines que sont la biologie, les mathématiques et l'informatique. Ces principaux champs d'applications sont : la bio-informatique des réseaux, la bio-informatique structurale et la bio-informatique des séquences. La puissance de calcul qu'offre la bio-informatique, a permis l'émergence d'un nouveau champ de la biologie qu'est la génomique.

Ce chapitre décrit la technologie puce à ADN récemment développée qui fournit un accès efficace à l'information génétique en utilisant des puces miniaturisées de haute densité d'ADN ou de sondes oligonucléotidiques. Ces puces à ADN sont des outils puissants pour étudier la base moléculaire des interactions à une échelle qui serait impossible en utilisant l'analyse conventionnelle. Le développement récent de la technologie des puces à ADN a considérablement accéléré les recherches sur la régulation des gènes. Les matrices sont principalement utilisés pour identifier quels gènes sont activés ou désactivés dans une cellule ou un tissu, et également pour évaluer l'étendue de l'expression d'un gène dans diverses conditions. En effet, cette technologie a été appliquée avec succès pour étudier l'expression simultanée de plusieurs milliers de gènes et la détection de mutations ou de polymorphismes, ainsi que pour leur cartographie et leur séquençage.

2. Génomique :

Le génome, mot formé à partir des mots « gène » et « chromosome », est l'ensemble de l'information héréditaire présente dans chaque cellule de chaque organisme vivant et nécessaire au développement et au fonctionnement de l'organisme. [1]

La génomique est l'étude exhaustive des génomes et en particulier de l'ensemble des gènes, de leur disposition sur les chromosomes, de leur séquence, de leur fonction et de leur rôle. [8]

La génomique permet par exemple de mieux comprendre la diversité du vivant, de construire des arbres phylogénétiques ou d'identifier des gènes associés à des maladies. [9]

2.1. Cellule :

Il s'agit d'un compartiment délimité par une membrane et rempli d'une solution concentrée d'éléments chimiques. C'est la plus petite unité capable de manifester les propriétés d'un être vivant, de vie autonome et de reproduction et également le véhicule de la transmission de l'information génétique. C'est l'unité fondamentale structurale et fonctionnelle tout être vivant. [10][11]

Il existe 2 grands types de cellule :

- *Eucaryote* : Cellule comportant un noyau.
- *Procaryote* : Cellule sans noyau (bactérie).

2.2.Acide désoxyribonucléique (ADN)

L'acide désoxyribonucléique (A.D.N) est une molécule présente dans le noyau de la cellule qui joue un rôle central dans la vie cellulaire. Il renferme l'ensemble des informations nécessaires au développement et au fonctionnement d'un organisme. Cette macromolécule a une structure en double hélice constituée de deux brins antiparallèles. Un brin simple est un polymère linéaire constitué de 4 nucléotides. Un nucléotide comprend une des bases : adénosine (A), cytosine (C), guanine (G), ou thymine (T). Les couples A-T et G-C sont appelés bases complémentaires par lesquelles les deux brins vont s'associer [13] (Figure 1.1).

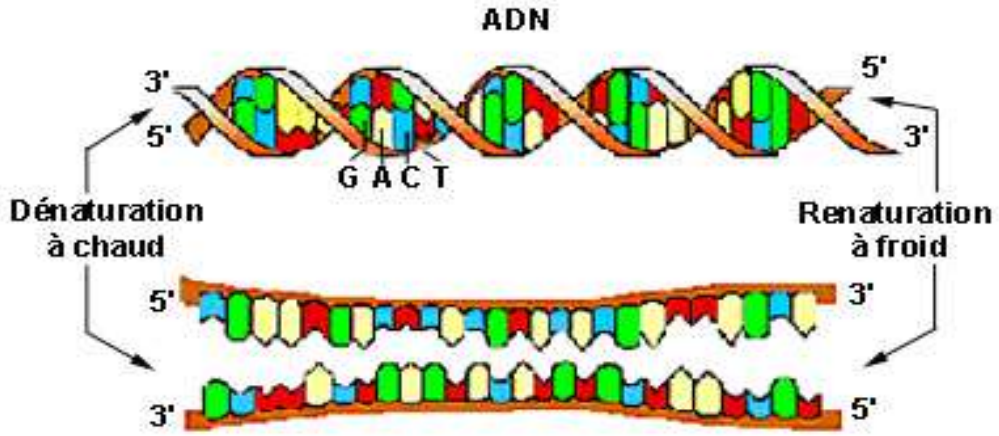

Figure 1 Structure d'une molécule d'ADN.

2.3.Transcriptome

Le gène, unité de base de stockage de l'information génétique, est une petite séquence d'ADN. Il y a environ 6000 gènes chez les levures par exemple et 30000 chez l'homme. L'ensemble du matériel génétique d'un individu ou d'une espèce encodé dans son ADN est appelé alors son génome.

En fonction de leurs besoins, les cellules utilisent à un instant donné une partie des gènes pour réaliser la synthèse des protéines nécessaires aux grandes fonctions cellulaires. Le passage du gène à la protéine se fait en deux grandes parties, la transcription et la traduction, à l'aide d'un agent essentiel l'ARNm, dit ARN messager. Le gène est transcrit (synthèse de l'ARNm) puis l'ARNm est conduit hors du noyau dans le cytoplasme ou il va servir de matrice pour la synthèse des protéines pour la traduction, c.-à-d. le transcriptome est l'ensemble des ARN messagers issus de l'expression d'une partie d'un génome, autrement dit des gènes exprimés.[12]

De manière générale, pouvoir comparer le transcriptome de différents types cellulaires, dans différentes conditions, ou pouvoir analyser l'ensemble du transcriptome d'une cellule à plusieurs phases de son cycle cellulaire ou dans diverses conditions pathologiques, doit permettre de mieux comprendre le fonctionnement cellulaire sur le plan fondamental.

Les méthodes d'analyse du transcriptome les plus utilisées reposent sur la technologie des puces à ADN car elles permettent de visualiser simultanément le niveau d'expression de plusieurs milliers de gènes dans un contexte physiologique ou pathologique particulier [13].

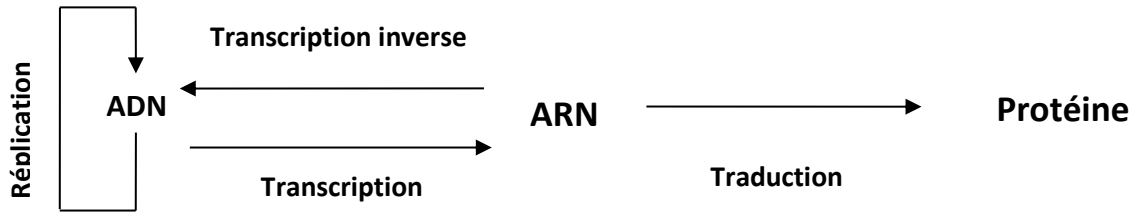

Figure 2 Dogme central de la biologie moléculaire (Source site ISIMA, auteur Vincent Barra)

3. Puce à ADN

Le concept de puce à ADN repose sur une technologie multidisciplinaire intégrant la biologie, la nanotechnologie, la chimie des acides nucléiques, l'analyse d'images et la bio-informatique.

Grace à cet outil, il est possible de mesurer le niveau d'expression de plusieurs milliers de gènes simultanément, et les applications (ex. détermination des familles de gènes Co-régulés, recherche de systèmes de régulation, …) sont en plein essor dans un grand nombre de domaine comme la pharmacologie, la médecine ou l'environnement.

L'analyse biologique des premières biopuces repose sur l'étude du transcriptome (étude des ARN messager) d'une cellule ou d'un organisme. Si de nouvelles applications pour les puces à ADN ont également été développées (biopuces pour le diagnostic, pour la génomique comparative, pour l'identification de régions d'ADN régulatrices après Immuno Précipitation de la chromatine (CHIPchips)).

Elles sont efficaces pour vérifier l'homologie entre deux brins d'ADN, pour détecter des polymorphismes ou des mutations génétiques au sein de l'organisme. Toutefois, dans une définition technique plus exacte, elle représente un arrangement ordonné de plusieurs milliers de gènes séquencés, identifiés et imprimés sur un support solide imperméable, généralement fait de verre, de silicium ou bien de membrane en nylon [2].

La fabrication d'une puce à ADN se décompose en trois étapes : la production des sondes (fragments courts d'une séquence d'ADN connus) et leur dépôt sur le support, la production et le marquage des cibles (fragments inconnus d'ADN que l'on cherche à identifier), enfin l'hybridation des sondes avec les cibles. Ces différentes étapes constituent les étapes de base pour la fabrication de toutes les puces, indépendamment de la technologie utilisées [3].

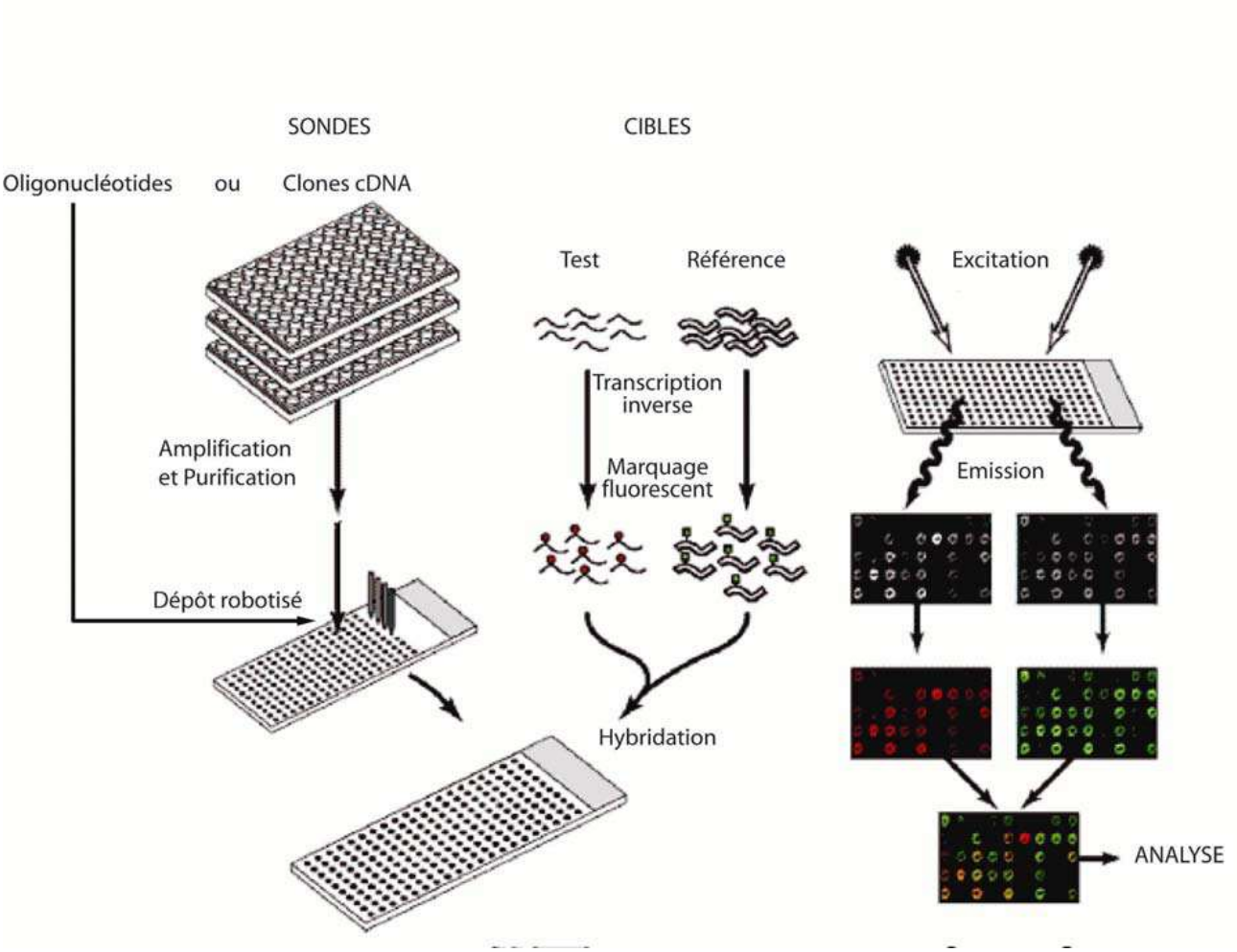

Figure 3 Principe d'utilisation de la puce à ADN.

3.1.Principe de fonctionnement d'une puce à ADN

Le but de **puce à ADN** était de mesurer chez l'homme l'expression simultanée d'un grand nombre de gènes, voire de l'ensemble des gènes contenus dans le génome. Il s'agit d'analyser des profils d'expression génique et à les corréler à des processus biologiques afin de définir des signatures d'expression. La confection des puces à ADN a permis d'étendre ce principe à la détection simultanée de milliers de séquences en parallèle.

Une puce comporte quelques centaines à plusieurs dizaines de milliers d'unités d'hybridation appelées « spots » (de l'anglais « spot » = tache), chacune étant constituée d'un dépôt (sondes) de fragments d'ADN ou d'oligonucléotides correspondant à des sondes de séquences données. L'hybridation de la puce avec un échantillon biologique, marqué par un radioélément ou par une molécule fluorescente, permet de détecter et de quantifier l'ensemble des cibles qu'il contient en une seule expérience. [15]

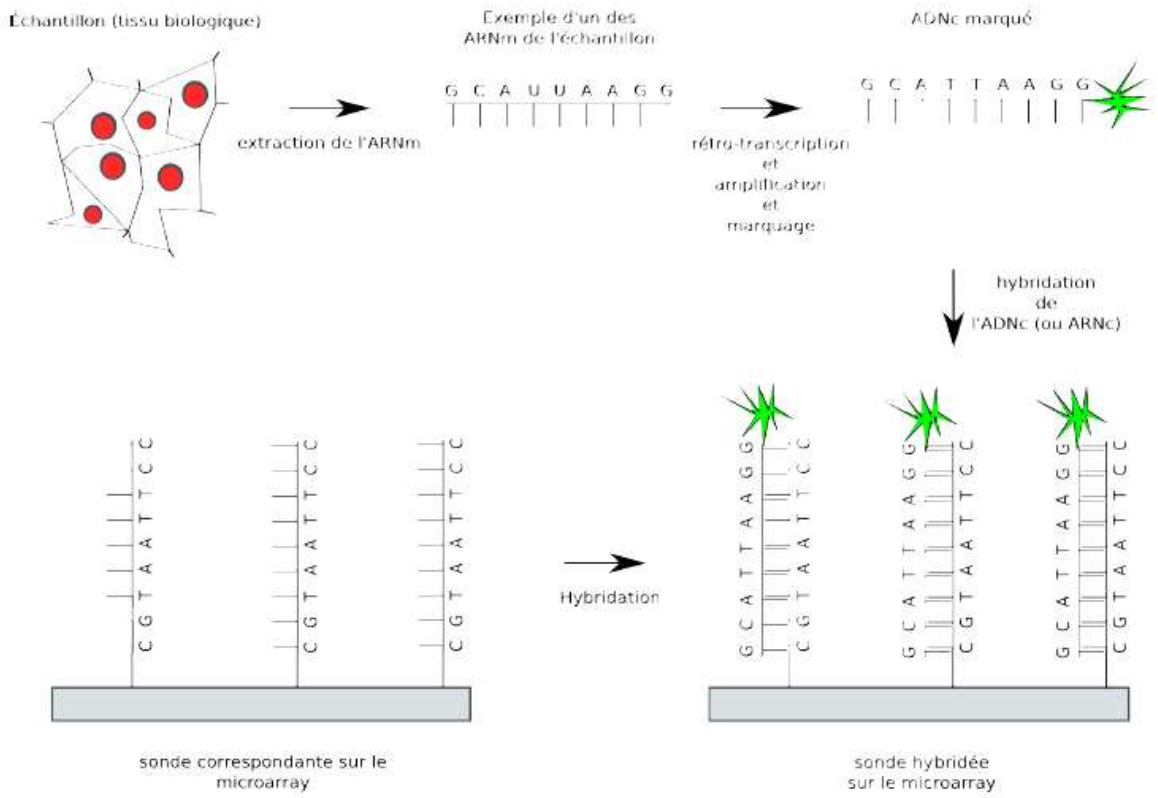

Figure 4 Principe générale des microarray dans le contexte de mesure de l'expression.

3.2. Sonde

Sont des fragments d'ADN synthétique représentatif des gènes dont on cherche à étudier l'expression. Les sondes (oligonucléotides ou clones d'ADNc purifiés et amplifiés) sont déposées mécaniquement sur une lame de verre. La zone de dépôt du gène est appelée spot. Les sondes sont dénaturées en un seul brin pour pouvoir être hybridées avec les cibles marquées dans la phase d'hybridation.

3.3. Cible

Les cibles sont couplées à des marqueurs fluorescents (parfois amplifiés) par transcription inverse. Par exemple, la cible test est marquée par une Cyanine 5 (Cy5) rouge et la cible de référence par une Cyanine 3 (Cy3) verte. Les cibles sont assemblées pour former un mélange complexe. Ce mélange pourra s'hybrider, dans des conditions d'astringence particulières, avec les sondes présentes sur la puce.

3.4. Acquisitions des données

La lecture est réalisée par un scanner muni d'un microscope convocable, couplé à deux lasers. Ces lasers possèdent des longueurs d 'ondes d'excitation spécifiques, correspondant à celles des deux marqueurs fluorescents. L'excitation et l'émission (amplifiée par des photomultiplicateurs) des fluorochromes permettent l'obtention de deux images (une pour chaque marqueur) en niveau de gris. Ces images sont ensuite converties en pseudo-couleur et fusionnées pour être analysées par un logiciel d'analyse d'images.

3.5.Plateformes

Il existe actuellement deux types de puces à ADN qui dominent le marché :

- Les puces à ADNc qui fonctionnent avec des micros points contenant des fragments d'ADN sur un support de verre. La société Agilent est l'une des plus grandes industries qui les commercialisent [2].
- Les puces à oligonucléotides qui reposent sur le principe de synthèse in situ de milliers de séquences distinctes d'oligonucléotides. La société Affymetrix est l'unique détenteur de cette technologie [2].

3.5.1. Technologie Agilent

Les solutions Agilent vont de l'analyse génétique à l'échelle du génome à la détection ciblée de mutations dans un ou plusieurs sites et à l'analyse des changements dans la régulation des gènes. Nous fournissons des puces à ADN, des réactifs cibles ciblés de nouvelle génération et un flux de travail d'hybridation par fluorescence in situ, puissants, personnalisables et faciles à mettre en œuvre dans votre laboratoire.

3.5.2. Technologie Affymetrix

fournit des produits, des outils et des ressources Affymetrix ™ innovants qui aident à faire progresser le travail des chercheurs grâce à l'analyse de puces à ADN. Les domaines d'application qui bénéficient d'une telle approche comprennent la génomique végétale et animale et la transcriptomique, y compris la recherche fondamentale et l'application industrielle des technologies pour la sélection, la diversité et la conservation de la population, l'analyse des caractères, etc. Des solutions uniques sont également disponibles pour la recherche sur le cancer, de la découverte à la recherche clinique et à la validation, comme la recherche en cytogénétique des troubles constitutionnels et du cancer. La génétique des traits complexes humains, les troubles mendéliens et les populations sont également des matrices d'application

qui peuvent être avancées avec l'analyse de puces à ADN Affymetrix, comme le génotypage optimisé en population et optimisé pour l'application afin de faciliter les processus de recherche en génétique humaine [5]

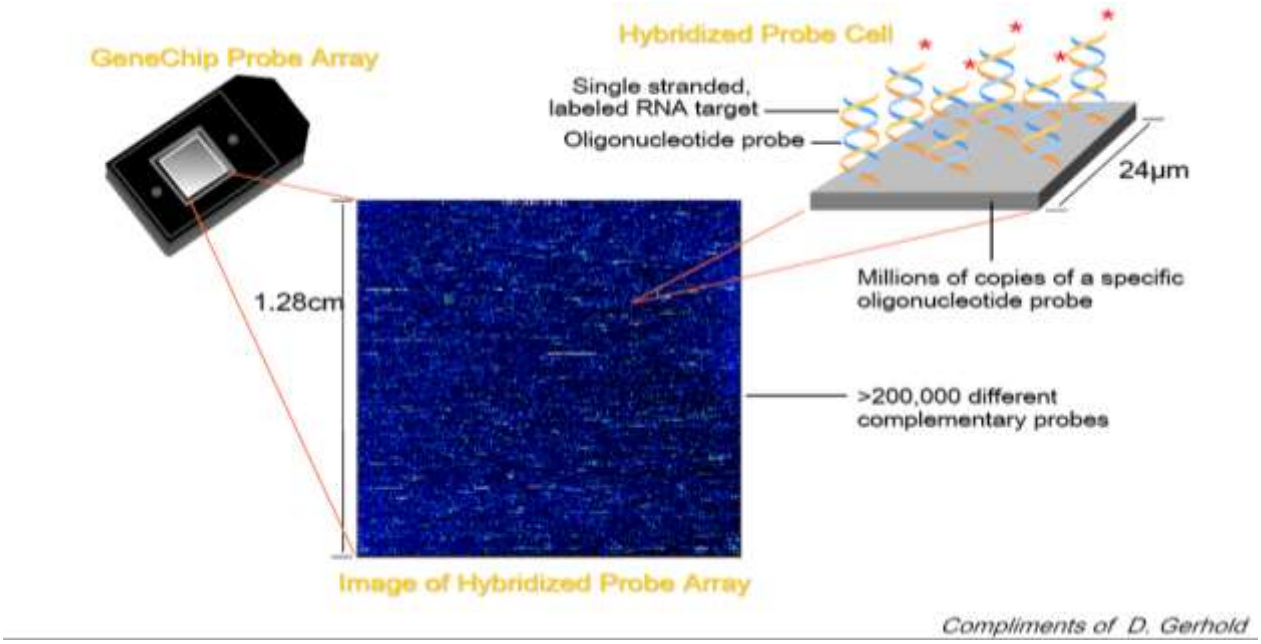

Figure 5 Technologie Affymetrix

3.6.Domaine d'application

La puce à ADN a été utilisées avec succès dans plusieurs processus biologiques variés : en cancérologie, recherche pharmaceutique, génotypage, diagnostic, contrôle agroalimentaire et industriel, bioterrorisme.

3.6.1. L'environnement

Les secteurs de la défense et de l'environnement espèrent aussi aux diverses applications des puces à ADN, notamment pour la détection rapide et à bas coût de substances organiques, principalement des agents pathogènes dilués dans l'environnement.

3.6.2. Diagnostic médicaux

La puce à ADN a encore un grand rôle à jouer dans une autre application des polymorphismes et de la détection banalisée de ceux-ci. Aussi, la Puce ADN peut faciliter dans l'identification des gènes neufs, ainsi qu'en étudiant leur fonctionnement et expressions dans différentes conditions. Elle comprend la détermination de la séquence génétique de tous les types d'organismes tels que des êtres humains, des souris, ainsi que des microbes. Elle aide

également dans des études de conduite dans le domaine de l'agriculture ; par exemple, elle peut être utilisée pour des études liées au contrôle des parasites [13].

3.6.3. Expertise médico-légale

Le but est l'identification humaine dans le cadre d'enquêtes policières ou judiciaires. Les analyses sur le terrain étant très souvent complexes ainsi que la confidentialité et le respect de la procédure judiciaire assez lourdes, il sera souhaitable d'avoir sur les lieux d'enquête des systèmes portables d'analyse de l'ADN. L'analyse immédiate sur le terrain permettrait d'affiner la recherche d'échantillons. La puce à ADN a là encore un avenir très prometteur.

3.7.Banques de données génomiques

Parmi les banques de données publiques, les banques données d'expression de gêne sont particulièrement importantes et intéressantes en termes de partage des connaissances. Ces banques de données se répartissent globalement en 2 catégories plus ou moins généralistes.

Les banques de données généralistes pour le dépôt des données d'expression de gènes (*dépôts*) ont été développées dans le but de partager les données d'expression de gènes (notamment issues des expériences de puces à ADN) au niveau de la communauté scientifique internationale. L'une de leur priorité est le respect par les biologistes du standard international MIAME [27] pour uniformiser les données et faciliter leur diffusion.

Les trois principales banques de données généralistes pour le dépôt des données d'expression de gènes sont ArrayExpress59 à l'EBI [28], GEO au NCBI [29] et Cibex au DDBJ [30]. Ces *repositories* sont d'importance grandissante puisque, aujourd'hui, la majorité des journaux scientifiques requièrent, pour toutes publications dans le domaine des puces à ADN, le dépôt des données d'expression dans au moins une des banques de données publiques conforme au standard international MIAME.

Les *dépôts* permettent de comparer les dessins expérimentaux réalisés pour répondre à diverses questions biologiques. Ils offrent la possibilité de confronter des matrices de données d'expression générées par différentes équipes, sur différents modèles et/ou différentes plates-formes. Les résultats de ces comparaisons permettent, entre autre, d'améliorer l'annotation et la connaissance sur les gènes dans les différentes conditions [31]. L'autre intérêt de ces banques de données généralistes est la mise à disposition des jeux de données aux communautés de chercheurs en bio-informatique, mathématiques et statistiques pour le développement de

nouvelles méthodologies d'analyse. La conférence internationale Pacific Symposium on Biocomputing est un exemple de réussite de ce type d'approche. De même, de nombreux articles, parus dans les journaux traitant de bio-informatique, présentent des algorithmes testés sur des jeux de données extraits des différentes banques de données publiques. Les résultats obtenus offrent même parfois un complément d'information sur les résultats biologiques [14].

3.7.1. Gene Expression Omnibus (GEO)

L'expression génique omnibus (GEO) [7] : est un dépôt public international qui archive et distribue librement les puces à ADN, le séquençage de nouvelle génération et d'autres formes d'ensembles de données génomiques fonctionnelles à haut débit. Environ 90% des données dans GEO sont des études d'expression génique qui étudient un large éventail de thèmes biologiques, y compris la maladie, le développement, l'évolution, l'immunité, l'écologie, la toxicologie, le métabolisme, et plus encore. Les données de non-expression dans GEO représentent d'autres catégories d'études génomiques et épi-génomiques fonctionnelles, y compris celles qui examinent la méthylation du génome, la structure de la chromatine, les variations du nombre de copies du génome et les interactions génome-protéine.

Les données du GEO représentent des recherches originales soumises par la communauté scientifique conformément aux conditions des subventions ou des revues qui exigent que les données soient disponibles dans un référentiel public, l'objectif étant de faciliter l'évaluation indépendante des résultats, l'analyse et un accès complet à toutes les parties du document à l'étude. La ressource prend en charge l'archivage de toutes les parties d'une étude, y compris les fichiers de données brutes, les données traitées et les métadonnées descriptives, qui sont indexées, interconnectées et consultables. Bien que le rôle principal de GEO soit de servir d'archive de données primaires, la ressource offre également plusieurs outils et fonctionnalités qui permettent aux utilisateurs d'explorer, d'analyser et de visualiser des données d'expression à la fois centrées sur les gènes et centrées sur l'étude.

En résumé, les principaux objectifs de GEO sont les suivants :

- Fournir une base de données d'archives de données primaires robuste et polyvalente dans laquelle stocker efficacement une grande variété d'ensembles de données génomiques fonctionnelles à haut débit.
- Offrez des procédures et des formats de soumission simples qui supportent des dépôts de données complets et bien annotés provenant de la communauté de recherche.

- Fournir des mécanismes conviviaux permettant aux utilisateurs de localiser, d'examiner et de télécharger des études et des profils d'expression génique d'intérêt.

3.7.2. Array Express

ArrayExpress [6] est une base de données publique d'expérience de puce à ADN et de profils d'expression des gènes[32].

Elle est constituée de 3 composantes :

– ArrayExpress repository : Qui est conforme au standard MIAME. Les expériences peuvent être soumises à cet entrepôt grâce à l'outil en ligne MIAM Express ou en chargeant des tableurs (MAGE-TAB de préférence)

– ArrayExpress Warehouse : qui est une base de données de gènes, sélectionnés à partir de l'ArrayExpress repository, dont les profils d'expression sont indexés.

– ArrayExpress Atlas : Qui est une nouvelle base de données résumée pour interroger les gènes d'expression organisés et classés à travers de multiples expériences et conditions.

Le scannage d'une puce permet de produire une image. Des repères sur la puce permettent de retrouver sur celle-ci, la localisation de chaque carré de sonde qui correspond à 1 secteur en ignorant les pixels externes. Un algorithme est utilisé pour calculer l'intensité de la cellule (secteur) à partir des pixels centraux (distribution des intensités par pixels : calcul du $75^{ème}$ centile= intensité du spot). L'intensité moyenne est égale à la valeur d'expression relevée par la sonde qui est égale au fichier brute de données. Pour les puces Affymetrix, l'image d'une puce est stockée dans un fichier à l'extension DAT et les intensités des sondes obtenues à partir de l'analyse des images sont stockées dans un fichier à l'extension CEL. Certaines informations additionnelles telles que l'identifiant associant une sonde ou paire de sondes à un ensemble de sondes est stocké dans un fichier CDF.

Le passage du fichier à l'extension DAT à celui à l'extension CEL nécessite l'utilisation de logiciel Affymetrix. A partir des données brutes récupérables dans de nombreux dépôts publiques, des prétraitements seront effectués sur celles-ci afin de les adapter à l'analyse souhaitée.

Un fichier CEL (Cell Intensity File) sauvegarde les données d'intensité pour chaque sonde sans traitement obtenues à partir d'un fichier DAT. Une valeur représentative de l'intensité est sauvegardée pour chaque cellule (pixel) de l'image. Les deux dernières versions de ce fichier,

la 3 et la 4 sont plutôt différentes. Dans la version 3, le format du fichier CEL est similaire à celui d'un fichier au format INI sous Windows. Il est divisé en section contenue entre une balise ouvrante et une fermante. Les différents noms de section sont "CEL","HEADER", "INTENSITY", "MASKS", "OUTLIERS" et "MODIFIED" et les données dans chaque section sont de la forme ETIQUETTE=VALEUR.

La version 4 du fichier est sous la forme binaire et les valeurs sont sauvegardées dans le format little-endian. Dans cette version il n'existe pas de sections mais des items stockant approximativement les mêmes données que la version 3 en utilisant les types de données integer, DWORD, float et short.

CDF (Chip Description File) est un fichier de description de puces Affymetrix qui décrit l'agencement d'un tableau GeneChip Affymetrix. Il contient les informations concernant les caractéristiques de conception du tableau de la sonde, l'utilisation et le contenu de la sonde, et les paramètres d'analyse et de scannage. Il existe 2 types de formats pour ce fichier. Le premier est un fichier texte au format ASCII utilisé par les logiciels MAS et GCOS1.0 et le second est un fichier au format XDA utilisé par les anciennes versions de GCOS. Le fichier texte au format ASCII est similaire à un fichier texte à l'extension INI sous Windows. Il est divisé en sections suivant le même principe que les fichiers CEL et les différentes sections sont : "CDF", "Chip", "QCI", "UnitJ" et "UNITJ_Block". Le format XDA quant à lui est un fichier binaire permettant un accès rapide aux données tout en minimisant l'espace de stockage. Il utilise le format little-endian pour stocker les valeurs dans ce fichier. Il a la même présentation que les fichiers CEL au format binaire et utilise les mêmes types de données.

3.8.Les étapes d'analyse des données des puces à ADN

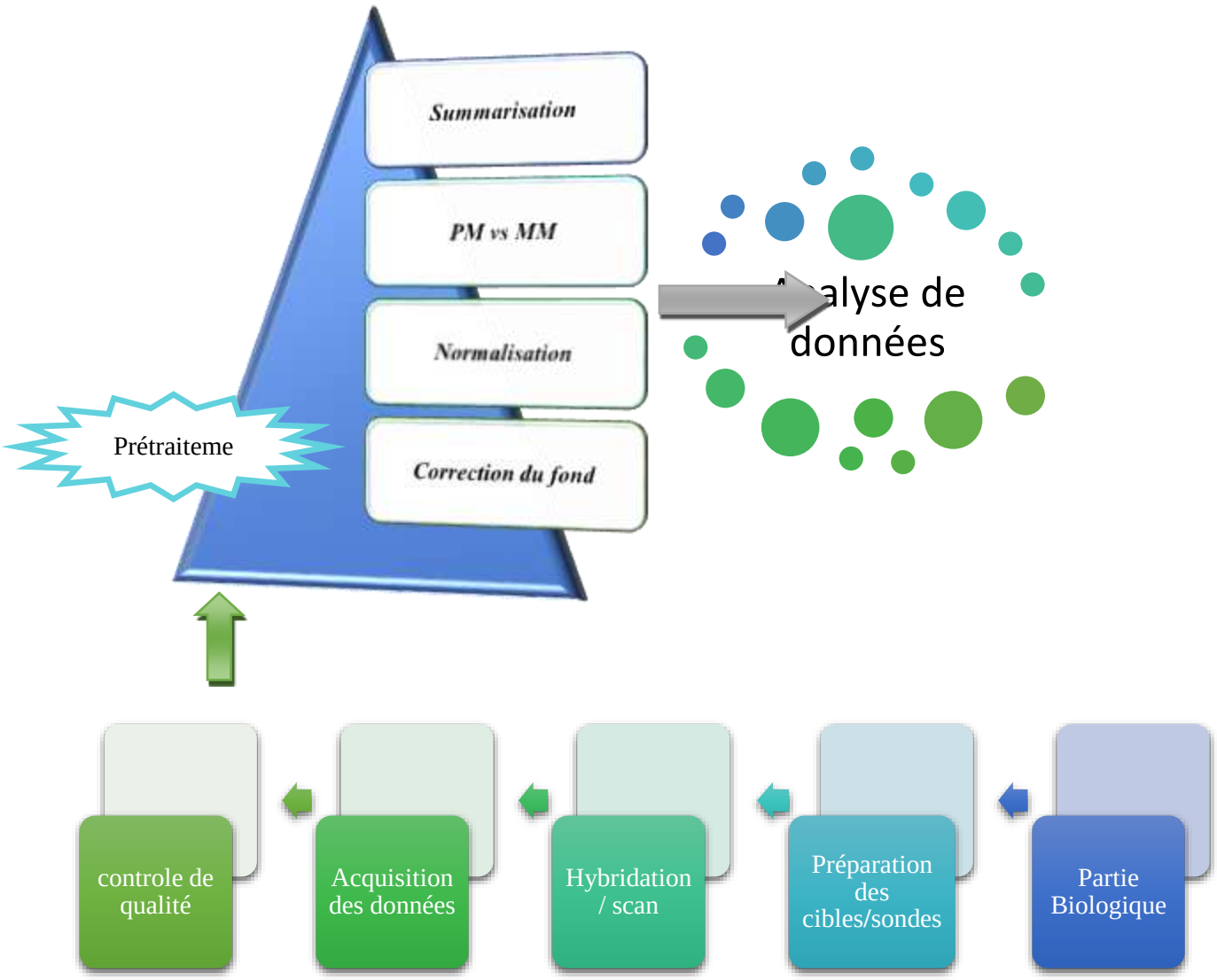

Figure 6 Déroulement typique d'une expérience de microarray sur la plateforme GeneChip.

Préparation des cibles/sondes : les cibles sont identifiées par un marquage radioactif ou fluorescent. Bien que moins sensibles que les marquages radioactifs, certains systèmes de marquages fluorescents présentent l'avantage de pouvoir identifier plusieurs cibles sur la même puce. Et Les sondes sont préparées séparément à partir de clones d'ADNc, soit de l'ARNm mature rétrotranscrit en ADN, puis clone, puis amplifie par PCR.

Hybridation : Une fois les sondes déposées, la lame est mise en contact avec le transcriptome du ou des tissus choisis. Pendant cette phase, les gènes exprimés dans le tissu vont s'hybrider

avec les séquences correspondantes déposées sur la lame, par le principe de complémentarité des bases.

Acquisition des données : le scanner produit une image avec des spots de différentes intensités en fonction du niveau d'expression des gènes correspondants. L'intensité des spots est alors quantifiée à l'aide d'un logiciel d'analyse d'images.

Contrôle de qualité : l'objectif étant de s'assurer de la qualité des données pour chaque biopuces, et d'obtenir des données comparables entre les différentes biopuces.

Prétraitement : est une étape primordiale dans le processus de découverte de connaissances. En fonction des données traitées, l'objectif de ce travail consistait à optimiser tout ce processus en proposant les méthodes et outils les mieux adaptés afin de s'assurer de la qualité des données.

Analyses des données : la dernière étape consiste à analyser les données pour en extraire de la connaissance. Différentes méthodes peuvent être appliquées aux données d'expression de gènes, la plus couramment utilisée dans le cadre de ce mémoire est la recherche des gènes différentiellement exprimés entre différentes conditions expérimentales grâce notamment aux tests d'hypothèse.

4. Conclusion

La bio-informatique est le domaine de la science où la biologie, l'informatique, les mathématiques et la technologie de l'information convergent comme une discipline unique. Reconnues tardivement dans la littérature scientifique, les techniques et les outils bio-informatiques d'analyse des données biologiques font aujourd'hui l'objet de nombreux articles et manifestations scientifiques.

Les puces à ADN sont des outils puissants pour l'analyse du transcriptome. Elles permettent, entre autres, de visualiser simultanément le niveau d'expression de plusieurs milliers de gènes dans un type cellulaire et un contexte physiologique et/ou pathologique particulier. Elles offrent des perspectives d'applications dans les domaines du diagnostic et pronostic médical.

Cette technologie est pluridisciplinaire. Elle intègre la biologie moléculaire, la chimie, l'informatique, l'électronique et la robotique. La production de données en masse, avec une fiabilité de plus en plus grande, ne cesse de s'accélérer. Le recours aux moyens informatiques pour gérer, exploité, analyser cette pléthore de données est devenu indispensable.

Chapitre 2 :

Introduction à l'analyse différentielle d'expression des gènes pour les puces à ADN

1. Introduction

Un des principaux objectifs des puces à ADN est de déterminer quels sont les gènes différentiellement exprimés entre différentes conditions expérimentales et déterminer quels gènes caractérisent un état particulier. Nous verrons dans ce chapitre, les principales méthodes permettant de répondre à ce type d'interrogations.

La détermination des gènes différentiellement exprimés se décompose en deux principales étapes, le prétraitement des données permettant d'assurer la fiabilité des résultats fournis puis l'analyse différentielle des données permettant de donner du sens à toutes ces informations. Diverses méthodes d'analyse de données sont particulièrement utilisées pour les données d'expression.

Ensuite, des méthodes de filtrage par *ebayes* permettent d'explorer les données, permettant ainsi de savoir quels gènes ou quels échantillons ont des profils d'expression similaires. Si des informations supplémentaires sont connues à priori sur les données et que nous souhaitons utiliser ces connaissances,

Le choix des méthodes dépend toujours de l'objectif global de l'étude. Divers logiciels dédiés aux données d'expression permettent de réaliser ces analyses, on peut citer GeneSpring GX (Agilent), Bioconductor...

2. Analyse absolue

L'analyse absolue regroupe les étapes de prétraitement des données, de la détection qualitative des gènes et de la recherche par profil. Elle donne en sortie un tableau gène expression dont les données sont nettoyées et adaptées à l'analyse envisagée.

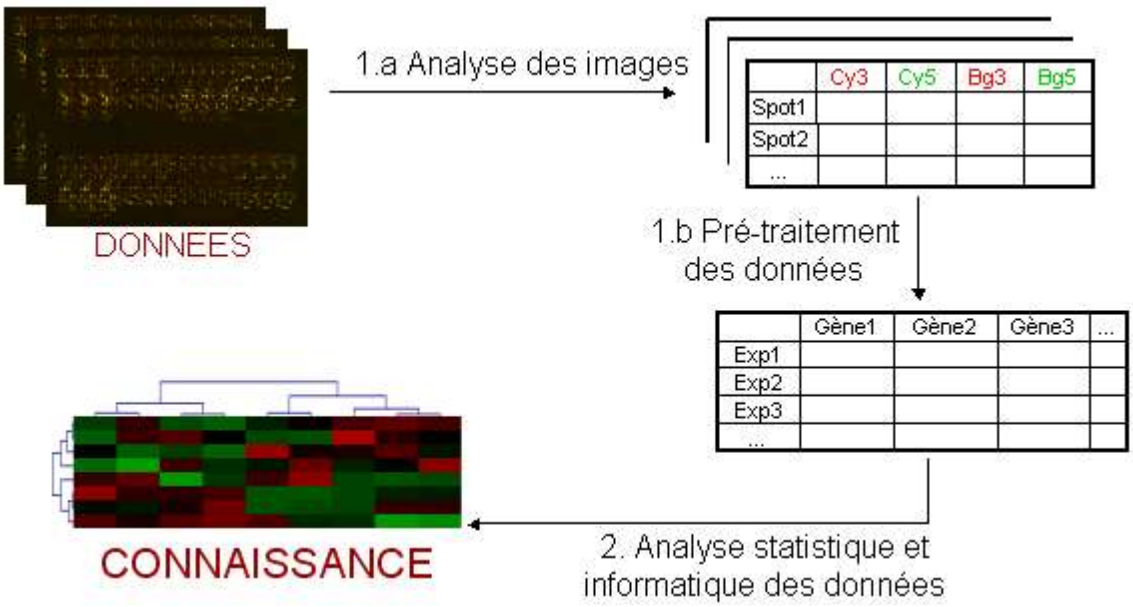

Figure 7 – Traitement des données d'expression

2.1.Prétraitement des données (Preprocessing)

Pour l'expression différentielle les données de microarray peuvent être analysés plusieurs étapes doivent être prises. Les données brutes doivent être évaluées pour assurer leur qualité intégrité. Les données brutes non traitées seront toujours soumises à une forme de variation et doit donc être prétraité pour éliminer autant de sources non désirées de variation autant que possible, pour s'assurer que les résultats sont du plus haut niveau possible de précision. Idéalement, les données analysées devraient être prétraitées en utilisant plusieurs différentes méthodes, dont les résultats devraient être comparés pour identifier méthode est du plus haut niveau de pertinence. La méthode la plus appropriée devrait ensuite être utilisé pour prétraiter les données brutes avant l'analyse d'expression différentielle.

2.1.1. Correction du bruit de fond (Background Correction)

La correction de bruit de fond est une étape de prétraitement importante pour les données de puce à ADN qui tente d'ajuster les données pour l'intensité ambiante entourant chaque caractéristique [16]. Après l'hybridation, une puce à ADN est scannée pour pouvoir générer des fichiers où les résultats de l'hybridation sont traduits numériquement (Fichiers CEL). On obtient dans ces fichiers une quantité énorme d'information. On a pour chaque gène : la moyenne des intensités de tous les pixels sur la zone correspondante au gène, la médiane de ces intensités, l'écart-type de ces intensités et le nombre de pixels dans la zone considérée.

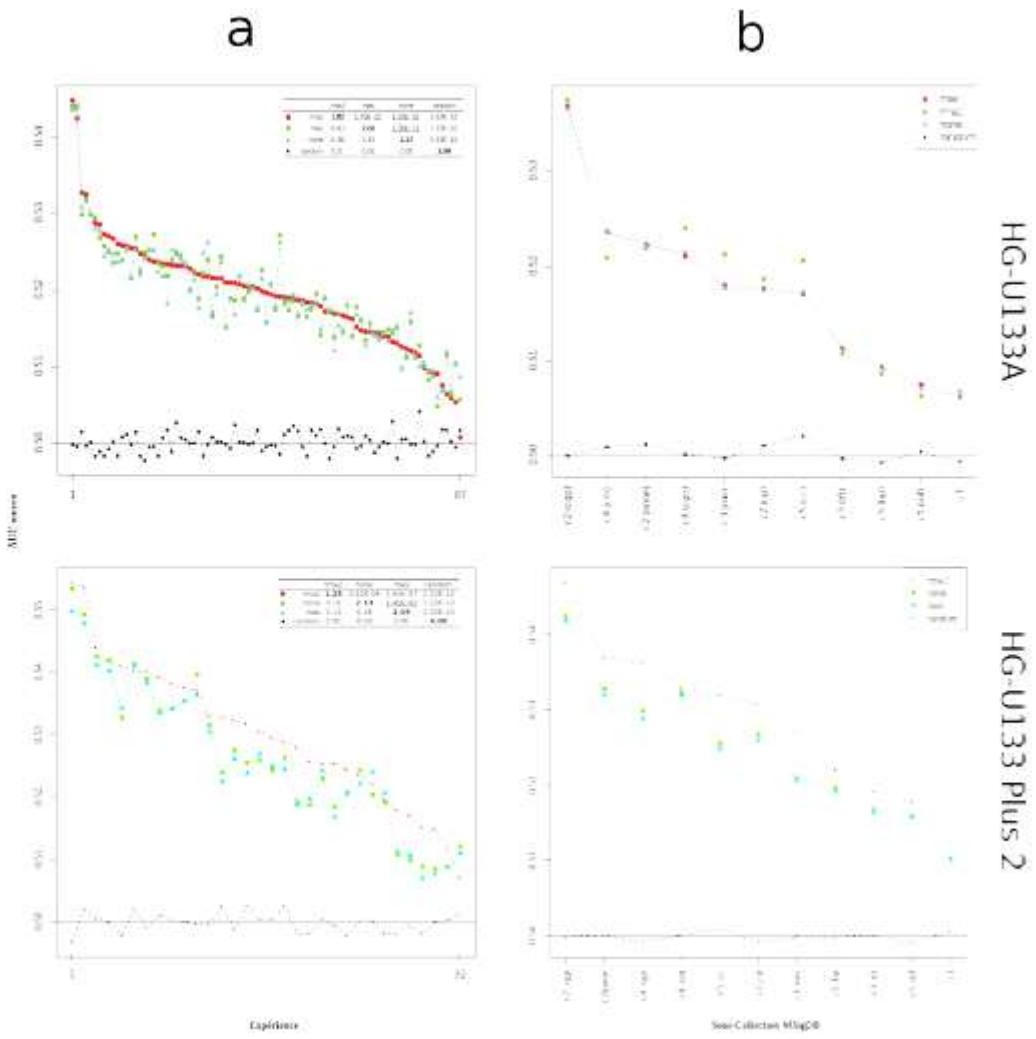

Figure 8 Exemple Correction de fond. Série d'AUC moyens par **(a)** expérience ou **(b)** sous-collection MSigDB, et ce pour les deux plateformes Affymetrix. Le tableau en encart pour **(a)** présente trois statistiques différentes sur les séries. En position i de la diagonale, la moyenne des rangs dans chaque expérience de la méthode i. Sous la diagonale, en position (i, j), la proportion des expériences pour lesquelles l'AUC moyen de la méthode i est supérieur à la méthode j. Sur la diagonale, la valeur-p d'un test de Wilcoxon pour échantillons apparies entre les méthodes i et j. Les expériences et sous-collections MSigDB en abscisse sont triées selon l'AUC moyen pour la méthode dont le rang moyen maximal.

2.1.1.1. **Correction par MAS5.0**

MAS5 est un algorithme développé par Affymetrix, cet algorithme le bruit de fond corrige à la fois les sondes PM et MM ; Les MM sont ensuite convertis en idéal discordances, où leurs valeurs sont toujours plus petites que leur PM correspondant valeurs. Rappelant que près de 30% du temps, les valeurs MM sont supérieures à leurs PM. Si MM<PM, la valeur MM reste inchangée. Un moyen robuste sur le log2 a transformé les différences entre les MP et l'idéal

déjà calculé la discordance est calculée. Les valeurs d'expression sont normalisées en réglant le trimmer moyen des signaux originaux de chaque puce à une valeur prédéfinie. Par conséquent, MAS5les données sont normalisées après la sommarisation, pas avant, comme dans beaucoup d'autres algorithmes [16].

2.1.1.2. **Correction par RMA**

Lors du développement de RMA [13], les auteurs ont jugé que les sondes MM posaient plus de problème qu'elles n'en solutionnaient et ont proposé de ne plus utiliser que les valeurs des sondes PM. Cette correction se fait en utilisant un modèle basé sur la distribution empirique des intensités de sondes. Le modèle observé (**Y**) est une somme de la composante « bruit »(**B**) et de la composante «signal » (**S**)

$$\boldsymbol{Y = B + S}$$

Avec **B** une distribution proche d'une distribution normale et **S** une composante exponentielle. Pour éviter la possibilité de valeurs d'expression négative, il est nécessaire de tronquer la distribution normale à zéro.

2.1.1.3. **Correction par GCRMA**

GCRMA est en grande partie basé sur RMA et ne diffère en réalité qu'en étape de correction de bruit de fond où il utilise des informations de séquence de sonde pour aider à estimer la Contexte. Cela conduit à une meilleure précision dans les changements de plis, mais au détriment de précision légèrement inférieure [16].

2.1.2. Normalisation :

Le but de la ***Normalisation*** est d'ajuster les données pour les variations, par opposition aux différences biologiques entre les échantillons. Là sera toujours un petit décalage entre les processus d'hybridation pour chaque tableau et ces variations ont tendance à conduire à des différences d'échelle entre l'ensemble niveaux d'intensité de fluorescence de diverse matrice. Par exemple la quantité d'ARN dans un échantillon, la durée pendant laquelle un échantillon s'hybride ou le volume d'un échantillon peut tous introduire une variance significative. Même physique subtile différences entre les tableaux ou entre les scanners utilisés pour lire les tableaux peuvent avoir un effet. En termes simples, la normalisation garantit que lorsque l'on compare les niveaux d'expression de différents tableaux, que nous sommes, autant que possible, en comparant comme avec comme. Des études ont montré que la méthode de normalisation utilisée a une différence significative sur les niveaux d'expression différentielle finale, il est donc essentiel de choisir une méthode appropriée [16].

2.1.2.1. Normalisation des puces Agilent :

Dans le cas des puces Agilent à ADNc, la normalisation [13] consiste à ajuster l'intensité globale des images acquises sur chacun des deux canaux vert et rouge, de manière à corriger des biais techniques systématiques qui tendent à déséquilibrer le signal de l'un des canaux par rapport à l'autre. La méthode a plus utilisée est la normalisation Lowess. La méthode Lowess est semblable aux méthodes de régression, mais diffère sur 3 aspects : La régression est effectuée sur les valeurs appelées MA (Mean Average) pour la sonde k et pour une régression entre les puces i et j on a :

Mk = log2 (xki=xkj) etAk = log2 (xki _ xkj)

Deuxièmement, la régression Lowess est locale selon la procédure Lowess, ce qui signifie que la courbe de régression n'est pas contrainte de suivre une forme fonctionnelle particulière. Troisièmement, cette méthode de normalisation est cyclique car plutôt que d'utiliser une des puces comme référence, l'ajustement total provient de la contribution de chacune des puces.

2.1.2.2. Normalisation des puces Affymetrix :

Dans le cas des puces à oligonucléotides [17], comme les puces Affymetrix, la normalisation est réalisée entre des répétitions de lames ou l'ensemble des lames d'une ou de plusieurs expériences. On parle souvent de normalization between-array. La normalisation la plus utilisée est la normalisation des quantiles (quantiles normalization).

Pour cela, il existe une méthode complète dite de centralisation permettant à la fois de normaliser et de calibrer les données de façon à permettre les comparaisons inter-lames. Cette méthode non paramétrique appelée aussi" normalisation des quantiles " suppose que la distribution de l'abondance des gènes est presque la même dans tous les échantillons. L'algorithme comporte plusieurs étapes :

1. On trie les gènes par colonnes selon leurs intensités.
2. On calcule la moyenne de chaque ligne
3. On remplace les valeurs de chaque élément ligne par la moyenne correspondante.
4. On redistribue les valeurs nouvelles selon l'ordre d'origine des intensités.

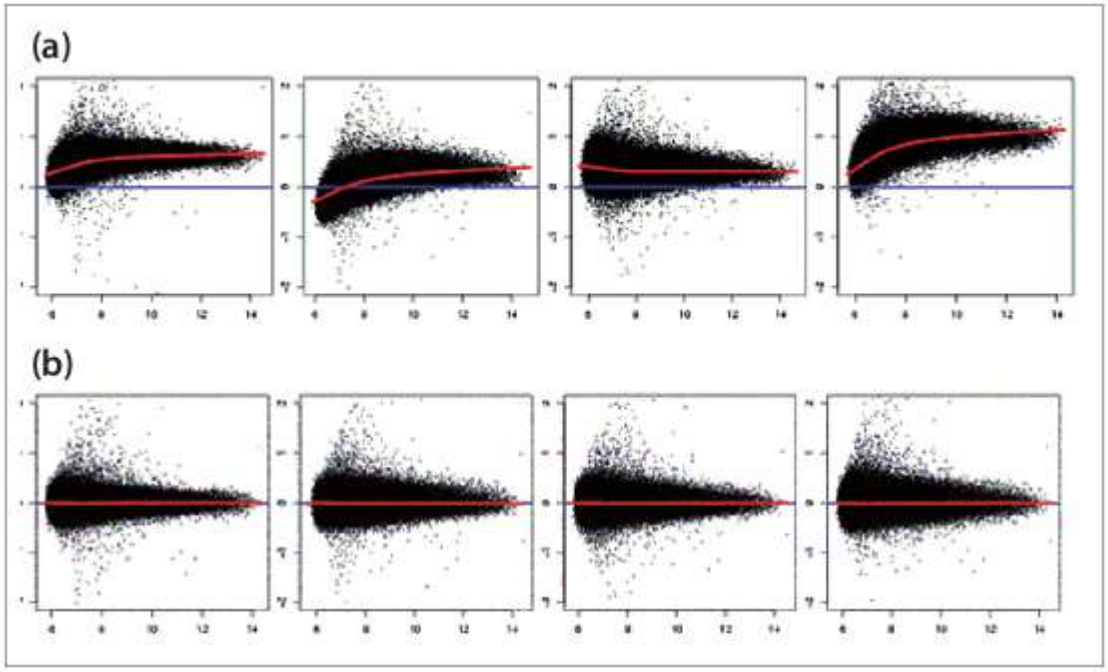

Figure 9 Nuage de points avant et après normalisation sur 4 puces Affymetrix.(a) Nuage de points avant normalisation (b) Nuage de points après normalisation.

2.1.3. Sommarisation :

C'est une étape propre à toute plateforme pour laquelle un même transcrit est sondé par plusieurs sondes que l'on doit résumer en une seule valeur d'expression, et consiste à obtenir une mesure qui représente le résumé collectif des différentes lectures pour le même gène.

2.1.3.1. ***MAS5 :***

Cette méthode utilise l'estimateur one-step Tukeybiweight pour calculer une moyenne robuste après correction des intensités des sondes Mismatch.

L'estimateur Tukeybiweight est calculé à l'échelle logarithmique. [13]

2.1.3.2. ***Avgdiff :***

Cette méthode de sommarisation ne consiste qu'à calculer la moyenne des sondes PM pour chaque probeset de chaque puce. Noter que la version employée ici n'effectue pas de transformation logarithmique préalable.

2.1.3.3. ***Farms :***

La sommarisation FARMS (*Factor Analysis for Robust Microarray Summarization*) est basée sur le modèle suivant pour les intensités PM :

$$\log(PM_{ij}) = z_i(\sigma + \tau_j) + \mu + \gamma_j + \epsilon_{ij}.$$

Ou z_i représente le score-z de la quantité réelle d'ARN dans l'échantillon hybride sur la puce i alors que cette quantité réelle est même de moyenne μ et variance σ. S'additionnent l'effet de

la sonde $\boldsymbol{j}$ $\boldsymbol{\gamma_j}$ sur la moyenne $\boldsymbol{\tau_j}$et sur la variance. Une procédure complexe d'analyse factorielle optimise ensuite les paramètres du modèle en usant d'une méthode a posteriori de maximum Bayésien, sous l'hypothèse d'erreur de mesure gaussienne :

$$\boldsymbol{x} = \boldsymbol{\lambda_z} + \boldsymbol{\varepsilon}$$

Ce modèle suppose qu'un facteur caché $\boldsymbol{z}$ est sous-jacent aux intensités des sondes $\boldsymbol{x}$ dont la matrice $\boldsymbol{\lambda}$ décrit la structure de corrélation. [18]

2.1.4. PM Correction :

2.1.4.1. ***MAS5 :***

Les versions initiales de l'algorithme de correction PM soustrayaient simplement la valeur de la sonde MM à celle de la sonde PM, ce qui mené pour certaines sondes a des valeurs négatives. Son successeur, la procédure *IdealMismatch* cherche à corriger le problème des valeurs négatives en calculant une valeur ajustée de MM, nommée IM. Soit $\boldsymbol{PM_{ij}}$ et $\boldsymbol{MM_{ij}}$ les intensités PM et MM pour la sonde $\boldsymbol{j}$ du probeset $\boldsymbol{i}$. Si $\boldsymbol{PM_{ij}} \leq \boldsymbol{MM_{ij}}$, alors MM est considéré comme une bonne estime du signal non spécifique et $\boldsymbol{IM_{ij}} = \boldsymbol{MM_{ij}}$. Dans le cas contraire, Affymetrix propose de se baser sur les rapports *PM/MM* des autres sondes du même probeset pour estimer une valeur raisonnable du signal non spécifique IM : $\boldsymbol{IM_{ij}} = \frac{\boldsymbol{PMij}}{\boldsymbol{2^{SB_i}}}$ où$\boldsymbol{SB}_i = T_{b_i}(\log_2(PM_{ij}) - \log_2(MM_{ij}))$, soit une moyenne robuste des rapports (écarts en échelle log). Affymetrix ajoute une condition supplémentaire aux valeurs de $\boldsymbol{SB}_i$ qui ne sera pas discutee ici. Le lecteur intéressé a toute cette procédure plutôt ad hoc peut se référer à Affymetrix. [18]

2.1.4.2. ***PMonly :***

Comme le nom l'indique, cette méthode correspond à n'effectuer aucune correction. Les valeurs brutes des sondes PM sont employées directement. [18]

2.1.5. transformation logarithmique

La transformation logarithmique n'apporte pas nécessairement l'effet stabilisateur recherche sur les données de microarray, particulièrement aux basses intensités ou cette dernière exagère la variance et mené à des foldChange dont la significativité n'est plus comparable à ceux des hautes intensités. En fait la transformation logarithmique est stabilisatrice seulement si la variance augmente linéairement avec la moyenne, ce qui pour les microarray n'est pas nécessairement le cas aux basses intensités.

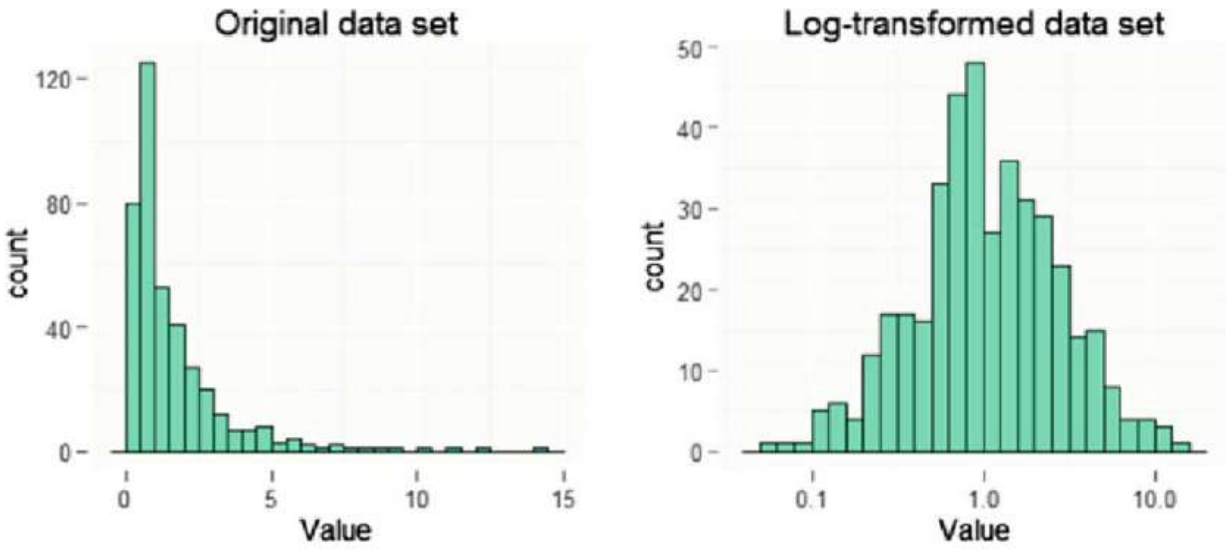

Figure 10 Exemple de l'effet d'une transformation logarithmique sur la distribution d'une base de données.

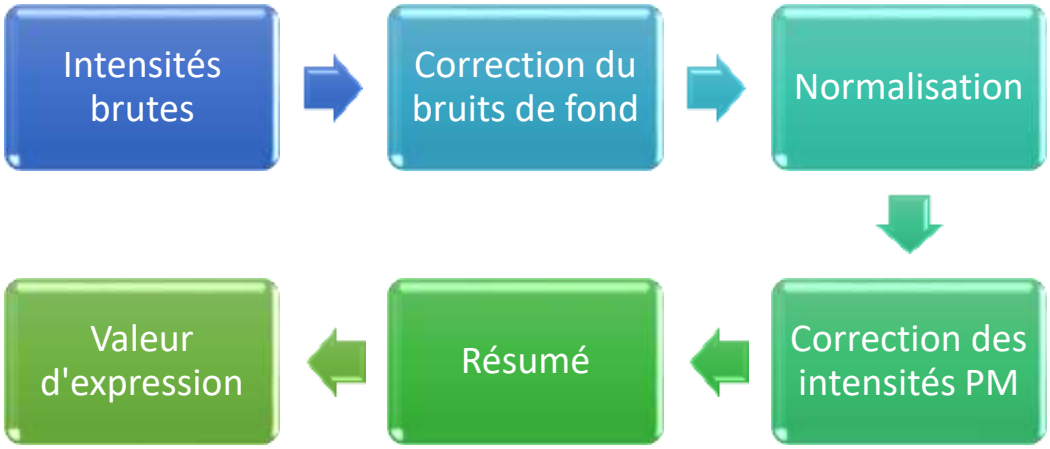

Figure 11 Diagramme des différentes étapes du prétraitement pour une puce à oligonucléotides.

2.2. Matrice d'expression des gènes

Après les différents prétraitements effectués sur les données des puces, les données d'expression peuvent être présentées sous la forme d'une matrice à M colonnes représentant les échantillons et à N lignes représentant les ensembles de sondes ou gènes. Chaque cellule du tableau représente le niveau d'expression M_{ij} d'un gène (ligne ***i***) dans un échantillon (colonne ***j***).

Gène id	Echantillon $_1$	Echantillon $_2$	……	Echantillon $_m$
Gène $_1$				
Gène $_2$				
. . . .				
Gène $_n$	M_{n1}	M_{n2}	……	M_{mn}

Table 1 Représentation standard des données transcriptomiques via une matrice d'expression des gènes.

Certaines méthodes de prétraitement associent les valeurs d'expressions des gènes aux indicateurs de mesure de fiabilité (Ex. détection calls), et ces indicateurs peuvent être directement utilisés pour éliminer les ensembles de sondes non fiables. En l'absence de ces indicateurs, le filtrage en fonction du niveau d'expression devra être utilisé.

2.2.1. Filtrage

Les méthodes de filtrage peuvent être utilisées pour réduire le nombre de tests et donc augmenter la puissance pour détecter les vraies différences. Une méthode de filtrage idéale permettrait d'éliminer les tests qui sont vraiment Nul (correspondant à des gènes exprimés de façon égale), tout en laissant ces tests correspondant à des gènes qui sont vraiment différentiellement exprimés. Plusieurs méthodes de filtrage ont été suggérées, y compris le filtrage par *foldChange*, par *le niveau d'expression moyen dans une classe…*

2.2.1.1. Filtrage basé sur Fold change

Pour la méthode du Fold Change [20], nous testons 2 classes (k = 2). Le principe est le suivant : le gène a_j est conservé si $moy_1^j / moy_2^j > \varepsilon$ ou $moy_1^j / moy_2^j < \varepsilon$.

On parle alors de ***ε***-Fold Change et une valeur ε = 2 est couramment utilisée comme dans la *figure 1* où en rouge sont donnés les gènes 2 fois plus exprimés dans la classe c_1 et en vert les gènes 2 fois plus exprimés dans la classe c_2.

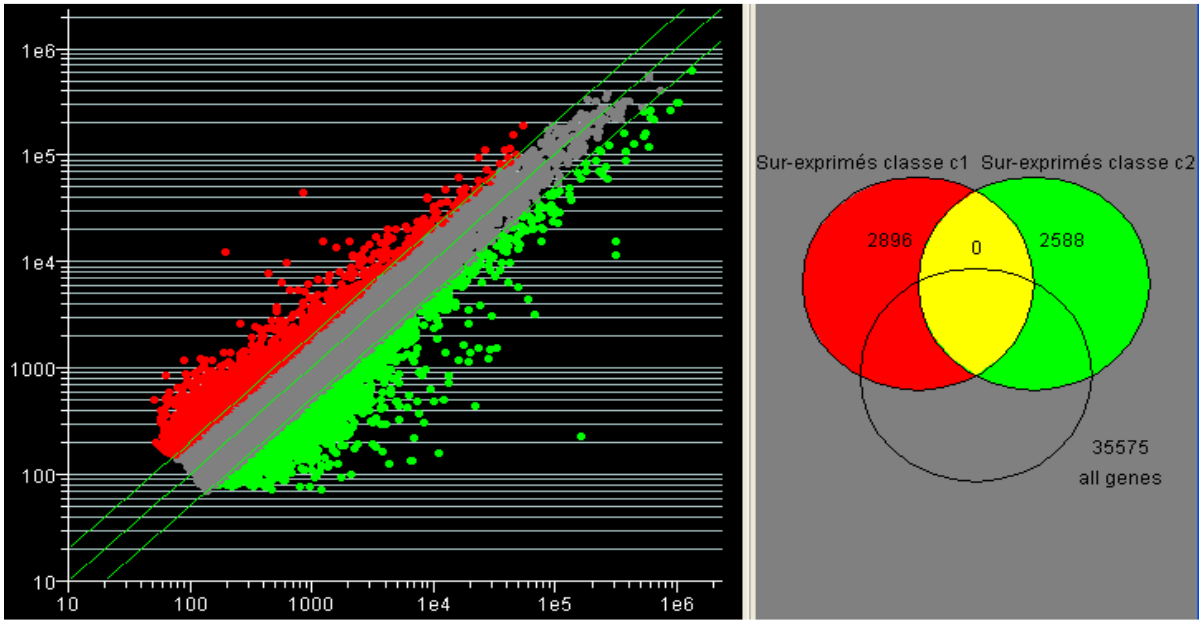

Figure 12 Méthode de Fold Change.

Cette méthode est très simple mais n'indique pas si la différence des moyennes est statistiquement significative. Pour tester cela, il convient d'utiliser des tests d'hypothèse présentés dans le paragraphe suivant. La méthode du Fold Change peut toutefois être utilisée comme un filtre préliminaire.

2.2.1.2. ***Filtrage basé sur le niveau d'expression moyen dans une classe***

Généralement, certains bruits expérimentaux plus ou moins de niveau arbitraire sont utilisés pour définir un seuil de filtrage. Ainsi les gènes n'ayant pas un niveau d'expression moyen supérieur à ce seuil dans certaines classes seront éliminés Filtrage basé sur le niveau d'expression maximal. Les gènes dont toutes les valeurs d'expressions, dans tous les échantillons, sont en dessous du seuil de bruit expérimental sont éliminés Filtrage basé sur l'amplitude des valeurs d'expression Les gènes avec une amplitude d'expression (max-min) inférieur au seuil de bruit expérimental prédéfini, seront retirés.

3. Analyse relative

L'analyse relative, qui est effectuée sur les données après le prétraitement et filtrage de ceux-ci est utilisée par les chercheurs pour se familiariser avec les données, pour apprendre

d'avantage sur certaines structures et pour vérifier si les données sont adaptées à l'étude envisagée.

3.1. Test Statique

La mise en évidence [21] des gènes « différentiels » par la simple analyse descriptive des amplitudes de variations d'expression (*Fold change)* est insuffisante. Des approches statistiques sont nécessaires afin d'estimer et de distinguer la variabilité intra et intergroupes. De nombreux tests statistiques ont ainsi été proposés allant du test *t* de Welche (lorsque les groupes ont des variances inégales) aux approches bayésiennes (Efron *et al.* 2001 ; Lönnstedt et Speed, 2002) en passant par les analyses de variance (Kerr *et al.* 2000).

L'application des tests dépend de plusieurs paramètres. Tout d'abord, il faut savoir si les données analysées sont indépendantes (Golub*et al.* 1999), appariées (Pérou *et al.* 2000) ou multi-variées (Khan *et al.* 2001). Ensuite, le mode de distribution des données doit être évalué : distribution gaussienne ou pas. En effet, les tests paramétriques tels que les tests *t* supposent une distribution normale des données. A l'inverse, les tests non paramétriques sont moins sensibles au mode de distribution des données et aux valeurs atypiques. Enfin, la variance intra et intergroupe doit être estimée. Les tests paramétriques sont moins sensibles à un écart à la normalité qu'à une mauvaise estimation de l'homogénéité des variances. Il est donc admis, sous condition d'un bon estimateur de la variance, de réaliser un test paramétrique même si le mode de distribution des données s'écarte légèrement de la normalité.

t-test :

Le test du student [22] ou *t-test* méthode permettant d'identifier des gènes exprimés différentiellement est l'utilisation du test de Student ou test t, sur les niveaux des intensités d'hybridation.

$$t = \frac{m_1 - m_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

$m_1$et m_2 : moyennes des intensités des signaux d'un gène donné dans chaque condition 1 et 2.

S_1^2et S_2^2: variances des intensités des signaux d'un gène donné dans chaque condition 1 et 2.

$n_1$et n_2 : nombre d'analyses pour les conditions 1 et 2 correspondant au nombre de réseaux (variabilité technique) ou au nombre d'individus (variabilité biologique), analysés par condition.

SAM :

La méthode SAM ou test *S* [20] est basée sur un test t gènes spécifique. Elle utilise une méthode de permutation consistant, pour un gène donné, à permuter entre les échantillons les mesures d'intensité des réplicas techniques. La statistique de t est alors calculée pour les données vraies (c'est-à-dire non permutées) et les données permutées. Les gènes déclarés différentiellement exprimés de façon significative avec les données permutées sont donc des faux-positifs. Cette méthode permet donc d'estimer le taux de fausse découverte (FDR pour "**False Discovery Rate**") (Tusher et al. 2001). Le principal avantage de cette approche est de calculer le risque d'erreur non plus par rapport à l'ensemble des gènes étudiés mais par rapport à ceux déclarés significativement différents. Ainsi on pourra parfaitement accepter de prendre un risque plus important. Par exemple, fixer le FDR à 10% n'induit qu'un seul faux-positif potentiel lorsque 10 gènes sont déclarés significativement différents, ce qui peut s'avérer tout à fait acceptable.

$$d = \frac{\bar{x}_1 - \bar{x}_2}{S_p \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} + S_0}$$

Il diffère du test statistique t pour les variances égales, par l'ajout de la constante S_0 pour minimiser la variation de*d*. Il permet de ne pas favoriser les gènes qui auraient une petite variance et du même coup une valeur de de grande.

4. Générations des gènes différentiellement exprimés

L'analyse de l'expression différentielle consiste à identifier quels gènes voient leur niveau d'expression varier *entre différentes conditions biologiques* (*figure 5*).

Plusieurs expériences de puces à ADN permettent de constituer une série d'expériences, permettant ainsi de suivre l'évolution de la quantité des transcrits en fonction de diverses conditions expérimentales.

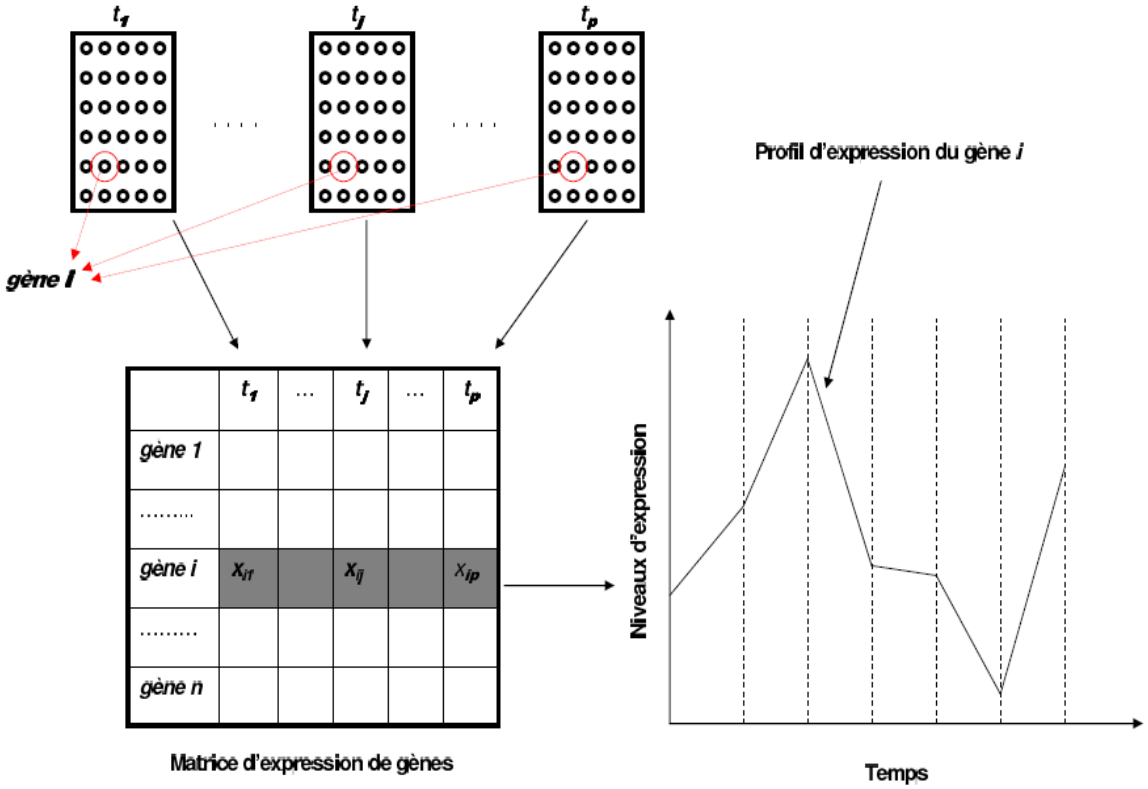

Figure 13 Variation du profil d'expression de gènes en fonction des conditions dans le temps.

A chaque gène est associé un profil d'expression, la recherche par profil d'expression consiste donc à trier ces profils en fonction de leur ressemblance et ceux à l'aide de méthodes statistiques. La recherche par profil se fait suivant 2 étapes principales :

- Quantification du degré de ressemblance entre les profils pris deux à deux (calcul d'une distance)
- Tri des profils en fonction de la distance qui les sépare.

La définition d'un critère de ressemblance est nécessaire pour permettre l'identification des gènes dont les profils d'expression sont similaires. Le calcul de la distance entre deux profils peut être effectué à l'aide de la distance euclidienne ou de la corrélation de Pearson. La distance euclidienne est dépendante de l'échelle des profils tandis que la mesure d'une corrélation permet de définir les profils corrélés et anti-corrélés.

5. Les méthodes génèrent les gènes différentiellement exprimés

Les méthodes utilisées pour évaluer un sous-ensemble de caractéristiques dans les algorithmes de sélection peuvent être classées en trois catégories principales : "Wrapper", "Embedded" et "Filter".

5.1. Approche *'wrapper'*

Les wrappers ont été introduits par John et al. en 1994. Leurs principe est de générer des sous-ensembles candidats et de les évaluer grâce à un algorithme de classification. Cette évaluation est faite par un calcul d'un score, par exemple un score d'un ensemble sera un compromis entre le nombre de variables éliminées et le taux de réussite de la classification sur un fichier de test. L'appel de l'algorithme de classification est fait plusieurs fois à chaque évaluation (c'est-à-dire à chaque sélection d'une variable, nous calculons le taux de classification pour juger la pertinence d'une caractéristique) car un mécanisme de validation croisée est fréquemment utilisé. Le principe de wrappers est de générer un sous ensemble bien adaptés à l'algorithme de classification. Les taux de reconnaissance sont élevés car la sélection prend en compte le biais intrinsèque de l'algorithme de classification. Un autre avantage est sa simplicité conceptuelle ; nous n'avons pas besoin de comprendre comment l'induction est affectée par la sélection des variables, il suffit de générer et de tester.

Cependant, trois raisons font que les wrappers ne constituent pas une solution parfaite. D'abord, ils n'apportent pas vraiment de justification théorique à la sélection et ils ne nous permettent pas de comprendre les relations de dépendances conditionnelles qu'il peut y avoir entre les variables. D'autre part la procédure de sélection est spécifique à un algorithme de classification particulier et les sous-ensembles trouvés ne sont pas forcément valides si nous changeons de méthode d'induction. Finalement, c'est l'inconvénient principale de la méthode, les calculs devient de plus en plus très longs, voir irréalisables lorsque le nombre de variables est très grand.

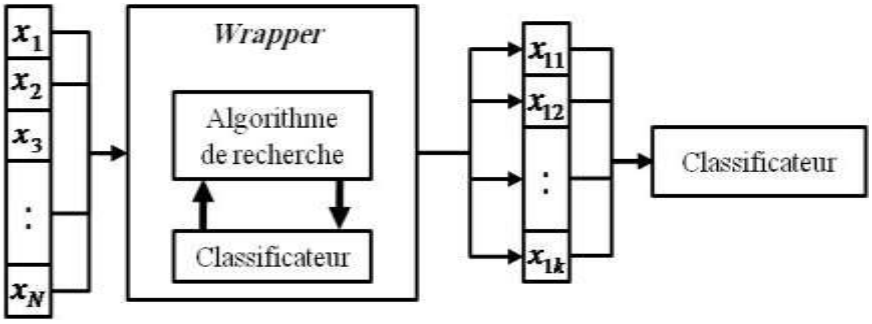

Figure 14 La procédure du modèle "wrapper"

5.2.Approche *'Embedded'*

Les méthodes Embedded intègrent directement la sélection dans le processus de l'apprentissage (f*igure 6*), les arbres de décision sont l'illustration la plus emblématique. Mais, en réalité, nous classons dans ce groupe toutes techniques qui évaluent l'importance d'une variable en cohérence avec le critère utilisé pour évaluer la pertinence globale du modèle.

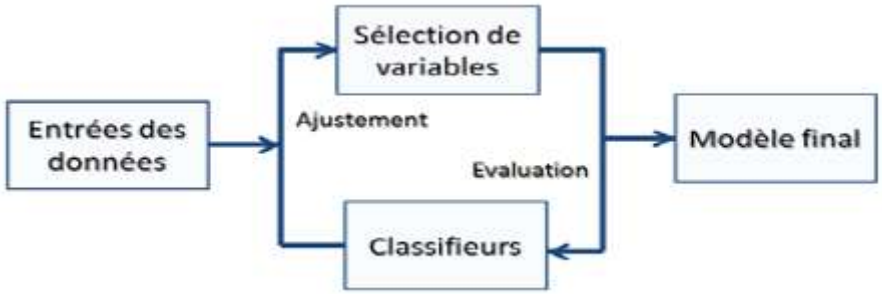

Figure 15 Procédure de l'approche Embedded.

5.3. Approche *'Filter'*

Le modèle "filter" a été le premier utilisé pour la sélection de caractéristiques. Dans celui-ci, le critère d'évaluation utilisé évalue la pertinence d'une caractéristique selon des mesures qui reposent sur les propriétés des données d'apprentissage. Cette méthode est considérée, davantage comme une étape de prétraitement (filtrage) avant la phase d'apprentissage. En d'autres termes, l'évaluation se fait généralement indépendamment d'un classificateur (John et al. [1994]). Les méthodes qui se basent sur ce modèle pour l'évaluation des caractéristiques, utilisent souvent une approche heuristique comme stratégie de recherche. La procédure du modèle "filter" est illustrée par la *figure 7*.

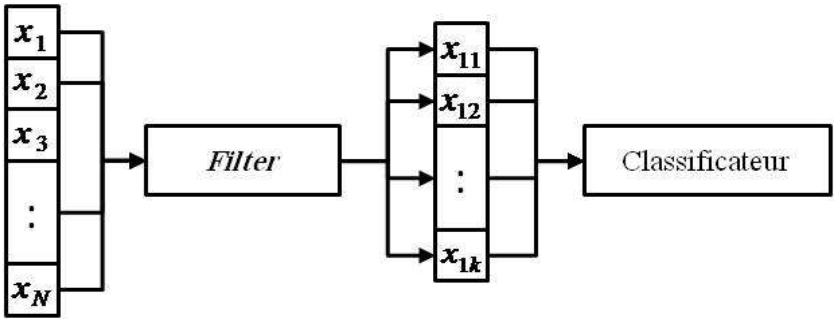

Figure 16 Procédure de l'approche filter.

Le simple test-t gagne par exemple beaucoup d'un filtre sur la variance alors que par exemple, *ebayes*, qui utilise la variance globale dans le calcul de régularisation, risque théoriquement d'être affectée par la procédure de filtrage. Huber et al. Recommandent soit d'effectuer un test-t suivit d'un filtre non spécifique ou d'employer directement une statistique-t régularisée comme *ebayes*

5.3.1. Réseau Bayésien

Les réseaux bayésiens [23] sont de graphes acycliques orientés où les variables sont représentées par des nœuds et les relations de dépendance ou de corrélation entre les variables sont représentées par des arcs directionnels.

Chaque variable est représentée par un tableau de probabilités qui sont déterminées en utilisant le théorème de Bayes

$$H(p|E) = \frac{P(E|H) * P(H)}{P(E)}$$

Si l'hypothèse H est basée sur des données observées et E est l'évidence (la preuve), alors P(H) est une probabilité a priori de l'hypothèse, en fait le degré initial de confiance dans l'hypothèse. P(E|H) est la vraisemblance des données observées, donc la mesure dans laquelle l'évidence a été observée quand l'hypothèse était vraie. P(H|E) est la probabilité a posteriori de l'hypothèse étant donné l'évidence ;

Le théorème de Bayes offre l'adaptabilité et la flexibilité qui permettent à l'utilisateur de réviser et changer les estimations et les prédictions si de nouvelles données pertinentes sont accueillies.

Les réseaux bayésiens sont la représentation de la dépendance entre un ensemble de variables. Un arc entre deux variables, disons de V à, W, dénote qu'il existe une relation de dépendance directe du V sur W et dans ce cas V est un parent de W. Les tableaux de probabilités associés aux nœuds sont la distribution de probabilités du nœud en considérant tous les parents de celui-là.

Définition :

Dawn Holmes et Lakshmi Jain définissent formellement le réseau bayésien :

Soit S un ensemble fini de sommets et ***A*** un ensemble des arcs entre ces sommets sans boucles de rétroaction, les sommets et les arcs forment un graph acyclique orienté $G = \{S, A\}$. Un ensemble d'événements est représenté par les sommets de ***G*** et donc aussi par ***S***. Soit chaque événement ait un ensemble fini de résultats mutuellement exclusifs, ou $\boldsymbol{E_i}$ est une variable qui peut prendre n'importe quelles issues e_i^j de l'évènement ***i*** ou ***j***=1,…, n. Soit ***P*** une distribution de probabilités sur les combinaisons d'événements. Soit ***C*** l'ensemble des contraintes suivantes :

- une distribution de probabilités somme à l'unité.
- pour chaque événement ***i*** et un ensemble de parents $\boldsymbol{M_i}$ il y a probabilités conditionnelles associées $P(E_i \cap_{j \in M_i} E_j)$ .pour chaque issue possible qui peut être assignée à E_i et E_j.

- ❖ Ces relations d'indépendance impliquées par d-séparation dans le graphe acyclique orienté. Alors N = (G, P, C) est un réseau causal si P doit satisfaire C.

Avantages et Inconvénients des réseaux bayésiens :

Avantage	Inconvénients
➢ Peuvent utiliser de données partielles ou incomplètes ➢ Permettent l'étude des relations causales et l'influence directe d'une variable sur l'autre ; ➢ Combinent l'estimation d'experts et les données statistiques pour mieux évaluer la casualité permettant de mettre ensemble toutes les sources de données disponibles, subjectives aux objectives ; ➢ Couvrent le raisonnement cause à effet de façon transparente et documentée ➢ Permettent l'analyse de type « what if » ; ➢ Permettent l'acquisition, la représentation et l'utilisation de connaissance ; ➢ Dans les dernières années, il y a une abondance d'outils et de logiciels qui permettent de saisir et traiter les réseaux bayésiens. Quelques exemples : Bayes server, Hugin, BayesNet, MSBNx.	➢ Complexité élevée d'intégration dans un cadre basé seulement sur l'opinion des experts ; ➢ Graphes et les algorithmes de calcul peuvent être lourds dans les réseaux complexes ; Difficultés à travailler avec les variables continues

Construction :

Plusieurs étapes sont à considérer dans la construction d'un réseau bayésien :

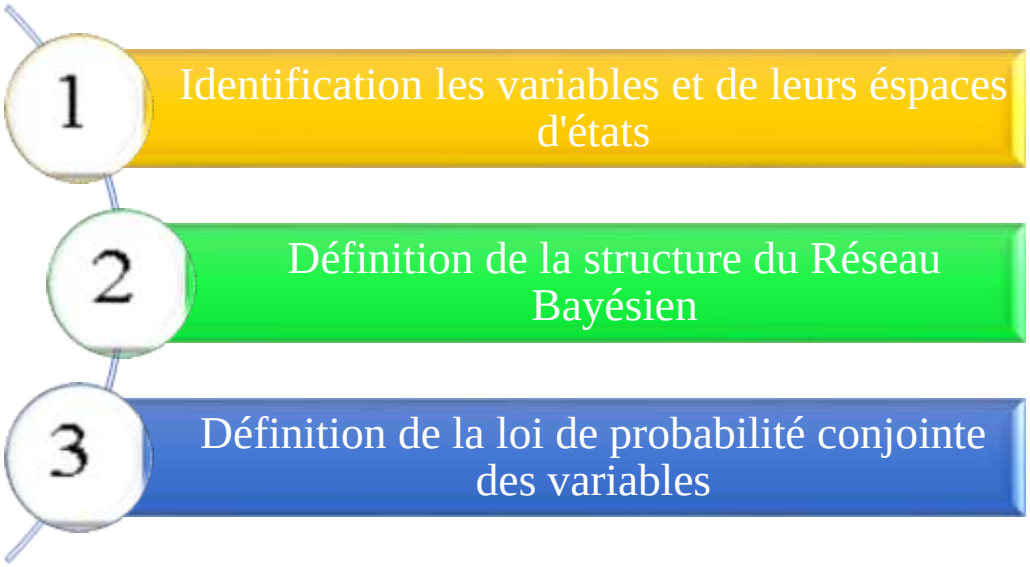

Figure 17 Étapes de construction d'un réseau bayésien

La première étape est l'identification de variables et pour chaque variable l'ensemble de ses valeurs possibles, pour cette étape l'intervention des experts du système est toujours nécessaire.

La deuxième étape est la définition de la structure du réseau bayésien, trouver les liens d'influence entre les variables tout en s'assurant qu'il n'y a pas de boucle ou cycle. Patrick Naim mentionne que : « *quelles que soient les dépendances stochastiques entre des variables, il existe toujours une représentation par réseau bayésien* »

La dernière étape vise la création des tableaux de probabilités pour les variables, soit de variables sans parentes pour lesquelles des probabilités marginales doivent être définies, soit de variables qui ont de variables partantes et, dans ce cas, de probabilités conditionnelles sont définies.

Conclusion :

Des approches statistiques sont nécessaires à la détermination des gènes différentiellement exprimés. Compte tenu de l'aspect bruité des données et des problèmes de dimensions des matrices, les tests statistiques « classiques » ne sont pas adaptés à l'analyse des données de puces à ADN

- il est nécessaire de corriger le résultat des tests statistiques pour tenir compte des comparaisons multiples,
- les tests non paramétriques sont plus robustes que les approches paramétriques face aux bruits expérimentaux. Les méthodes par permutations sont, actuellement, les approches les plus puissantes.

L'analyse de données multifactorielles peut nécessiter des approches bayésiennes ou l'application de modèles linéaires généralisés.

La sélection des gènes d'intérêts est une étape clef du traitement des données de puces à ADN. Elle est notamment préliminaire à toute analyse par des techniques de classification, supervisée ou non, ainsi qu'à l'inférence de réseaux de régulation génétique. Les méthodes wrapper sont intéressantes selon plusieurs points de vue. D'une part, elles permettent de considérer l'ensemble de l'information présente dans le jeu de données pour procéder à la sélection des gènes. D'autre part, elles peuvent présenter des caractéristiques comparables aux méthodes filter.

Chapitre 3 :

Conception Et Implémentation

1. Introduction

Un des principaux objectifs des puces à ADN est de déterminer quels sont les gènes différentiellement exprimés entre différentes conditions expérimentales et déterminer quels gènes caractérisent un état particulier.

Dans ce chapitre nous présentons la conception de notre système en commençant par sa conception générale puis sa conception détaillée en spécifiant les différents composants de notre système et précisant son fonctionnement.

2. Conception du système

2.1. Architecture et fonctionnement de notre système

Notre système consiste à découper le processus d'analyse et de visualisation en plusieurs étapes. Les différentes étapes seront le prétraitement qui prendra en entrée un dossier contenant les fichiers CEL que l'on souhaite traiter et le fichier CDF associé (aux caractéristiques de fichier .CEL) décrivant le type de puce utilisé pour l'expérimentation. Cette étape donnera en sortie une matrice gène/expression.

2.2. Conception globale

Notre objectif est de concevoir un système qui prendre en entrée une base de données généré par une expérience de puce à ADN et de générer comme résultat les gènes différentiellement exprimés après l'application des plusieurs étapes d'analyse de données biopuces.

Le système est vu à travers ses entrées et ses sorties qui sont spécifiées dans le schéma suivant :

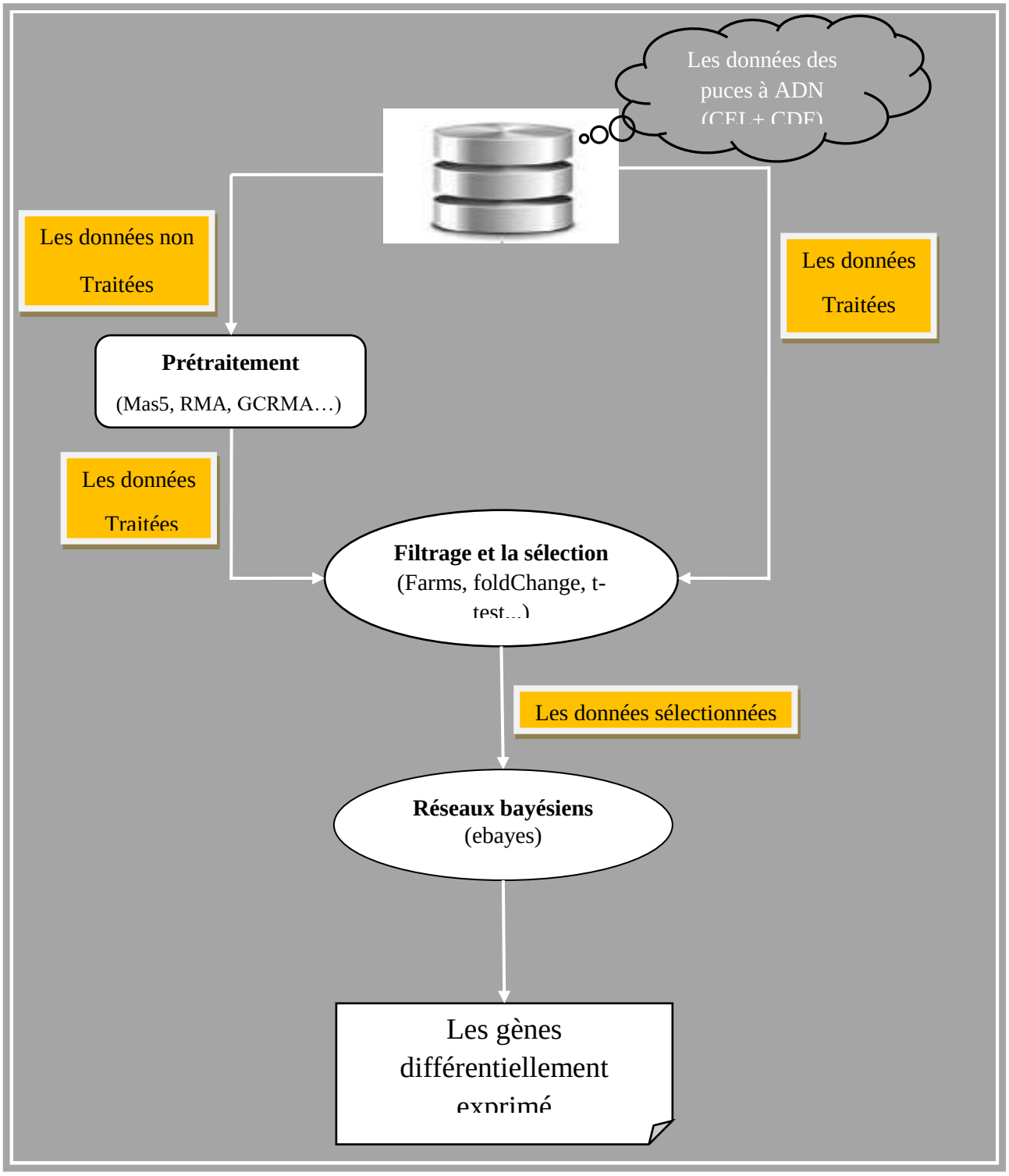

Figure 18 Architecture générale de notre système.

Acquisition des données

Notre base de données compose de :

- Fichier Celles : fichier texte (ou binaire) contenant des intensités de sonde brutes pour une seule puce créée par le logiciel MAS (GCOS).
- Fichier CDF : Le fichier CDF est un fichier de description de puces Affymetrix. Ce fichier de bibliothèque contient des informations sur les sondes qui appartiennent à quel ensemble de sondes.

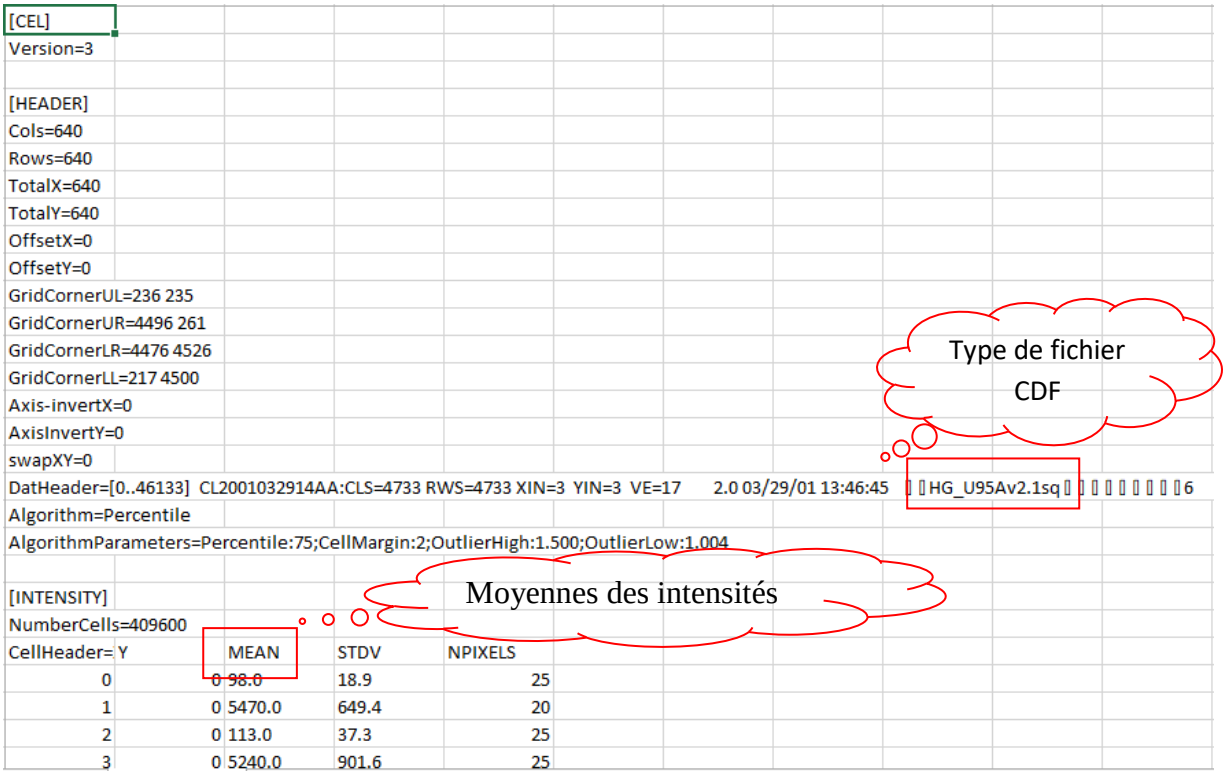

Figure 19 Les données de fichier CEL et le fichier CDF.

- **Description de la base :**

Nom : cancer prostate

Type de puce : HG-U95Av2

Nombre d'échantillons : 102

Type-val : Affymetrix4.0

Mots-clés : prostate|cancer

BD-source : MIT

Lien : Cancer Program Legacy Publication Resources

Nombre des gènes : 12625

Niveau-sécurité : public

Nom-échantillon:

N27__normal|N01__normal|N02__normal|N03__normal|N04__normal|N05__normal|
N06__normal|N10__normal|N11__normal|N13__normal|N14__normal|N15__normal|
N16__normal|N17__normal|N18__normal|N19__normal|N20__normal|N21__normal|
N23__normal|N24__normal|N25__normal|N26__normal|N28__normal|N30__normal|
N31__normal|N32__normal|N33__normal|N34__normal|N35__normal|N36__normal|
N37__normal|N38__normal|N39__normal|N40__normal|N41__normal|N42__normal|
N43__normal|N44__normal|N45__normal|N46__normal|N47__normal|N50__normal|
N53__normal|N54__normal|N55__normal|N58__normal|N59__normal|N60__normal|
N61__normal|N62__normal|T38__tumor|T01__tumor|T02__tumor|T03__tumor|T04_
_tumor|T05__tumor|T06__tumor|T10__tumor|T11__tumor|T13__tumor|T14__tumor|
T15__tumor|T16__tumor|T17__tumor|T18__tumor|T19__tumor|T20__tumor|T21__tu
mor|T22__tumor|23__tumor|T24__tumor|T25__tumor|T26__tumor|T27__tumor|T28_
_tumor|T29__tumor|T30__tumor|T31__tumor|T32__tumor|T33__tumor|T34__tumor|
T36__tumor|T37__tumor|T39__tumor|40__tumor|T41__tumor|T42__tumor|T43__tum
or|T45__tumor|T46__tumor|T47__tumor|T49__tumor|T50__tumor|T53__tumor|T54__
tumor|T55__tumor|T56__tumor|T57__tumor|T58__tumor|59__tumor|T60__tumor|T62
__tumor.

Prétraitement : permet de supprimer les bruits et de nettoyer et normaliser les données.

Filtrage et Sélection : permet de filtrer (sélectionner) des gènes à partir d'un ensemble de données de microarray.

Réseaux bayésiens : Ces fonctions sont utilisées pour classer les gènes par ordre de preuve pour l'expression différentielle. Ils utilisent une méthode bayésienne empirique pour resserrer les variances résiduelles dans le sens de la chaîne vers une valeur commune (ou vers une tendance globale).

2.3.Conception détaillées

2.3.1. La phase de prétraitement des données

La première étape de c'est le prétraitement. Cette étape est considérée comme une première phase pour assurer la qualité intégrité de données brutes. Dans cette étape, plusieurs tâches ont été effectuées dont l'objectif est éliminer autant de sources non désirées de variation autant que possible. Afin d'extraire les données importantes, les données brutes sont d'abord prétraitées pour annuler les données les moins importantes.

	N01__normal.CEL	N02__normal.CEL	N03__normal.CEL	T01__tumor.CEL	T02__tumor.CEL	T03__tumor.CEL
1	98.0	124.0	152.0	158.0	116.0	373.0
2	5470.0	6125.8	5024.5	5666.0	5955.5	6664.3
3	113.0	147.0	151.0	163.0	129.0	259.0
4	5240.0	5989.0	5017.0	5747.0	5791.0	6393.0
5	82.5	112.0	205.5	146.5	90.5	127.8
6	92.0	127.0	159.0	134.0	103.0	145.0
7	5012.0	5191.0	4299.0	5109.0	5307.0	6795.0
8	111.0	116.0	131.0	163.0	117.3	464.5
9	4612.0	4837.0	4202.0	5269.0	5329.0	7855.0
10	115.0	127.0	153.0	128.0	130.0	646.0
11	4676.5	5048.0	4361.3	5343.0	5233.5	7171.3
12	110.0	132.5	161.0	150.0	112.0	152.0
13	4736.0	4859.0	4282.0	5183.0	5104.0	6166.0
14	99.3	122.5	156.3	139.0	109.3	122.3
15	4678.8	4844.3	4373.0	5473.5	5450.0	6340.0
16	110.0	134.3	123.0	127.0	111.0	126.0
17	4902.3	5070.5	4570.0	5833.5	5588.3	6324.8
18	126.5	129.5	149.0	131.3	135.0	135.0
19	5058.3	5290.8	4747.3	5932.5	5666.3	6357.5
20	119.3	152.0	140.3	144.3	123.0	120.3
21	5068.5	5079.0	4790.0	5784.8	5597.0	6511.8
22	113.3	126.5	147.3	141.3	115.3	101.0
23	5105.0	5165.0	4654.8	5699.0	5527.0	6610.0
24	115.0	151.0	151.0	130.0	118.0	111.0
25	5025.0	5127.8	4554.5	5899.3	5489.0	6768.5
26	120.0	134.0	172.0	145.0	127.0	135.0
27	4638.0	4830.0	4292.0	5179.0	5279.0	6426.3
28	102.3	139.5	140.8	141.3	122.0	134.0
29	4622.0	4604.0	4296.0	5904.0	5058.0	6146.0
30	113.0	136.0	149.0	134.0	118.0	117.0

Figure 20 Matrix Gene Expression

Les lignes représentent les gènes et les colonnes représentent les échantillons et Chaque valeur de M ij est la mesure du niveau d'expression de l'i−ème gène dans le j−ème échantillon.

Plusieurs méthodes de prétraitement décrit dans littérature, cette phase comporte 03 sous phases **Background correction,** ***Normalisation, sommarisation…***

1. Background correction :

Permet d'ajuster les données pour l'intensité ambiante entourant chaque caractéristique, et cela dépend de plusieurs méthodes :

- MAS 5.0
- RMA

- GCRMA

Nous appliquons la méthode MAS et RMA pour la correction de fond à notre base de données (prostate) :

MAS 5.0

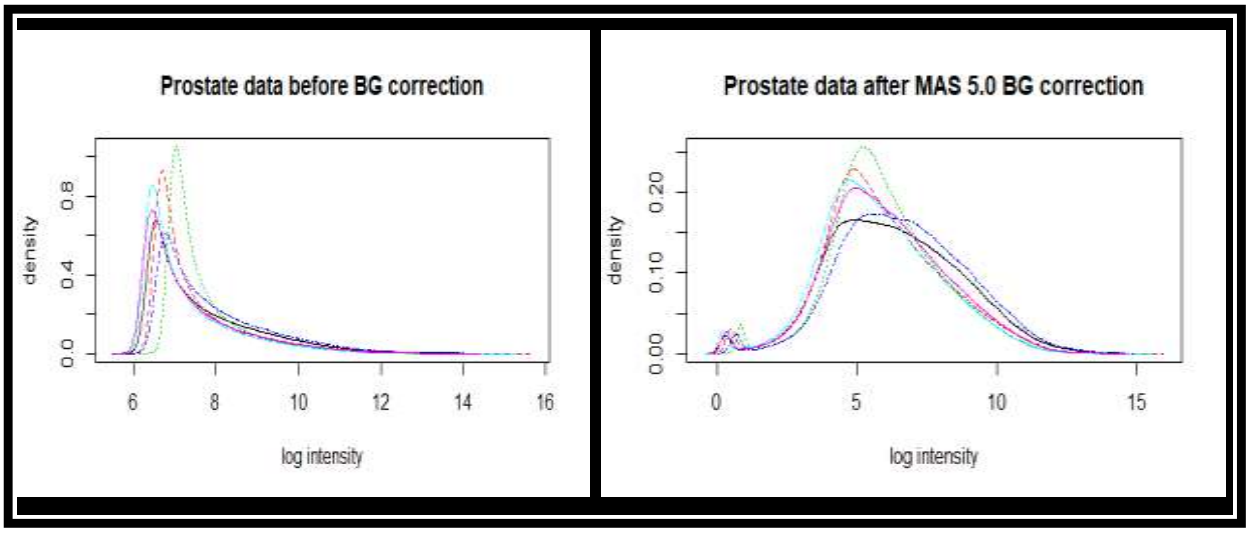

Figure 21Les données de la base prostate cancer Avant/Après BG correction par méthode MAS5.

RMA

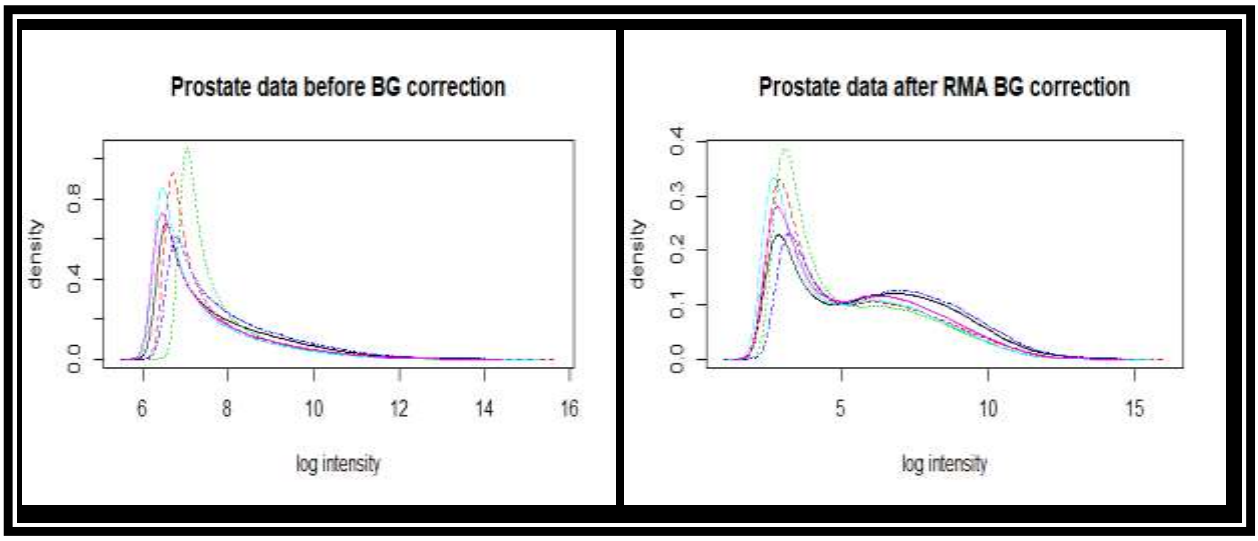

Figure 22 Les données de la base prostate cancer Avant/Après BG correction par méthode RMA.

2. Normalisation

La seconde étape du processus de prétraitement des données est la normalisation. La normalisation des données vise à minimiser (voire corriger) les biais techniques, systématiques ou liés au hasard. Théoriquement, dans une expérience de puces à ADN, la majorité des gènes reporters n'est pas différentiellement exprimée et la distribution des ratios est généralement centrée sur 0 (log2), Des études ont montré que la méthode de normalisation utilisée a une différence significative sur les niveaux d'expression différentielle finale, il est donc essentiel de choisir une méthode appropriée, parmi ces méthodes, nous avons vu :

MAS5

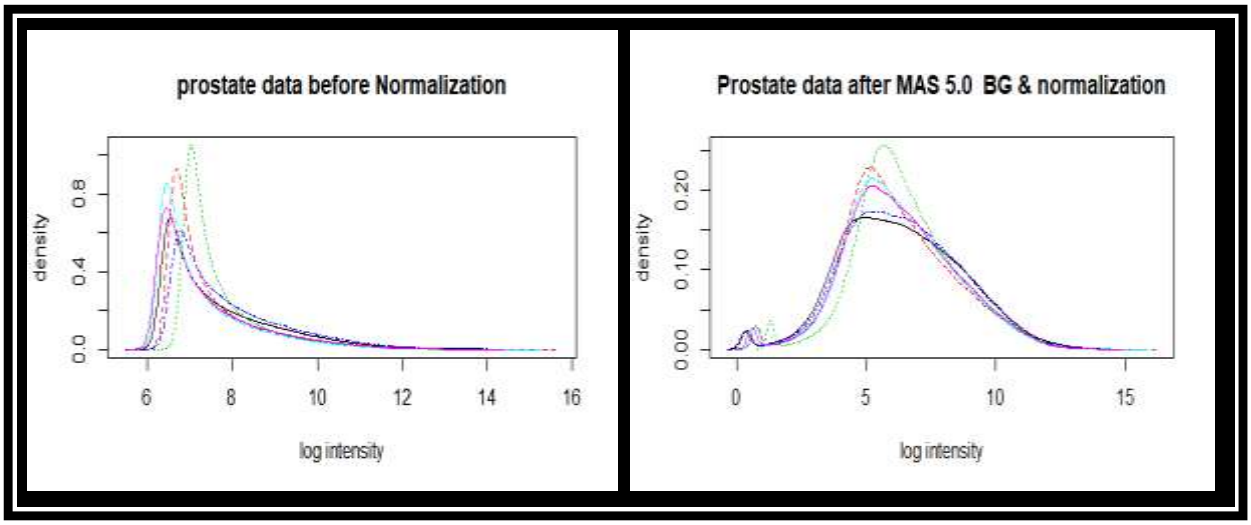

Figure Les données de la base prostate cancer Avant/Après Normalisation par méthode MAS5.

RMA

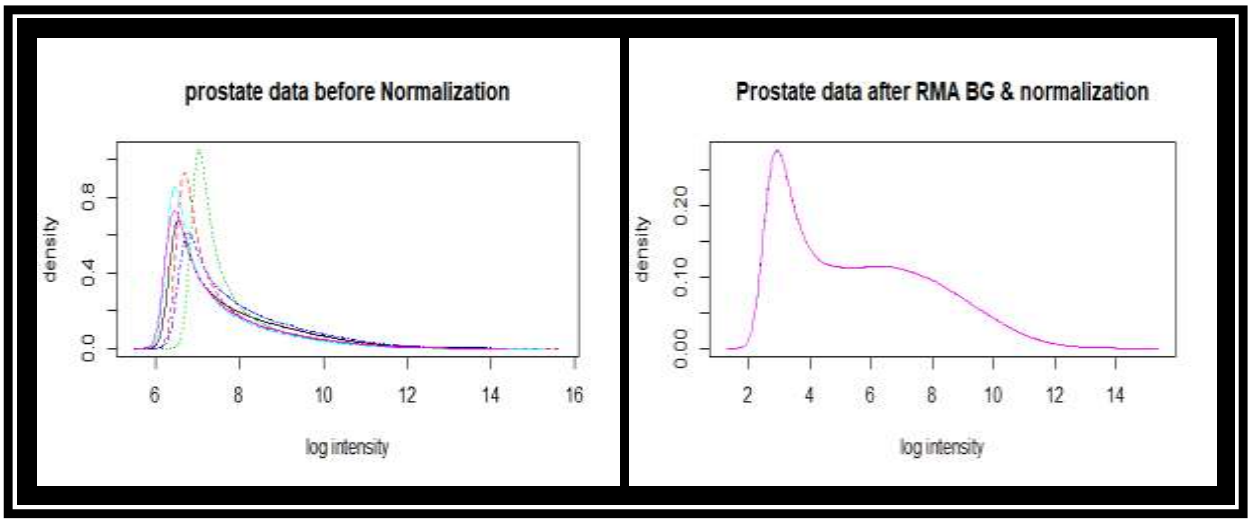

Figure 23 Les données de la base prostate cancer Avant/Après Normalisation par méthode RMA.

3. Sommarisation :

C'est une fonction générique utilisée pour produire des résumés de résultats des résultats de diverses fonctions d'ajustement de modèle, Le résultat du prétraitement est une matrice gène/expression qui est stocké dans un fichier texte. La fonction appelle des méthodes particulières qui dépendent de la classe du premier argument.

- **Farms** : peut être augmenté de la procédure de filtrage I/NI (informative/non informative) qui est un filtre non informé du design expérimental qui exploite l'architecture en probeset (ensemble de sondes) des GeneChip d'Affymetrix. . Les auteurs ont par ailleurs démontre que cette méthode fonctionne généralement bien si le nombre total de puces de l'expérience analysée est de six ou plus.

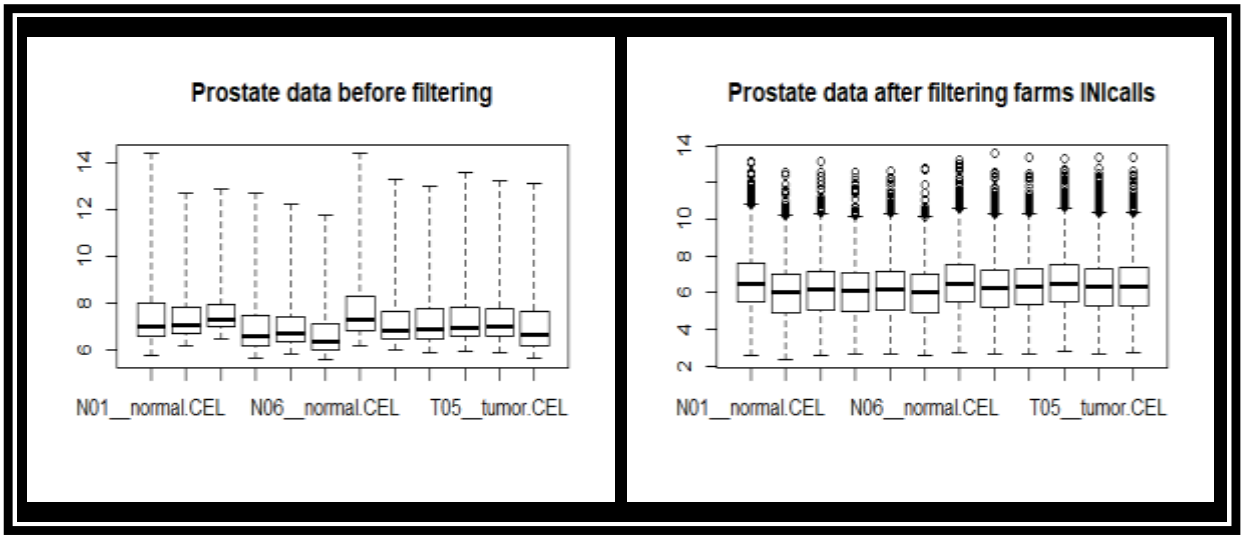

Figure 24 Les données de la base prostate cancer Avant/Après Filtrage par méthode Farms.

- MAS5
- Avgdiff

PM Correction

Nous utilisons le PM correction pour Le but donc d'une méthode de correction PM est de combiner les intensités des sondes perfect match(PM) et mismatch(MM) en une valeur proportionnelle au nombre réel de transcrits cible hybridés à la sonde perfect match(PM), plusieurs méthodes ont été suggérées :

- MAS5
- Phonily

Transformation log

Il convient de considérer les logarithmes des ratios qui ont une distribution normale. Par exemple, le passage en log2 permet d'avoir une symétrie entre les gènes sous- et sur-exprimés.

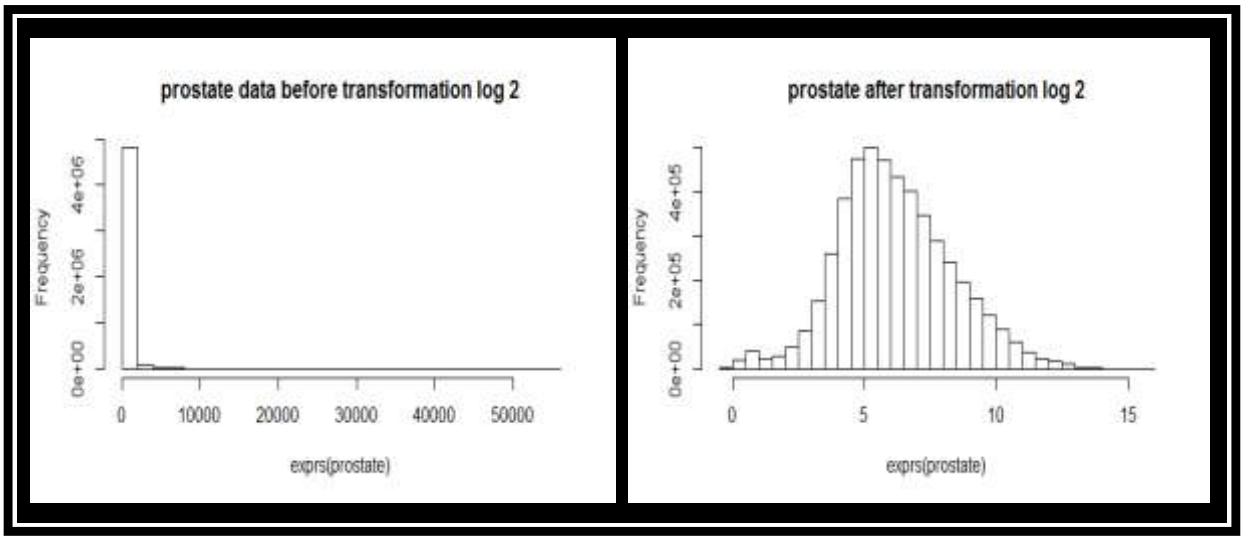

Figure 25 Les données de la base prostate cancer Avant/Après Transformation logarithmique.

2.3.2. La phase de filtrage

Nous appliquons donc un filtrage pour l'élimination ou au moins la minimisation de d'un tel éventuel bruit et après ça appliquons la sélection des gènes informant (pertinents), plusieurs méthodes de filtrage et sélection ont été suggérées :

- **Fold change**

Nous avons utilisé cette fonction pour calculer le facteur changement pour deux ensembles de valeurs. Logratio2foldchange convertit les valeurs des log-ratios en foldchanges.

Les facteurs de changements sont couramment utilisés dans les sciences biologiques comme un mécanisme pour comparer la taille relative de deux mesures.

Figure 26 Les données de la base prostate cancer Avant/Après Filtrage par méthode FoldChange.

- **t-test**

Elle permet de déterminer les gènes différentiellement exprimés entre 2 conditions (Expérimentale et de contrôle). Elle implémente le test de student pour les variances égales et non égales et ces tests.

- **SAM**

Pour chaque gène, calcul d'un score qui quantifie la différence d'expression du gène par rapport à 0.

2.3.3. Phase des Gènes différentiellement exprimés 'ebayes'

Est une autre statistique de l'expression différentielle fortement populaire en pratique, spécialement conçue pour l'ajustement de modèles linéaires aux données de microarray.

Qualitativement, l'idée des méthodes de Bayes empiriques s'articule ainsi :

- On estime d'abord la distribution des variances réelles à partir des données mêmes. Cette distribution est le Prior. Le Prior est souvent estime en supposant que seule une fraction des gènes sont véritablement différentiellement exprimes.
- On dérive un po*sterior* de la précédente distribution, c'est à dire la densité de probabilité de la variance réelle *étant donnée* la variance échantillonnalle observée.

Il est alors possible de remplacer la valeur observée de la variance échantillonnalle par la valeur attendue du *posterior*. Par exemple, on peut comprendre intuitivement qu'il est peut-être plus probable qu'une valeur élevée de la variance échantillonnalle provienne d'une variance réelle moyennement élevée que d'une variance réelle elle-même élevée, cette dernière étant elle-même peu probable selon le *Prior*.

Cette façon de raisonner peut sembler étrangement circulaire, mais il semblerait qu'elle fonctionne bien en pratique, c'est-à-dire que les estimes de la variance sont en moyenne plus rapproches de la variance réelle. La modération s'effectue en quelque sorte en « empruntant » de l'information à l'ensemble de gènes pour aider dans l'inférence sur un gène particulier.

3. Implémentation

3.1.L'objectif de Notre travail

Les données que nous allons étudier correspondent à une étude de l'effet des hormones sur l'expression des gènes au cours du temps dans des cellules cancéreuses (cancer de la prostate). Le but de l'étude est d'dentier les gènes qui répondent à cette hormone.

3.2.Bio-informatique appliquée à l'analyse

Les techniques d'analyse des données à haut débit produisent une quantité phénoménale de données brutes. L'analyse de ces données requiert l'utilisation d'ordinateurs et d'algorithmes sophistiqués visant à mieux en apprécier la signification biologique.

Pour développer notre application et valider notre proposition, nous avons utilisé le langage R et l'environnement WinDev pour écrire les programmes de traitement des données et d'extraction des caractéristiques.

3.2.1. Environnement de développement

3.2.1.1. R et Bioconductor

Bioconductor est une plateforme d'analyse bio-informatique basée sur le langage R [299]. Disponible gratuitement et développé par la communauté scientifique, Bioconductor comprend des modules d'analyse couvrant la plupart des scénarios imaginables - il est toujours possible de programmer un module répondant à un besoin particulier. L'adoption de cette plateforme est facilitée par le grand nombre de modules de visualisation et de statistique développés par la communauté R. Il est également possible d'accéder aux modules de Bioconductor à partir d'autres langages de programmation tels que Perl ou Python, ce qui ouvre la porte à la création d'interfaces interactives, plus facile d'accès pour un non-initié. Le langage R et Bioconductor sont tous deux plutôt difficiles d'accès, demandant une période d'apprentissage non-négligeable. La syntaxe du langage et le modèle de programmation vectoriel sont deux obstacles majeurs.

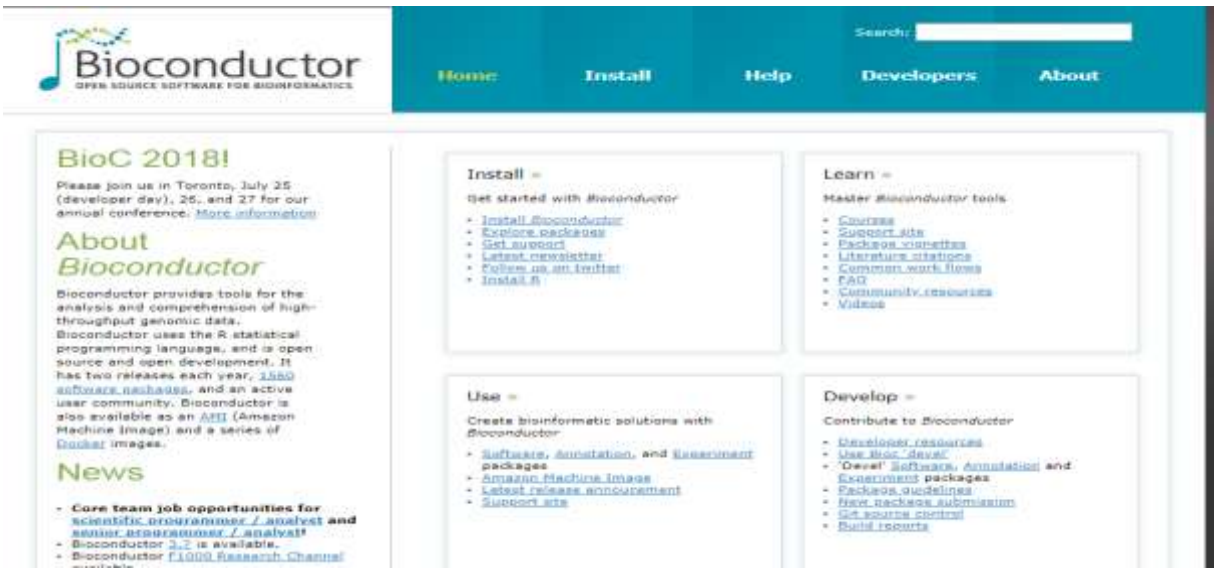

Figure 27 Page d'accueil du projet Bioconductor.

La nature ouverte et communautaire du logiciel est à la fois une force et une faiblesse. Le code source de chaque module est disponible et peut être consulté, voir modifié au besoin. D'un autre côté, chaque module est développé et maintenu par une ou deux personnes en moyenne - la qualité des modules varie beaucoup et certains responsables sont plus faciles à rejoindre et ouverts aux suggestions / rapport d'erreurs que d'autres. La communication avec la communauté (demande d'aide, suggestions, rapport d'erreurs) passe par l'entremise d'une liste

d'envoi de courriels - l'efficacité de celle-ci est variable selon le sujet. La documentation de certains modules laisse parfois à désirer et il est rarement possible de connaître les changements d'une version à l'autre sans comparer les codes sources, qui ne sont rarement commentés et utilisent des noms de variables non-descriptifs, ce qui complique la compréhension du code.

R / Bioconductor représente plus une boîte à outils qu'une solution accessible - cette solution offre la possibilité de bâtir un pipeline d'analyse sophistiqué et puissant, mais cette tâche demande un effort de programmation certain. Par contre, un grand avantage de la plateforme est d'avoir un accès pratiquement immédiat aux tous derniers algorithmes publiés, Bioconductor / R étant la plateforme d'implémentation de choix dans les publications bio-informatiques.

Une version 64 bits de R est récemment devenue disponible, abolissant les limites de mémoires de la version 32 bits, dans laquelle certains types d'analyse ne pouvaient être réalisés. Cependant, l'utilisateur doit être familier avec la compilation de code source, de nombreux modules n'étant pas disponibles nativement pour la version 64 bits.

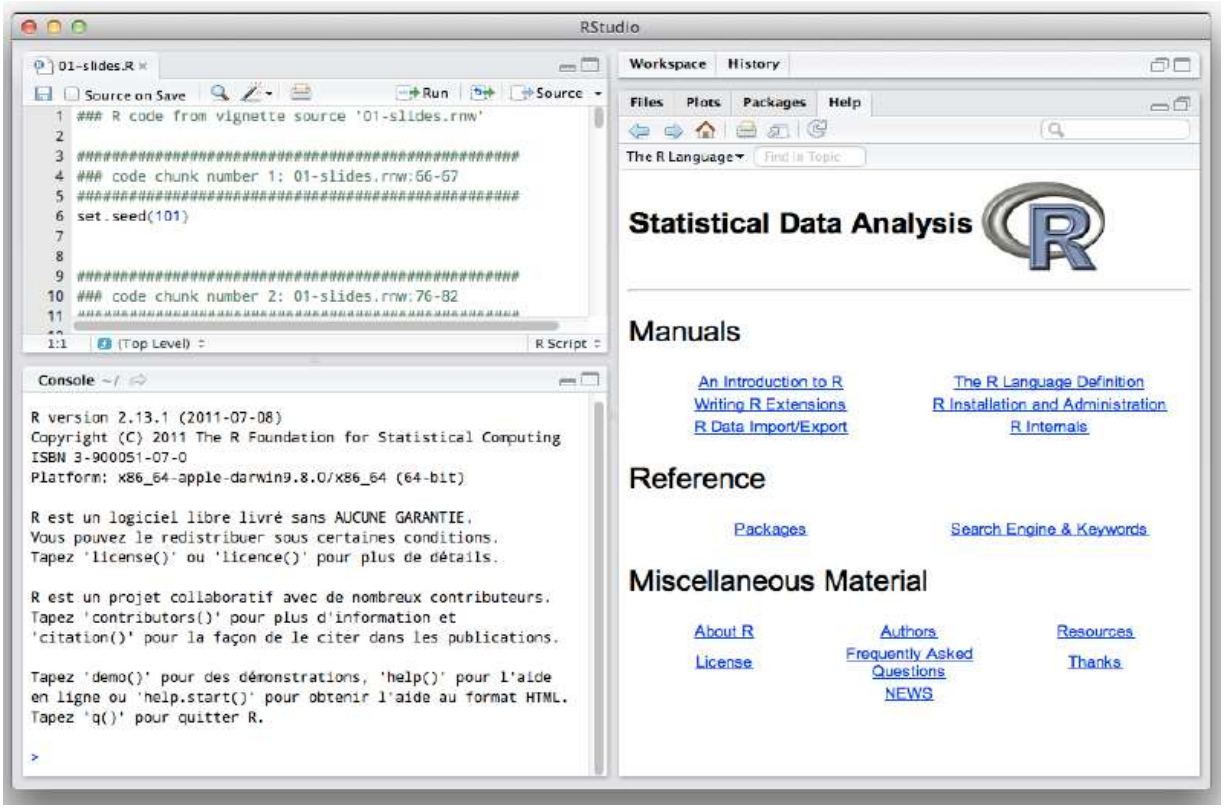

Figure 28 Langage R studio.

3.2.1.2. WinDev :

WinDev [25] est un atelier de génie logiciel (AGL) édité par la société française PC SOFT et conçu pour développer des applications, principalement orientées données pour Windows 10, 8, 7, Vista, XP, 2008, 2003, 2000, et également pour Linux, .NET et Java. Il propose son propre langage : le WLangage. La première version de l'AGL est sortie en 1993.

WinDev inclut en standard un ensemble d'éditeurs qui composent l'Atelier de Génie Logiciel : éditeur d'analyse (description des données), éditeur de fenêtres, éditeur de requêtes SQL, éditeur d'états, éditeur de tests automatisés, éditeur d'aide, éditeur d'images, éditeur UML, éditeur de code, éditeur de télémétrie, robot de surveillance, audit d'application, éditeur de dossier RGPD…

WinDev fonctionne selon un mode différent des autres langages : les fenêtres et états sont créés à l'aide d'un éditeur visuel. Les différents champs sont créés sous l'éditeur, et leurs paramètres sont définis à l'aide d'assistants de paramétrage visuels nommés « 7 onglets ». Chaque champ dispose en moyenne d'une centaine de paramètres. Cet éditeur ne génère pas de code mais crée un objet WinDev (fenêtre ou état).Cet objet sera ensuite utilisé par l'application. Ces objets effectuent directement un grand nombre de traitements : masque, tests de saisie, lien avec les bases de données, gestion des différentes langues, effets visuels, messages d'aide, etc.

WinDev utilise son propre langage de programmation, le WLangage, ressemblant beaucoup à du pseudo-code par son côté langage naturel qui peut faciliter la lecture du code par un débutant.

L'éditeur d'interface graphique permet de créer des IHM par glisser-déplacer. Il permet également de choisir un modèle de charte graphique parmi un ensemble proposé et d'en créer de nouveaux.

WinDev gère de nombreux systèmes de gestion de base de données, que ce soit par l'intermédiaire des protocoles ODBC ou OLE DB ou par accès natif.

3.3.Système de détection les gènes différentiellement exprimé proposé :

Nous avons développé une application implémentant la proposition présentée dans le chapitre précédent. L'application est composée de trois parties représentant chacune une phase de la méthode proposée :

- Prétraitement de données.
- Filtrage et test.
- Détection des gènes différentiellement exprimées.

3.4.Présentation des interfaces

L'application que nous avons développée comprend une interface permettant d'accéder à une étape du système :

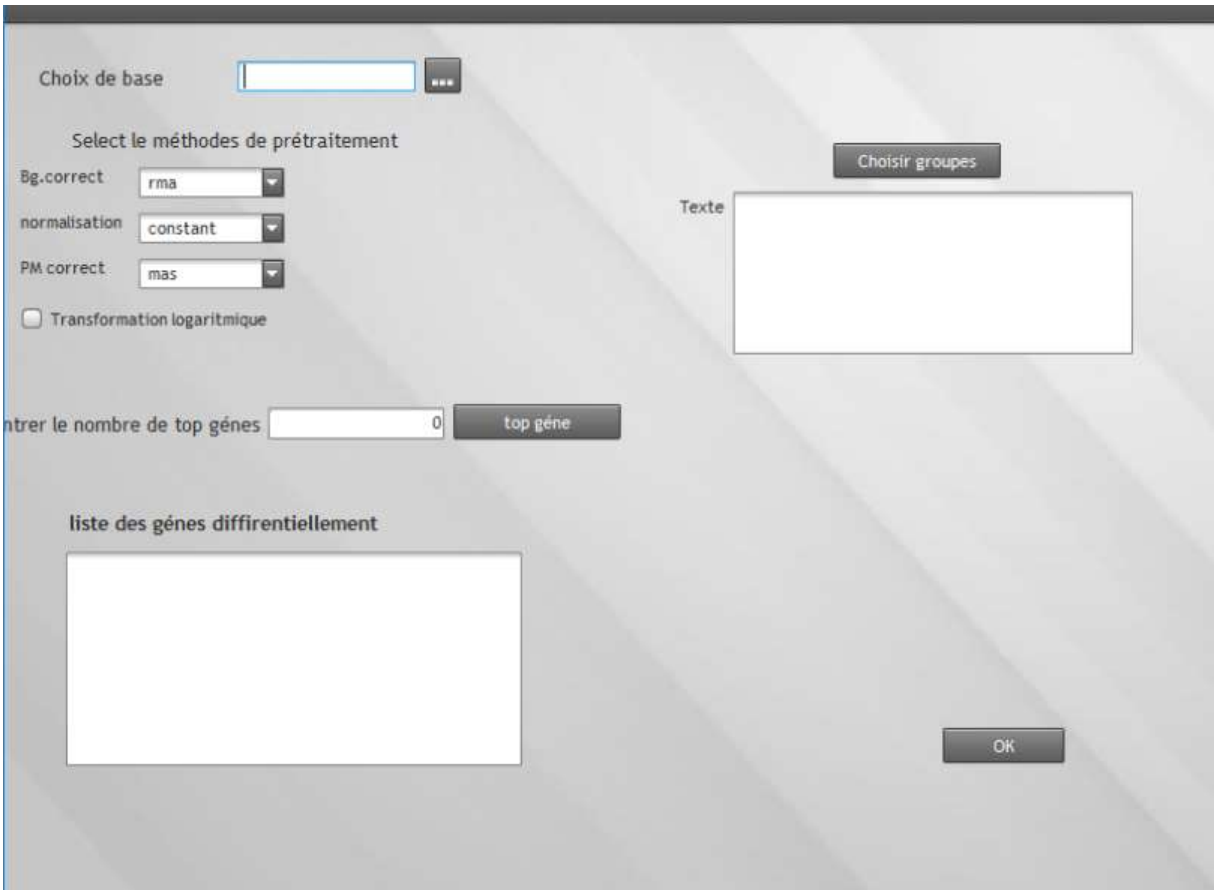

Figure 29 Interface graphique de notre système.

Cette interface permet de :

- Sélectionner la base que nous allons utiliser.
- Sélectionner les paramètres de prétraitement :
  - Correction de fond.
  - Normalisation.
  - Correction PM.
- Utiliser la transformation logarithmique.
- Choisir les groupes qui nous allons les étudier.
- Afficher les groupes a été utilisé.
- Enter le nombre des Top gènes.
- Sélectionner à le Botton top gènes.
- Afficher la liste des gènes différentiellement exprimés.

Expérimentations et résultats

Pour tester notre méthode, nous avons utilisé l'expérimentation suivante :

Dans cette expérimentation nous avons utilisé 12 échantillons des cellules de cancer prostate, 6 échantillons de type « Normal » et 6 échantillons de type « Tumor ». Puis nous avons appliqué la méthode de filtrage ***ebayes*** pour obtenues à les gènes différentiellement exprimé et pour les déterminer utilisons la fonction ***topTable*** (c.-à-d. Top Gene).

Nous avons utilisé les paramètres suivants :

- Fit : devrait être un objet de classe MArrayLM tel que produit par lmFit et eBayes.
- Sort.by : chaîne de caractères spécifiant une statistique pour classer les gènes par. Les valeurs possibles pour topTable et toptable sont "logFC", "AveExpr", "t", "P", "p", "B" ou "none".
- Number : nombre maximal de gènes à répertorier

Les résultats obtenus sont présentés dans le tableau de Top Gene de la figure ci-dessous. Le tableau de Top Gene présente les Top Gene (gènes différentiellement exprimé) qui nous avons les cherchons.

```
> output
             logFC   AveExpr         t      P.Value    adj.P.Val        B
38076_at  8.196174  8.196174 83.747490 8.416940e-20 3.175093e-17 32.552931
32752_at  7.931217  7.931217 80.061087 1.547859e-19 3.175093e-17 32.228521
40140_at  7.435434  7.435434 77.649139 2.341480e-19 3.175093e-17 31.999734
556_s_at  8.631771  8.631771 77.031117 2.608830e-19 3.175093e-17 31.938857
41696_at  7.845846  7.845846 76.589638 2.819799e-19 3.175093e-17 31.894786
33436_at  6.402201  6.402201 74.657519 3.984183e-19 3.738492e-17 31.696003
38969_at  7.271559  7.271559 70.599270 8.485255e-19 6.824569e-17 31.244924
35773_i_at 6.380566  6.380566 69.437926 1.061898e-18 7.473108e-17 31.106803
38780_at  7.644138  7.644138 67.311231 1.617168e-18 8.686118e-17 30.842579
32601_s_at 6.132873  6.132873 67.250727 1.636951e-18 8.686118e-17 30.834841
820_at    7.455896  7.455896 67.071450 1.697110e-18 8.686118e-17 30.811836
39113_at  7.138668  7.138668 65.442577 2.366267e-18 1.110173e-16 30.597644
945_at    6.336436  6.336436 64.234263 3.044201e-18 1.318373e-16 30.432519
1030_s_at 5.962568  5.962568 63.420162 3.616991e-18 1.454547e-16 30.318145
34352_at  5.780029  5.780029 62.606453 4.306715e-18 1.616454e-16 30.201231
35278_at 11.158826 11.158826 61.070143 6.025088e-18 2.120078e-16 29.973155
32819_at  5.557451  5.557451 59.518865 8.529939e-18 2.679714e-16 29.732663
38450_at  6.786031  6.786031 59.499529 8.567468e-18 2.679714e-16 29.729598
39741_at  7.769736  7.769736 58.769481 1.012249e-17 2.999453e-16 29.612647
39154_at  5.884859  5.884859 58.167733 1.163234e-17 3.074864e-16 29.514400
1804_at  12.827293 12.827293 58.072523 1.189261e-17 3.074864e-16 29.498700
31962_at 12.078612 12.078612 57.900764 1.237809e-17 3.074864e-16 29.470267
37639_at  8.728933  8.728933 57.837712 1.256161e-17 3.074864e-16 29.459794
37393_at  6.769356  6.769356 57.475202 1.367502e-17 3.207933e-16 29.399213
1248_at   5.628907  5.628907 57.122939 1.485902e-17 3.329748e-16 29.339733
41250_at  5.105872  5.105872 56.978152 1.537717e-17 3.329748e-16 29.315110
38472_at  5.478764  5.478764 56.074261 1.908389e-17 3.979344e-16 29.159030
549_at    5.257221  5.257221 54.822150 2.588796e-17 5.205328e-16 28.935903
41535_at  5.322360  5.322360 54.361605 2.901107e-17 5.632149e-16 28.851745
1521_at   5.855444  5.855444 53.281953 3.803420e-17 6.927618e-16 28.649895
34027_f_at 5.739158  5.739158 53.207812 3.875585e-17 6.927618e-16 28.635794
297_g_at  5.577841  5.577841 53.145328 3.937545e-17 6.927618e-16 28.623886
38057_at  7.000437  7.000437 52.991304 4.094863e-17 6.986085e-16 28.594438
```

Figure 30 Table des Top Gene.

P. Value : valeur de coupure pour les p-value ajustées. Seuls les gènes ayant des p-value plus faibles sont listés.

LogFc : minimum absolue log2-fold-change requis.

B : type de chaîne de caractères spécifiant une statistique pour classer les gènes

AveExpr : moyen de niveau d'expression.

Adj.p.val : Étant donné un ensemble de p-value, retourne les p-value ajustées en utilisant quelques méthodes.

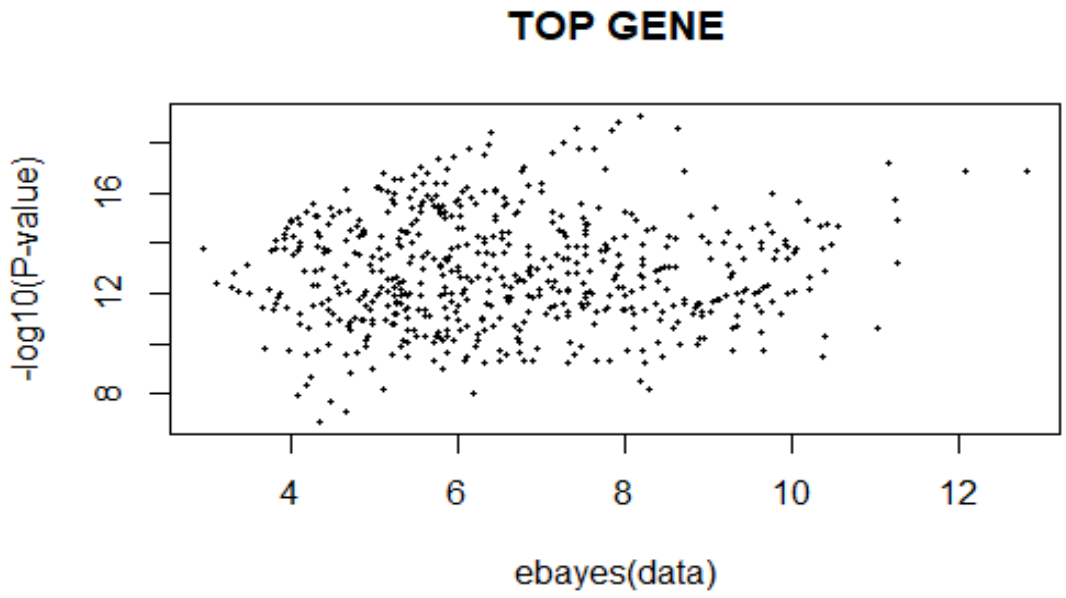

Figure 31 Nuage des points des Top Gene.

4. Conclusion :

Dans ce chapitre, nous avons présenté la conception de notre méthode, la conception globale des 4 phases (Collection des données, prétraitement, filtrage et apprentissage bayésiens). Après ca nous avons détaillé la conception de chaque phase.

Aussi, nous avons décrire l'implémentation du système. En outre, nous avons présenté les différents détails de l'implémentation tels que les outils et langages de programmation que nous avons utilisés. Nous avons présenté également les résultats obtenus.

Conclusion Générale

L'analyse des données des puces à ADN pour produire des listes de gènes différentiellement exprimés comporte plusieurs étapes qui peuvent différer en fonction du type de données analysées. Cependant, toutes les données suivent le même pipeline général qui implique la lecture des données brutes, la qualité évaluant les données, supprimant les mauvaises taches / tableaux d'une analyse plus poussée, prétraitant les données et calculant l'expression différentielle par analyse statistique.

Cette liste de gènes différentiellement exprimés peut ensuite être annotée avec des informations utiles qui expliquent la fonction des différents gènes, par exemple, l'ontologie des gènes.

En conclusion, l'utilisation de la technique des micropuces à ADN nous a prouvé sa puissance quant à l'impressionnante quantité d'informations qu'elle a été capable de nous fournir concernant les gènes qui seraient directement ou indirectement impliqués dans la détection les maladies cancer. Cette technique a mené à une meilleure caractérisation et à une meilleure compréhension des mécanismes qui seraient responsables de l'identification des gènes. La poursuite de ce projet par des analyses fonctionnelles permettrait de mieux identifier les gènes différentiellement exprimés

Bibliographie

[1] site web : http : //www.genomequebec.com/genomique-101.html, visité le 16-01-2018

[2] Genome Resource Facility GRF, Microarray section, London School of Hygiene and Tropical, Article technique, Medicine. 2006

[3] Diallo A., Classification des profils d'expression gènes : application à l'étude de la régulation du cycle cellulaire chez les eucaryotes, thèse de doctorat, Université de Grenoble, Juin 2010.

[5] Site web : https://www.thermofisher.com/dz/en/home/life-science/microarray-analysis.html, Visité le 28-10-2017.

[6] Parkinson H., et al, Array Express—a public database of microarray experiments and gene expression Profiles 2007.

[7] site web : https://www.ncbi.nlm.nih.gov/books/NBK159736/, visité le 07-12-2017.

[8] Pr Maier, cours UE Algorithmes pour la bio-informatique, Master recherche Informatique Maude Pupin.

[9] site web : https ://www.futura-sciences.com/sante/definitions/adn-mitochondrial-genomique-156/, visité le 16-01-2018.

[10]Pr Maier, cours généralité sur la cellule, 2014-2015.

[11] site web : http://frankpaillard.perso.infonie.fr/infirmier_generalites_cellule.html, visité le 16/01/2018.

[12] site web : https://www.futura-sciences.com/sante/definitions/genetique-transcriptome-5883, visité le 16-01-2018.

[13] Hassam.A, et Ouattara. IA, Construction d'un workflow d'analyse de données issues de puces à ADN, projet fin d'études, université d'Oran, juin 2014.

[14] khabzaoui.Med, Modélisation et résolution multi-objectifs des règles d'association : Application à l'analyse de données biopuces, Thèse de doctorat, université des Sciences et Technologies de Lille, Novembre 2006.

[15] Nguyen H.Tuong, Puce à ADN, Mémoire du Master 2 ECD 2007 – 2008, Ecole Polytechnique de l'Université de Nantes, France.

[16] cours DNA Microarray Data Analysis, chapter1.

[17] Mr MOUSSATI Omar, Classification des données de biopuces, Mémoire en vue de l'obtention du Diplôme de Magistère, université d'Oran, 2015/2016.

[18] François Lefebvre, Comparaison des méthodes d'analyse de l'expression différentielle basée sur la dépendance des niveaux d'expression, Université de Montréal, Mars 2011.

[20] Marie Agier. De l'analyse de données d'expression à la reconstruction de réseau de gènes, Bio-informatique, Université Blaise Pascal - Clermont-Ferrand II, 2006. Français.

[21] LE MEUR Nolwenn, De l'Acquisition des Données de Puces à ADN vers leur Interprétation : Importance du Traitement des Données Primaires, THESE DE DOCTORAT, Ecole Doctorale CHIMIE BIOLOGIE, Le 13 juin 2005.

[22] Bernard states, cours Acquisition et traitements statiques des données de génomique, Cah. Techn. I.N.R.A., 2004,52, 29-44.

[23] Cristian Chirca, Un modèle de réseau bayésien pour la gestion des risques dans la gouvernance d'un centre de services informatiques, thèse de doctorat, Faculté des sciences Université de Sherbrooke, Juin 2014.

[24] Hassan CHOUAIB, Sélection de caractéristiques : méthodes et applications, thèse de doctorat, Faculté de Mathématiques et Informatique Université Paris Descartes, juillet 2011.

[25] Site web: https://fr.wikipedia.org/wiki/WinDev. visité le 10-06-2018

[26] MICHAEL IMBEAULT, Etude Transcriptomique de l'effet du VIH-& sur le système immunitaire, Faculté de médecine, université LAVALQUEBEC, 2010.

[27] Brazma *et al.* 2001.

[28] Brazma *et al.* 2003.

[29] Edgar *et al.* 2002.

[30] Ikeo *et al.* 2003.

[31] Stuart *et al.* 2003 ; MC Carroll *et al.* 2004.

[32] Européen Bio-informatiques Institute(EBI), 2002.