Université Paris-Nord – Paris 13

Aide à l'exploration et à la découverte de relations dans des données de Génomique Médicale Fonctionnelle

THÈSE

présentée et soutenue publiquement le 11 décembre 2009

pour obtenir le titre de

Docteur de l'Université Paris-Nord - Paris 13

(Spécialité Informatique Médicale)

par

Arriel BENIS

Composition du jury

Président :	Pr. Younès Bennani	Professeur des Universités	Université Paris 13
Rapporteurs :	Dr. Alain Duhamel	Maître de Conférences Praticien Hospitalier	Université Lille 2
	Pr. Pascal Roy	Professeur des Universités Praticien Hospitalier	Université Lyon 1
Examinateurs :	Pr. Younès Bennani	Professeur des Universités	Université Paris 13
	Pr. Alain Venot	Professeur des Universités Praticien Hospitalier	Université Paris 13
Directeur de thèse :	Pr. Jean-Daniel Zucker	Directeur de Recherches	Institut de Recherche pour le Développement
Encadrante :	Dr. Mélanie Courtine	Maître de Conférences	Université Paris 13

Laboratoire d'Informatique Médicale et de Bioinformatique

Remerciements

Par ces quelques lignes, je souhaite exprimer mes remerciements :

à Jean-Daniel Zucker pour m'avoir fait découvrir le Monde de l'Extraction de Connaissances dans les Données ainsi que pour avoir accepter de diriger et commenter mes recherches ;

à Mélanie Courtine pour avoir accepter en cours de route d'encadrer mes travaux, pour son aide, sa patience, sa disponibilité, nos discussions sur les Sciences et le Monde, nos multiples collaborations pédagogiques et scientifiques, et pour son amitié ;

à Alain Venot pour d'une part m'avoir acceuilli dans son laboratoire alors que celui-ci était à ses débuts, pour ses conseils tout au long de mon doctorat et d'avoir accepter de faire partie de mon jury en qualité d'examinateur ;

à Younes Bennanni pour avoir accepter d'être examinateur lors de ma soutenance de thèse et d'avoir assure la Présidence de mon jury ;

à Alain Duhamel et Pascal Roy pour avoir été les rapporteurs des travaux que j'ai présenté dans mon manuscrit et d'avoir des membres actifs de mon jury ;

aux membres du LIM&Bio pour leurs conseils lors des réunions de laboratoire ;

à Karine Clement et aux membres de son équipe (« INSERM U872 equipe 7 Nutriomique ») pour notre collaboration Medico-Scientifique et pour avoir été la source des données qui ont permis mes travaux de recherches ;

à l'Institut Benjamin Delessert pour le prix de projet de Recherche qu'il m'a décerné en 2005 et qui m'a permit de me concentrer sur ma travaux et d'obtenir des résultats ;

à l'Association Française d'Etudes et de Recherches sur l'Obésité et la filiale française de la société pharmaceutique ABBOTT pour le prix de poster qu'elles m'ont décernée en 2004 ;

aux différents responsables des établissements aux seins desquelles j'ai enseigné pour m'avoir accorder leur confiance en me donnant de pleines responsabilités pédagogiques ;

aux étudiants et aux stagiaires, avec qui j'ai développé mon goût de l'enseignement et mes capacités à expliquer ce qui est difficile par des mots simples ;

à mes amis pour avoir accepter mes silences et mes absences pendant de longues périodes au cours de cette thèse ;

à tous ceux qui se reconnaîtront et que je n'ai pas cité ici ;

enfin à ma Famille pour sa présence, son soutien et sa patience pendant ces longues années d'études.

Une personne qui n'a jamais commis d'erreurs n'a jamais tenté d'innover.
Albert Einstein

Table des matières

Table des figures

Table des algorithmes

Introduction

Des données à la Découverte de Connaissances

Dans les entreprises, quelque soit leur domaine d'activité, le volume de données disponible est en constante augmentation depuis une vingtaine d'années [Wolfe 2000, Myatt 2006, Nisbet *et al.* 2009]. Cette croissance est due à plusieurs facteurs. Tout d'abord, les évolutions « informatiques » (au sens large) ont permis aux utilisateurs de pouvoir stocker de plus en plus de données sur des supports de plus en plus variés [Morris et Truskowski 2003], de les traiter de plus en plus rapidement [Mahajan *et al.* 2000] et de les échanger de plus en plus facilement *via* les réseaux de télécommunications [Meyn 2007]. De plus, l'arrivée de l'ère du numérique fait que les capteurs et les systèmes de numérisations sont présents en grand nombre ; les données sont donc présentes partout autour de nous. Par exemple, dans le secteur commercial, de nombreuses enseignes ont mis en place des cartes de fidélité, qui enregistrent à chaque passage en caisse des données plus ou moins complexes sur le client et sur ses habitudes (la fréquence des achats, leurs montants, la liste des produits achetés, leurs marques, ...).

Toutes ces données sont stockées sous la forme de fichiers ou dans des bases de données plus ou moins complexes et plus ou moins centralisées. Les dimensions de ces espaces de stockage sont telle que la plus grande de leur richesse réside dans leur exploitation [Wolfe 2000, Myatt 2006, Nisbet *et al.* 2009]. Le but est ainsi de rechercher dans ces données des informations cachées afin d'améliorer une méthode, un système ou un processus décisionnel [Christopher 2004, Uguen *et al.* 2007, Ohsawa et Yada 2009]. Par exemple, dans le secteur commercial, les analystes chercheront à découvrir les habitudes de consommation des clients afin d'orienter leurs campagnes publicitaires vers ceux les plus susceptibles d'y répondre [Giudici et Passerone 2002, Watada et Yamashiro 2006]. C'est pour atteindre ces objectifs que l'Extraction des Connaissances à partir des Données (ECD) *(ou Knowledge Discovery in Databases - KDD)* [Piatetsky-Shapiro et Frawley 1991, Han et Kamber 2006] et plus particulièrement la Fouille de Données (FDD) *(ou Data Mining - DM)* [Frawley *et al.* 1992, Cios *et al.* 2007] se sont développées.

L'Extraction des Connaissances à partir des Données a été initialement définie par Frawley [Frawley *et al.* 1992] puis citée par Piatetski-Shapiro [Piatetsky-Shapiro et Frawley 1991] comme suit : « *L'essence de l'ECD est l'extraction non triviale, à partir de données d'une information potentiellement utile, implicite et inconnue auparavant* ». Elle est ainsi définie comme le processus qui étant données des bases de données génère automatiquement des informations et/ou de la Connaissance. Ces résultats se doivent d'être exprimés de manière concise et facilement compréhensible pour les experts sans nécessiter de nouvelles interprétations, comme c'est généralement le cas lorsque l'on utilise des méthodes d'analyses statistiques [Motulsky 1999].

Le processus d'Extraction des Connaissances à partir des Données [Fayyad *et al.* 1996] correspond à un flux de traitements constitué de cinq étapes successives, comme le montre la figure 1. Ces cinq étapes sont les suivantes :

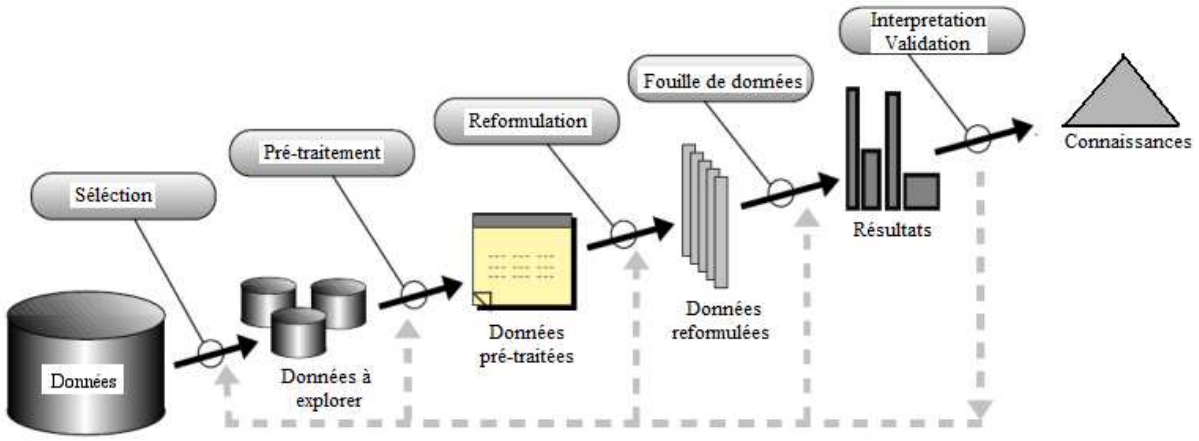


FIGURE 1 – Étapes du processus d'Extraction de Connaissances à partir de Données [Fayyad *et al.* 1996].

Sélection À partir de données disponibles stockées sous une forme structurée (base de données, fichiers tabulaires) ou non-structurée (textes, images, ...) une sélection des sous-ensembles à explorer est réalisée ;

Pré-traitement/Reformulation Les données ont, généralement, ensuite besoin d'être préparées, en vue des traitements, qui seront réalisés dans l'étape suivante. Cette préparation correspond, par exemple à la fusion de plusieurs fichiers « sources » ou à la binairisation de certaines variables ;

Fouille de Données L'étape suivante correspond à la Fouille de Données proprement dite. Elle s'appuie sur des approches issues de l'Intelligence Artificielle, mais aussi de la Statistique et des Mathématiques [Fayyad *et al.* 1996] et son but est de mettre en place des méthodes automatiques ou semi-automatiques pour découvrir des régularités, des motifs et/ou des règles dans un ensemble de données [Cornuéjols *et al.* 2002]. La Fouille de Données permet ainsi de réaliser des classifications, des estimations, des prédictions et/ou des regroupements. Même si l'ECD et la FDD sont souvent utilisées comme synonymes par abus de langage, la Fouille de Données n'est en réalité qu'une des étapes clés du processus complexe d'Extraction des Connaissances à partir des Données ;

Interprétation/Évaluation/Visualisation Cette dernière étape est essentielle, c'est celle où l'expert du domaine dont sont issues les données va apporter ses connaissances personnelles pour évaluer et valider les résultats obtenus pendant l'étape de Fouille de Données. Il est important que lors de cette étape, l'expert puisse avoir une vision condensée et précise des résultats afin d'accéder facilement aux nouvelles Connaissances et de les valider.

Cette thèse s'inscrit dans le cadre général de la Découverte de Connaissances dans les Bases de Données et dans un contexte spécifique d'application : le domaine biomédical et l'étude des maladies multifactorielles.

Un cas particulier : le domaine biomédical

Les données issues du domaine médical connaissent la même croissance que dans les autres domaines. En effet, le développement des microtechnologies et des nanotechnologies [de Rosnay

1975] ont permis des avancées significatives dans le domaine de la Santé, mais elles ont aussi été sources de très nombreuses données qu'il faut aujourd'hui analyser. Par exemple, l'apparition des biopuces dans les années 1990 a permis de développer un nouveau champ de recherche relatif à l'expression de gènes. La masse de données générées par un ensemble de biopuces mises en œuvre pour un protocole donné est très importante. En effet, une biopuce, ou plus particulièrement une puce à ADNc « pangénomique », utilisée pour étudier l'expression génique chez l'Homme correspond à 40 000 valeurs (données numériques) pour un individu donné à un instant donné. Ainsi, si l'on souhaite comparer plusieurs individus ou plusieurs conditions expérimentales, cela multiplie d'autant la quantité de données générées et donc à traiter.

Par ailleurs, une autre source de l'accroissement du volume des données en Sciences Biomédicale est le passage à l'informatisation dans un très grand nombre de structures médicales qu'elles soient dédiées à la clinique ou à la recherche. En effet, la mise en place du « Dossier Médical Informatisé » [Degoulet et Fieschi 1998, Cour des comptes 2009] devrait permettre d'acquérir des données physiologiques (taille, poids, ...), biologiques (taux de glycémie, de cholestérol, ...), socio-économiques et professionnelles (profession, revenus, ...), psychologiques (état de santé mentale, ...), thérapeutiques (médicaments prescrits, opérations subies, ...), d'explorations fonctionnelles (clichés radiographiques, tracés d'électrocardiogramme, ...), ...
Les données ainsi collectées pour un seul individu seront nombreuses, stockées dans différentes applications et leur volume devrait croître inévitablement au fil des années et du suivi médical du patient. À ce jour, ce projet n'a pas réellement encore abouti d'un point de vue nationale [Cour des comptes 2009] ; cependant les praticiens ont vu l'intérêt de l'informatisation des données patients et de nombreux systèmes indépendants de gestion ont vu le jour dans les services hospitaliers et les cabinets privés [O'Carroll *et al.* 2003, Benis *et al.* 2003e, Bagaria et Bagaria 2007].

Même si, les données biologiques et médicales sont d'un point de vue « informatique » des données comme celles issues de n'importe quel autre domaine d'application (tels que l'accidentologie, l'aéronautique, l'agriculture, ...), c'est-à-dire des données booléennes, symboliques ou numériques, elles possèdent néanmoins certaines particularités qui font qu'elles ne peuvent pas être traitées avec des méthodes classiques de Fouille de Données. En effet, les données issues du Monde de la Santé ont un statut particulier car elles concernent le Vivant et toute sa complexité. Pour simplifier, il est possible de dire que l'Homme peut aujourd'hui être étudié suivant deux perspectives [de Rosnay 1975] :

- la première, la plus commune, « macroscopique », c'est-à-dire en s'appuyant sur ce qui est visuellement observable (aspects physiques, exploration fonctionnelle, ...) ou simplement quantifiable (taille, poids, bilan sanguin, ...) ;
- la seconde « microscopique », c'est-à-dire en mettant en œuvre des techniques de laboratoire plus ou moins complexes : par exemple, les biopuces pour connaître les niveaux d'expression des gènes [Schena *et al.* 1995], l'amplification de fragments précis de l'ADN pour permettre la détection d'une mutation particulière sur un gène [Abou 2002], ...

Des points de vue biologique et médicale, les données macroscopiques correspondent à des « *phénotypes* » et les microscropiques aux « *génotypes* ».

Enfin, contrairement à d'autres domaines [Breton 2002, Hugueney 2003] où l'expert ne peut avoir qu'un rôle consultatif, l'expert a un rôle primordial lorsque l'on traite de données biomédicales. De manière générale, l'expert est une personne, qui peut être considérée comme un « référent » dans un domaine particulier car il possède une base de connaissances « démontrée » dans ce domaine et il est familier avec les données étudiées. Dans le cas de la Génomique Fonctionnelle, l'expert est un biologiste ou un médecin ayant une connaissance approfondie de la pathologie étudiée. Par exemple, la Génomique Fonctionnelle des Obésités requiert des connais-

sances particulières du métabolisme et des maladies qui sont les causes et/ou les conséquences d'une obésité. Il permet d'orienter l'analyse des données en formulant des hypothèses, en interprétant les résultats obtenus grâce à ses connaissances médicales et à sa propre expérience et en mettant en évidence sous la forme de découvertes, des résultats qui semblent insignifiants pour un non-expert. Il est donc important de limiter les *a priori* qu'il pourra « apporter » lors du traitement des données afin de lui fournir ensuite le maximum de résultats, et ainsi de faciliter leur exploitation pour lui ouvrir des perspectives de recherche sans pour autant l'emmener sur de fausses pistes. Il est donc important de trouver un compromis entre le degré de contrôle que l'on donne à l'expert lors d'une analyse et la volonté d'avoir une procédure totalement automatique pour gérer et traiter les données.

L'analyse de ces bases de données biomédicales a un but précis : l'aide à l'étude des maladies et la découverte de nouveaux moyens diagnostiques et thérapeutiques. Dans le cadre de cette thèse, nous nous intéresserons plus particulièrement aux maladies multifactorielles.

La Découverte pour la Recherche sur les maladies multifactorielles

Les maladies multifactorielles [van Rossum *et al.* 2005], tels que l'Obésité [National Institutes of Health 2000], le Diabète [Katz *et al.* 2000] et le Syndrome de Fatigue Chronique (SFC) [Whistler *et al.* 2003], représentent aujourd'hui un enjeu considérable en terme de Santé Publique. En effet, ces pathologies sont à l'origine de près de 60% des décès à travers le Monde [Organisation Mondiale de la Santé 1995]. Ces maladies sont dues à l'interaction entre des facteurs génétiques et environnementaux [Wadden *et al.* 2002]. Les travaux, en recherche fondamentale et en recherche appliquée, relatifs à ces maladies sont réalisés principalement sous des perspectives biologiques et médicales d'une part et psychologiques, sociologiques et économiques d'autre part [Organisation Mondiale de la Santé 2008]. Afin de permettre une prise en charge à terme plus efficace des patients atteints de maladies multifactorielles, des protocoles de recherche clinique transversaux se sont développés pour étudier les patients sous ces différents aspects simultanément.

Les données générées dans ces protocoles de recherche correspondent à un grand volume de données, pour les raisons que nous avons citées précédemment, mais aussi pour des raisons scientifiques et technologiques. En effet, d'un point de vue scientifique, les études s'appuient sur des cohortes de plus en plus grandes [Mullins *et al.* 2006] dans la perspective de spécialiser ou généraliser le plus possible les résultats obtenus [Vazquez Martinez *et al.* 1998] pour une pathologie donnée au sein d'une population donnée. De plus, au niveau des technologies de laboratoire, les avancées récentes ont permis de développer et de mettre en œuvre du matériel permettant de collecter de plus en plus de données avec un nombre réduit de manipulations telles que les biopuces pour l'expression des gènes [Lander 1999, Waltz 2001, Meur *et al.* 2004] ou les enquêtes épidémiologiques *via* des formulaires disponibles sur Internet (`http://www.etude-nutrinet-sante.fr`, [Hercberg 2009]). Enfin, les données obtenues grâce à ces nouvelles technologies ont une qualité de plus en plus élevée car elles respectent des normes de plus en plus strictes [Brazma *et al.* 2001, Spellman *et al.* 2002].

Dans le cadre de cette thèse, nous travaillerons principalement avec des données issues de travaux en Génomique « Médicale » Fonctionnelle et plus particulièrement sur des données biocliniques (« macroscopiques ») et des données d'expression génique (« microscopiques »).

Problématique

L'Extraction de Connaissances à partir de données médicales n'est pas une tâche triviale, car elle doit prendre en compte à la fois la diversité des données, l'hétérogénéité des connaissances du domaine d'application ainsi que les connaissances implicites et les hypothèses de travail de l'expert. De nouvelles thématiques de recherche en Informatique Biomédicale se sont développées avec pour objectif de proposer un cadre unifié pour l'acquisition, le stockage, l'exploration et l'exploitation des données aussi bien biologiques que des données médicales qu'elles soient « macroscopiques » ou « microscopiques ». L'objectif de cette thèse n'est pas de proposer un algorithme « innovant » pour traiter ces données. Mais notre objectif est de montrer que la construction d'un flux de données avec des méthodes appropriées et bien paramétrées ainsi qu'une visualisation adaptée permet d'analyser de manière efficace ces données. Le but final est que l'expert puisse découvrir avec un investissement en « temps expert » réduit des relations dans ces données et « comprendre » certains phénomènes biologiques causes de pathologies [Marton *et al.* 1999, Debouk et Goodfellow 1999].

Cette analyse va être basée sur la mise en évidence de relations entre différentes données, c'est-à-dire entre des données d'expression génique et des données biocliniques. Cette recherche de relations s'avère complexe et coûteuse en temps et en ressources, car les données sont nombreuses, les connaissances associées multiples provenant de différentes sources d'information et le processus de découverte semble avoir besoin d'être « contrôlé » fréquemment par un expert-biologiste. Stratégiquement, cette recherche est « valorisable », car elle permet la découverte de biomarqueurs. La connaissance de ces biomarqueurs [Symmans *et al.* 2007] va permettre de déterminer de manière rapide et efficace le statut d'un individu vis-à-vis d'un état ou d'une maladie et ainsi les médecins vont pouvoir engager une thérapeutique préventive ou curative adaptée aux patients.

D'un point de vue informatique, découvrir des biomarqueurs revient à découvrir des attributs particuliers mettant en relation des données génomiques et biocliniques pour un ensemble d'individus donnés dans des conditions expérimentales définies. La définition d'une relation entre deux ensembles de données consiste, en général, à calculer une « distance » entre ces deux ensembles [Labart *et al.* 2000]. La définition de cette distance n'est pas triviale, car elle impose de prendre en compte les contraintes et les limitations particuliers de données biomédicales. Les protocoles de recherche biomédicale, ayant pour but la découverte de biomarqueurs, s'appuient sur des technologies génératrices d'un grand nombre de données pour un nombre réduit d'individus. Cela va poser des problèmes pour la généralisation des résultats à l'ensemble d'une population, d'où l'importance d'avoir des mesures permettant d'analyser la pertinence des résultats. Par ailleurs, dans le cadre particulier des biopuces, il n'est pas rare d'obtenir des données bruitées et lacunaires [Pevsner 2005] liées au nombre important de manipulations au cours de leur production, de leur stockage et de leur mise en œuvre. Cela induit deux grands problèmes : le premier est lié à l'identification des données dites « anormales » et le second à leur prise en compte lors des processus d'analyse. Ces problèmes se retrouvent aussi au niveau des données biocliniques qui subissent les mêmes contraintes. Enfin, la découverte de relations entre deux ensembles de données dans le contexte biomédical doit prendre en compte l'expert. Cet expert doit pouvoir intervenir tout au long du processus d'exploitation des données afin de guider la découverte en fonction de ses propres connaissances sans pour autant induire trop de contraintes sur les résultats.

Ainsi, ces travaux s'articulent autour de deux grandes axes. Le premier consiste à définir un flux de données permettant de quantifier les relations entre les ensembles de données disponibles, ainsi que leurs qualités, en s'appuyant sur des approches connues issues de la Statistique et de la Fouille de Données. Le seconde a pour objectif de proposer une visualisation adaptée de ces

résultats afin de faciliter le travail de l'expert lors de leur exploration.

Plan de lecture

La première partie de cette thèse est consacrée à la présentation du cadre général de nos travaux. Nous commençons dans le chapitre 1 par présenter le cadre applicatif de ceux-ci : la Génomique Fonctionnelle et les différentes sources de données existantes dans ce domaine. Nous allons travailler avec des données issues de protocoles de recherche clinique mettant en œuvre des puces à ADNc et des données biocliniques. Dans le second chapitre, nous présentons un état de l'art d'approches permettant la recherche de relations, entre différents ensembles de données numériques, en Statistique et en Intelligence Artificielle. Ces données sont fortement impactées par des valeurs anormales. Nous présentons, dans le chapitre 3, différentes approches permettant de détecter et de traiter ce type de valeurs par des méthodes statistiques et issues de l'Intelligence Artificielle. Dans le chapitre 4, nous explicitons la problématique de la découverte de relations dans des données biomédicales et l'importance du rôle de l'expert dans ce processus automatique et notamment dans la cadre de l'exploration des résultats. La seconde partie de cette thèse présente les aspects théoriques de nos recherches, c'est-à-dire toutes les étapes de notre processus. Le chapitre 5 propose une approche exploratoire des découvertes de corrélations linéaires entre des données génomiques et des données biocliniques. Nous abordons ce problème dans une perspective expert, où le but est à la fois de réduire le nombre d'*a priori* lors du calcul des relations, mais aussi de reformuler les résultats numériques sous forme symbolique et visuelle afin de simplifier leur interprétation. Dans le chapitre suivant, nous proposons une méthode permettant de réduire le risque de fausses découvertes dans les résultats précédents. Cette approche, nommée PAMout, permet de détecter les valeurs singulières sans *a priori* dans les données que nous traitons.

La dernière partie de cette thèse est consacrée à la présentation de notre système DiscoClini et aux expérimentations. Dans le chapitre 7, nous présentons le système DiscoClini. Ce système permet à l'expert de soumettre des données puis d'explorer les résultats générés de manière guidée par un outil de visualisation en-ligne simple et intuitif. Le chapitre 8 est consacré aux expérimentations sur des données réelles dans le domaine de l'étude des Obésités. Ces données ont permis de montrer la faisabilité de l'approche, d'évaluer en pratique sa complexité, de montrer l'intérêt d'un tel outil pour les experts, mais aussi les limites de notre approche. De plus, des études d'utilisabilté du système ont permiss de monter la facilité d'utilisation de l'outil et de sa généricité. Ces résultats sont présentés dans le chapitre 9.

Enfin, nous concluons sur ce travail dans le dernier chapitre et proposons différentes perspectives de recherche à court, moyen et long termes.

Première partie

Etat de l'art et problématique

Chapitre 1

Sources de données en Génomique Fonctionnelle

Le contexte applicatif de cette thèse s'inscrit dans le cadre de la Génomique Fonctionnelle (GF) [Pevsner 2005] des pathologies complexes tels que l'Obésité et les maladies métaboliques [Basdevant *et al.* 1993, Organisation Mondiale de la Santé 1995, National Institutes of Health 2000, Clément *et al.* 2002, Forga *et al.* 2002] et les Cancers [Golub *et al.* 1999, Qiao *et al.* 2004].

Nous allons, dans ce chapitre, exposer le contexte d'application principal de nos travaux. Nous allons successivement présenter la Génomique Fonctionnelle puis les puces à ADNc, technologie de criblage haut-débit de laquelle sont issues nos données, puis les autres données dont nous disposons lors des protocoles de recherche clinique.

1.1 La Génomique Fonctionnelle

La *Génomique* [Gibson et Muse 2003] est l'étude des génomes et en particulier de l'ensemble des gènes, de leurs dispositions sur les chromosomes, de leurs séquences, de leurs fonctions et de leurs rôles. Les gènes, c'est-à-dire les parties d'Acide DésoxyriboNucléique (ADN) porteuses d'une information génétique, ne constituent qu'une partie du génome. Le reste du génome est composé de parties non codantes dont le rôle reste encore un mystère.

Un gène est formé de longs brins d'ADN (figure 1.1(f)) qui forment les chromosomes (figure 1.1(e)). Ils sont logés au sein du noyau (figure 1.1(c/d)) de chaque cellule (figure 1.1(b)). L'ADN est une chaîne structurée formée de quatre types de nucléotides (appelées aussi bases) : l'Adénine (A), la Cytosine (C), la Guanine (G) et la Thymine (T) (figure 1.1(g)). La forme la plus courante de l'ADN dans une cellule est dans une structure en double hélice, dans laquelle deux brins d'ADN individuels tournent l'un autour de l'autre sous forme d'une spirale [Watson et Crick 1953] (figure 1.1(f)).

La transcription (figure 1.1(j)) est la synthèse de l'ARN (Acide RiboNucléique, voir figure 1.1(k)) à partir de l'ADN et plus particulièrement des gènes (figure 1.1(i)). Ces deux séquences d'acides nucléiques utilisent le même langage (seules les thymines (T) sont remplacées par des uraciles(U)), et l'information est tout simplement transcrite, ou copiée, d'une molécule à l'autre (ADN en ARN). Dans le cas des séquences d'ADN codant des protéines (figure 1.1(m)), la transcription est la première étape, qui conduit généralement à l'expression des gènes, par la production de l'ARNm (Acide RiboNucléique messager, voir figure 1.1(l)). En effet, l'unité de transcription (partie d'ADN traduite en protéine) contient des séquences qui dirigent et régulent la synthèse des protéines [The Royal Swedish Academy of Sciences 2006].

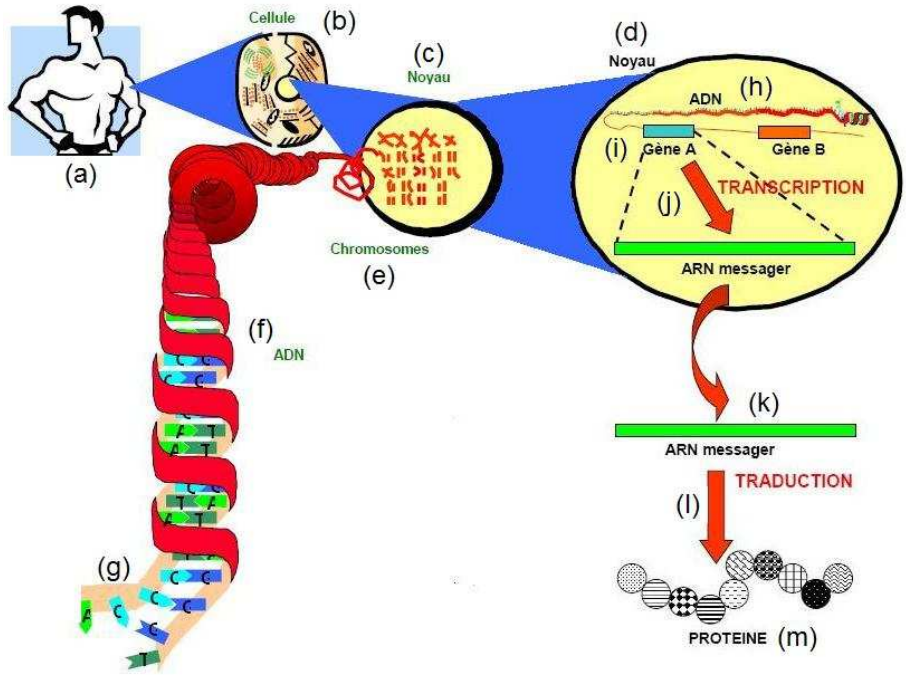


FIGURE 1.1 – L'homme, la cellule, l'ADN et la protéine (d'après [Abou 2002]).

Dans des conditions normales, l'expression d'un gène est en général le fruit d'interactions responsables de l'équilibre (homéostasie) cellulaire. Connaître le contexte macroscopique (et microscopique) au cours duquel un gène est transcrit (ARN) puis traduit (protéine) permet de définir sa fonction biologique et son implication physiologique. Des mutations dans le code génétique et/ou des erreurs au cours de la transcription et/ou de la traduction peuvent avoir un impact important sur le mécanisme physiologique lié à un gène et être la source d'un comportement pathologique, c'est-à-dire responsable de l'apparition d'une maladie.

Contrairement à la Génomique et à la Protéomique classiques, la *Génomique Fonctionnelle* [Pevsner 2005] se concentre sur les aspects dynamiques du génome et du protéome. Ainsi, elle va étudier la transcription des gènes, la traduction et les interactions gène-gène, protéine-protéine et gène-protéine, les relations entre les gènes et les fonctions métaboliques et physiologiques... Son but est ainsi de définir le rôle des gènes sur le métabolisme [Takase *et al.* 2000, Higami *et al.* 2004], de modéliser les réseaux de régulation génique [de Lichtenberg *et al.* 2005, Ashrafi *et al.* 2006, Missal *et al.* 2006], de caractériser les effets d'un traitement médical sur l'expression de gènes [Scherf *et al.* 2000, Lonning *et al.* 2007]...

En raison du grand nombre de séquences d'ADN, gènes et/ou protéines étudiées simultanément (quelques centaines, voire quelques milliers), la Génomique Fonctionnelle nécessite le recours à des outils informatiques puissants et à des méthodes automatisées en constante évolution pour analyser de plus en plus rapidement et efficacement les données [Guthke *et al.* 1997, McClure et Wit 2004].

Au cours des dernières années, la mise en œuvre des puces à ADNc [Schena *et al.* 1995] s'est développée et est devenue l'une des technologies les plus utilisées en Génomique Fonctionnelle.

1.2 La biologie à haut-débit

1.2.1 De nouvelles technologies

La *Génomique Fonctionnelle* s'appuie sur les outils issues de la Biotechnologie. Ces outils sont en permanente évolution et s'inscrivent dans le contexte, entre autres, des technologies dites de *criblage à haut débit* (en anglais *High Throughput Screening*). Dans le cadre général de la Biologie Moléculaire et plus particulièrement de la Génomique, un nombre grandissant de techniques à haut-débit ont vu le jour [Gibson et Muse 2003]. Ils permettent notamment d'étudier et d'identifier des molécules ayant des propriétés intéressantes d'un point de vue biologique et/ou médicale, tels que les gènes. Ces techniques s'appuient aussi bien sur la miniaturisation des technologies de laboratoire ainsi que sur les technologies issues de l'Informatique et de l'Électronique. Parmi les biotechnologies mises en œuvre en Génomique Fonctionnelle, les biopuces sont des outils incontournables aujourd'hui.

1.2.2 Les biopuces

De manière générale, les *biopuces* [Schena *et al.* 1995, Waltz 2001, Abou 2002] sont des systèmes miniaturisés dédiés à l'analyse biologique. Elles font intervenir des technologies multidisciplinaires issues de la Biologie, de la Physique, de l'Électronique, de l'Informatique, de l'Imagerie et des Mathématiques. Elles ont pour objectifs l'automatisation et la parallélisation de différentes étapes des protocoles d'analyses biologiques et médicales.

Les domaines d'applications envisageables (recherches biologiques et médicales, pharmaceutiques, agro-alimentaires) pour les biopuces sont tels que cette technologie ouvre des perspectives scientifiques et industrielles importantes (criblage moléculaire, génotypage, diagnostic, contrôle-qualité en agronomie,...). D'un point de vue biomédicale, elles deviennent une voie quasi-obligatoire lors de l'étude de maladies multi-factorielles [Abou 2002].

Elles sont couramment utilisées comme un outil d'exploration du génome. Elles se présentent sous la forme d'un support en verre ou en silicium de quelques cm^2 sur lequel sont fixés des milliers de fragments d'ADNc, d'ARN ou de protéines. Une biopuce permet ainsi d'identifier en un temps record un grand nombre de gènes (puces à ADN(c)/ARN) ou de protéines (puces à protéines) afin d'en étudier leurs fonctionnements.

Les biopuces peuvent se répartir en plusieurs catégories tels que les puces à ADN, les laboratoires sur puce (*Lab-On-Chip*, en anglais) et les puces à cellules (*Cell-On-Chip*, en anglais). Nos travaux s'appuyant sur la technologie des puces à ADN, nous allons présenter plus en détails cette technologie dans le paragraphe suivant.

Les puces à ADNc

Les puces à ADNc [Schena *et al.* 1995, Waltz 2001] (figure 1.2(a)) permettent de quantifier le niveau d'expression des gènes dans une cellule d'un tissu donné (tissu adipeux, foie, tumeur,...) à un instant donné et dans un contexte expérimental donné (individu sain ou malade, traité par une technique « A » ou « B »,...). Elles peuvent être en verre, en silicium ou bien encore en plastique et à leur surface, sont fixés des fragments d'ADN, aussi appelés sondes. Ces sondes sont capables de s'apparier à des brins d'ADNc (ADN complémentaire) dans une préparation biologique.

Les puces à ADNc permettent ainsi de mesurer puis de visualiser l'expression d'un très grand nombre de gènes. Par exemple, dans le cas des puces pangénomiques de type Stanford, qui sont utilisées pour le génome humain, ±40 000 expressions de gènes sont mesurées simultanément.

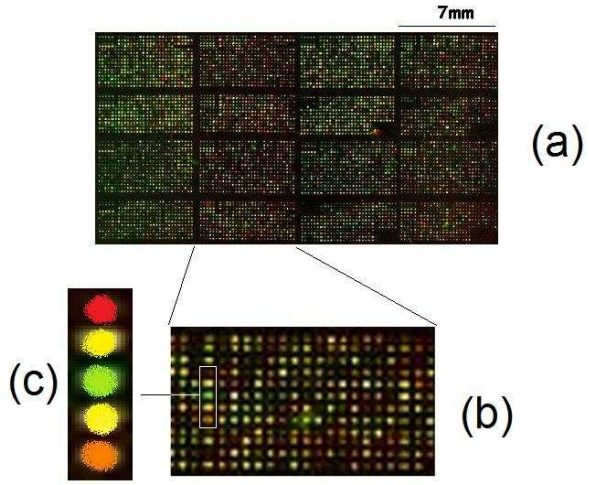


FIGURE 1.2 – Photographie d'une puce à ADNc.

Pour mettre en œuvre une puce à ADNc, le biologiste doit réaliser un certain nombre de manipulations, comme le montre la figure 1.3. Tout d'abord, des prélèvements de cellules (figure 1.3(a et b)) sont effectués sur deux sujets (humains, animaux ou végétaux) ou dans deux conditions expérimentales données (avant et après traitement, par exemple). Ensuite, l'ARNm est extrait de chacun des échantillons et marqué par des fluorochromes qui vont permettre de visualiser l'intensité de l'expression des gènes lors de la quantification de celle-ci (figure 1.3(c et d)). Les deux produits sont ensuite hybridés, c'est-à-dire qu'ils sont déposés sur les puces à ADNc afin de se combiner avec des fragments qui y sont fixés (figure 1.3(e)). La dernière étape est l'étape dite de « numérisation » (figure 1.3(f)) ; elle permet d'acquérir une image pour chacune des fluorescences (figure 1.3(g et i)). Ces deux images sont ensuite superposées afin d'obtenir une image unique (figure 1.3(h)). Les résultats de cette étape correspondent au niveau d'expression de chaque gène qui est matérialisé par une « couleur artificielle » prise par la sonde (figure 1.2). Cette couleur va du vert au rouge en passant par le jaune, si les fluorochromes utilisés lors du marquage sont le vert et le rouge :

- lorsque la sonde est rouge, ceci indique que le gène est *sur-exprimé*, en d'autre terme le gène s'exprime plus dans la « solution » marquée de rouge que dans celle marquée de vert ;
- lorsque la sonde est verte, ceci indique que le gène est *sous-exprimé*, en d'autre terme le gène s'exprime plus dans la « solution » marquée de vert que dans celle marquée de rouge ;
- lorsque la sonde est jaune, ceci indique que le gène s'exprime de la même manière dans la situation marquée rouge et dans celle marquée vert donc il n'y a pas de différence d'expression entre les deux solutions.

Ainsi, il est important de souligner que les notions de « sur-expression » et de « sous-expression » sont relatives à une situation donnée et aux marquages faits. Il est essentiel d'avoir des informations sur ces aspects afin de pouvoir analyser de manière efficace les puces à ADNc.

Ces données sont converties sous forme numérique afin d'être utilisées lors de traitements informatiques et/ou statistiques. De manière synthétique, elles sont souvent représentées sous la forme d'une matrice [Causton *et al.* 2003, McClure et Wit 2004]. Chaque colonne correspond à une condition expérimentale (qui peut correspondre à un individu, un état ou une date de protocole) et chaque ligne aux valeurs d'expression d'un gène dans l'ensemble des conditions expérimentales étudiées dans le protocole. L'intersection de chaque ligne et de chaque colonne

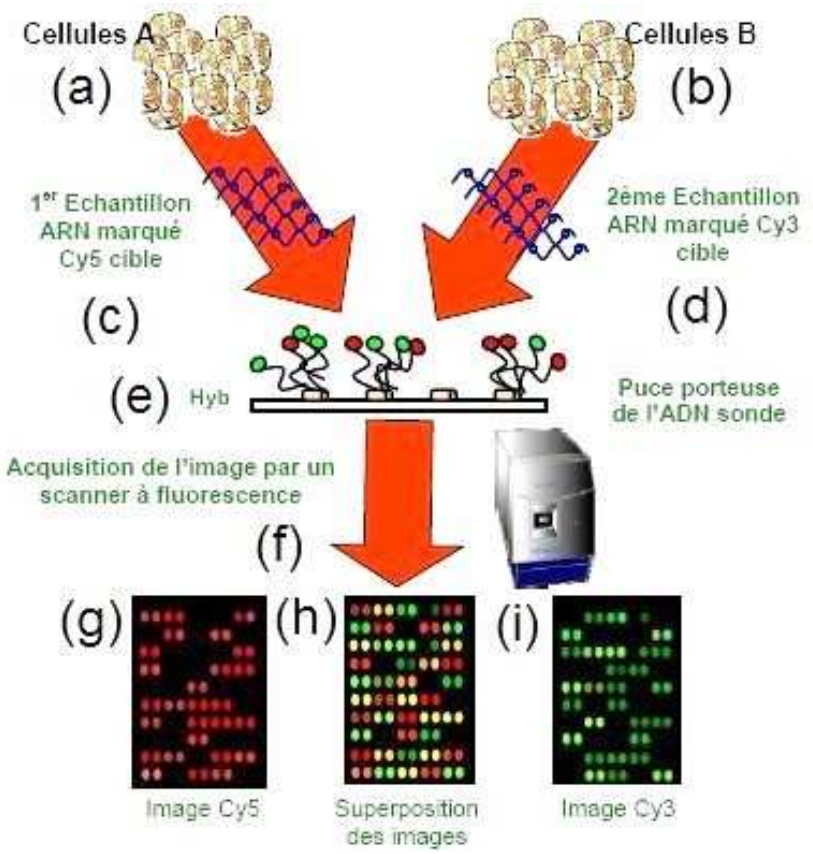


FIGURE 1.3 – Fabrication d'une puce à ADNc (d'après [Abou 2002]).

correspond à l'expression d'un gène pour une condition donnée. Un exemple d'une telle matrice est donné dans le tableau 1.4.

L'hétérogénéité des informations recueillies au cours de la mise en œuvre des puces à ADNc est conséquente. En effet, elle ne se limite pas seulement aux valeurs d'expression. Elle concerne aussi les différentes interventions humaines et automatiques qui ont eu lieu au cours de leur production et de leur traitement, et aussi le stockage du matériel biologique qui a été fait. Afin de permettre un stockage structuré des puces à ADNc et comprenant des informations communes au plus grand nombre des modèles de bases de données ont été proposés. Le schéma standard développé par la *MGED Society* est le modèle *MAGE* [Spellman *et al.* 2002]. Ce schéma consensuel de structuration des informations issues de puces à ADNc, il tend à permettre à tout laboratoire de stocker et d'échanger de manière homogène avec d'autres laboratoires les données qu'il a générées. Ce modèle s'appuie sur *MIAME* (Minimum Information About a Microarray Experiment) [Brazma *et al.* 2001, Whetzel *et al.* 2006], qui est un ensemble de recommandations visant à définir les données minimum à recueillir afin de permettre une interprétation non ambigüe des données issues de l'utilisation de biopuces. Il devient de plus en plus important de prendre en compte ces standards dans la perspective du développement croissant des réseaux thématiques de collaborations internationales [Brazma *et al.* 2003, Mukherjee *et al.* 2005, Rayner *et al.* 2006]. De plus, les éditeurs scientifiques commencent à demander le respect de ces standards dans les publications des chercheurs.

	Condition 1	*Condition* 2	...	*Condition* m
Gene 1	a_{11}	a_{12}	...	a_{1m}
Gene ...	...	...	...	...
Gene i	a_{i1}	a_{i2}	...	a_{im}
Gene ...	...	...	...	...
Gene n	a_{n1}	a_{n2}	...	a_{nm}

FIGURE 1.4 – Exemple d'une matrice représentant les valeurs d'expression génique pour une expérimentation donnée (d'après [Madeira et Oliveira 2004]).

1.2.3 Les annotations génétiques

Les travaux réalisés que ce soit en Génétique ou en Génomique permettent d'enrichir les connaissances relatives aux gènes, à leur expression, à leurs interactions et leurs implications dans des mécanismes biologiques aussi bien à l'échelle moléculaire que physiologique. Les annotations concernent aussi bien le nom commun du gène, ses noms « synonymes », ses noms courts (abrégés ou abréviations), ses identifiants dans les différents systèmes de référencement (comme dans *Entrez Gene* `http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene` [Maglott *et al.* 2005], *Gene Ontology* `http://www.geneontology.org/` [Gene Ontology Consortium 2000], *Gene Expression Omnibus* (GEO) `http://www.ncbi.nlm.nih.gov/geo/` [Barrett et Edgar 2006] ou SOURCE `http://smd.stanford.edu/cgi-bin/source/sourceSearch` [Diehn *et al.* 2003]) ainsi que des données génétiques (telle que la localisation chromosomique). Il est aussi possible d'obtenir d'autres informations issues d'autres sources de données comme ses références dans la littérature (PubMed : `http://www.pubmedcentral.nih.gov/`), ses fonctions moléculaires, biologiques et cellulaires, les pathologies dans lesquelles ce gène est impliqué,... L'ensemble de ces données est généralement très important en terme de volume et demande une bonne expertise des données étudiées afin d'en extraire les informations pertinentes pour l'étude en cours. La description du gène THRAP3 (*Thyroid Hormone Receptor Associated Protein 3*) grâce au site SOURCE [1] est donnée en exemple à la figure 1.5. Cette figure illustre la complexité des informations aujourd'hui disponible pour décrire chaque gène.

1.3 Les autres données biomédicales

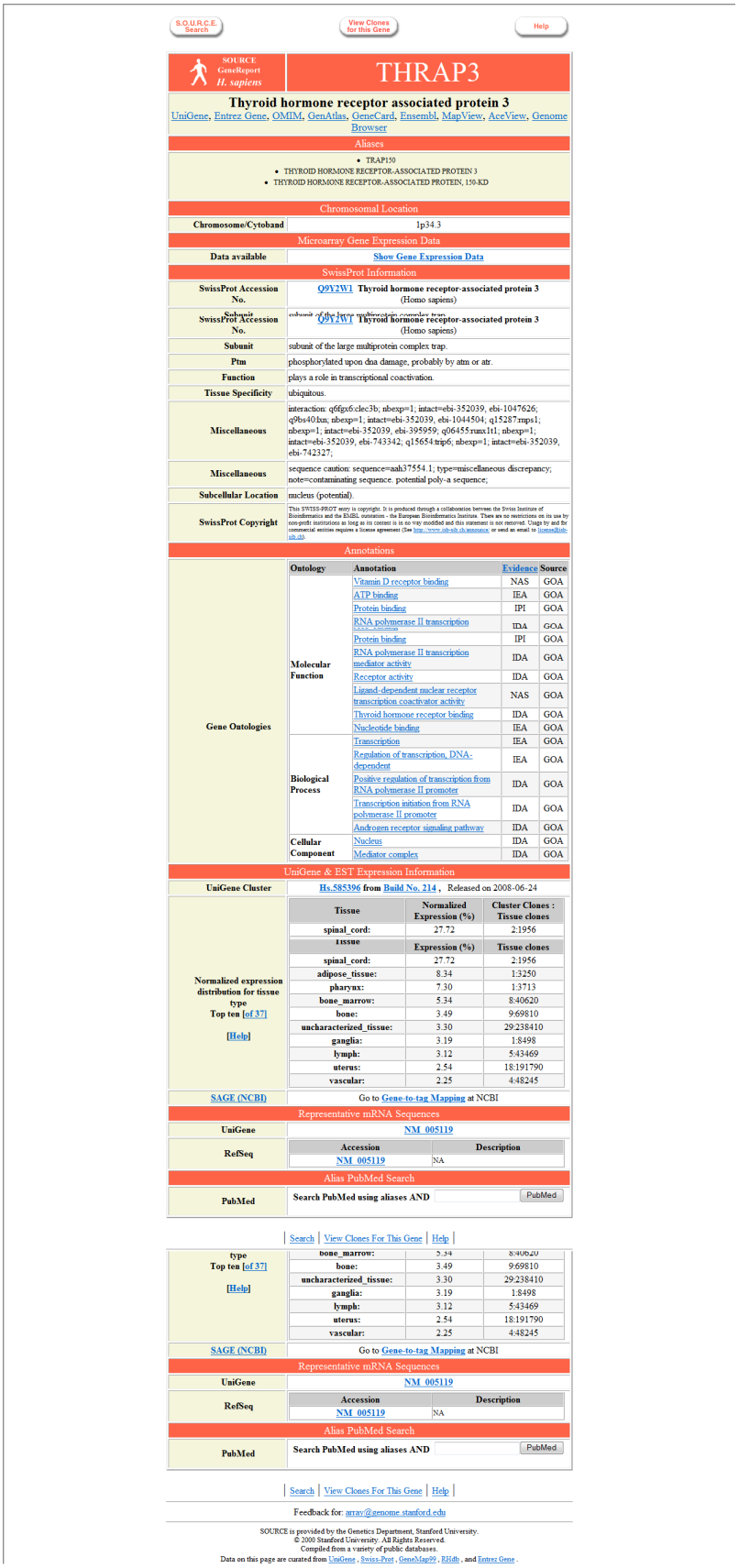
Les données biomédicales désignent à la fois toutes les données dont on vient de parler mais aussi deux autres grandes familles de données :

Les données biocliniques : elles peuvent être issues d'examens cliniques (taille, poids, tour de taille, tour de hanches,...), d'analyses biologiques (mesures de la glycémie, de l'insulinémie, des cholestérols, présence de mutations sur un ou plusieurs gènes,...). Elles peuvent aussi être le résultat de la mise en œuvre de méthodes de calcul permettant de synthétiser un ensemble de mesures (Indice de Masse Corporelle, sensibilité à l'insuline,...).

Les données épidémiologiques : elles correspondent à des données issues de questionnaires d'enquêtes socio-économiques (type et durée de l'activité professionnelle,...), de tests psychologiques et comportementaux (états de santé mentale, habitudes de consommation,...),...

L'ensemble de ces valeurs est représenté suivant le même formalisme que l'expression des gènes, à savoir une matrice à deux dimensions, où chaque colonne correspond à une condition (individu, état, date, ...) et chaque ligne à un paramètre bioclinique ou épidémiologique.

1. `http://smd.stanford.edu/cgi-bin/source/sourceSearch`



S.O.U.R.C.E. Search | View Clones for this Gene | Help

SOURCE GeneReport *H. sapiens*

THRAP3

Thyroid hormone receptor associated protein 3

UniGene, Entrez Gene, OMIM, GenAtlas, GeneCard, Ensembl, MapView, AceView, Genome Browser

Aliases

- TRAP150
- THYROID HORMONE RECEPTOR-ASSOCIATED PROTEIN 3
- THYROID HORMONE RECEPTOR-ASSOCIATED PROTEIN, 150-KD

Chromosomal Location

Chromosome/Cytoband	1p34.3

Microarray Gene Expression Data

Data available	Show Gene Expression Data

SwissProt Information

SwissProt Accession No.	Q9Y2W1 **Thyroid hormone receptor-associated protein 3** (Homo sapiens)
SwissProt Accession No.	Q9Y2W1 **Thyroid hormone receptor-associated protein 3** (Homo sapiens)
Subunit	subunit of the large multiprotein complex trap.
Ptm	phosphorylated upon dna damage, probably by atm or atr.
Function	plays a role in transcriptional coactivation.
Tissue Specificity	ubiquitous.
Miscellaneous	interaction: q6fgx6:clec3b; nbexp=1; intact=ebi-352039, ebi-1047626; q9bs40:txn; nbexp=1; intact=ebi-352039, ebi-1044504; q15287:rnps1; nbexp=1; intact=ebi-352039, ebi-395959; q06455:runx1t1; nbexp=1; intact=ebi-352039, ebi-743342; q15654:trip6; nbexp=1; intact=ebi-352039, ebi-742327;
Miscellaneous	sequence caution: sequence=aah37554.1; type=miscellaneous discrepancy; note=contaminating sequence. potential poly-a sequence;
Subcellular Location	nucleus (potential).
SwissProt Copyright	This SWISS-PROT entry is copyright. It is produced through a collaboration between the Swiss Institute of Bioinformatics and the EMBL institute - the European Bioinformatics Institute. There are no restrictions on its use by non-profit institutions as long as its content is in no way modified and this statement is not removed. Usage by and for commercial entities requires a license agreement (See http://www.isb-sib.ch/announce or send an email to licensing@isb-sib.ch).

Annotations

Gene Ontologies

Ontology	Annotation	Evidence	Source
Molecular Function	Vitamin D receptor binding	NAS	GOA
	ATP binding	IEA	GOA
	Protein binding	IPI	GOA
	RNA polymerase II transcription cofactor activity	IDA	GOA
	Protein binding	IPI	GOA
	RNA polymerase II transcription mediator activity	IDA	GOA
	Receptor activity	IDA	GOA
	Ligand-dependent nuclear receptor transcription coactivator activity	NAS	GOA
	Thyroid hormone receptor binding	IDA	GOA
	Nucleotide binding	IEA	GOA
Biological Process	Transcription	IEA	GOA
	Regulation of transcription, DNA-dependent	IEA	GOA
	Positive regulation of transcription from RNA polymerase II promoter	IDA	GOA
	Transcription initiation from RNA polymerase II promoter	IDA	GOA
	Androgen receptor signaling pathway	IDA	GOA
Cellular Component	Nucleus	IDA	GOA
	Mediator complex	IDA	GOA

UniGene & EST Expression Information

UniGene Cluster	Hs.585396 from Build No. 214 , Released on 2008-06-24

Normalized expression distribution for tissue type
Top ten [of 37]
[Help]

Tissue	Normalized Expression (%)	Cluster Clones : Tissue clones
spinal_cord:	27.72	2:1956

Tissue	Expression (%)	Tissue clones
spinal_cord:	27.72	2:1956
adipose_tissue:	8.34	1:3250
pharynx:	7.30	1:3713
bone_marrow:	5.34	8:40620
bone:	3.49	9:69810
uncharacterized_tissue:	3.30	29:238410
ganglia:	3.19	1:8498
lymph:	3.12	5:43469
uterus:	2.54	18:191790
vascular:	2.25	4:48245

SAGE (NCBI)	Go to Gene-to-tag Mapping at NCBI

Representative mRNA Sequences

UniGene	NM_005119

RefSeq	Accession	Description
	NM_005119	NA

Alias PubMed Search

PubMed	Search PubMed using aliases AND [ ] PubMed

Search | View Clones For This Gene | Help

Normalized expression distribution for tissue type
Top ten [of 37]
[Help]

Tissue	Expression (%)	Tissue clones
bone_marrow:	5.34	8:40620
bone:	3.49	9:69810
uncharacterized_tissue:	3.30	29:238410
ganglia:	3.19	1:8498
lymph:	3.12	5:43469
uterus:	2.54	18:191790
vascular:	2.25	4:48245

SAGE (NCBI)	Go to Gene-to-tag Mapping at NCBI

Representative mRNA Sequences

UniGene	NM_005119

RefSeq	Accession	Description
	NM_005119	NA

Alias PubMed Search

PubMed	Search PubMed using aliases AND [ ] PubMed

Search | View Clones For This Gene | Help

Feedback for: array@genome.stanford.edu

FIGURE 1.5 – Description du gène THRAP3 *via* le site *SOURCE*

Il est possible d'associer à des données d'expression génique des données relatives aux individus étudiés afin d'améliorer les analyses qui vont être faites [Jensen *et al.* 2006]. En effet, les données génétiques ne suffisent pas pour l'étude de la plupart des pathologies, d'autres facteurs, tels que l'environnement, la nutrition et le stress, peuvent influencer une pathologie et son développement [Clément *et al.* 2002, van Rossum *et al.* 2005].

1.4 Qualité de puces à ADNc

Comme nous l'avons exposé plus haut, la conception et la mise en œuvre des puces à ADNc peuvent induire un ensemble de biais d'origine :

- humaine (par exemple, dégradation du matériel biologique par surchauffe lors de son traitement),
- technique (par exemple, dégradation du matériel biologique suite à un mauvaise stockage (trop chaud et/ou trop longue) ou encore une défaillance ou une usure excessive d'une ou plusieurs têtes du « robot spotteur » utilisé pour le dépôt des ADNc sur les lames de verre).

Différents tests de contrôle-qualité sont effectués par les laboratoires produisant des puces à ADNc ou à oligonucléodides. Néanmoins, les résultats de ces tests sont considérés comme acceptables si moins de 10% des spots sont manquants ou présentent une intensité trop élevée d'un signal d'hybridation alors qu'il s'agit normalement de spots non-porteurs de matériel biologique. Ces contrôles sont d'autre part réalisé sur une ou deux lames que l'on considère comme l'échantillon représentatif d'un lot de plusieurs dizaines.

D'après Michaud [Michaut 2006], la qualité des puces à ADNc tend à s'améliorer à la condition de standardiser les pratiques expérimentales. En effet, d'après Tan [Tan *et al.* 2003], l'absence de standardisation des processus de conception, de fabrication et de mise en œuvre des puces à ADNc et des biopuces en général conduit à une irréplicabilité des résultats obtenus dans un nombre élevé de situations. Ainsi outre MIAME et MAGE desquels nous avons précedement parlés, le protocole *MicroArray Quality Control (MAQC)* [Shi *et al.* 2005] a été mis en place pour permettre de structurer et d'harmoniser le contrôle-qualité des biopuces. Les expérimentations relatives à la mise en pratique du MAQC a conduit *via* l'implication de 51 centres de recherche à démontrer que l'homogénéisation des procédures de production implique que les seules différences notables entre les résultats proviennent de la qualité des échantillons biologiques et du protocole d'analyse. Cela, nous permet donc de souligner que les données issues de puces à ADNc ne sont pas de qualité « élevée » de part les biais présent non pas au cours de leur production mais de leur mise en œuvre [Huber *et al.* 2006].

La qualité des données biocliniques est très variable. En effet, les valeurs issues des analyses de laboratoire présentent dans la majorité des cas une bonne qualité. Néanmoins, les données issues de l'exercice clinique ou collectées auprès du patient sont praticien- et/ou patient-dépendantes. Ainsi, ces données ne sont pas de qualité parfaite bien que des guides de bonnes pratiques existent [Degoulet et Fieschi 1998].

1.5 Conclusion

Le besoin de l'utilisation de techniques statistiques ou d'Intelligence Artificielle dans l'exploitation des données produites lors de protocole de recherche clinique est justifié par la complexité des données étudiées. Ainsi, dans ce contexte d'étude, on dispose de peu d'exemples pour un grand nombre d'attributs (quelques individus avec chacun beaucoup d'attributs : expression génique, données biocliniques,...) ce qui posent de nombreux problèmes [Meur *et al.* 2004], notamment

en terme de généralisation des résultats. D'autre part, les données et les systèmes d'information qui les gèrent, sont hétérogènes dans leurs structures et leurs sémantiques, ceci induit l'existence d'un grand nombre de connaissances cachées et/ou privées et/ou inconnues.

Dans ce manuscrit, nous utiliserons le terme de « Génomique Médicale Fonctionnelle » pour définir notre domaine d'application. Ce terme permet de définir la médecine clinique comme le contexte applicatif des études de Génomique Fonctionnelle [2]. Cela implique que les données utilisées sont issues de protocoles de recherche clinique et de l'exercice médicale de manière plus générale. Ces données peuvent être uniquement des valeurs d'expression génique ou ces mêmes valeurs associées à d'autres types de données telles que des valeurs biocliniques. Ainsi, dans le cadre de nos travaux, nous nous intéresserons plus particulièrement aux liens pouvant exister entre deux types de données : les données issues des puces à ADNc et les données biocliniques.

2. Une équipe de recherche danoise (University of Copenhagen, Faculty of Health Sciences, Department of Medical Biochemistry and Genetics, *Medical Functional Genomics Group*) définit son activité de cette même manière.

Chapitre 2

Découverte de relations en Biologie

La mise en relation de données, quelque soit le domaine d'application, a pour objectif de permettre la définition d'attributs qui présentent des caractéristiques dépendantes, pour tout ou partie de deux ensembles de données. Il est possible à l'aide de méthodes statistiques et informatiques de calculer et de rechercher ces liens au sein des bases de données comportant un très grand nombre d'attributs de types différents que ce soit par exemple en *Sciences Biomédicales* [Arredondo-Vega *et al.* 1998, Stratowa *et al.* 1999, Bhardwaj et Lu 2005], en *Intelligence Économique* [Cecil *et al.* 2002, Chiang *et al.* 2005] ou en *Documentation* [Kovalerchuk 2001a, Hardoon *et al.* 2006].

L'aide à la découverte de relations entre des données d'expression génique et des données biologiques et/ou cliniques est un domaine abordé dans la littérature suivant différentes perspectives. Les objectifs sous-jacents à la définition de ces relations pour des gènes et des paramètres cliniques donnés à un instant précis, au sein d'une population définie ayant des caractéristiques particulières sont multiples. Ils consistent aussi bien à identifier des biomarqueurs, qu'à aider aux diagnostics et aux prédictions/pronostics thérapeutiques. Un biomarqueur est une caractéristique, biologique et/ou clinique, mesurable liée à un processus ou à un état défini. La valeur relative à sa mesure permet de savoir si le processus ou l'état, auquel il est lié, est dans une configuration normale ou non. Par exemple, un biomarqueur peut être une protéine, tel qu'un marqueur tumoral, dosable dans le sang et permettant de diagnostiquer un cancer ou un type d'obésité [Butte 2002, Higgins 2004, Taleb *et al.* 2005].

Nous allons dans ce chapitre présenter un état de l'art relatif à différentes approches permettant d'établir l'existence de relations entre des données d'expression génique et des données biocliniques. Nous allons successivement présenter de manière générale les approches généralement utilisées par les biologistes. Puis, nous présenterons brièvement des méthodes issues de l'Intelligence Artificielle et de la Statistique. Nous ne nous limiterons pas de manière stricte au domaine de la Génomique Fonctionnelle dans l'ensemble de cet état de l'art afin d'enrichir notre analyse.

2.1 Mise en relation des données biomédicales

Dans le contexte de la recherche biomédicale, la mise en relations de données est une étape couramment réalisée au cours du processus de recherche de biomarqueurs en Génomique Fonctionnelle. Un des exemples de l'importance de la mise en œuvre de ces approches est celle exposée par [Goldner et Messier 2002]. En effet, d'après les auteurs, il est aujourd'hui clair que les interactions entre les agents biologiques, les pathologies et leur hôte (un sujet atteint ou porteur) sont

fortement dépendantes du potentiel génétique de la pathologie et influencées par l'environnement de la personne atteinte. L'étude des corrélations entre l'expression du génome d'un individu et les paramètres biocliniques d'une pathologie est un élément clé dans la compréhension de ses mécanismes et de ce qui l'influence.

Les données utilisées dans ce contexte sont :

des *données cliniques* obtenues au cours d'échanges verbaux entre soignants et patients, via des questionnaires d'enquêtes et d'évaluation de la qualité de vie et/ou socio-économiques et/ou psychologique et/ou des habitudes alimentaires et/ou des modes de consommation d'alcool, de tabac..., pendant les examens infirmiers et médicaux ;

des *données biologiques* « standards » obtenues à l'aide de mesures réalisées en laboratoire d'analyses biologiques et médicales, tels la Numération de la Formule Sanguine qui est l'examen hématologique complet quantitative et qualitative des globules rouges, des globules blancs, des plaquettes, l'insulinémie ou les cholésterols, ... ;

des *données génétiques et génomiques* obtenues par la mise en œuvre de technologies telles que :

- les *puces à ADNc* permettant de mesurer l'expression d'un très grand nombre de gènes ou de fragments de gènes simultanément,
- le séquençage dans la perspective de détecter des « *Single Nucleotide Polymorphims* » (S.N.P.) correspondant à des variations interindividuelles (polymorphismes) au niveau des gènes chez un individu donné [Levy *et al.* 2007].

Dans le cadre de nos travaux, notre objectif est de proposer une méthodologie permettant de guider l'expert vers la découverte de biomarqueurs génomiques et/ou génétiques d'états physiologiques. Ces biomarqueurs, à l'échelle moléculaire, peuvent permettre de connaître l'état d'un individu vis-à-vis d'une situation physio-pathologique (une maladie) et de définir son statut par rapport à celle-ci. Par exemple, un gène dont l'expression est liée au développement d'une maladie permettra d'indiquer en fonction de cette valeur si un patient développe ou non la maladie et de faire des pronostics sur l'évolution de cette pathologie.

Plusieurs approches sont envisageables pour définir une relation entre une gène (plus particulièrement son expression) et un état physio-phatologique. Nous avons choisi de baser notre approche sur une mesure statistique, le coefficient de corrélation des rangs de Spearman [Ancelle 2002, Motulsky 2002, Beuscart *et al.* 2009]. Nous allons justifier dans la partie suivante notre choix.

2.2 Relations entre deux ensembles de données

Les données issues de protocoles expérimentaux de recherche clinique (environnements desquels sont issues les données avec lesquelles nous travaillons) peuvent être de type continu, ordinal, binaire ou textuel. Afin de définir la relation qu'il existe entre deux ensembles de données, différentes métriques existent en fonction des types de données concernées. Ainsi, en nous limitant aux données continues (taille, poids, glycémie, expression génique, ...), ordinales (état d'avancement d'un cancer, ...) et binaires (présence/absence d'un antécédent, ...), nous allons présenter quelques unes des mesures possibles en décrivant les avantages et les inconvénients de chacune d'entre elles dans notre contexte d'application.

Il existe de manière générale deux grandes familles de mesures de relation : celles issues des Mathématiques et celles issues de la Statistique. Dans tous les cas, le choix de cette mesure de similarité dépend du type de données que l'on va avoir à traiter. Dans les sections suivantes, nous allons en décrire quelques unes (cette liste n'est pas exhaustive, mais elle reflète celles utilisées couramment en Biologie).

2.2.1 Mesures de relations en Mathématiques

En Mathématiques, la notion de relation entre les données est souvent quantifiée sous la forme d'une distance. En Algèbre, plus particulièrement, il existe différentes mesures de distances qui permettent de mettre en relation deux ensembles de données. La mesure de distance la plus connue est la distance Euclidienne [Lebart *et al.* 2000]. Cette approche est principalement applicable sur des données continues, discrètes et ordinales. Elle permet de calculer la distance entre deux points de coordonnées (x_1, y_1) et (x_2, y_2). Dans le cas de la recherche de relations entre deux ensembles de données, cette mesure n'est pas pertinente, tout comme la distance de Manhattan (aussi appelée aussi « métrique absolue ») ou celle de Cherbychev (aussi appelée distance « métrique maximum »), car elles ne permettent pas le calcul de la distance entre les ensembles, mais uniquement entre les individus décrits par les données de chaque ensemble.

Afin de calculer la distance entre deux ensembles de données binaires, la distance de Hamming est couramment utilisée en Informatique, car elle permet de connaître le nombre de couples de données divergents. Plus cette distance est petite, plus les ensembles sont identiques. Cette approche est envisageable dès lors que l'on souhaite étudier des données relatant des variations entre deux dates pour deux paramètres (par exemple, perte ou prise de poids au cours d'un régime mis en rapport avec l'expression d'un gène dans les deux situations). Dans le cadre de nos recherches, cette distance n'a pas d'intérêt, car elle ne permet pas de définir le sens et l'intensité de la variation de la relation entre les deux paramètres.

La distance de Mahalanobis [Lebart *et al.* 2000] est une mesure de distance basée sur la corrélation entre plusieurs variables. Elle est basée sur l'analyse de modèles mathématiques et statistiques. Elle permet, par exemple, de définir la similarité entre deux ensembles de données dont l'un est connu et l'autre non. De plus, elle affecte un poids à chaque composante (attribut) de l'ensemble afin de réduire l'impact des valeurs (*in extenso* des vecteurs) bruitées. La distance de Mahalanobis n'est pas adaptée à nos travaux car elle s'applique sur des valeurs de corrélations pour définir les distances au sein d'ensembles de données ayant un nombre de dimensions supérieures à deux dimensions. Pour deux dimensions, les résultats sont similaires à ceux obtenus par le calcul d'une simple valeur de corrélation ou de distance entre deux ensembles de données.

Un des formalismes mathématiques permettant de mettre en relation deux ensembles de données est celui des fonctions arithmétiques, comme les fonctions linéaires et affines [Azaïs et Bardet 2006]. Elles correspondent aux équations des droites de *régression linéaire* [Ancelle 2002, Motulsky 2002, Beuscart *et al.* 2009] des ensembles de données qu'elles décrivent. Ces approches permettent d'estimer les paramètres de ces droites. Leur utilisation est intéressante pour avoir des informations précises sur la relation existante. Néanmoins, dans le cadre de l'exploration de nombreux résultats, comme dans la recherche de relations en Génomique Fonctionnelle, la formulation de ces résultats (une fonction mathématique) n'est pas facilement et rapidement exploitable par les experts car elles se révèlent trop abstraites dans la pratique pour des non-mathématiciens.

Certaines relations peuvent être définies avec d'autres types de fonctions telles que les fonctions polynômiales, exponentielles, trigonométriques... Mais comme pour les précédentes, elles posent un problème d'interprétation. Même s'il est simple de visualiser une fonction exponentielle ou hyperbolique d'un point de vue mathématique, lui associer une interprétation biologique est souvent beaucoup plus difficile. Ainsi, plus la fonction mathématique est complexe, moins un expert non-mathématicien pourra se la représenter facilement mentalement et plus l'interprétation biologique associée sera une tâche ardue, pour ne pas dire impossible.

L'utilisation de la régression linéaire est fortement influencée par le volume et la répartition spatiale des données. Elle se révèle donc actuellement peu adaptée à la recherche de relations entre

des données d'expression génique et des données biologiques et cliniques. En effet, ces données inclues en général très peu d'individus (au maximum quelques dizaines) et leur volume et leur répartition spatiale sont très variables et difficilement réplicables. Les coefficients de régression linéaire définis dans ce contexte ne peuvent avoir qu'une valeur informative et elles n'apportent donc qu'une information partielle et subjective.

D'autres mesures de distance utilisées notamment en Intelligence Artificielle [Malandain 2006, Slimani *et al.* 2007, Rajman et Lebart 2008] sont dédiées à des données particulières comme les données symboliques ou structurées. Nous ne ferons pas mention de ces mesures ici, car elles sont trop éloignées des données que nous souhaitons traiter.

2.2.2 Mesures de relations en Statistique

Pour palier aux inconvénients des mesures précédentes, la mesure de corrélation semble être une bonne solution pour obtenir des informations relatives à la fois au sens, à l'intensité et à la distance entre deux ensembles de données. Différentes mesures de corrélation existent et plusieurs d'entre elles sont couramment utilisées en Biologie. Il s'agit des coefficients de corrélation linéaire de Bravais-Pearson (dit de Pearson), de corrélation de rangs de Spearman et de corrélation de rangs de Kendall [Nelson 2004]. Ces mesures se différencient par la nature des données qu'elles traitent [Zou *et al.* 2003]. Le coefficient de Pearson s'applique sur des ensembles constitués de données continues. Les coefficients des rangs de Spearman et de Kendall s'appliquent tous les deux sur des données ordinales et elles fournissent des résultats équivalents. Dans la pratique, en Biologie et en Médecine, le coefficient des rangs de Spearman reste le plus utilisé [Griffiths 1980, Beuscart *et al.* 2009].

Il existe d'autres mesures plus complexes telles que les mesures tétrachoriques, polychoriques, bisérielles ou de Mantel [Shannon *et al.* 2002, Nelson 2004]. Les mesures tétrachoriques et polychoriques permettent d'avoir la valeur de corrélation pour des données issues d'une table de contingence uniquement. Le coefficient de corrélation bisérielle ne s'applique que sur des ensembles de types binaire *vs.* continue. Ces mesures ne s'appliquent donc pas dans notre contexte d'étude. Le test de corrélation de Mantel [Shannon *et al.* 2002] est une statistique évaluant la corrélation entre deux matrices, les matrices devant être de même dimension. Il n'est donc pas non plus adapté à notre problème, car si nous utilisions ce coefficient pour corréler des ensembles de données à une dimension, le résultat serait équivalent à celui des approches « classiques » (e.g. *Pearson*, *Spearman* ou *Kendall*) avec des temps de calcul beaucoup plus longs.

2.2.3 Choix d'une mesure de relation

Dans le cas de données paramétriques (données quantitatives -continues ou discrettes-) ou non (données qualitatives -ordinales ou nominales-) et de données paramétriques de qualité relative en terme de précision, comme il est possible de considérer les données issues des puces à ADNc ou les données biologiques et cliniques, le coefficient de corrélation des rangs de Spearman, noté ρ_s, est la méthode qui nous semble la plus adaptée. Sa valeur est basée sur le calcul des rangs des valeurs des ensembles qui la composent et non sur les valeurs elles-mêmes. Cela permet donc, au niveau des résultats, de réduire l'impact de la qualité peu constante des données. C'est cette mesure que nous allons utiliser dans notre approche et sur laquelle nous allons concentrer le reste de notre état de l'art. Nous reviendrons sur ce choix dans le chapitre 5.

2.3 Approches empiriques en Génomique Fonctionnelle

Dans le cadre de la mise en œuvre de biopuces, les biologistes ont l'opportunité de mesurer simultanément l'expression de plusieurs milliers voir dizaines de milliers de gènes. Pour exploiter les résultats ainsi obtenus, différentes approches ont été développées. Ces approches sont :

- soit empiriques et développées au cours des expérimentations en fonction des résultats obtenus ;
- soit théoriques et basées sur des méthodes mathématiques et/ou informatiques complexes.

Les approches développées par les biologistes appartiennent en général à la première catégorie et nous allons commencer par les présenter.

Depuis quelques années, la littérature abonde d'articles relatant la découverte de relations et plus précisément de corrélations entre des données génomiques et phénotypiques lors d'études de maladies complexes et ce quelque soit le domaine de la médecine étudiée :

- Cancérologie [Ramaswamy et Golub 2002, Rosen *et al.* 2005, Labrecque *et al.* 1999] ;
- Maladies métaboliques : Diabétologie [Engeli *et al.* 2003], Enzymologie [Eisensmith *et al.* 1996] ... ;
- Neurologie [Mirnics *et al.* 2004] ;
- Ophtalmologie [Hargitai *et al.* 2005] ;
- Immunologie et virologie [Hargitai *et al.* 2005, Porter *et al.* 1994, Labrecque *et al.* 1999, Arredondo-Vega *et al.* 1998]...

Dans ces travaux, les données utilisées sont hétérogènes (plusieurs types de données), complexes (issues d'environnements aux échelles microscopique (génome) et macroscopique (individu dans son ensemble) à la fois) et les approches qui y sont présentées sont basées sur les connaissances des chercheurs. Leurs *a priori* expérimentaux sont nombreux et contraignants [Zou *et al.* 2003, Jensen *et al.* 2006]. Ils limitent leurs découvertes à l'étude de quelques gènes et/ou de quelques paramètres biocliniques. Les études menées dans ces articles peuvent se résumer en deux types d'approches :

- Les premières consistent à s'appuyer dans un premier temps sur la littérature afin d'initier ou de valider de nouvelles voies de recherche [Jensen *et al.* 2006]. À partir des connaissances recueillies par l'intermédiaire d'une étude bibliographique, dont l'abondance est relative en fonction du domaine de recherche, le génomicien réalise *a posteriori* une sélection de gènes potentiellement intéressants à étudier. Cette sélection permet de faire l'étude statistique des relations entre les 2 ensembles. Ces études sont réalisées par des logiciels dont les capacités de calculs sont limitées à des ensembles de données de dimensions réduites (tel que Microsoft Excel) ou dont la complexité demande une expertise telle, que le biologiste n'utilise qu'un nombre limité des fonctionnalités disponibles dans ces logiciels (comme dans S+ [Venables et Ripley 2002], SPSS [Beddo et Kreuter 2004], R [Ihaka 1996, R Development Core Team 2006],...). Les résultats obtenus sont au final validés par la littérature et la mise en place de nouvelles expérimentations en laboratoire.
- Le second type d'approches est basé sur la notion de présélection de gènes potentiellement « intéressants ». A partir des ensembles de données d'expression, des tests statistiques de significativité permettent de ne retenir que ceux qui sont les plus significatifs par rapport aux critères de l'expérimentateur. En général, les biologistes s'intéressent à la sur- ou à la sous-expression des gènes, donc deux groupes de gènes particuliers. Ils utilisent des logiciels comme SAM [Tusher *et al.* 2001] pour sélectionner ces gènes d'intérêt. Ces gènes sont ensuite souvent mis en relation avec d'autres types de données, comme les données biocliniques ou les annotations géniques pour découvrir des particularités dans chacun des groupes. Les résultats sont au final confrontés à ceux de la littérature afin d'être

validés, invalidés ou assimilés à une nouvelle découverte, voir expérimentés dans un contexte particulier.

Ces deux types d'approches sont manuelles et non automatiques. Les *a priori* y sont très présents de part les connaissances issues de la littérature et de l'expérience clinique des médecins-cliniciens qui participent aux recherches. Ces *a priori* permettent généralement de réduire le nombre de données à étudier afin de n'avoir au maximum que quelques gènes et qu'une dizaine de paramètres biocliniques et de permettre à ces approches essentiellement nouvelles d'être efficaces. Aujourd'hui dans les protocoles de recherche clinique, le nombre de patients reste encore inférieur à une centaine, alors que le nombre de gènes par puce à ADNc est de l'ordre de 40000 et le nombre de paramètres biocliniques est d'une centaine. Il est évident que le biologiste ne peut humainement pas étudier manuellement ce grand nombre de données. Il est donc essentiel d'automatiser cette tâche et de permettre une étude à grande échelle de ces données. Cette étude devrait idéalement étudier systématiquement toutes les relations existantes entre les données testées afin de favoriser la découverte de nouvelles connaissances. Cependant, il est important de noter que l'expert ne doit pas être mis à l'écart de cette étude par la mise en place d'un système totalement automatique, mais il doit être inclus au sein du processus de développement, de Fouille des Données et d'analyse des résultats [Langley et Nordhausen 1986, Kovalerchuk 2001b].

2.4 Approches issues des travaux en Apprentissage

Peu d'approches sont proposées en Fouille de Données pour la mise en relation de données d'expression génique et des données biocliniques au sens large. Cependant, certaines approches de classification ont été appliquées avec succès dans le domaine de la Biologie dans des contextes particuliers, comme nous allons le voir dans les paragraphes suivants.

En Génomique Fonctionnelle, les biologistes disposent de nombreuses données. Les approches totalement manuelles, telles que nous avons pu les citer précédemment, deviennent inefficaces car les dimensions des espaces des données à étudier sont de plus en plus grands. Afin d'explorer ces données, il est donc nécessaire de disposer d'outils permettant de traiter de grande quantité de données et de pouvoir découvrir des nouvelles connaissances [Hanczar *et al.* 2004]. Les outils issus de l'Intelligence Artificielle offrent de nombreuses approches, qui se différencient par le type des résultats qu'ils proposent et/ou par le mode de visualisation de leurs résultats. Plus particulièrement, l'Intelligence Artificielle offre des approches d'apprentissage supervisé et non supervisé dans le cadre de la construction de modèles de prédiction ou de classification. Dans chacune de ces familles, des approches se sont développées pour mettre en relation des ensembles de données, comme nous allons le voir dans les paragraphes suivants.

2.4.1 Approches supervisées

La première famille des méthodes de classification sont les supervisées [Caruana et Niculescu-Mizil 2006, Han et Kamber 2006, Kotsiantis 2007]. Elles nécessitent des connaissances *a priori* sur les données que l'on souhaite traiter [Slonim 2002]. En effet, ces méthodes permettent la construction d'un modèle à partir de données dites d'apprentissage. Ce modèle est ensuite testé sur un autre ensemble de données (les données tests). Parmi les approches supervisées appliquées aux domaines de la Biologie, pour la recherche de relations, les plus courantes sont les k-plus proches voisins (k-ppv), les machines à vecteurs de support et les arbres de décision. Beaucoup d'autres méthodes existent, mais elles ne sont pas décrites ici [Han et Kamber 2006].

L'approche k-plus proches voisins (k-ppv) est une méthode simple, qui consiste à affecter un individu à la classe dont les autres individus ont des propriétés qui lui sont les plus proches. Cette

méthode permet de regrouper rapidement des données dans des groupes distincts. L'un de ses inconvénients majeurs est que l'utilisateur doit avoir une bonne connaissance des données qu'il a à traiter afin de déterminer le nombre de groupes que l'algorithme devra utiliser. Dans le cas des données que nous avons à traiter, nous ne connaissons pas cette information. De plus, comme nous devons traiter des données en deux dimensions (expression génique *vs.* données biocliniques), les groupes de gènes que l'on va découvrir auront une cohérence « numériquement » parlante, car proche du point de vue de la distance euclidienne, et non une « évolution » commune.

Les machines à vecteurs de supports (*Support Vector Machine,* SVM) [Vapnik 1998] permettent de séparer, en au moins deux classes, des données définies dans un espace à n dimensions à l'aide d'une droite (2 dimensions) ou d'un hyperplan (n dimensions). La séparation des groupes s'effectue en maximisant la distance entre les classes et l'hyperplan. En Génomique, les SVM [Berrar 2003, Temanni *et al.* 2005] sont souvent utilisées dans un but prédictif par la création de classes d'individus sains et non sains. Les SVM ne sont pas adaptées pour l'aide à la découverte de relations entre les données d'expression génique et les données biocliniques, car les relations que nous cherchons ne sont pas des relations de proximité entre les individus mais entre les paramètres eux-mêmes.

L'Analyse Discriminante Linéaire [Balakrishnama et Ganapathiraju 1998, Sengur 2008], n'est pas de manière issue des travaux en Apprentissage ; elle est issue de la Statistiques mais est très proche des SVM qui sont eux issus des travaux en Apprentissage. Elle consiste à calculer une ligne droite ou un hyperplan afin de séparer au mieux deux classes connues. Cette séparation est réalisé de telle sorte que la variation intra-classe soit minimale et la variation inter-classe soit maximale. Un individu inconnu est affecté à la classe dont il est le plus proche en terme de caractéristiques. Cette approche n'est pas adéquate dans le cadre de nos travaux, car elle permet de « classer » des individus, donc de « trouver » des relations de proximité en terme d'individus, elle ne peut donc être performante que dans le cas où les données sont séparables linéairement [Dudoit *et al.* 2002].

Les travaux de Tan et Gilbert [Tan et Gilbert 2003] s'intéressent aux approches d'études systématiques de recherche de relations et de corrélations entre les profils d'expression de gènes et les différents états d'une pathologie, comme par exemple l'évolution d'une tumeur dans le cadre d'un cancer. Comme nous l'avons nous-même noté sur nos données [Benis 2003], les auteurs soulignent la bonne classification des patients en fonction de données biologiques avec des arbres de décision tel que C 4.5 [Quinlan 1993, Quinlan 1996] d'une part, mais aussi à partir de l'expression de leurs gènes d'autre part. Les résultats de ces travaux montrent que la combinaison de techniques, tels que C 4.5, le « bagging », et le « boosting », permet d'obtenir de meilleurs résultats de classifications pour la discrimination des individus sains et non sains. Ces résultats soulignent l'importance du calcul des corrélations pour définir des relations entre des valeurs d'expression génique et un attribut « binaire » modélisant un statut vis-à-vis d'une maladie. Les arbres de décision sont ainsi utilisés en routine en clinique dans le cadre du diagnostic différentiel d'une pathologie. Cependant, cette approche est fortement dépendante de la qualité des données initiales. Or, les données issues des puces à ADNc et les données biocliniques sont de qualité très variable, cela peut donc avoir un fort impact sur la qualité des arbres construits.

Les approches d'apprentissage supervisé ont permis de montrer l'intérêt de la mise en corrélation des données d'expression génique et des données biocliniques dans le cadre de l'étude de pathologies complexes [Tan *et al.* 2003, Temanni *et al.* 2005]. Cependant, dans le cadre de nos expérimentations, nous n'avons pas d'informations sur l'appartenance à une catégorie ou à un état binaire des individus. Nous ne pouvons donc pas appliquer ce type d'approches.

2.4.2 Approches non supervisées

Les méthodes non supervisées [Cornuéjols *et al.* 2002, Kotsiantis et Pintelas 2004, Han et Kamber 2006] ne nécessitent pas de connaissances préalables pour classer les données. On trouve parmi elles notament : les approches telles que les cartes auto-organisatrices et les classifications conceptuelles. Nous allons brièvement présenter ces approches.

Suivant le principe des k-plus proches voisins, il y a aussi les algorithmes des k-moyennes [MacQueen 1967, Wu 2008] et des k-médoïdes [Kaufman et Rousseeuw 1990]. Les k-moyennes permettent de classer les données en minimisant la distance entre les objets (les individus) qui appartiennent à une même classe et en maximisant la distance entre les classes. Le nombre de classes, k, doit être défini *a priori* par l'utilisateur. Ces approches sont très utilisées en Génomique car elles permettent de découvrir des gènes qui ont le même comportement, c'est-à-dire s'expriment de la même manière. Cependant, elles ne sont pas performantes dans l'absolu, car elles sont très sensibles aux minima locaux. En effet, le centre de chacune des k classes est défini de manière aléatoire et donc les membres de chaque classe ne sont pas forcément les mêmes à chaque application de l'algorithme, ce qui introduit une incertitude dans les résultats obtenus. Les k-médoïdes ont pour centre de leurs groupes des données, ce qui leur permet d'être moins sensible à cette phase d'initialisation. Bien que les k-moyennes et les k-médoïdes permettent de classer les bi-ensembles (données d'expression génique vs. données biocliniques) et donc de manière indirecte de faire apparaître des types de relations, elles ne répondent pas à notre problématique car k n'est pas connu à l'avance [Tanay *et al.* 2002]. En effet, comme nous l'avons expliqué précédemment, il est nécessaire de définir en amont la valeur de k.

Les cartes auto-organisatrices de Kohonen [Kohonen 1997, Lebbah 2003] correspondent à un autre type d'approches de classification non supervisée. Elles s'appuient sur le concept de réseaux de neurones et dérivent de l'algorithme des k-moyennes sur lequel sont ajoutées des contraintes spatiales pour permettre de réduire l'espace des données à explorer. Elles permettent d'une part de définir des ensembles d'individus qui se ressemblent et d'autre part de proposer un mode de visualisation accessible aux non-informaticiens et aux non-mathématiciens. Bien que cette approche classe, elle aussi, les individus en fonction de leurs « ressemblances », elle ne permet pas de définir de manière directe des relations (et plus précisément le type de celles-ci). Il est nécessaire d'étudier de manière spécifique chaque cellule (classe au sein des cartes auto-organisatrices de Kohonen) et son voisinage afin de comprendre les relations découvertes.

Plus proches des techniques d'analyses multidimensionnelles de données, les classifications ascendantes hiérarchiques (CAH) [Struyf *et al.* 1997, Cornuéjols *et al.* 2002, Rousseeuw et Leroy 2003, Pakhira 2008], peuvent être assimilées à des approches d'Appentissage, de part l'absence d'intervention extérieur lors de processus de classification. Les biologistes l'utilisent couramment et en connaissent un version sous le nom de « méthode d'Eisen » [Eisen *et al.* 1995, Eisen 1999]. Elle a l'avantage de mettre en évidence visuellement des ensembles de gènes ayant le même profil d'expression. Ainsi, le génomicien dispose d'un outil simple de visualisation des profils et de l'expression des gènes. L'inconvénient majeur d'une CAH, dans notre cas d'application, est que le résultat fourni prend uniquement en compte un type de données (l'expression génique), aucun lien n'est fait avec d'autres sources d'informations. Néanmoins, le fait de disposer d'un mode de visualisation intuitif et ludique [Kovalerchuk 2001b] a fait de cette approche son succès. Les biologistes se sont appropriés l'outil en lui faisant indirectement « perdre » sa référence initiale à l'Intelligence Artificielle, domaine au sein duquel cette approche a initialement été développée.

Les travaux de Courtine [Courtine 2002] s'intéressent à la construction automatique non supervisée de classifications dans des treillis de Galois. Cet espace permet de mettre en évidence toutes les relations existantes entre les données, quelles soient numériques ou symboliques.

Parmi les expérimentations réalisées dans ces travaux, on retrouve la mise en relation de données génomiques (*Single-Nucleotide Polymorphism* ou SNP) et des données biologiques et cliniques (IMC, Glycémie,...). L'avantage de cette approche est la mise en évidence de toutes les relations « symboliques » existantes entre les données, ce qui peut permet de simplifier la discussion des résultats obtenus entre des chercheurs de domaines différents [Mondada 2005]. Néanmoins, les inconvénients sont multiples : la taille de l'espace généré est grand (de l'ordre de 2^n où n est le nombre d'objets), les relations existantes entre les différentes mutations d'un même SNP ne sont pas prises en compte et il n'y a pas de mise en relation réelle avec les données biocliniques. L'ensemble des données, quelles que soit leur nature, ont le même rôle dans la classification et les liens biologiques entre les deux types de données sont donc difficilement interprétables/identifiables.

L'approche proposée par Besson [Besson 2005] est proche de celle que nous venons de décrire. Néanmoins, elle pose le problème de la discrétisation des données afin de découvrir des motifs avec le même comportement. Ces travaux se déroulent en deux temps : le premier consiste à préparer les données grâce à une discrétisation binaire des informations ; la seconde correspond à la phase d'extraction des motifs sous contraintes, c'est-à-dire répondant à des critères de formes et de structures. L'un des avantages de cette approche est de permettre la découverte de sous-ensembles de données présentant des caractéristiques proches. Cependant, les inconvénients sont multiples. Tout d'abord, il s'agit d'une représentation binaire d'informations complexes, le langage de représentation des résultats n'est pas intuitif pour les experts du domaine. De plus, l'expert doit intervenir régulièrement au cours du processus pour valider les choix du système, ce qui accroît le risque d'introduire des « biais de connaissance » qui peuvent être importants dans l'analyse.

La famille des méthodes de classification non supervisée que nous venons de présenter sont quelques unes des approches utilisées couramment en Biologie et en Génomique Fonctionnelle. Nous avons volontairement limité notre présentation à ces approches. Ces méthodes ont la particularité de permettre l'aide à la découverte de relations entre des données d'expression génique et des données biologiques ou cliniques. Cependant, elles permettent uniquement de classer les individus qui ont un comportement proche, c'est-à-dire qui varient presque de la même manière pour un ensemble de paramètres donnés, et elles ne décrivent pas la relation découverte entre les deux ensembles de données.

2.5 Approches statistiques

Les approches statistiques sont couramment utilisées pour définir les relations entre des données issues des puces à ADNc et des données biocliniques. Il existe deux grandes catégories d'approches. Dans la première, nous trouvons d'une part les approches permettant de calculer directement une statistique et d'autre part celles basées sur l'Analyse en Composantes Principales (ACP). Dans la seconde catégorie, s'inscrivent des approches qui ne sont pas de manière explicite dédiées à la définition de relations, mais qui ont pour but de définir la significativité de celles-ci. Nous concentrons nos propos sur une mesure statistique particulière : la corrélation, puisqu'il s'agit de la mesure que nous avons choisi d'appliquer dans notre contexte (voir section 2.2).

2.5.1 Approches basées sur les corrélations en Biologie

Une puce à ADNc permet la mesure en simultané, dans un contexte expérimental donné, de l'expression de plusieurs dizaines de milliers de gènes ou de fragments de gènes. Les biologistes

qui utilisent cette technologie mettent en œuvre différentes approches statistiques et empiriques leur permettant de filtrer les données avant de les utiliser dans leur analyse.

Dès 1999, les premières recherches systématiques d'association entre des données d'expression génique issues de puces à ADNc et des données cliniques ont eu lieu. Leur but était de découvrir des corrélations entre ces deux ensembles de données [Stratowa *et al.* 1999]. Les auteurs anticipaient les résultats et ils étaient positivement concluant. Ainsi bien que l'article suppose l'étude systématique d'un grand nombre de données, il est important de souligner qu'en réalité le nombre de gènes étudiés n'était que de 1024, alors que les puces étudiées en contenaient 10 000.

Une seconde approche très utilisée par les biologistes pour mettre en relation des données d'expression génique et des données biocliniques est l'Analyse en Composantes Principales (ACP) [Jiang *et al.* 2004]. L'objectif de l'ACP est de proposer une mise en relation n *(*$n \geq 2$*)* d'ensembles de « points » à l'aide du calcul par paire de corrélations. Si l'on souhaite étudier les relations entre

- 3 ensembles X_1, X_2, X_3, on doit calculer les corrélations (X_1, X_2), (X_1, X_3), (X_2, X_3).
- 4 ensembles X_1, X_2, X_3, X_4, on doit calculer les corrélations (X_1, X_2), (X_1, X_3), (X_1, X_4), (X_2, X_3), (X_2, X_4), (X_3, X_4).

Ainsi, plus le nombre d'ensembles initiaux est important, plus le nombre de corrélations à calculer est grand ; cela n'est pas un inconvénient en soi, au vue des puissances de calculs que nous avons aujourd'hui. Néanmoins, dans le cadre de nos travaux, le but est d'aider à la découverte de corrélations entre des données d'expression génique et des données biocliniques. Or, l'ACP dépasse cet objectif en terme de nombres de calculs nécessaires à effectuer. En effet, l'ACP calculera en plus les corrélations gènes *vs.* gènes et paramètres biocliniques *vs.* paramètres biocliniques, alors que seules les corrélations gènes *vs.* paramètres biocliniques ont de l'intérêt pour nos études. Cependant, l'avantage de l'ACP est d'obtenir des variables explicatives (composantes principales). Les n premiers sont les plus intéressantes, c'est-à-dire celles qui permettent le mieux d'expliquer les relations. Par rapport à notre problématique, cette approche pose un problème extrême d'interprétation des résultats, car les variables explicatives sont des variables composées et complexes qui ne sont pas toujours facilement et rapidement compréhensibles par des novices.

L'étude des relations à partir de puces à ADNc a aussi été faite avec des données biocliniques, mais aussi avec des données relatives aux réseaux de régulation biologique et génique et à la Génomique (comme le positionnement des gènes sur les chromosomes) [Yamanishi *et al.* 2003, Bhardwaj et Lu 2005]. Par exemple, Yamanishi et son équipe ont proposé deux approches basées sur l'analyse canonique des corrélations [Yamanishi *et al.* 2003]. La première, nommée « Multiple Kernel Canonical Correlations Analysis », traite des relations entre n types de données (c'est-à-dire Réseaux de régulations *vs.* Génome *vs.* Expression). La seconde « Integrated Kernel Canonical Correlations Analysis » tend à définir des relations entre n ensembles au regard d'un autre (c'est-à-dire Réseaux de régulations *vs.* Génome et expression génique). Ces approches sont intéressantes dans le cadre de la découverte de nouvelles connaissances car elles permettent d'avoir une vision globale des relations entre réseaux de régulations biologiques et/ou géniques, données génétiques et d'autres types de données. Cependant, cette méthode n'est pas applicable sur de très grands ensembles de données, comme ceux dont nous pouvons disposer. En effet, un problème d'explosion combinatoire en temps de calculs et en nombre de relations potentiellement découvertes apparaît dès que les ensembles de données sont trop grands.

Bhardwaj et Lu [Bhardwaj et Lu 2005] présentent une approche de mise en relation entre des données d'expressions géniques et des interactions protéines-protéines. Ils proposent son utilisation dans le cadre d'études phylogénétiques en calculant la distance de corrélations entre les espaces. Ces travaux présentent l'avantage de s'intéresser à des relations entre de données complexes qui intègrent aussi bien des données issues de la Génomique que de la Protéomique.

Néanmoins, ils exigent au minimum une double expertise pour l'analyse des résultats, ce qui n'est pas aujourd'hui chose courante dans les laboratoires de Biologie.

Pinto [Pinto *et al.* 2005] propose une approche d'analyse intégrant des données d'expression et des annotations « Gene Ontology » (GO) [Gene Ontology Consortium 2001, Gene Ontology Consortium 2004]. Leur objectif est d'interpréter des valeurs géniques grâce aux annotations et vice versa. La méthode proposée consiste à déterminer les valeurs des coefficients de corrélations locaux correspondant à l'intensité des relations entre les distances de valeurs d'expressions et celles existants entre les fonctions GO. Les auteurs montrent qu'une valeur de corrélation négative est intéressante au sens de cette méthode, car elle permet de définir de nouvelles d'hypothèses biologiques. Les résultats obtenus via cette approche sont plus intéressants que ceux obtenus par des méthodes d'enrichissement classiques [Al-Shahrour *et al.* 2004]. L'inconvénient de cette approche est que seules les données issues des puces à ADNc et de l'ontologie GO sont analysables. Néanmoins, cette approche démontre l'intérêt de pratiquer une exploration à grande échelle des données issues des biopuces en les mettant en relation avec d'autres types de données.

Il existe dans la littérature d'autres approches permettant, dans le cadre de l'exploration de données, de définir des relations pouvant exister entre deux ensembles de données, telles que des données génomiques et des données biocliniques. Néanmoins, il n'existe pas à notre connaissance, pour ce type de données, d'approches permettant la découverte automatique de corrélations potentiellement intéressantes à grande échelle.

2.5.2 Approches basées sur les corrélations dans d'autres domaines

Afin d'élargir notre état de l'art, nous allons présenter quelques approches utilisant le calcul de corrélations sur des données issues d'autres domaines d'application dans les paragraphes suivants.

Des travaux ayant pour but d'étudier la qualité des signaux transmis au-dessus des océans ont cherché à définir des relations linéaires entre le niveau des océans et la pluviométrie [Natarajakumar *et al.* 2004]. Des résultats intéressants ont été obtenus avec un très grand nombre de mesures (plusieurs milliers) pour chaque paramètre sur quelques sites. Cette approche ne peut s'appliquer directement aux données issues des puces à ADNc, puisque nous avons un nombre réduit de mesures pour chaque paramètre, mais elle nous montre que ce type d'algorithme peut être utilisé avec profit.

Les Sciences Économiques et plus particulièrement le domaine de la gestion des relations clients ont été le terrain de prédilection pour développer des approches de « découverte de corrélations linéaires ». Ce domaine [Chiang *et al.* 2005] a pour but de mettre en évidence des corrélations globales ou locales « cachées » dans de grands ensembles de données. Le principe de l'approche consiste à « associer » par paire, automatiquement, tous les attributs d'une base de données afin de détecter des groupes d'attributs ayant des corrélations linéaires potentiellement intéressantes entre eux. C'est en nous inspirant de cette approche, que nous avons développé notre méthodologie d'aide à la découverte de corrélations en Génomique Fonctionnelle, comme nous allons le voir dans le chapitre de problématique.

2.6 Conclusion

Les différentes approches que nous avons présentées dans ce chapitre montrent un échantillon de la diversité des méthodes permettant de mettre en relation des données. Néanmoins, nous avons pu souligner l'importance de la prise en compte du domaine applicatif des données que l'on souhaite traiter dans de nombreux cas. Les données issues de la Génomique Fonctionnelle sont particulières et requièrent un choix judicieux de la mesure de distance à utiliser. En effet,

trop de précision risque de ne pas permettre de découvrir des résultats « intéressants » et des biomarqueurs, tout comme l'utilisation d'un outil trop laxiste risque de « masquer » les vraies découvertes.

Chapitre 3

Valeurs singulières

Dès lors que nous sommes amenés à traiter des données réelles, la question des valeurs singulières se pose. En effet, ces valeurs peuvent influencer les résultats issues du traitement des données initiales et donc les conclusions que les experts peuvent en extraire. Le développement et la généralisation des moyens informatiques et plus particulièrement ceux relatifs aux bases de données ont permis à de nombreux domaines de la Recherche et de l'Industrie de stocker en masse des données. Afin d'exploiter au mieux ces données, l'utilisation de procédures automatiques d'analyse est primordiale et se révèle de plus en plus répandue même si cela pose des problèmes au niveau de la qualité des données initiales [Planchon 2005]. De manière générale, la recherche et la définition de valeurs suspectes s'inscrit dans le processus de pré-traitement des données qui doit être pris en compte lors de l'analyse des résultats.

Dans le contexte de la Génomique Médicale Fonctionnelle, les valeurs singulières peuvent permettre d'identifier aussi bien des individus ne présentant pas de caractéristiques communes avec la population étudiée que des résultats incohérents et/ou inconsistants pour un type de mesures données (examen biologique, réponse à un questionnaire,...). Il est donc essentiel de les prendre en compte.

Dans ce chapitre, nous présenterons différentes approches existantes pour traiter les valeurs singulières au sein d'ensembles de données. Ces approches permettent de prendre en compte les valeurs suspectes au cours, avant ou après, l'analyse d'un ensemble de données.

3.1 Définition de valeurs singulières

La notion de « valeur singulière » est très vaste, elle désigne à la fois des *valeurs aberrantes*, des *valeurs suspectes* et du *bruit*. Les *valeurs aberrantes* sont des valeurs qui ne semblent pas être cohérentes par rapport au reste d'un ensemble de données. Elles ont un impact important sur l'analyse des données car elles influencent fortement les valeurs issues de tests statistiques ou d'algorithmes de Fouille de Données. Les *valeurs suspectes*, quant à elles, sont des données qui semblent ne pas être totalement cohérentes par rapport à un ensemble de données, car les valeurs divergent mais trop peu pour être considérées comme des valeurs aberrantes. Les valeurs suspectes peuvent être vues comme des valeurs qui influencent de manière très modérée les résultats. Le *bruit* est l'information issue d'un ensemble de données qui est induit au cours de la génération ou du traitement sur les données. Il se distingue des valeurs aberrantes et des valeurs suspectes par le fait qu'il est « noyé » dans les données et qu'il n'est que difficilement détectable. Le bruit correspond à des valeurs présentes de manière diffuse dans le reste des données. Il peut donc avoir une influence plus ou moins importante sur les données en fonction de son intensité.

La figure 3.1 présente un exemple simple de la notion de valeur singulière. Dans cet exemple, les données regroupées en haut à droite sont des données dites « normales » et alors que celle en bas à gauche est une « valeur singulière ». De manière plus précise, il s'agit d'une valeur aberrante, car elle est fortement divergente du modèle qui forme les données dites « normales ».

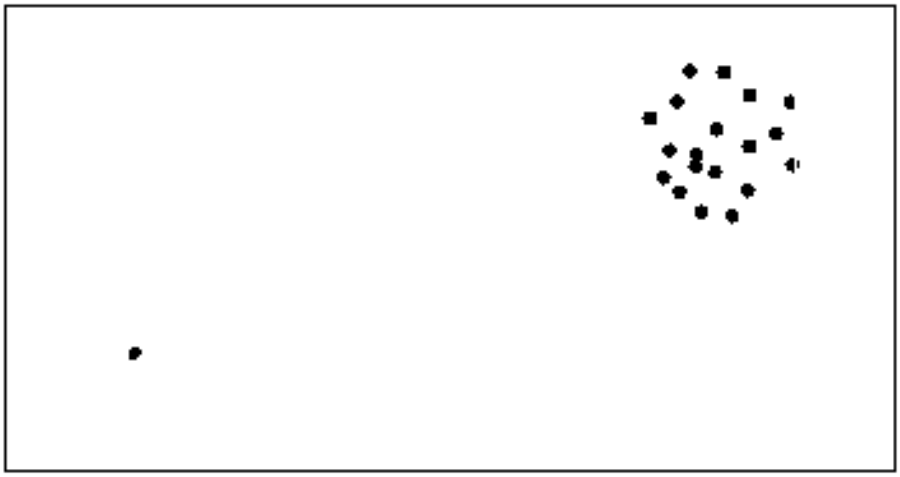


FIGURE 3.1 – Données bivariées contenant une « valeur singulière » de type « valeur aberrante » (en bas à gauche).

Peirce [Peirce 1852] est le premier à s'intéresser au problème des valeurs singulières (*outliers* en anglais).

> *« Il y a des observations qui diffèrent tellement des autres qu'elles servent uniquement à rendre l'expérimentateur perplexe et à l'induire en erreur. »*

Au fur et à mesure des années, la définition de la notion de valeur singulière a évolué :

Grubbs [Grubbs 1969] considère une valeur singulière comme étant une observation qui semble dévier par rapport à l'ensemble des autres membres d'un échantillon ;

Carletti [Carletti 1989] propose de considérer les valeurs singulières comme étant des données s'écartant d'une façon « importante » des autres valeurs de l'ensemble étudié ou ne semblant pas respecter une relation définie ;

Munoz-Garcia, Moreno-Rebollo et Pascual-Acosta [Munoz-Garcia *et al.* 1990] rajoutent une dimension contextuelle et proposent d'éviter la subjectivité en ajoutant la condition que l'observation dévie nettement du comportement général vis-à-vis du critère sur lequel l'analyse est effectuée ;

Barnett et Lewis [Barnett et Lewis 1994] ont défini, pour leur part, un ensemble de réflexions à avoir avant de faire une analyse de valeurs singulières. Il faut ainsi, selon eux, se questionner sur la distinction :

- entre les causes déterministes ou aléatoires d'apparition de ces valeurs, c'est-à-dire si leurs causes et leurs localisations peuvent être « prédites » ou non ;
- entre les différents objectifs à atteindre lors de l'étude de ces valeurs, c'est-à-dire si l'on cherche à les abstraire par d'autres opérations sur les données ou à les signaler à l'expert ;
- entre les différents modèles de probabilités spécifiques, c'est-à-dire que si les données sont distribuées suivant un modèle particulier, une valeur singulière est alors une valeur qui ne respecte pas ce modèle et donc en diverge ;

- entre les données univariées et multivariées, c'est-à-dire si les données sont étudiées suivant un ou plusieurs attributs avec une ou plusieurs dimensions ;
- entre les valeurs singulières simples ou multiples, c'est-à-dire si l'on cherche à définir une ou plusieurs valeurs singulières simultanément ou seulement une seule.

Everitt [Everitt 2002] précise la définition s'appuyant sur les modèles de probabilité. Les valeurs singulières sont des observations déviant de manière importante du reste de la population. Il s'agit de valeurs inconsistantes par rapport à un modèle probabiliste supposé connu.

Toutes ces définitions généralistes s'accordent donc à dire qu'une valeur singulière est, dans un ensemble d'observations, une observation semblant être inconsistante par rapport aux autres données. Elle est caractérisée par son impact sur l'observateur et/ou sur l'analyste, et non son impact sur les résultats d'analyse eux-mêmes.

3.2 Sources et effets des valeurs singulières

Un élément important dans la définition d'une valeur singulière est la détermination et la caractérisation de son origine. D'après Barnett et Lewis [Barnett et Lewis 1994], les valeurs singulières peuvent avoir différentes origines :

des variabilités inhérentes aux caractéristiques naturelles de la population étudiée, chaque individu a naturellement des comportements divergents des autres, c'est ce qui fait la diversité d'une population (diversité génétique, par exemple) ;

des manipulations ou des observations expérimentales incorrectes ou imprécises réalisées par un opérateur humain ou un système automatique ;

une acquisition erronée des résultats d'expérimentations ou d'observations réalisées par un opérateur humain ou un système automatique ;

de mauvaises sélections des individus inclus dans la population étudiée par manque de précision de leurs caractéristiques dans le plan d'expérience ou le protocole de recherche ;

des évènements exceptionnels ou nouveaux, par exemple, des modifications ponctuelles dans le comportement d'un individu ou une mutation génique.

Ces différentes sources qui induisent l'apparition de valeurs singulières sont de nature différente et montrent le niveau de complexité de l'analyse de ces valeurs. De plus, ces sources peuvent avoir différents types d'effets [Werner 2003] :

additif qui sont des valeurs ponctuelles qui n'influent que ponctuellement sur les données, c'est-à-dire qu'elles ne vont modifier les résultats d'une analyse que sur un intervalle de données restreint ;

innovant qui sont des valeurs qui peuvent sembler « singulières » et qui influencent ponctuellement les données mais qui ont une validité au sens où elles correspondent à des données correctes et nouvelles, c'est-à-dire qu'elles sont assimilables aux valeurs additives à la nuance près que ces valeurs sont vraies ;

de changements d'échelle ou de référence qui sont des valeurs pouvant influer toutes les données en y introduisant un décalage. Ce type de valeurs se retrouve généralement dans des ensembles temporels ou d'appareils de mesures ou d'opérateurs différents. Elles correspondent aussi bien à des valeurs singulières additives, qu'innovantes et elles induisent un changement localisé dans la distribution des valeurs. Ces valeurs singulières peuvent être vues comme des résultats contextuels.

Le rejet inconsidéré des valeurs singulières peut avoir des conséquences non négligeables pour l'analyse ultérieure de l'échantillon puisque ce dernier n'est plus considéré comme « aléatoire » mais il devient un *échantillon censuré* [Ben-Gal 2005]. Nous retrouvons donc ce phénomène lorsqu'il y a remplacement des données rejetées par des équivalents statistiques (moyenne, médiane,...). Dans tous les cas, les résultats des analyses seront fortement influencés par les valeurs manquantes ou substituées, ce qui pourra conduire à de « fausses » nouvelles découvertes.

Afin de prendre en compte la présence de valeurs singulières dans les données étudiées, l'analyse de celles-ci doit donc faire preuve d'« adaptation », de « réserve » et de « mesure ». Il peut être nécessaire d'utiliser des méthodes d'analyses robustes [Hampel *et al.* 1986] pour minimiser l'impact des valeurs singulières [Cleveland 1979]. Cela revient à utiliser des procédures statistiques qui ne recherchent pas les valeurs singulières mais qui vont chercher à réduire leurs impacts lors du calcul de valeurs statistiques [Rousseeuw et Leroy 2003].

Il est important de noter que les valeurs singulières ne sont pas obligatoirement des valeurs fausses ou erronées et ne sont pas non plus obligatoirement des sources d'erreurs. Leur traitement ne doit donc pas être le sujet d'une attitude radicale qui tend à les éliminer ou à les conserver systématiquement, mais il faut autant que possible éviter toute perte de données pouvant apporter une information sur la qualité des données ou toute contamination pouvant masquer des informations « pertinentes ». Il est donc important de comprendre la source des valeurs singulières que l'on peut avoir afin de les gérer au mieux. Ainsi, dans le contexte applicatif de nos travaux, la Génomique Médicale Fonctionnelle, il est donc important de prendre en considération la qualité « relative » du matériel biologique et des manipulations qui sont faites lors de la mise en œuvre des puces à ADNc.

3.3 Types de valeurs singulières

Ils existent deux grands types de valeurs singulières : les *univariées* et les *multivariées*. Ces données ne peuvent pas être traitées de la même manière car elles présentent une complexité différente [Shneiderman et Plaisant 2004]. Les données univariées correspondent à des données sur une dimension alors que les données multivariées correspondent à des données sur plusieurs dimensions.

3.3.1 Valeurs singulières univariées

D'après Barnett et Lewis [Barnett et Lewis 1994], les valeurs singulières univariées peuvent être de deux types : les valeurs aberrantes et les valeurs suspectes de type contaminant.

Soit x_1, x_2,..., x_n un échantillon aléatoire univarié de taille n provenant d'une distribution F, et soit $x_{(1)}$,$x_{(2)}$,..., $x_{(n)}$ les données ordonnées dans l'ordre croissant ; les valeurs $x_{(1)}$ et $x_{(n)}$ sont les observations extrêmes inférieures et supérieures. Une observation extrême est une *valeur singulière* si et seulement si elle diverge du modèle F. Une valeur *aberrante* est toujours une valeur extrême de l'échantillon. Si toutes les observations ne proviennent pas de la distribution F mais d'une distribution G, de moyenne plus élevée que F, les observations de G sont considérées comme des *contaminants*. De tels contaminants peuvent apparaître comme étant extrêmes mais ce n'est pas forcément le cas. Néanmoins, si $x_{(n)}$ est extrême et contaminant, il n'est pas forcément une valeur aberrante en soi. Une valeur semblant aberrante peut donc être la manifestation de la présence d'un contaminant. Ces diverses situations indiquent la complexité de l'étude de valeurs singulières et la difficulté de définir le type d'observations rencontrées de manière précise. Une *valeur suspecte* correspond, selon Barnett et Lewis [Barnett et Lewis 1994], à une valeur moins extrême qu'une valeur jugée aberrante de manière statistique.

Parmi les valeurs singulières univariées, il y a aussi les *observations influentes* [Everitt 2002]. Ce sont des observations qui ont une influence disproportionnée sur un ou plusieurs estimateurs de paramètres, en particulier, sur les coefficients de régression. Cette influence peut être due à des différences par rapport aux autres observations de la variable explicative ou à une valeur extrême de la variable à expliquer. Everitt souligne que les valeurs aberrantes sont souvent des observations influentes. Selon Cook et Weisberg [Cook et Weisberg 1980], les observations influentes sont celles pour lesquelles les caractéristiques de l'analyse sont altérées de manière considérable quand elles sont supprimées.

3.3.2 Valeurs singulières multivariées

La valeur singulière multivariée, contrairement au cas univarié, est un concept difficile à définir. Elle est moins apparente intuitivement, car elle est effectivement cachée dans une masse de données et se trouve en général en périphérie d'un nuage de points [Afifi et Azen 1979]. Ainsi, les points qui ne se trouvent pas à l'intérieur du nuage de points sont des valeurs potentiellement *aberrantes*.

Un exemple d'ensemble de données multivariées contenant des valeurs singulières est présenté à la figure 3.2. Les points considérés ici, dans un contexte multivarié (figure 3.2 C), comme des valeurs singulières, sont issus de données qui respectivement ne sont pas des valeurs singulières chacune dans leur dimension respective, comme le montre les figures 3.2 A et B.

D'après Gnanadesikan et Kettering [Gnanadesikan et Kettering 1972], les conséquences de la présence de valeurs singulières dans un échantillon multivarié sont bien plus complexes que dans le cas univarié. En effet, la valeur singulière multivariée peut déformer non seulement les mesures de position et d'échelle mais également les relations entre les variables. Il est donc aussi important de rechercher les valeurs singulières dans les dimensions univariées qu'en multivariées (combinaisons de dimensions) afin de réduire le risque de leurs « non-détections » et donc de leurs influences.

3.4 Identification de valeurs singulières

Pour caractériser des données et définir les valeurs singulières, il est possible de s'appuyer aussi bien sur des approches visuelles [Caussinus *et al.* 2003] que sur des approches calculatoires issues de la Statistique ou de la Fouille de Données [Ben-Gal 2005]. Nous allons dans cette partie faire un tour d'horizon des différentes approches classiquement utilisées pour la découverte de valeurs singulières et voir si ces approches peuvent s'adapter aux données que nous souhaitons traiter, c'est-à-dire des données dans lesquelles on a peu d'exemples et beaucoup d'attributs.

3.4.1 Analyse des distributions

Approches paramétriques

Les approches paramétriques [Nelson 2004] permettent de supprimer les valeurs singulières de données en se basant sur la *distribution* statistique de l'ensemble des données. Les données sont considérées dans une perspective univariée et s'appuyant sur une distribution « standard » (c'est-à-dire selon, par exemple, une loi Normale ou une loi de Poisson). Dans ce contexte, les valeurs singulières sont définies comme des valeurs qui divergent des autres données, c'est-à-dire qui ne s'inscrivent pas dans la variation générale des données, car elles ne suivent pas la loi de distribution qui correspond à la majorité des données. Cette divergence peut être quantifiée

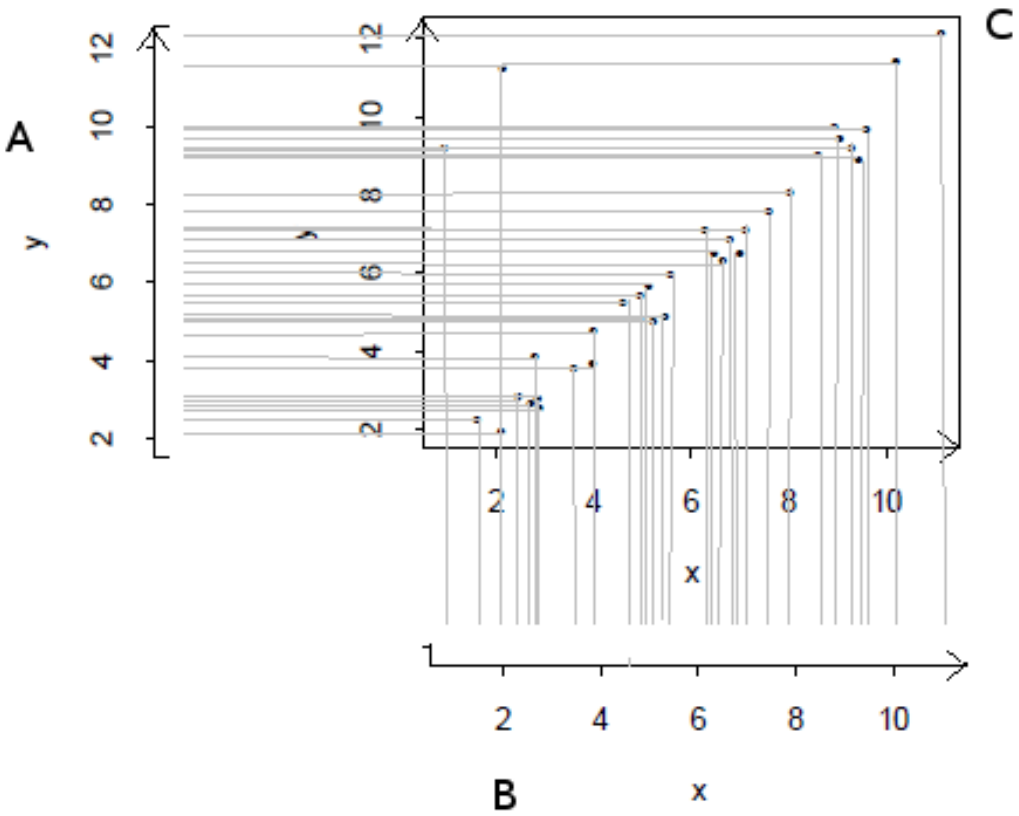


FIGURE 3.2 – Données contenant des « valeurs singulières » multivariées [Werner 2003].

par une valeur numérique dépendante de la loi de distribution utilisée. Par exemple, dans le cas d'une loi Normale, le test de Kolmogorov-Smirnov [Chakravarti *et al.* 1967] permet de vérifier l'adéquation des données à une distribution normale (figure 3.3). Le *z-score* est un autre exemple de test permettant de définir les seuils à partir desquels les valeurs sont considérées comme singulières, donc exclus. La figure 3.3 montre comment le choix d'une seule peut influencer sur l'exclusion de valeurs lors d'une distribution selon une loi Normale.

Pour résumer, avec les approches basées sur les lois de distributions, les observations susceptibles d'être déclarées comme des valeurs singulières sont toujours des valeurs extrêmes. Or, dans le monde réel, ces dernières ne sont pas forcément des valeurs singulières. En Génomique Médicale Fonctionnelle, c'est cependant ce type d'approches qui est le plus souvent mis en œuvre.

Approches non paramétriques

Généralement, les données issues du monde réel ne reposent pas sur des lois de distribution classiques telles que la loi Normale ou celle de Poisson. Pour pallier à ce type de problème, la mise en œuvre d'approches non paramétriques s'avère nécessaire [Rousseeuw et Leroy 2003]. Parmi celles-ci, la plus simple consiste à calculer le premier et le troisième quartiles, $Q1$ et $Q3$,

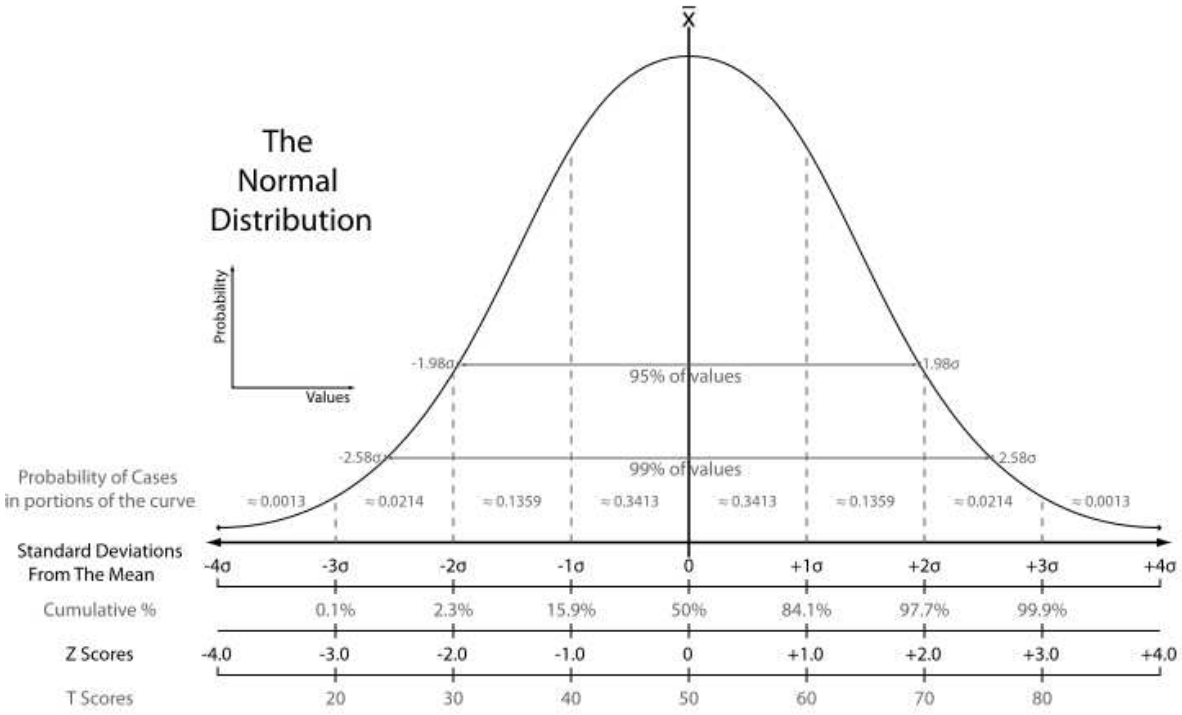


FIGURE 3.3 – Comparaison des différentes mesures liées à la distribution normale : les écarts types, les pourcentages cumulés, les z-scores et les T-scores (d'après Wikipedia).

de l'ensemble des données (l'intervalle $(Q1, Q3)$ noté IQR, Intervalle Inter-Quartiles. Les valeurs singulières sont celles qui sont supérieures à $Q3 + [1.5 \times (IQR)]$ et inférieures à $Q1 - [1.5 \times (IQR)]$ (voir figure 3.4), donc comme précédément des valeurs extrêmes.

Il existe beaucoup d'autres méthodes permettant de détecter des valeurs singulières dans des ensembles multivariés avec des approches basées sur les lois de distribution [Georgiev 2007]. Ces méthodes ne sont pas adaptées à la Fouille de Données en Génomique Médicale Fonctionnelle car, dans la majorité des situations, les données ne sont pas basées sur un modèle de distribution connu mais sur un modèle complexe qu'il est difficile de caractériser précisément. L'utilisation de ce type d'approches implique, en général, une qualité faible des résultats car les données singulières ne sont définies que par rapport à des distributions connues du système. De plus, la recherche du modèle de distribution le plus adapté aux données est très souvent longue et n'aboutit soit à aucun résultat, soit à un résultat non pertinent.

3.4.2 Étude des données marginales

Un autre ensemble d'approches permettant d'identifier les valeurs singulières dans un ensemble de données sont celles basées sur la *marginalité des données*. Contrairement, à la famille d'approches précédentes, celles-ci sont plus stringentes car elles ne s'appuient pas sur une valeur statistique mais sur un nombre-seuil de valeurs à extraire [Nelson 2004]. Ainsi, dans le contexte d'ensembles de données univariées, on parlera respectivement de rognage (*trimming*, en anglais) et de lissage (*winsorisation* en anglais). Formellement, le rognage consiste à abstraire d'un ensemble de données un sous ensemble joint ou non de données marginales ; moins formellement, le rognage (voir figure 3.5) consiste à supprimer de l'ensemble des données un nombre (ou un

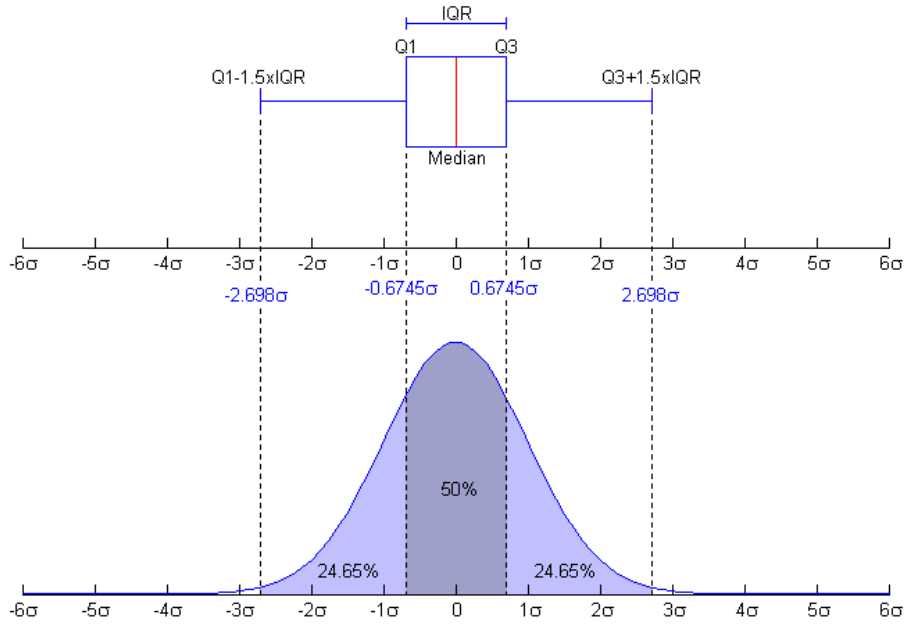


FIGURE 3.4 – Exemple de calcul de l'Intervalle Interquartiles *(*en Anglais, Inter-Quartil Range (IQR)) dans le cas d'une distribution selon une loi Normale.

pourcentage) de valeurs sur l'une des deux extrémités, voir sur les deux. Le lissage (voir figure 3.6), quant à lui, consiste à affecter une valeur v_1 égale à la $(n^{ieme}/2) + 1$ valeur d'une première extrémité d'un ensemble de données univariées aux $n/2$ données qui tendent vers cette extrémité. Pour l'extrémité opposée, le processus est identique. Dans ces deux types d'approches, la valeur de n est définie *a priori* par l'utilisateur.

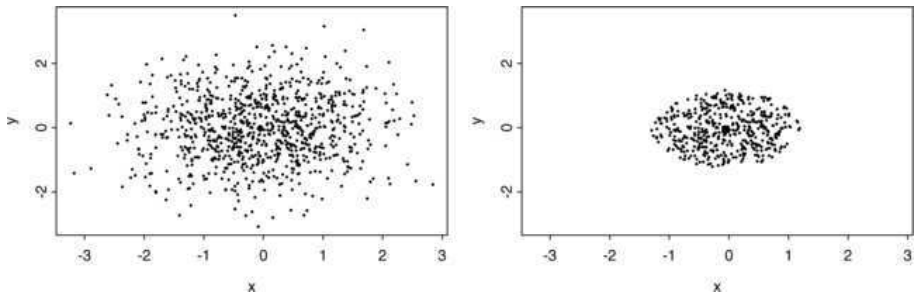


FIGURE 3.5 – Exemple d'application d'un rognage sur des données bivariées.

Ces deux approches ont été généralisées aux ensembles de données multivariées [Zuo 2006]. Dans ce contexte, le traitement des valeurs singulières consiste respectivement soit à rogner, soit à lisser les données qui sont les moins « profondes » en terme de positionnement par rapport à un nuage formé par l'ensemble des données multivariées prises en considération.

Ces approches, bien que ne tenant pas compte de la distribution des données, ne sont pas intéressantes dans le cadre général de la Fouille de Données (*a fortiori*) dans le contexte de la Génomique Médicale Fonctionnelle et ce pour deux raisons. D'une part, le nombre de données exploitées dans la suite des traitements est réduit de manière plus ou moins drastique par rapport à la dimension des données et d'autre part, les données rognées ou lissées ne sont pas forcément des valeurs singulières, il peut s'agir tout simplement de données manquantes, qui vont influencer

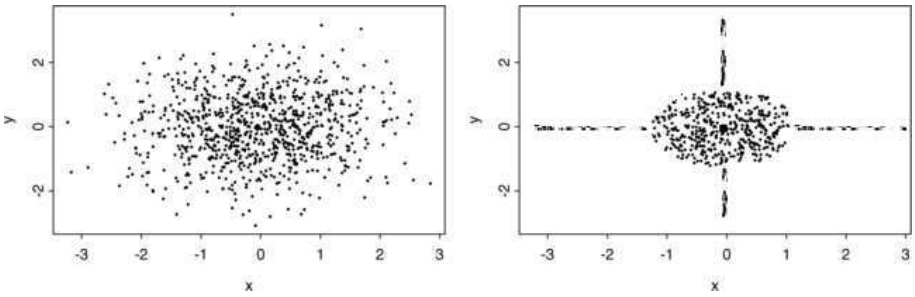


FIGURE 3.6 – Exemple d'application d'un lissage sur des données bivariées.

de manière conséquente les résultats des analyses faites sur les données par la suite.

3.4.3 Approches basées sur la distance

La troisième famille d'approches permettant de définir des valeurs d'un ensemble de données comme des valeurs singulières sont les approches basées la notion de *distance* [Knorr *et al.* 2000, Knorr 2002]. De manière générale, pour ces approches, plus une donnée est distante des autres, plus elle est considérée comme pouvant être une valeur singulière. Ce type d'approches est applicable aussi bien sur les ensembles de données univariées et multivariées. Elles ne s'appuient que sur la définition d'une *distance-seuil* au-delà de laquelle les données peuvent être considérées comme des valeurs singulières. La part d'*a priori* de ces approches réside dans le fait qu'il est nécessaire de « bien » définir cette valeur-seuil. Ainsi, si la *distance-seuil* est trop petite, un grand nombre de valeurs sont définies comme des valeurs singulières ; inversement si la *distance-seuil* est trop grande, peu, voir pas, de valeurs seront définies comme singulières. Cela pose un problème majeur selon nous, car le risque de suggérer la présence ou l'absence de valeurs singulières est fonction d'une définition subjective de la *distance-seuil*. La définition de cette dernière n'est pas simple, elle nécessite une bonne connaissance des données et de leur répartition dans l'espace. Ce type d'approche n'est pas pertinent pour des données de Génomique Médicale Fonctionnelle, dans lesquelles les variations sont fonction des conditions expérimentales et la notion de *valeur seuil* peut varier d'un gène à l'autre, la variation de l'expression du gène étant relative au gène lui-même.

3.4.4 Analyse des densités

Parmi les approches issues de l'Intelligence Artificielle permettant de générer des sous-ensembles de données homogènes, les méthodes basées sur la *densité des données* sont une autre famille de méthodes permettant de détecter des valeurs singulières : DBSCAN [Ester *et al.* 1996], OPTICS [Ankerst *et al.* 1999] et DENCLUE [Hinneburg et Keim 1998].

Dans ces algorithmes, la notion de densité est définie via la notion de nuages de points. Un nuage de points correspond à la présence d'un nombre minimum de points dans un espace donné et cet espace a une distance prédéfinie minimale avec les autres sous-ensembles de points. Ces algorithmes tendent à ignorer les valeurs singulières dans les résultats qu'ils restituent, puisque les points isolés sont considérés comme du bruit dans les données, elles sont généralement donc écartées des analyses et rarement notifiées à l'utilisateur.

STING [Wang *et al.* 1997] et CLIQUE [Agrawal *et al.* 1998] sont des approches basées sur le regroupement par décomposition de l'espace des données initiales en espaces plus restreints organisés sous la forme d'une grille. Cette dernière, multidimensionnelle, forme des cellules au

sein desquelles les données sont positionnées. Les données sont regroupées par rapprochements successifs des cellules localisées dans la même région. Les sous-ensembles de données sont par la suite définis par ces regroupements. Ces approches peuvent permettre de détecter des valeurs singulières, mais les résultats sont fortement dépendants du type de maillage utilisé. En effet, si le maillage est trop fin, un grand nombre de données va être considéré comme des valeurs singulières et inversement, si le maillage est trop gros, moins de données seront considérées comme isolées.

Breunig [Breunig *et al.* 2000] a proposé un algorithme dédié pour la détection des valeurs singulières : l'algorithme LOF. Il est basé sur la détection des valeurs singulières en fonction de la densité locale des données. Ainsi, en définissant un nombre de voisins minimaux dans une distance maximale à un point, l'algorithme permet d'identifier des données isolées. Cette approche est très efficace lorsque l'on dispose de beaucoup de données. La figure 3.7 montre un exemple de résultats issus de la mise en œuvre de LOF : C_1 et C_2 sont respectivement deux ensembles de données homogènes. o_1 est détecté comme une valeur singulière par rapport à l'ensemble de données C_1 et o_2 est détecté comme une valeur singulière par rapport à l'ensemble de données C_2. Dans les approches par regroupement, les ensembles de données formés respectivement par $\{C_2, o_2\}$, $\{C_2, o_2, o_1\}$, $\{C_1, o_1\}$ et $\{C_1, o_1, o_2\}$ auraient été considérés comme des ensembles de valeurs singulières vis-à-vis des autres ensembles de l'espace de données.

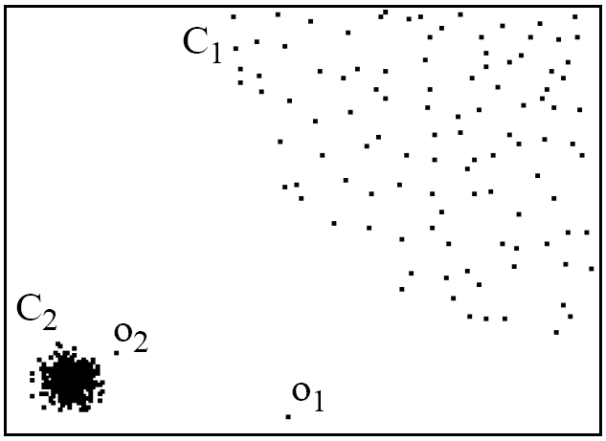


FIGURE 3.7 – Valeurs singulières définies localement (o1 et o2) dans un espace à 2 dimensions [Breunig *et al.* 2000].

Cette approche est intéressante et fournit des résultats pertinents sur des grands ensembles de données. Néanmoins, dans le contexte de la Génomique Médicale Fonctionnelle, le nombre de valeurs disponibles pour un attribut (expression d'un gène ou d'un paramètre bioclinique) au cours d'un protocole de recherche donné, sont, aujourd'hui, trop réduit (quelques dizaines d'individus au maximum). Ceci ne permet pas une définition suffisamment « fine » du nombre de voisins minimaux et d'autre part elle pose le problème de la connaissance *a priori* des données.

3.4.5 Recherche de singletons en apprentissage

Dans de nombreuses approches d'Intelligence Artificielle, la notion de valeur singulière apparaît notamment dans le cadre des méthodes de *classification* et de *regroupement* [MacQueen 1967, Cornuéjols *et al.* 2002, Datta 2003, Han et Kamber 2006]. Une des différences principales entre les notions de classification et de regroupement réside dans le fait que la première s'appuie sur des méthodes issues de l'*apprentissage supervisé*, alors que la seconde concerne plutôt des

approches issues de l'*apprentissage non supervisé*. En général, dans l'ensemble de ces travaux les valeurs singulières sont assimilées à une classe de données particulières ou à un singleton (valeur isolée). Nous allons dans les paragraphes suivants présenter quelques méthodes d'apprentissage [Han et Kamber 2006] qui nous semble pertinentes pour comprendre la notion de valeurs singulières dans le cadre de l'Intelligence Artificielle.

Approches en classification

L'une des tâches de l'apprentissage supervisé est la *classification*. Le principe est le suivant : l'expert affecte l'ensemble des données à des classes, dont une qui va contenir l'ensemble des valeurs singulières (donc en général inexplicables ou incohérentes aux yeux de l'expert), puis l'algorithme va apprendre à caractériser chacune de ces classes de manière la plus précise possible. Lorsque l'utilisateur va avoir une nouvelle donnée, le système va automatiquement lui indiquer à quelle classe cette donnée « semble » appartenir.

Les arbres de décision [Quinlan 1993, Quinlan 1996] sont une des nombreuses approches de classification utilisées pour la détection de valeurs singulières. Ils permettent de modéliser sous une forme facilement exploitable (arborescence ou règles) un ensemble de données défini par différents attributs et pour lequel on souhaite prédire la valeur de l'un des attributs ou la classe d'appartenance de la donnée (en mettant à l'écart les valeurs singulières). En général, l'arbre décisionnel est construit à partir d'un échantillon des données (dit « ensemble d'apprentissage ») puis testé sur le reste des données (dit « ensemble de test ») avant d'être mis en application sur de « nouvelles » données. L'un des inconvénients des arbres de décision est leur forte sensibilité aux bruits et à la topologie des données. L'utilisation de l'arbre issue d'un apprentissage avec des valeurs singulières fortement réparties dans l'espace, induit de nombreuses ramifications et une définition diffuse de l'ensemble des classes.

Une seconde classe de méthodes de classification utilisées pour la détection des valeurs singulières sont les machines à vecteurs de support (en anglais *Support Vector Machine, SVM*). Ces méthodes sont destinées à résoudre des problèmes de discrimination et de régression [Burges 1998]. Dans le cas de la recherche de valeurs singulières, les SVM [Escalante 2005, Unnthorsson *et al.* 2003] ont pour objectif de définir deux groupes de points (qu'il s'agisse de données univariées ou multivariées) : le premier correspond aux données dites *normales*, c'est-à-dire ayant un comportement homogène et le second est qualifié de *valeurs singulières*. Ces groupes de données vont être définis à l'aide de lignes séparatrices qui ont l'avantage de pouvoir être ou non linéaires. La limitation de ce type d'approches vient, comme dans le cas des arbres, du fait que dans l'espace, il devient alors très difficile de les projeter au sein de l'hyperplan dans lequel sont les données. Ainsi, avec ce type d'approches, la détection des valeurs singulières n'est efficace que dans le cas particulier où les données sont « regroupées » d'un point de vue spatiale.

De manière générale, la détection des valeurs singulières par des approches d'apprentissage supervisé correspond à une utilisation particulière des algorithmes de classification et elles se révèlent sur les données réelles souvent peu efficaces dans le monde réel. De plus, il est rare que l'expert est une très bonne connaissance des données qu'il a à traiter, c'est-à-dire une connaissance suffisante pour pouvoir identifier de manière certaine la classe des valeurs singulières.

Approches en regroupement

Les approches basées sur le *regroupement* sont des approches issues de l'apprentissage non-supervisé [Cornuéjols *et al.* 2002, Datta 2003, Han et Kamber 2006]. Contrairement à la classification, le but du regroupement n'est pas de caractériser des classes existantes, mais de regrouper

dans des sous-ensembles homogènes les données en respectant les contraintes suivantes :

minimiser l'inertie intra-classe pour obtenir des groupes les plus homogènes possibles ;

maximiser l'inertie inter-classe afin d'obtenir des sous-ensembles bien différenciés.

Nous allons nous intéresser ici plus particulièrement à deux types d'approches : les approches *hiérarchiques* et les approches *par partitionnement*.

Les approches de *regroupement hiérarchique* [Kaufman et Rousseeuw 1990, Cornuéjols *et al.* 2002] sont basées sur deux types de méthodologies. D'une part, il y a les méthodes agglomératives, comme CLARA (Clustering Large Applications) [Pakhira 2008], qui partent des données initiales et assemblent progressivement les données (respectivement les groupes) proches (en terme de distance) ensemble jusqu'à n'obtenir qu'un seul groupe contenant toutes les données. D'autre part, il y a les méthodes divisives, comme AGNES (Agglomerative Nesting) [Struyf *et al.* 1997], qui commencent par regrouper l'ensemble des données dans un groupe, puis qui divise ce groupe progressivement en sous-groupes en fonction de leurs différences (en terme de distance) jusqu'à n'obtenir que des groupes contenant une seule et unique donnée. Dans le cadre des approches hiérarchiques (agglomératives ou divisives), les valeurs singulières correspondent à des singletons (groupe à une donnée), qui fusionnent que très tardivement avec les autres données de la classification, c'est-à-dire très haut dans la hiérarchie, donc avec des grands groupes de données. En général, ces éléments sont facilement identifiables visuellement sur les dendrogrammes représentant les classifications, comme le montre la figure 3.8, où les données nommées *LAC*, *LAI* et *PSI* sont des valeurs singulières.

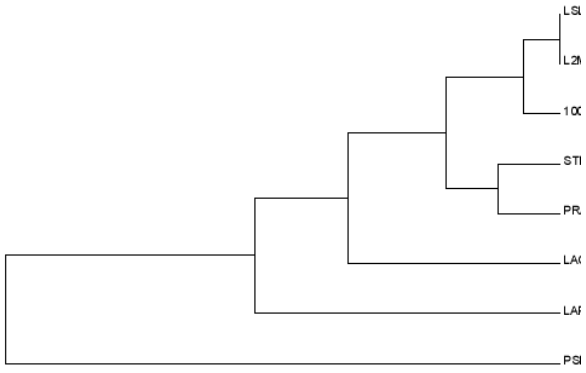


FIGURE 3.8 – Exemple de dendrogramme présentant des potentielles valeurs singulières.

Contrairement aux approches hiérarchiques, qui construisent progressivement les groupes, les approches de partitionnement [Eisen 1999, Labart *et al.* 2000] vont « directement » construire des sous-groupes à partir des données initiales. Le nombre de sous-groupes, k, est défini à l'avance par l'utilisateur. Les algorithmes de partitionnement les plus simples sont les algorithmes de type « k-centroïdes », comme les k-moyennes (*k-means* en anglais) [MacQueen 1967] et les k-médoïdes (*k-medoids* en anglais) [Kaufman et Rousseeuw 1990, van der Laan *et al.* 2003].

L'algorithme *k-moyennes* [MacQueen 1967] est considéré comme l'un des plus simples algorithmes d'apprentissage non supervisé. Il permet de séparer les données dans k groupes distincts, qui respectent les deux propriétés d'inertie mentionnées précédemment. Chaque groupe est identifié par son centroïde, c'est-à-dire par un point « artificiel » correspondant à son centre. Cette méthode permet de facilement trouver des données isolées dans des ensembles, à la condition de bien définir k. Cependant, l'une de ses limitations est sa sensibilité à la phase d'initialisation des groupes. En effet, deux exécutions successives de la méthode sur le même ensemble de données

peut donner des répartitions totalement différentes. Plus précisément, la localisation initiale des centroïdes varient d'une initialisation à l'autre, puis d'une itération à l'autre de l'algorithme. Il en résulte donc une répartition différente des points et une définition différente des centroïdes. Une façon de surmonter ce problème consiste à rechercher non pas une donnée artificielle représentant au mieux chaque sous-ensemble, mais des données « réelles » (des prototypes). L'algorithme des k-médoïdes, nommé aussi PAM, [Kaufman et Rousseeuw 1990, van der Laan *et al.* 2003] fonctionne sur ce principe. Une autre limitation des algorithmes de partitionnement est liée au fait que l'utilisateur doit définir le nombre de groupes k *a priori*. Si k n'est pas adapté à la structure de l'ensemble de données, les potentielles valeurs singulières peuvent être considérées comme des données à part entière. Nous reviendrons sur ces 2 derniers points dans les chapitres suivants car nous allons nous baser sur cette famille de méthodes pour découvrir les valeurs singulières dans nos données.

3.4.6 Une approche multi-algorithme

Brodley et Friedl [Brodley et Friedl 1996] ont proposé de combiner plusieurs méthodes pour améliorer la qualité des résultats obtenus pour la détection de valeurs singulières. Leur approche consiste à assimiler une donnée à une valeur singulière dès lors qu'elle est définie comme telle par deux des trois méthodes utilisées (arbre de décision C4.5 avec élagage, k-plus proches voisins et régression linéaire). Le problème posé par ce type d'approche réside dans le choix des algorithmes à utiliser. En effet, ce choix est fortement dépendant de la nature des données que l'on traite. De plus, ces algorithmes doivent nécessiter le minimum d'intervention de l'expert afin de réduire les risques dus à un mauvais étiquetage des données initiales.

3.5 Conclusion

La définition que nous retiendrons de la notion de valeur singulière est la suivante : une valeur présentant une distance élevée avec les autres valeurs de l'ensemble des données est une valeur singulière. Cette définition est floue en soit de part le terme de « distance élevée » dont l'interprétation est « analyste-dépendante » ou fixé consensuellement. Ainsi, pour considérer des données comme étant des valeurs singulières, il est nécessaire de caractériser les données dites « normales ».

L'ensemble des approches et des méthodes que nous venons de présenter sont de manière générale principalement dédiées et optimisées pour permettre le traitement de données singulières, mais elles ne permettent que très rarement de les détecter. La plupart des méthodes statistiques cherchent, avant tout, à minimiser l'impact de celles-ci en proposant des méthodes notamment des méthodes robustes [Cleveland 1981]. Les approches issues de l'Intelligence Artificielle ont la particularité de pouvoir détecter les données isolées, à partir soit de nombreuses données, soit en approchant au maximum la topologie des données.

Dans nos travaux, les données qui nous intéressent sont composées de peu d'exemples et de nombreux attributs. Il est donc essentiel de ne pas minimiser l'impact des valeurs singulières, mais de pouvoir les détecter de manière simple et efficace pour signaler à l'utilisateur que certains résultats sont basés sur des données de mauvaise qualité (problèmes humain(s) et/ou matériel(s) rencontrés au cours du protocole expérimentale) ou plus particulièrement divergentes et pouvant être des indicateurs de comportements et/ou d'états physiologiques particuliers (individus présentant des mutations géniques, des dysfonctionnements métaboliques,...).

Chapitre 4

Problématique

Aujourd'hui les protocoles de recherche clinique sont de plus en plus complexes, en terme de type de données qu'ils récoltent. Même si beaucoup d'entre eux sont dédiés à des études en Génomique Fonctionnelle (donc sur le rôle des gènes) ou en Clinique (liens entre les données médicales), de plus en plus de protocoles « mixtes » apparaissent. Ils traitent d'informations dans les deux domaines et visent à découvrir des liens entre toutes ces données. Cette nouvelle thématique de recherche est définie comme étant la Génomique Médicale Fonctionnelle définit par le groupe de Génomique Médicale Fonctionnelle de l'Université de Copenhague, comme nous l'avons vu à la fin du chapitre 1. Ce terme est particulièrement bien adapté pour décrire notre domaine d'application, car notre travail se situe aussi bien dans le contexte de la Génomique Fonctionnelle que dans le contexte de la Médecine Clinique.

L'un des objectifs en Génomique Médicale Fonctionnelle va être de trouver des relations entre des données d'expression génique et des données biocliniques. Ces relations peuvent permettre de découvrir de nouveaux biomarqueurs qui sont la base du diagnostic ou de la prédiction du statut d'un individu par rapport à une pathologie et à son évolution, comme nous l'avons présenté au chapitre 2.

La recherche de relations s'inscrit notamment dans la Fouille De Données (FDD) et l'Aide à la Découverte Scientifique (ADS). Pat Langley [Langley 1999] définit la différence entre ces deux domaines au niveau de la Représentation des Connaissances :

> *There exist two computational paradigms for discovering explicit knowledge from data :*
> – *Data mining generates knowledge cast as decision trees, logical rules, or other notations invented by AI researchers ;*
> – *Computational scientific discovery instead uses equations, structural models, reaction pathways, or other formalisms invented by scientists and engineers.*

Le premier, la Fouille de Données, fournit des résultats suivant un formalisme défini par l'informaticien (ou tout au moins par le concepteur du système d'aide à la découverte) et le second la Découverte Scientifique, fournit des résultats dans un formalisme familier et propre au domaine de l'expert.

Nos travaux s'inscrivent dans ces deux domaines parce que notre but est de permettre la découverte de relations entre deux ensembles de données avec des formalismes facilement compréhensibles par les biologistes-génomiciens, à savoir des données de statistiques descriptives et des représentations graphiques « simples » du point de vue perceptif [Brunet 2002, Bertrand et Garnier 2005]. De notre point de vue et dans le cadre de nos travaux, la Découverte Scientifique s'appuie sur la Fouille de Données pour conduire à des découvertes dans les bases de données en

utilisant un formalisme « scientifique ».

La notion d'expert qui fouille des données en Génomique Médicale Fonctionnelle doit être définit au sens large : il peut être bioinformaticien, informaticien médicale, biologiste « généraliste » ou ayant des connaissances et des capacités d'interprétation dans le domaine d'étude. *In extenso*, l'expert tel que nous le définissons est tout utilisateur de la méthode et du système que nous proposons, c'est-à-dire toute personne capable d'utiliser notre système.

Dans ce chapitre, nous allons présenter notre problématique en traitant successivement de la définition de relations entre deux ensembles de données en Génomique Médicale Fonctionnelle, puis de la significativité de résultats obtenus et de la prise en compte des valeurs singulières dans ces données.

4.1 Relations en Génomique Médicale Fonctionnelle

Différentes approches ont été proposées dans le cadre de la mise en relation de données de Génomique Fonctionnelle. L'étude de Fu [Fu *et al.* 2005] s'intéresse aux fréquences alléliques dans les populations et montrent qu'elles sont corrélées quand les populations ont une histoire partagée. Les auteurs remarquent que les mesures traditionnelles sur la structure des populations tendent à surestimer la quantité de différentiation génétique existante lorsque la corrélation est négligée. L'intérêt de ces travaux est d'une part de montrer l'importance de l'étude des corrélations entre différents ensembles d'individus afin de valider et de généraliser une potentielle découverte et d'autre part de monter indirectement que les populations sont décrites par des caractères phénotypiques, génétiques et génomiques dont il est nécessaire d'étudier les relations et plus particulièrement les corrélations afin de les caractériser.

D'un point de vue générale, la recherche de relations entre des données issues de la pratique clinique ou d'examens de biologie de laboratoire et des données issues de mesures d'expression génique par l'intermédiaire de biopuces, est un domaine pour lequel il existe des travaux s'appuyant sur différents types d'approches. Ces approches ont des avantages et des limites, comme nous allons le résumer dans les paragraphes suivants.

La grande partie des approches et des travaux que nous avons exposés dans le chapitre 2 présentent des résultats soit obtenus en s'appuyant sur des *a priori* forts, en terme de connaissances dans le domaine d'application, soit sur des *a priori* basés sur des concepts statistiques définis empiriquement, soit sur la mise en œuvre d'approches d'Analyse et de Fouille de Données complexes et fonctionnant comme des « boîtes noires » vis à vis de l'expert du domaine [Mondada 2005]. Le problème de la confiance dans les résultats peut ainsi être posé : « les algorithmes utilisés sont-ils pertinents vis à vis des données du domaine et les résultats obtenus sont-ils réalistes ? ».

Ainsi, la principale limite de ces travaux est qu'ils travaillent sur un nombre réduit de paramètres (gènes ou données biocliniques). Cette sélection est faite à partir de connaissances *a priori* de l'expert telles que celles concernant les gènes et leurs fonctions, les paramètres biocliniques et leurs significations, les individus étudiés, la qualité des données... Ce type d'approches réduit de manière considérable la probabilité de découvrir de nouvelles découvertes puisque toutes les relations possibles (d'un point de vue quantitatif) ne sont pas explorées.

Ainsi, il est important d'approcher le problème de l'aide à la découverte de relations entre des données d'expression génique issues de biopuces et des données biocliniques sous un angle large et bien défini. L'approche proposée doit permettre d'automatiser la recherche des relations parmi toutes les données dont nous disposons et de prendre en compte le rôle prédominant de l'expert dans ce domaine d'application.

La recherche de relation au sein d'attributs décrivant des populations peut s'appuyer sur des

calculs de régression. Néanmoins, ce type de méthodes de calcul est sensible aux unités de chacune des variables étudiées. Comme nous l'avons exposé au chapitre 1, nos données sont hétérogènes en terme d'échelle de mesures (e.g. taille en *cm*, poids en *kg*, âge en *années*, expression génique en *valeur relative*). Cette caractéristique en fait un critère de choix pour les études des données en Génomique Médicale Fonctionnelle. Il est donc plus pertinent de baser cette notion de recherche de relations sur une mesure de corrélation que sur un calcul de régression, car une corrélation est moins sensible, voir n'est pas sensible du tout aux unités.

Différentes mesures de corrélation existent. Nous les avons présentées de manière générale au chapitre 2. En nous appuyant sur la définition des différentes corrélations et sur la description des données de Génomique Fonctionnelle (chapitre 1), nous avons pris le partie d'utiliser le coefficient de corrélation de rang de Spearman ρ_S.

Le coefficient de corrélation de Spearman ρ_S est une statistique non-paramétrique qui ne prend pas en compte la valeur exacte des données mais leur rang ce qui permet d'atténuer les effets de biais dus aux manipulations expérimentales et aux variations inter-individuelles. Le coefficient de corrélation de rang de Spearman ρ_S est d'autre part une statistique connue (plus que le coefficient de corrélation de rang de Kendall τ) des biologistes ce qui leur permet une compréhension plus facile des résultats. De plus, le coefficient de corrélation de rang de Spearman ρ_S permet de s'abstraire de connaissances préalables sur les lois de distribution de chacune des variables étudiées. En effet, dans le cas des données que nous souhaitons étudier, nous ne connaissons pas ces informations.

Cependant, il est important de noter qu'une erreur courante d'interprétation des résultats consiste à penser qu'une valeur de corrélation élevée (qui tend vers -1 ou $+1$) correspond à des relations décrivant des liens de causes à effets entre les deux ensembles. Ces deux ensembles de données peuvent ne pas être liés directement mais en réalité corrélés à un même phénomène-source dont ils dépendent tous les deux.

La découverte de relations en Biologie et plus particulièrement en Génomique Fonctionnelle, est liée aujourd'hui sur la notion de leur significativité statistique. Différentes approches ont été proposées pour soutenir et renforcer cette philosophie, comme nous allons le voir dans la partie suivante.

4.2 De la significativité aux valeurs singulières

Les biologistes culturellement ne sont pas intéressés par la valeur d'une corrélation mais les statistiques relatives à la significativité de cette corrélation.

Lorsque l'on cherche à travailler sur des données qui sont sujettes à un grand nombre de tests statistiques, il est nécessaire de réaliser des tests de multiplicité, qui permettent de définir, parmi les résultats statistiques, ceux qui sont les plus réellement significatifs. Ainsi, différentes approches sont possibles et on peut distinguer deux grandes familles de tests de multiplicité. La première, la plus stringente est dite *« FWER » (Familly Wise Error Rate)*. Le test le plus connu de cette famille est le test dit de Bonferroni [Benjamini *et al.* 2001, Abdi 2007]. Cette approche consiste à définir un seuil de significativité que l'on souhaite contrôler. Ce seuil est défini en divisant α (le seuil permettant de considérer un test comme significatif) par le nombre de tests réalisés. Plus le nombre de tests de significativité est élevé, plus le seuil du test de multiplicité va être réduit. Ceci n'est pas intéressant dans le cadre d'études comme l'exploration de données issues de puces à ADNc, car le seuil du test de Bonferroni risque d'être très faible au vue du nombre de tests qu'il faut effectuer et le nombre de tests significatifs peut être nul.

Ainsi pour pallier à ce problème une autre famille de tests moins stringent a été proposée, le

« FDR » (False Discovery Rate) [Benjamini et Hochberg 1995, Benjamini *et al.* 2001, Benjamini et Yekuteli 2001]. Son objectif est de définir le taux de fausses découvertes, c'est-à-dire que l'on cherche à définir si par rapport à un ensemble de tests de significativité réalisés pour une étude, un test pris individuellement est probablement une fausse découverte. Cette valeur correspond au rapport entre le nombre de tests rejetant incorrectement l'hypothèse nulle et l'ensemble des tests rejetant l'hypothèse nulle. Ce rapport permet de pondérer la valeur initiale de chaque test de significativité. De manière simple, le FDR permet d'ajuster les valeurs des tests de significativité entre elles. Une variante (*local FDR*) de ce test existe afin de permettre de prendre en compte dans le calcul du taux de fausses découvertes, les valeurs voisines d'une valeur de significativité que l'on cherche à contrôler [Efron 2006b].

Les résultats du calcul du FDR peuvent être rendus plus accessibles en utilisant la *q-valeur* [Storey 2002, Storey et Tibshirani 2003], qui est la valeur de la significativité d'un test de significativité, autrement dit de manière simple la p-valeur d'une p-valeur.

Ces différentes statistiques, et à commencer par la *p-valeur*, ont pour avantage de permettre un filtrage sur des bases probabilistes de relations (dans notre cas). Le coût des expérimentations biologiques, *in vivo* et *in vitro*, est beaucoup plus élevé que des expérimentations *in silico*, ou tout au moins, elles permettent de ne retenir que des expérimentations pertinentes à réaliser *in vivo* ou *in vitro*. Néanmoins, le choix de l'une de ces statistiques pour évaluer la significativité et/ou le taux de fausses découvertes est important et dépend du contexte expérimental. Certaines de ces méthodes sont plus stringentes et peuvent conduire à éliminer par excès des relations que d'autres considèrent comme de potentiels candidats relations.

L'utilisation de l'ensemble de ces tests a été remise en cause dans le cadre de l'étude de corrélations. En effet, sans avoir pour objectif de généraliser les propos d'Efron [Efron 2006a], l'utilisation du *FDR* dans le cadre de l'évaluation de la significativité d'une corrélation, surtout lorsqu'elle est forte, n'est pas adaptée. En effet, une mesure de corrélation entre deux ensembles (dans le cadre des approches que nous utilisons) prend en compte l'ordre des ensembles étudiés ainsi que leur effectif. Plus une population a un effectif réduit, plus la valeur de sa corrélation peut être élevée pour qu'elle soit significative. De manière identique, plus une population a un effectif important, moins la valeur de sa corrélation doit être élevée pour qu'elle soit significative. Les tests d'évaluation du taux de fausses découvertes, ou plus généralement les méthodes d'ajustement multiple, permettent d'ajouter un filtrage sur les relations à explorer(pour nous des corrélations existantes entre des valeurs d'expression d'un gène chez les différents individus et des valeurs biologiques d'un paramètre chez ces mêmes individus).

En plus du problème que soulève l'étude de la significativité des résultats, le coefficient de corrélation (comme une majorité des tests statistiques) est sensible à la présence de valeurs singulières, comme nous allons le développer dans la partie suivante.

4.3 Découverte de valeurs singulières

Les approches du chapitre 2 peuvent être victimes lors de leur mise en œuvre de l'effet de valeurs dont le comportement diverge des autres. Ces valeurs peuvent modifier de manière plus ou moins substantielle les résultats et leur qualité. Nous avons présentés au chapitre 3 des théories et des travaux pour de détecter ces valeurs singulières. Les *valeurs singulières* permettent d'une part de pouvoir améliorer les résultats obtenus lors de l'analyse des données et de mettre en avant celles qui ont réellement un intérêt potentiel pour l'expert ; et d'autre part de détecter des individus ayant des comportements suspects vis-à-vis d'attributs ce qui permet de faire apparaître des particularités qui peuvent être intéressantes à étudier par l'expert. Dans le contexte

des données de la Génomique Médicale Fonctionnelle, chaque donnée univariée peut être source de valeurs singulières. Ces dernières doivent être considérées suivant une double origine. D'une part, les données issues des puces à ADNc sont des données sujettes à un très grand nombre de manipulations durant leur « fabrication » (cf. 1.3). Chacune d'entre elles peut conduire à l'apparition de valeurs singulières due aussi bien à la qualité du matériel biologique qu'à des erreurs de manipulations,... De plus, les données issues des mesures biologiques et cliniques sont sujettes à des variations inter-individuelles aussi bien du côté du sujet expérimental (d'origine biologique ou clinique) que de l'opérateur (d'origine technique). Ainsi, les données issues de la Génomique Médicale Fonctionnelle peuvent inclure des données singulières aberrantes ou contaminantes de manière plus ou moins importante. Leur prise en compte au cours du processus de Fouille de Données est pertinente afin de pondérer les résultats obtenus. De plus, la détection de valeurs singulières multivariées est une tâche qui n'est que peu, voir pas, du tout prise en compte. En effet les processus d'analyse étant majoritairement fondé sur des *a priori* sur les données, les valeurs singulières sont rarement prises en compte et, lorsqu'elles le sont elles le sont uniquement au cours de l'analyse univariée des données. La manière la plus simple et efficace de traiter les valeurs singulières dans des données multivariées consiste donc à considérer les échantillons marginaux de ces données, c'est-à-dire d'étudier chacune de leurs composantes univariées [Dodge et Rousson 1999]. Ce type d'approches est aujourd'hui privilégié en Génomique Médicale Fonctionnelle comme nous l'avons souligné précédemment. Néanmoins, la suppression de valeurs extrêmes sur quelques dimensions que ce soit peut avoir des conséquences néfastes sur l'analyse des données et la détection des valeurs singulières. En effet, les valeurs extrêmes de chaque ensemble ne sont pas forcément des valeurs singulières, elles se révèlent souvent être dans la parfaite continuité des données (comme le montre l'exemple de la figure 3.2). De plus, ce type d'approche masque des valeurs singulières qui seraient le produit des valeurs de différents attributs pour un même individu présentant des caractéristiques particulières.

Parmi l'ensemble des approches que nous avons présenté au chapitre 3 pour détecter les valeurs singulière, nous avons choisi de baser notre méthodologie sur un algorithme de partitionnement : l'algorithme PAM (Partitining Around Medoids) [Kaufman et Rousseeuw 1990, van der Laan *et al.* 2003]. Cet algorithme est très similaire à l'algorithme des k-moyennes que nous avons présenté au chapitre 2. Mais contrairement à celui-ci, un groupe ne va pas être représenté par un point artificiel, mais par une donnée réelle (la donnée au centre du groupe). Cette subtibilité permet de le rendre moins sensible que l'algorithme des k-moyennes à la phase d'initialisation des groupes et donc plus résistant aux bruits. Les deux algorithmes ont une limitation commune basée sur leur paramétrage de base : la définition du nombre de groupes k. Cette valeur doit être donnée *a priori* par l'utilisateur et peut avoir des conséquences importantes si elle est mal choisie. Afin de pallier à ce problème, nous allons définir de manière automatique cette valeur k en utilisant les propriétés d'inertie des partitions et adapter l'algorithme PAM.

4.4 Visualisation des résultats

Afin de permettre à l'expert d'être partie prenante dans le processus de découverte, il est important de conjuguer automatisation et interaction avec l'expert-explorateur. Ainsi, à partir des résultats obtenus lors de l'étude des corrélations, l'utilisateur doit disposer d'outils simples lui permettant :

- de visualiser les nombres de relations potentiellement intéressantes ;
- de visualiser en parallèle plusieurs relations définies afin de les comparer ;
- de visualiser une relation (en deux dimensions) en particulier pour valider son intérêt d'un

point de vue « numérique » ;
- de se documenter (*via* des ressources locales ou sur Internet) sur des relations pour valider leurs intérêts d'un point de vue « biomédicale ». Une relation peu ou pas documentée dans la littérature peut présenter un intérêt pour l'expert, car il peut ainsi chercher la fonction et l'implication biologique d'un gène en fonction des gènes documentés et présentant les mêmes tendances « numériques ».

4.5 Synthèse de la problématique

La problématique de notre thèse s'articule donc autour de trois grandes questions. Le but est de proposer aux Biologistes et aux médecins de potentiels biomarqueurs permettant de diagnostiquer et/ou de caractériser une pathologie dans une perspective d'y adapter au mieux une thérapeutique. Ainsi :

Quel flux de Fouille de Données est adapté à la découverte sans* a priori *de relations linéaires entre des données d'expression génique issues de biopuces et des données biocliniques ?

Comment prendre en compte les valeurs singulières et leurs influences dans ce flux de Fouille de Données ?

Comment permettre à l'expert d'explorer et d'exploiter simplement et rapidement les résultats issus de ce Fouille de Données dans un contexte aussi complexe que celui de la Génomique Médicale Fonctionnelle ?

Afin de répondre à ces différentes problématiques, nous allons proposer une méthodologie d'aide à la découverte de relations entre des données d'expression génique issues de puces à ADNc et des données biocliniques. Cette méthodologie doit répondre à différentes contraintes qui prennent en compte aussi bien des aspects relatifs :
- au volume des données explorées (plusieurs milliers de données « sources », de données de résultats et de données externes) ;
- au déséquilibre entre le nombre d'individus pour lesquels on peut disposer de données (quelques dizaines au maximum) et le nombre d'attributs mesurés pour chaque individu (quelques milliers au minimum) ;
- à la qualité relative des données explorées (mesures imprécises ou inexactes) ;
- au besoin de guider l'exploration des résultats de l'analyse des données sans pour autant introduire d'*a priori* en terme de connaissance ;
- à la nécessité d'aider l'expert à visualiser les relations « potentiellement intéressantes ».

Il est très intéressant de remarquer dès à présent que les algorithmes de mise en relation entre deux ensembles de données et que les algorithmes de détection de valeurs singulières ont des origines et des méthodologies communes issues aussi bien de la Statistique que de l'Intelligence Artificielle. À partir de cette observation, nous allons essayer de montrer dans les chapitres suivants la complémentarité de ces deux thématiques dans la définition de notre problématique. Notre but principal est d'utiliser les approches les plus adaptées issues de ces deux domaines afin de définir et d'explorer rapidement et sans *a priori* des relations linéaires potentiellement intéressantes existantes entre des données issues de puces à ADNc et des données biocliniques.

Dans le chapitre 5, nous allons nous concentrer sur la mise en relations des données issues de la Génomique Médicale Fonctionnelle. Nous allons montrer que les statistiques se révèlent adapter mais qu'elles donnent des résultats très globaux. Ainsi, nous allons présenter un second algorithme qui par fenêtrage améliorera les descriptions des relations découvertes.

Le chapitre 6 s'intéressera à la détection des valeurs singulières grâce à un algorithme d'apprentissage automatique simple que nous allons adapter. Ce nouvel algorithme permet la détection automatique de la répartition des données dans l'espace et donc la découverte de potentiels valeurs isolées. Afin de montrer l'intérêt de cette nouvelle approche, nous la testerons sur un ensemble d'exemples « jouets ».

La mise en œuvre de ces algorithmes dans le système DiscoClini est présenté dans le chapitre 7 et évalués sur des données réelles dans le chapitre suivant. Dans un premier temps, notre évaluation se fera étape par étape dans le cadre de protocoles de recherche clinique sur les Obésités dans le chapitre 8. Dans un second temps, de manière plus globale par des tests d'utilisabilité dans le chapitre 9.

Deuxième partie

Aide à la découverte de relations en Génomique Médicale Fonctionnelle

Chapitre 5

Mesure des relations entre les données

L'expression seule d'un gène ne permet pas de découvrir des connaissances sur les pathologies étudiées. Des études d'associations entre l'expression génique et les paramètres biologiques et cliniques sont apparues [Courtine 2002, Tan *et al.* 2003, Yamanishi *et al.* 2003, Bhardwaj et Lu 2005]. Nos travaux concernent ce domaine et plus particulièrement la découverte de relations entre les données d'expressions géniques et les données biocliniques. Plusieurs approches sont envisageables pour définir de telles relations. Nous avons choisi de fonder notre approche sur le coefficient de corrélation de rang de Spearman. Nous allons montrer dans ce chapitre comment cette mesure statistique, classiquement utilisée en Biologie, va être utilisée dans le cadre de notre flux de données de fouille de données. Notre méthodologie s'appuie sur le concept de *Découverte de Corrélations Linéaires* tel qu'il a été introduit par Cecil et Chiang [Cecil *et al.* 2002, Chiang *et al.* 2005]. Dans leurs articles, ils étudient des données de Gestion des Relations Clients et d'Intelligence Économique pour lesquelles ils disposent d'un grand nombre de valeurs pour chaque attribut (au minimum, quelques centaines de très bonne qualité). En Génomique, il est rare de disposer de plus de quelques dizaines de valeurs par attribut et les données sont souvent incomplètes avec des valeurs singulières.

Nous allons donc montrer dans ce chapitre comment calculer des corrélations dans le cas particulier de nos données [Benis *et al.* 2003a, Benis *et al.* 2003c, Benis *et al.* 2003b, Benis *et al.* 2003d] et proposer des méthodes de visualisation originales et utiles pour explorer efficacement les résultats obtenus automatiquement et *in extenso* dans le processus de décision conduisant à la définition de nouveaux biomarqueurs [Benis 2005, Benis 2007, Benis et Courtine 2009b, Benis et Courtine 2009a].

5.1 Étude de corrélations globales

5.1.1 Calcul d'une corrélation de rang de Spearman

Le coefficient de corrélation de rang de Spearman ρ_S [Beuscart *et al.* 2009] est un cas particulier du coefficient de corrélation paramétrique de Pearson r_P. En effet, les valeurs (X_i et Y_i) de chacun des ensembles étudiés sont converties en la valeur de leurs rang respectifs (x_i et y_i) dans chacun des ensembles avant le calcul de ρ_S. Le coefficient de corrélation de rang de Spearman ρ_S est un coefficient de corrélation non paramétrique. Il est défini de la façon suivante :

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{5.1}$$

où : d_i est $x_i - y_i$, c'est-à-dire la différence entre les rangs des valeurs correspondant à X_i et Y_i, et n est le nombre de couples de données existant dans l'ensemble des données étudiées.

Il est important de noter que pour que le calcul du coefficient de corrélation de Spearman, tel qu'il est présenté ci-dessus, soit applicable, il ne faut pas qu'il existe de couple de valeurs identiques (tel que $X_i = X_j \vee Y_i = Y_j$) ce qui n'est pas toujours le cas avec nos données. Néanmoins, afin de dépasser cette limitation, il est possible de mettre en œuvre l'ajustement de la corrélation de Spearman défini tel que :

$$\rho = \frac{(n^3 - n) - 6 \sum_{i=1}^{n} d_i^2 - (T_x + T_y)/2}{\sqrt{(n^3 - n)^2 - (T_x + T_y)(n^3 - n) + T_x T_y}} \tag{5.2}$$

avec

$$T_x = \sum_{g=1}^{G} (t_g^3 - t_g) \tag{5.3}$$

où : G est le nombre de valeurs distinctes parmi les rangs moyens de X pour le facteur de correction T_x (idem pour T_y), t_g étant le nombre d'occurrences pour chaque valeur du rang moyen [Siegel et Castellan 1988, Beuscart *et al.* 2009].

De manière générale, la valeur du coefficient de corrélation est comprise entre -1 et $+1$. Elle est *égale à* $+1$ lorsque les deux variables étudiées évoluent dans le même sens et que leur représentation graphique correspond à une *droite affine croissante*. Lorsque deux variables évoluent dans un sens opposé, la valeur du coefficient de corrélation est *égale à* -1 et la représentation graphique de cette relation est une *droite affine décroissante*. Plus le coefficient est proche des valeurs extrêmes -1 et $+1$, plus la corrélation entre les variables est *forte*. Les valeurs intermédiaires (c'est-à-dire situées entre -1 et $+1$) renseignent sur le degré de dépendance linéaire entre les deux variables, *in extenso* l'intérêt de cette relation. Une corrélation égale à 0 signifie que les variables sont linéairement indépendantes mais ne signifie pas l'« absence » de relation car le coefficient de corrélation indique uniquement une dépendance linéaire ayant un comportement pouvant être globalement monotone (*e.g.* corrélation de manière exponentielle ou sous forme de puissance).

5.1.2 Valeurs statistiques annexes

En plus, des calculs des valeurs de corrélations, d'autres traitements sont effectués afin de mettre à disposition de l'expert de informations complémentaires, tels que des tests statistiques ou la valeur de l'Information Mutuelle des relations étudiées. L'expert peut ainsi prendre en compte ces valeurs lors de la définition d'un gène comme potentiel biomarqueur.

Les tests statistiques

L'objectif général d'un test statistique est de tester une *hypothèse* (noté H_0) définie et relative à un ensemble de données. Cette hypothèse consiste à poser comme une *affirmation* un état de fait et à le considérer comme *vrai* ; c'est cette notion de vérité qui est testée. Fondamentalement, les tests statistiques permettent de définir la probabilité de commettre une erreur est fortement dépendante de l'interprétation qui en est faite. Ces tests peuvent être vus comme des mesures subjectives. Néanmoins, il existe deux manières de se tromper lors de la réalisation et de l'interprétation d'un test statistique. Ainsi, il est possible de :

rejeter à tort une hypothèse H_0 lorsqu'elle est vraie. Ce risque est dit de première espèce et est noté α : c'est la probabilité d'avoir un *faux-négatif* ;

accepter à tort une hypothèse H_0 lorsqu'elle est fausse. Ce risque est dit de deuxième espèce et est noté β : c'est la probabilité d'avoir un *faux-positif* ;

Information Mutuelle

Afin de produire une information reflétant globalement l'intérêt d'une relation même si elle ne présente pas d'intérêt en terme de corrélation linéaire globale ou partielle, il est possible de calculer l'*Information Mutuelle* (noté I) de cette relation. I est une « quantité » mesurant la dépendance statistique entre (au moins) deux variables. L'Information Mutuelle I d'un ensemble de données (X, Y) correspond au degré de dépendance, au sens probabiliste, entre ces deux ensembles. Comme le coefficient de corrélation, de manière générale, I n'implique pas de relation de causalité entre X et Y. Plus I tend vers 0, plus les variables X et Y sont indépendantes. Plus I tend vers $+\infty$, plus les variables X et Y sont dépendantes. Les notions de *corrélation* et d'*information mutuelle* sont fortement liées. Notre approche n'utilise pas directement la notion d'Information Mutuelle. Cette valeur est calculée pour être utilisée au cours du tri des résultats de la Fouille de Données. Elle permet de donner à l'expert une indication supplémentaire sur l'intérêt d'une relation.

5.1.3 Approche globale

Il est possible d'intégrer les différentes mesures que nous avons présentées, afin de rechercher des relations dans des données. L'approche globale que nous proposons dans le cadre de notre méthodologie consiste à calculer automatiquement toutes les corrélations « globales » (et les statistiques associées) existantes entre tous les gènes (pour lesquels on a l'expression génique) et chaque paramètre bioclinique disponible, comme décrit dans les algorithmes 5.1, 5.2 et 5.3.

ALGORITHME **5.1** Algorithme de calcul des valeurs statistiques univariées : StatistiqueUnivariee

Soit $D = [d_1, d_2, \ldots, d_p]$: l'ensemble des données à étudier
$moy \leftarrow \text{mean}(D)$ {Moyenne de D}
$med \leftarrow \text{median}(D)$ {Médiane de D}
$kurt \leftarrow \text{kurtosis}(D)$ {Coefficient d'amplitude de D}
$skew \leftarrow \text{skewness}(D)$ {Coefficient d'aplanissement de D}
$stat \leftarrow [moy, med, kurt, skew]$
Retourner *stat*

Le résultat de la mise en œuvre de cet algorithme est un ensemble de corrélations décrivant de manière simple les liens existant entre deux ensembles de données étudiées.

ALGORITHME **5.2** Algorithme de calcul des valeurs statistiques bivariées : StatistiqueBivariee

Soit $c = [cl_1, cl_2, \ldots, cl_m]$: valeurs pour un paramètre bioclinique
Soit $g = [ge_1, ge_2, \ldots, ge_n]$: valeurs pour un gène
$nb \leftarrow$ le nombre de couples (c, g)
$\rho_S \leftarrow$ corrélation de rang de Spearman pour (c, g)
$p \leftarrow$ significativité de ρ_S
$univG \leftarrow$ StatistiqueUnivariee(g)
$univC \leftarrow$ StatistiqueUnivariee(c)
$regCG \leftarrow$ regression (c, g) {Paramètres a et b de la droite de régression $g = a.c + b$}
$regGC \leftarrow$ regression (g, c) {Paramètres a' et b' de la droite de régression $c = a'.g + b'$}
$IM \leftarrow$ InformationMutuelle (g, c) {Calcul de la valeur de l'Information Mutuelle entre c et$g$}
$stat \leftarrow [nb, \rho_S, p, univG, univC, regCG, regGC]$
Retourner *stat*

ALGORITHME **5.3** Algorithme de calcul des corrélations « globales » : CorrelationGlobale

Soit $C = [c_1, c_2, \ldots, c_m]$: l'ensemble des paramètres biocliniques
Soit $G = [g_1, g_2, \ldots, g_n]$: l'ensemble des gènes
$stat \leftarrow$ NULL
Pour Tout $c_i \in C$ **Faire**
  $cvs \leftarrow$ NULL {Les statistiques pour un couple}
  $tabCVS \leftarrow$ NULL {Les statistiques pour tous les couples contenant c_i}
  Pour Tout $g_j \in G$ **Faire**
    $cvs \leftarrow$ StatistiqueBivariee(c_i, g_j)
    Ajouter cvs à $tabCVS$
  $Q \leftarrow$ calcul du FDR pour tous les p de $tabCVS$
  Ajouter Q dans $tabCVS$
  Ajouter $tabCVS$ à $stat$
Retourner *stat*

Il est important de noter que le nombre de valeurs numériques que nous allons obtenir après application de l'algorithme est considérable. Par exemple, si on dispose de données relatives à 30 individus décrits chacun par une puce de 40000 expressions géniques et 20 attributs, le nombre de résultats obtenus sera de 800 000 valeurs de corrélations et autant pour chaque valeur statistique complémentaire (test de significativité p, de fausses découvertes q, paramètres de la droite régression linéaire affine,...). Bien que le nombre de résultats soit conséquent, l'intérêt et l'originalité de cette approche est qu'elle réduit aux maximum les *a priori* dans le calcul des relations ; elle offre donc la possibilité de présenter à l'expert toutes les relations globales existantes dans les données initiales.

La complexité du calcul du coefficient de corrélation de Spearman qui est égale à $O(n^3 log(n))$ [Croux 2000]. La complexité de cet algorithme 5.3 est liée au nombre d'individus n, de gènes g et de paramètres biocliniques p pour lesquels on dispose de valeurs. La complexité globale de cet algorithme est donc au minimum égale à $O((n^3 log(n))gp)$; nous faisons ici abstraction de la complexité du calcul des valeurs statistiques annexes. Ceci en fait, de manière générale, un algorithme exigeant en ressources afin de fournir des résultats dans un temps raisonnable.

5.1.4 Interprétation des corrélations globales

L'approche que nous avons développée génére à un très grand nombre de corrélations. Or, l'interprétation d'une valeur de corrélation n'est pas quelque chose de triviale. Elle est fortement dépendante du domaine d'application. Plusieurs guides d'interprétation ont été proposés dans différents domaines, comme le montre la figure 5.1. Dans le cadre d'études de Psychologie, Cohen [Cohen 1988] propose l'interprétation suivante : si $|r| \in [0,10;0,29]$ alors la corrélation est faible, si $|r| \in [0,30;0,49]$ alors elle est moyenne et si $|r| \in [0,50;1,00]$ la corrélation est forte. En Physique, Hopkins [Hopkins 2004] propose une interprétation plus fine applicable à l'interprétation de résultats issus d'expérimentations. Lorsque $|r| \in [0,00;0,09]$ la corrélation est quasi-nulle ; $|r| \in [0,10;0,29]$, elle est faible ; $|r| \in [0,30;0,49]$ elle est moyenne ; $|r| \in [0,50;0,69]$, elle est forte ; $|r| \in [0,70;0,89]$ elle est très forte ; $|r| \in [0,90;0,99]$, elle est quasi-parfaite ; $|r| = 1,00$, la corrélation est parfaite. D'autres guides ont aussi été proposés dans d'autres domaines en fonction des besoins des utilisateurs [Dodge et Rousson 1999, Grawe et Mulligan 2002].

Ces guides d'interprétation nous permettent de montrer l'importance de la prise en compte du domaine d'applications pour l'interprétation des valeurs de corrélation. Ainsi pour interpréter leurs résultats les experts, de part leur pratique, ont l'habitude (inconsciente) d'utiliser une échelle de corrélations. Nous nous sommes donc intéressés à la définition empirique de cette échelle. Nous proposons aussi un second guide d'interprétation qui dans la pratique semble plus adapté car basée sur la Statistique.

Ainsi, le premier guide d'interprétation des valeurs de corrélation pour la Génomique Médicale Fonctionnelle que nous avons défini est fondé sur la base de discussions avec des biologistes et d'observations empiriques de leurs méthodes de travail. Ainsi, si $|r_S| \geq [0,66;1,00]$ alors la corrélation est potentiellement intéressante, sinon elle ne l'est pas. On note ce seuil r_{Se}.

Le second guide d'interprétation des valeurs de corrélation est plus stringent, il est fondé sur la valeur statistique r_S^2 qui correspond au coefficient de détermination d'une relation, c'est-à-dire au pourcentage de valeurs de la relation qui permet de la « justifier ». Ainsi, pour qu'au moins 50% d'une relation soit determinée par ses données, il est nécessaire que $r_S \geq \sqrt{0.50}$ soit $r_S \geq 0,71$. Ainsi, si $|r_S| \in [0,71;1,00]$, alors la corrélation est potentiellement intéressante d'un point de vue statistique. On note ce seuil r_{Ss}.

Ces deux guides sont graphiquement présentés par la figure 5.2. Par défaut, il nous semble plus pertinent de conseiller aux experts le premier guide, celui qui utilise le seuil d'intérêt ρ_{Se}, car il est

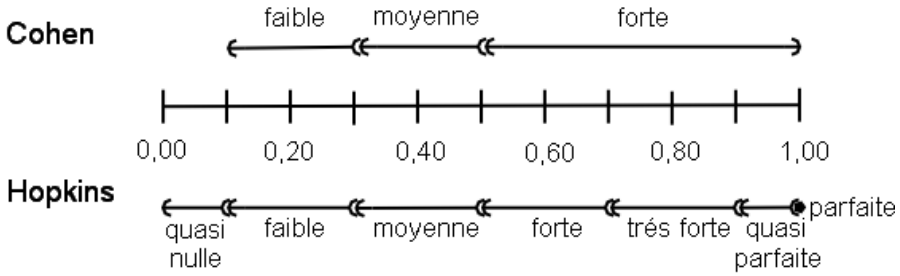


FIGURE 5.1 – Exemples de guides d'interprétation de valeurs de corrélation : le guide de Cohen en Psychologie [Cohen 1988] et le guide de Hopkins en Physique [Hopkins 2004].

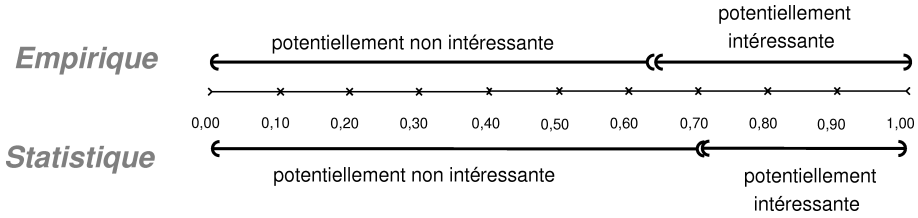


FIGURE 5.2 – Guides d'interprétation des valeurs de corrélation en Génomique Fonctionnelle.

plus proche de leur pratique. Néanmoins, le second guide est intéressant dans la pratique, car il est plus contraignant, il permet donc d'avoir plus rapidement accès aux relations potentiellement les plus intéressantes. L'originalité et l'intérêt de ces deux guides d'interprétation résident d'une part dans leur définition car il n'en n'existe pas, à notre connaissance, dans le domaine de la Génomique Médicale Fonctionnelle ; d'autre part, ils sont définis de manière à être stringents dans la perspective de mettre en avant plus facilement de potentiels biomarqueurs.

5.1.5 Vers une approche locale

L'étude des corrélations globales occupent une place majeure dans les études réalisées par les biologistes et de manière plus générale par les utilisateurs des outils statistiques. Néanmoins, il est nécessaire de préciser que la présence de sous-populations dans une population principale justifie l'intérêt de définir une approche permettant de prendre ces cas de figure en compte.

Les corrélations partielles [Labart *et al.* 2000, Everitt 2002] permettent de regrouper des données de deux ensembles pour lesquelles on cherche à définir la valeur du coefficient de corrélation sur un intervalle de valeurs de données issues d'un troisième ensemble. Ce type d'approche n'est efficace et ne fournit de résultats précis que si le nombre de valeurs étudiées est suffisant, c'est-à-dire qu'il comporte au moins 10 individus [Nelson 2004]. Cette condition ne peut pas être respectée dans le contexte de la Génomique Médicale Fonctionnelle. En effet, dans certaines études il est possible de se retrouver dans cette situation (en raison des données manquantes, notamment). De plus, les connaissances actuelles des biologistes/génomiciens ne permettent pas de comprendre ou de justifier facilement des relations complexes avec peu de données de qualité relative (beaucoup de variabilités inter-individuelles et inter-expérimentales).

Nous proposons, dans le cadre de notre approche, l'étude de *corrélations locales* [Papadimitriou *et al.* 2006]. Ce terme caractérise une approche principalement utilisé dans le domaine de

la Finance et de la Géologie. Son étude en Génomique Médicale Fonctionnelle est originale car elle n'est pas étudiée dans ce domaine de manière systématique et automatique. La *corrélation locale* est une approche conjuguant les spécificités de la *corrélation globale* et de la *corrélation partielle*. En effet, la *corrélation locale* n'est calculée que sur deux variables comme dans le cas de la *corrélation globale* et le résultat de ce calcul se limite à un intervalle restreint de valeurs de l'ensemble des données initiales, comme dans le cas de la *corrélation partielle*. L'intérêt et l'utilité de l'étude des corrélations locales sont qu'elles permettent de définir des sous-populations pouvant présenter des relations particulières. Son étude en Génomique Médicale Fonctionnelle est originale car elle n'est pas étudiée dans ce domaine que ce soit de manière systématique ou automatique.

Dans le paragraphe suivant, nous allons proposer une méthode pour étudier des corrélations locales et préciser les résultats obtenus par les corrélations globales.

5.2 Étude de corrélations locales

5.2.1 Définition par l'exemple de la corrélation locale

Une corrélation locale est une corrélation qui est définie sur un sous-ensemble de données à deux dimensions.

Les figures 5.3 et 5.4 présentent différents exemples de valeurs de corrélations obtenues sur des ensembles comprenant un grand nombre de données bidimensionnelles avec des topologies spatiales très différentes. Il permet de justifier l'intérêt de la recherche de corrélation locale, même sur des ensembles avec un effectif réduit.

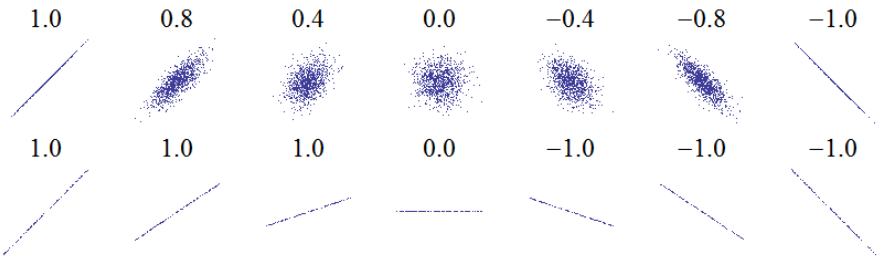


FIGURE 5.3 – Exemples de valeurs de corrélations pour des ensembles bivariées (d'après Wikipedia).

FIGURE 5.4 – Exemples de distrubutions spatiales d'ensembles bivariées dont la valeur de corrélation est égale à zéro (d'après Wikipedia).

La figure 5.3 présente des résultats qu'il est possible d'obtenir avec l'approche globale. Les résultats égaux à $|1|$ (ou $|0,8|$) sont obtenus sur des données ayant une topologie spatiale idéale (ou quasi-idéale) [Jiang et Omer 2007] : c'est-à-dire avec un nuage de points regroupés de manière à former une droite ou une courbe monotone. Les autres résultats sont obtenus sur des données diffusent dans l'espace de représentation.

La figure 5.4 présente des relations dont la valeur de corrélation est égale (ou proche de) 0. Le premier et le second exemples sont des cas qu'il n'est pas possible de rencontrer dans la nature (dans le domaine de la Biologie), car les données étudiées ne sont jamais aussi « parfaites » en terme de topologie, nous considérons donc qu'il s'agit d'un cas classique de corrélation nulle comme ceux présentés à la figure 5.3. Le troisième et le quatrième exemples montrent des exemples dans lesquels la relation existante entre les deux ensembles est composée successivement de segments croissants et décroissants. Ces enchaînements font qu'en terme de valeurs de corrélations, d'un point de vue globale, les valeurs propres à chaque segment s'annulent. Il en résulte que la valeur de corrélation décrivant la relation est nulle et ne reflète en rien l'existence de segments d'intérêt : une corrélation proche de $|1|$ sur chacun des segments peut présenter un intérêt du point de vue biologique car un premier segment peut définir un effet dans un contexte particulier et son contraire au-delà d'un certain seuil. Les trois derniers exemples sont composés de sous-ensembles de données qu'il n'est possible de détecter qu'à la condition de décomposer leur espace de projection suivant deux dimensions. Ces situations sont plus faciles à étudier en utilisant des techniques de regroupement, notamment les méthodes de grilles que par des approches statistiques. Nous ne nous intéresserons pas ici à ce type de situations, car elles ne présentent pas d'intérêt du point de vue de la Génomique Médicale Fonctionnelle puisqu'elles ne correspondent ni à des situations « réelles », ni à des situations interprétables biologiquement par les experts.

L'intérêt du calcul du *coefficient de corrélation locale*, qu'il le soit suivant le coefficient de corrélation de Pearson, de Spearman, de Kendall ou bisériel, peut être mis à profit dans le cadre de la recherche de biomarqueurs en Génomique Médicale Fonctionnelle. L'utilité et l'avantage du calcul du coefficient de corrélation locale résident dans le fait de définir et de découvrir :

- d'un point de vue informatique, des sous-ensembles de données présentant des caractéristiques particulières telle qu'une valeur de coefficient de corrélation qui soit « optimale » sur un sous-ensemble de données, c'est-à-dire ayant la meilleure valeur possible sur un intervalle donné ;
- d'un point de vue biomédicale, des sous-populations d'individus ayant un comportement similaire ou variant linéairement dans un contexte particulier vis-à-vis d'un facteur particulier (par exemple, variation de l'expression d'un gène au cours d'un régime chez des individus en état d'obésité et non chez les individus dits « sains »).

L'approche que nous proposons pour la recherche de corrélations locales s'appuie sur le concept de fenêtrage afin de permettre la découverte, au sein des données explorées, de segments d'intérêts définis par leur valeur de corrélation.

5.2.2 Les algorithmes de fenêtrage

Le fenêtrage permet de manière générale de définir un intervalle de données sur lequel on va effectuer des calculs dans la perspective de détecter un motif au sein d'une série temporelle ou un comportement particulier qui se distingue du reste de l'ensemble des données. Plusieurs grandes approches de fenêtrage sont possibles [Hugueney 2003].

La plus simple est fondée sur une taille minimale de la fenêtre, t, est consiste à effectuer des calculs sur les t premiers points de l'ensemble étudié puis sur les t points suivants et ainsi de suite jusqu'à ce que tous les points aient été inclus dans une « fenêtre ». Cette approche ne présente pas d'intérêt en Génomique Médicale Fonctionnelle car nous ne connaissons pas la topologie des données et donc la taille de la fenêtre optimale à utiliser. De plus, ces données ne sont pas des données répondant à un critère de motif qui se répèterait à intervalle régulier. La probabilité de trouver des sous-populations dans ce contexte est donc très faible.

Une seconde approche consiste à fixer une taille minimale t de fenêtrage et de calculer les

valeurs statistiques souhaitées sur ces données. Par la suite, on fait glisser cette fenêtre de p points et on répète les calculs sur ces t nouveaux points. On répète le glissement de la fenêtre jusqu'à avoir parcouru l'ensemble des données. Dans cette approche, nous cherchons la fenêtre ayant les meilleurs résultats sur l'ensemble de données.

Une troisième approche consiste à fixer une taille de fenêtre t, à calculer les valeurs statistiques souhaitées puis à élargir la fenêtre de p points afin de réaliser les calculs statistiques sur ce nouvel ensemble de données. L'élargissement de la fenêtre s'effectue jusqu'à ce que tous les points soient inclus dans l'ensemble. Cette approche permet de rechercher le sous-ensemble de points présentant la meilleure corrélation sur les n premières données.

Une quatrième approche correspond à mixer les deux approches précédentes. Elle consiste à appliquer la « troisième » méthode puis à faire glisser la fenêtre de p points (p étant le nombre de points qu'on ajoute à chaque nouvelle itération du « troisième » algorithme) et d'appliquer de nouveau la « troisième » méthode. Ce processus est réitéré tant que la fenêtre « initiale » n'inclut pas les derniers points de l'ensemble des données étudiées.

Dans notre contexte applicatif, le but de fenêtrage va être de mettre en évidence des sous-populations présentant des relations définies par une bonne valeur de corrélation (la meilleure possible pour un nombre minimal d'individus). Nous allons donc nous appuyer sur la quatrième approche.

5.2.3 Construction d'un algorithme de fenêtrage pour les corrélations

Les différentes approches exposées précédemment expliquent le principe du fenêtrage. Dans le cadre de l'utilisation du fenêtrage dans notre problématique, un double problème général se pose. En effet, il est nécessaire de définir la valeur de t *a priori*. La solution la plus simple et la plus théoriquement pertinente est de s'appuyer sur les bases de la Statistique [Dodge et Rousson 1999, Nelson 2004] : les résultats d'une analyse statistique bivariée sont pertinents s'ils incluent au moins 10 données. Nous allons donc poser $t = 10$ dans le cadre de notre algorithme.

De plus, il faut définir la notion de « bonne » fenêtre. Dans notre cas, une « bonne » fenêtre est une fenêtre qui a une très bonne corrélation. Cette très bonne corrélation peut être définie par rapport soit par l'utilisateur ou de manière relative par rapport à d'autres valeurs calculées précédemment lors du processus de fenêtrage. Afin de ne pas biaiser nos résultats, nous avons choisi la seconde solution et donc une très bonne corrélation est une corrélation qui sera meilleure que la précédente et que la suivante lors des calculs successifs pour des fenêtres consécutives.

L'algorithme 5.4 permet de calculer les corrélations locales grâce à notre approche de fenêtrage. L'idée de cet algorithme consiste à prendre en compte progressivement les couples de données (valeurs d'expression génique *vs.* valeurs biocliniques) et à calculer à chaque fois la corrélation associée. On répète l'opération jusqu'à trouver notre maxima local (en terme de corrélation) et on définit l'ensemble des données associées à cette corrélation comme étant le premier segment. On renouvelle l'opération jusqu'à ce qu'il n'y ait plus de potentiel segment à découvrir, c'est-à-dire plus d'individus à traiter pour le couple étudié. Il est possible de faire *glisser* la fenêtre afin de détecter des segments d'intérêt qui ne débutent pas avec les premières données. Cet algorithme est efficace sur des ensembles de données d'effectifs restreints (quelques dizaines de tuples) comme c'est le cas en Génomique Médicale Fonctionnelle. Néanmoins, sur des ensembles avec des effectifs plus importants (quelques centaines de tuples) sa principale limitation est le temps de calculs nécessaire à son utilisation car il croît de manière exponentiel avec le nombre de tuples à traiter.

ALGORITHME **5.4** L'algorithme de calcul des corrélations « locales » : CorrelationLocale

Soit $C = [c_1, c_2, \ldots, c_m]$: l'ensemble des paramètres biocliniques
Soit $G = [g_1, g_2, \ldots, g_n]$: l'ensemble des gènes
Soit $t = 10$: le nombre minimum d'individus par segment
$stat \leftarrow$ NULL
{Parcours de toutes les relations}
Pour Tout $c_i \in C$ **Faire**
  Pour Tout $g_j \in G$ **Faire**
    $nb \leftarrow$ le nombre de couples dans la relation (c_i, g_j)
    $rtmp \leftarrow$ NULL {la liste des segments découverts}
    $cmp \leftarrow 0$ {le nombre d'individus traités}
    Tant Que $cmp < (nb - t + 1)$ **Faire**
      $cmp \leftarrow cmp$+1
      $F \leftarrow$ NULL {fenêtre en cours}
      $RS \leftarrow$ NULL {liste des corrélations calculées}
      {Remplir la fenêtre avec les t premiers individus}
      Pour $k = 1$ to t **Faire**
        Ajouter ($c_{i,cmp}$, $g_{j,cmp}$) à F
        $cmp \leftarrow cmp$+1
      $rsp \leftarrow$ CorrelationSpearman(F)
      Ajouter rsp à RS
      {Ajouter un individu supplémentaire}
      Ajouter ($c_{i,cmp}$, $g_{j,cmp}$) à F
      $cmp \leftarrow cmp$+1
      $rsp \leftarrow$ CorrelationSpearman(F)
      Ajouter rsp à RS
      {Recherche du meilleur segment}
      Si $RS[0] > RS[1]$ **Alors**
        Répéter
          Ajouter ($c_{i,cmp}$, $g_{j,cmp}$) à F
          $cmp \leftarrow cmp$+1
          $rsp \leftarrow$ CorrelationSpearman(F)
          Ajouter rsp à RS
        Jusqu'à ($cmp == nb$) OU ($|RS[cmp{-}t{-}3]| < |RS[cmp{-}t{-}2]|$ ET $|RS[cmp{-}t{-}2]| > |RS[cmp - t - 1]|$)
      Sinon
        $cmp \leftarrow cmp$+1
      {Mise à jour des variables}
      Si cmp != nb **Alors**
        $cmp \leftarrow cmp$-3
        Enlever les 2 derniers couples de F
      {Sauvegarder le segment}
      $cvs \leftarrow$ StatistiqueBivariee(F)
      Ajouter cvs à $rtmp$
    Ajouter $rtmp$ à R
Retourner $stat$

5.3 Classification et visualisation des corrélations

Les notions de classification et de visualisation sont essentielles pour assister l'utilisateur dans l'exploration de nombreux résultats obtenus par des approches automatiques. D'après Kovalerchuk [Kovalerchuk 2001a], les méthodes de visualisation des corrélations sont un des éléments fondamentaux de la prise de décision. Trois niveaux de visualisation existent pour leur analyse et la prise de décision :

1. un premier où les résultats sont présentés de manière à faire ressortir uniquement les résultats les plus intéressants pour faciliter l'interprétation ;
2. le second, qualifié de niveau moyen, permet une interactivité avec l'utilisateur mais ne fait pas ressortir de manière exhaustive les résultats les plus intéressants ;
3. le dernier niveau ne fait intervenir que des mécanismes de perception visuelle et la subjectivité de l'observateur.

Du point de vue de la Génomique Fonctionnelle, ces différentes approches sont importantes dans la perspective d'aide à la découverte de relations entre des paramètres biocliniques et les valeurs d'expression génique.

Afin de permettre à l'utilisateur, plus précisément au biologiste, d'exploiter au mieux les résultats obtenus de manière automatique, il est nécessaire de lui présenter visuellement les résultats par ordre d'intérêt potentiel. Pour répondre à ce problème double, nous proposons deux langages de reformulation :

– d'une part, un langage symbolique de description des résultats L_S
– d'autre part, un langage de visualisation des données L_V.

Ainsi, notre préoccupation majeure est ici plus cognitive qu'algorithmique [Varela 1997], c'est-à-dire que nous ne souhaitons pas innover en créant un nouvel algorithme mais en tentant d'améliorer la perception de l'information par l'utilisateur.

Reformuler pour décrire

Le langage L_S est un langage symbolique qui va aider à simplifier l'interprétation des valeurs de corrélation. Il s'appuie sur une observation des méthodes de travail des biologistes, avec lesquels nous collaborons. Un gène peut être corrélé positivement, négativement ou non corrélé avec un paramètre clinique. Dans une perspective d'aide à la découverte de potentiels biomarqueurs, il est nécessaire que les valeurs des corrélations considérées comme positives ou négatives soient suffisamment élevées pour avoir un intérêt potentiel. En s'appuyant sur nos guides d'interprétation (cf. 5.2), nous avons donc défini un langage symbolique qui permet l'interprétation des corrélations sous la forme d'une chaîne de caractères simple :

– une corrélation positive est représentée par un « / » ;
– une corrélation négative par un « \ » ;
– une corrélation nulle, potentiellement sans intérêt par un « ◯ ».

Ainsi, plutôt que de présenter les résultats des méthodes de calcul des coefficients de corrélations par des nombres, nous allons les présenter à l'utilisateur par des symboles. Ainsi, toutes les corrélations positives seront facilement identifiables par le symbole « / ». L'utilisateur n'aura donc plus à se poser la question du seuil. Le tableau 5.5 montre un exemple d'un tel changement de représentation pour des données mono- et multi-segments.

Comme nous l'avons mentionné précédemment, une corrélation potentiellement sans intérêt peut tout de même être « intéressante ». Ainsi, nous avons choisi d'étendre notre langage L_S avec deux autres symboles :

– si l'ensemble des points peut être assimilé à une ellipse dont le grand axe est quasiment parallèle à l'axe des abscisses, il sera symbolisé par « _ » ;
– si l'ensemble des points est quasiment parallèle à l'axe des ordonnées, il sera symbolisé par « | ».

La définition d'une ellipse est fonction de la discrétisation des ensembles des données d'expression génique et des paramètres cliniques réalisés au cours de la mise en œuvre du langage de reformulation L_S ; une ellipse sera un sous-ensemble de données dont l'étendue des valeurs (en terme d'intervalles) sur l'un des axes sera au minimum égale à la moitié de l'étendue sur l'autre axe. Le tableau 5.5 montre dans la dernière colonne la reformulation dans le langage enrichi. Ce langage permet de classer les différents types de relations découvertes en fonction de leur reformulation, ce qui en facilite leur exploitation par l'utilisateur.

$\rho_{globale}$	L_{S0}	$\rho_{Seg.1}$	$\rho_{Seg.2}$	$\rho_{Seg.3}$	$\rho_{Seg.4}$	L_{S1}	L_{S2}
$+0,70$	/	$+0,70$				/	/
$-0,70$	\	$-0,70$				\	\
$+0,10$	◯	$+0,10$				◯	|
$+0,10$	◯	$+0,10$				◯	_
$+0,53$	◯	$+0,68$	$+0,69$			◯	//
$+0,24$	◯	$+0,68$	$-0,69$			◯	/\
$+0,24$	◯	$+0,68$	$-0,15$			/◯	/_
$-0,32$	◯	$-0,08$	$-0,71$			◯\	| \
$+0,67$	/	$+0,68$	$+0,75$	$-0,20$		//◯	//_
$+0,20$	◯	$+0,68$	$-0,75$	$-0,20$	$+0,68$	/\◯	/_/
$-0,20$	◯	$+0,68$	$-0,75$	$-0,20$	$+0,68$	/\ ◯ /	/\ | /

FIGURE 5.5 – Exemples de relations définies dans les langages L_s standard et étendu. $\rho_{globale}$ est la valeur de la corrélation globale d'un ensemble de données et L_{S0} sa reformulation ; $\rho_{Seg.1}$, $\rho_{Seg.2}$, $\rho_{Seg.3}$ et $\rho_{Seg.4}$ sont les valeurs des corrélations locales correspondantes à l'exploration de l'ensemble de données et L_{S1} et L_{S2} sont leurs reformulations respectivement avec les langages L_S standard et étendu.

Reformuler pour visualiser

Le langage L_V est un langage visuel dont le but est de représenter simultanément de manière simple et facilement compréhensible un ensemble de relations. L'idée maîtresse de ce langage est de visualiser, dans un espace à une dimension, à l'aide de gradients de couleurs, les relations « expression génique *vs* paramètres biocliniques ». Chaque point va correspondre à un couple de données : sa position correspondant à la valeur de l'un des deux attributs et sa couleur à la valeur de l'autre attribut. Plus précisément, dans le cadre de notre contexte applicatif, les couleurs utilisées pour la projection des valeurs du paramètre bioclinique sont le résultat d'une discrétisation de celles-ci en un nombre d'intervalles défini préalablement par l'utilisateur.

Un exemple d'utilisation du langage L_V est présenté à la figure 5.6. On y considère une relation $G1$ entre l'expression d'un gène $g1$ et un paramètre bioclinique (un phénotype). Les données de la relation sont initialement projetées dans un repère à 2 dimensions. Ensuite, les valeurs d'expression sont projetées sur l'axe des abscisse et les valeurs biocliniques correspondantes à chaque valeur d'expression sont représentées par une couleur. Plus le point est rose foncé, plus

la valeur du paramètre bioclinique (le phénotype) à une valeur élevé (ou tend vers $+\infty$) ; plus le point bleu est foncé, plus la valeur du paramètre bioclinique à une valeur faible (ou tend vers $-\infty$) ; un point blanc correspond à une valeur bioclinique nulle.

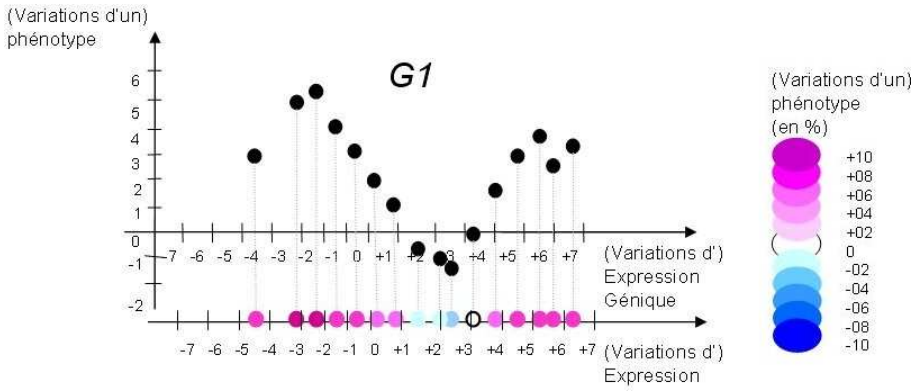


FIGURE 5.6 – Reformulation d'une relation « expression génétique *vs.* phenotype » dans le langage L_V [Benis 2005].

L'avantage du langage L_V est d'offrir la possibilité de comparer rapidement les relations pouvant exister entre l'expression génique de plusieurs gènes et la variation d'un paramètre clinique, comme il est possible de le constater à la figure 5.7.

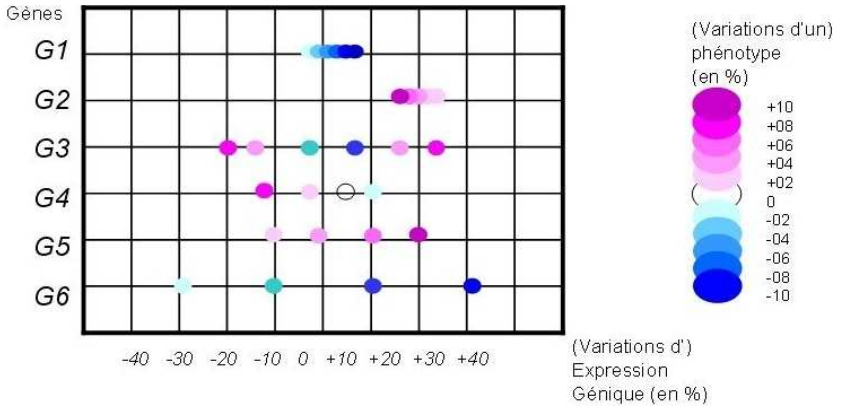


FIGURE 5.7 – Comparaison visuelle de plusieurs relations « gène *vs.* paramètre bioclinique » grâce au langage L_V.

Filter pour explorer

Il est ainsi une « porte d'entrée » vers une description plus précise des relations concernées par les nœuds sélectionnés.

En observant les méthodes de travail des biologites, nous nous sommes aperçus qu'ils analysaient les résultats de la manière suivante : ils trient les données en fonction de ρ_S, p et q et ils ne s'intéréssent qu'à celles qui sont maximales. Parfois, ils regardent aussi les relations avec beaucoup d'individus. En se basant sur cette observation, nous avons mis en place une méthode visuelle de filtrage des résultats. L'idée repose sur l'utilisation d'un diagramme de Hasse [Cour-

tine 2002]. Ce diagramme va permettre de regrouper dans des nœuds (figure 5.8) les résultats qui ont une valeur définie comme « pertinentes » pour ρ_S (noté r dans le diagramme) et/ou p et/ou q et/ou n.

Un diagramme de Hasse est une représentation visuelle d'un ordre fini E incluant un nombre N d'éléments (ici $E = \{r, p, q, n\}$ et $N = 4$). Les nœuds de ce diagramme représentent toutes les parties existantes dans notre ensemble (soit $2^N = 2^4 = 16$ nœuds). Il existe une relation d'ordre partiel entre les nœuds. Ainsi, l'ensemble des éléments qui sont dans le nœud $\{r, p, q\}$ sont aussi dans les nœuds $\{r, p\}$, $\{r, q\}$ et $\{p, q\}$. Ces derniers ayant la particularité d'être moins stringents que le nœud $\{r, p, q\}$, ils vont donc couvrir plus d'éléments. De la même manière, les éléments communs aux nœuds $\{n\}$ et $\{p\}$ se retrouveront dans le nœud $\{n, p\}$ et uniquement cela.

Dans le cadre de nos travaux, le diagramme de Hasse est utilisé à la fois pour filtrer les résultats, mais aussi pour savoir combien de données sont concernées par ce filtrage. Son utilisation permet ainsi à l'utilisateur d'être guidé dans son exploration.

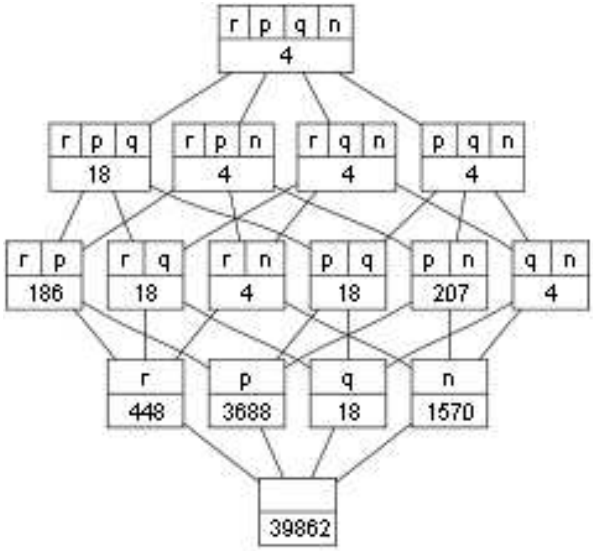


FIGURE 5.8 – Diagramme de Hasse représentant la classification d'un ensemble de relations globales calculées entre des valeurs d'un paramètre bioclinique et des données d'expression génique issues de puces à ADNc.

5.4 Conclusion

Les deux approches de calculs et de recherche de corrélations que nous avons présentées, permettent de détecter, sans *a priori*, des gènes pour lesquels il existe une relation potentiellement intéressante avec un paramètre bioclinique parmi l'ensemble de toutes les relations possibles. L'algorithme « global » (voir algorithme 5.3) permet une exploration automatique d'un grand ensemble de données et ainsi que la mise en exergue des relations qui peuvent conduire à la définition d'un biomarqueur sur une population « générale ». L'algorithme « local » (voir algorithme 5.4) comme le précédent permet une exploration automatique d'un grand ensemble de données ; il permet néanmoins de rechercher des relations plus précises sur des sous-ensembles de données. Ce type de relations ne sont pas « classiquement » recherchées par l'expert. Cet algorithme permet de définir de potentiels biomarqueurs pour une sous-population spécifique. L'utilisation de ces

deux approches est complémentaire car l'une permet de s'intéresser à toute la population étudiée et la seconde de découvrir des sous-populations ayant des « comportements » particuliers.

Les langages de reformulation L_S et L_V permettent de réduire le volume de résultats présenter à l'utilisateur. Néanmoins, les résultats « intéressants » peuvent être trop nombreux. Il peut s'avérer utile de mettre en œuvre des approches permettant de ne mettre en avant que les relations les plus pertinentes aussi bien en terme de statistiques que de visualisation [Kulkarni et Simon 1988] en réduisant le taux de fausses découvertes et en détectant les relations comprenant des valeurs singulières. Nous développerons ces points dans le chapitre suivant.

Chapitre 6

Réduire le risque de fausse découverte

> *L'empirique, semblable à la fourmi, se contente d'amasser et de consommer ensuite ses provisions. Le dogmatique, telle l'araignée ourdit des toiles dont la matière est extraite de sa propre substance. L'abeille garde le milieu ; elle tire la matière première des fleurs des champs, puis, par un art qui lui est propre, elle la travaille et la digère. (...) Notre plus grande ressource, celle dont nous devons tout espérer, c'est l'étroite alliance de ces deux facultés : l'expérimentale et la rationnelle, union qui n'a point encore été formée.* [Bacon 1620]

Un des problèmes majeurs, qui se pose lors de l'analyse de données réelles est la qualité des données traitées. Comme nous l'avons abordé précédemment, les données issues du processus expérimentales en Génomique Médicale Fonctionnelle peuvent être bruitées pour différentes raisons. Afin de pallier à ce problème, différentes approches existent [Benjamini et Hochberg 1995, Dudoit et van der Laan 2007]. Elles ont pour objectif d'informer l'expert de la pertinence des résultats obtenus. Les principales approches utilisées en Biologie sont les *tests statistiques de significativité* [Schervish 1996] et depuis quelques années les *tests de multiplicité* [Dudoit et Shaffer 2003, Dudoit et van der Laan 2007].

Un autre problème majeur auquel il est nécessaire de faire face lors d'un processus de Fouille de Données est celui de la présence de valeurs singulières dans les données (et non pas seulement de bruit) et qui peuvent influencer les valeurs des résultats d'analyse. Différentes approches permettant de pallier à ce problème ont été présentées au chapitre 3. Notre objectif n'est pas de proposer une méthode « robuste » aux valeurs singulières, mais une approche permettant d'informer l'expert de leur présence dans les résultats obtenus.

Dans ce chapitre, nous allons présenter deux approches que nous proposons pour réduire le risque de fausse découverte. La première correspond aux limites d'utilisation des *valeurs de significativité* et *de multiplicité*. La seconde correspond à une approche de détection des *valeurs singulières* grâce à une méthode d'apprentissage, PAMout, afin d'indiquer leurs présences à l'expert lors de l'analyse des résultats.

6.1 Significativité et multiplicité

Dans le contexte de nos travaux, la prise en compte du risque de fausse découverte est essentielle. En effet, les coûts de la validation biologique d'un potentiel biomarqueur sont élevés. Ainsi, les biologistes utilisent les valeurs issues de tests de significativité et de multiplicité afin de réduire un partie du risque de fausse découverte. Néanmoins, comme nous le présentons dans les lignes qui suivent, l'utilisation des valeurs issues de ces tests peut s'avérer problématique.

6.1.1 Mesures de significativité

Comme nous l'avons exposé précédemment à la section 5.1.2, l'objectif général d'un test statistique est de tester une *hypothèse* (noté H_0) définie et relative à un ensemble de données. Idéalement, les seuils α et β, utilisés pour ces tests, devraient être nulles. Néanmoins, en raison de l'imperfection du monde réel et par extension de la qualité non absolue des données, ce n'est que très rarement le cas. Il est donc nécessaire de définir pour chacun de ces risques un seuil d'acceptabilité du taux d'erreur. Classiquement, on calcule la significativité d'une valeur statistique par rapport à α et cette valeur de significativité est noté p. Ainsi, si $p \leq \alpha$, on rejette l'hypothèse H_0, c'est-à-dire que la valeur testée est très probablement un *vrai-positif* donc un résultat valide ; dans le cas inverse (pour $p > \alpha$), le résultat n'est pas intéressant au regard de α et donc la valeur n'est pas un résultat valide.

La principale limite des tests de significativité vient du fait que l'utilisateur doit fixer un seuil d'intérêt, une valeur pour α (ou pour β). La communauté biomédicale (comme la majorité des communautés scientifiques) est d'accord pour accepter dans leur analyse 5% de taux de faux-négatif (soit $\alpha = 5\%$ ou $p < 0,05$), c'est-à-dire qu'il accepte un risque de commettre une erreur de l'ordre de 5%. Le problème posé par cette « règle » est qu'elle ne prend pas en considération les effets dus à la taille de la population sur laquelle est réalisée le test, ni la qualité globale des données [Friedman et Bitterer 2009]. Par conséquent, ce *consensus* nous fait nous poser la question sur sa pertinence si l'on est dans des situations telles que, par exemple, $\rho_{s1} = 0,80$ avec $p_1 < 0,05$ pour une population d'effectifs $n_1 = 10$ et $\rho_{s2} = 0,80$ avec $p_2 < 0,05$ pour une population d'effectifs $n_2 = 50$. Il ne nous semble pas pertinent, bien que cela soit fait par nombre d'utilisateurs, de considérer ces résultats comme équivalents. La prise en compte de la taille de la population est quelque chose d'essentiel, surtout lorsque l'objectif est de généraliser les résultats obtenus à une plus large population.

Les utilisateurs de tests de corrélations ont tendance à accepter un résultat « médiocre » (r, la valeur du coefficient de corrélations, qui tend vers 0) si celui-ci à une valeur p inférieure à α. Cela pose un problème conséquent car ce résultat *significatif* ne correspond pas à un résultat *pertinent*. En effet, comme nous l'avons déjà mentionné une corrélation dont la valeur est proche de 0 n'est pas forcément le reflet d'une relation sans intérêt.

6.1.2 Problème de la multiplicité

Le concept de test de multiplicité a été développé dans la perspective de faire ressortir d'un très grand nombre de tests de significativité, ceux qui sont, ou tout du moins qui semblent être, les plus intéressants quand ils sont ajustés les uns par rapport aux autres. L'ajustement des valeurs de tests de significativité par les tests de multiplicité consistent à mettre en évidence les valeurs de significativité qui tendent les plus vers 0 et qui sont les moins redondantes.

Un certain nombre de méthodes ont été proposées pour ajuster les valeurs de p en fonction du nombre de tests statistiques effectués [Benjamini et Hochberg 1995, Tusher *et al.* 2001, Benjamini *et al.* 2001, Benjamini et Yekuteli 2001, Storey 2002, Storey et Tibshirani 2003, Storey 2005].

L'approche proposée par Bonferonni [Benjamini *et al.* 2001, Abdi 2007], dite *correction de Bonferonni*, consiste dans un premier temps à ordonner les valeurs de significativité p_i de manière croissante. Ensuite, chaque valeur p_i est comparée à $\alpha \times (i/n)$, où i désigne le rang de p_i dans la liste ordonnée et n est le nombre total de valeurs de la liste. Une valeur p_i est considérée comme significatif si $p_i \leq \alpha \times (i/n)$. Par exemple, soit une biopuce permettant de mesurer l'expression de 40000 gènes ; on déclare les n premiers gènes significatifs si la n^{eme} valeur p la plus faible est inférieure ou égale à $\alpha \times (n^{eme}/40000)$. Le problème, ici, est que le résultat est dépendant de α et

que seuls les résultats ayant une valeur de p très faible seront pris en considération. Si la valeur de α est trop faible, les chances de dégager un résultat significatif sont très faible, ce qui induit un test trop stringent et trop sélectif pour le type de données que nous souhaitons étudier.

Un autre test, couramment utilisé aujourd'hui en Biologie, est le *taux de fausse découverte* (en anglais *False Discovery Rate*, noté *FDR*). Ce test consiste à ajuster les valeurs d'un ensemble de tests de significativité p afin de mettre en évidence la proportion possible de faux positifs. L'ajustement des valeurs de significativité p s'appuie avec cette approche sur la distribution de ces valeurs. Cette approche est considérée comme moins stringente que d'autres approches d'ajustement et en particulier que la *correction de Bonferonni*. Le problème que soulève le *FDR* tient du fait que les résultats obtenus sont dépendants d'une distribution spécifique. Les valeurs de p soumises à l'ajustement doivent majoritairement tendre vers 0 pour que les valeurs ajustées tendent, elles aussi, vers 0. D'autre part, la distribution des valeurs p doit être *décroissante* et *monotone* ; ainsi, quand ce n'est pas la cas, les résultats de l'ajustement ne reflètent pas l'intérêt réel des données.

6.1.3 Conclusion

Notre objectif dans les lignes précédentes était de mettre en évidence de manière *non calculatoire* les problèmes posés par l'utilisation des tests de significativité et de multiplicité en tant que paramètres définis comme fortement discriminants par les utilisateurs. De manière générale, les statistiques telles que les *tests de significativité* et les *tests de multiplicité* servent d'indicateurs de probabilité d'intérêt d'une autre valeur statistique. Dans notre cas, cette dernière est un coefficient de corrélation.

Les résultats des tests de significativité et de multiplicité ne peuvent servir que de filtres dans le cas des corrélations. Ainsi, Efron [Efron 2006a] souligne que dans le cas des corrélations, ce qui compte n'est pas la valeur du résultat d'un test de significativité ou de multiplicité mais la valeur du coefficient de corrélation calculée. *Significativité* ne signifie pas « intérêt certain » mais une valeur de corrélation qui tend vers $|1|$ reflète une relation linéaire monotone et qui par extension est une relation qu'il est pertinent d'étudier. Par exemple, une *relation A* avec une corrélation $\rho = 0,30$ ayant une valeur de significativité $p \ll 0,05$ et dont l'ajustement $q \ll 0,05$ (le *FDR*) correspondent à une relation de n valeurs visualisables sous la forme d'un nuage fortement étalé sur ces deux dimensions et légèrement croissant, ne peut pas être selon nous (et *in extenso* selon Efron) considéré comme pertinente. Par contre, une *relation B* ayant un nombre identique de valeurs n mais avec une valeur de corrélation $\rho = 0,70$, $p \approx 0,10$ et $q \approx 0,10$, visualisable sous la forme d'un nuage fortement étalé sur une seule de ses deux dimensions et fortement croissant, est une relation pertinente car elle montre une liaison forte entre les deux variables qui la composent.

En conséquence, ces statistiques (*tests de significativité et de multiplicité*) sont donc à utiliser avec précaution car elles ne reflètent pas la structure des données. Afin de proposer aux biologistes, et plus largement à tout utilisateur de notre méthodologie des résultats potentiellement pertinents, nous conjuguons les approches ayant pour but de réduire le risque d'erreur et de fausses découvertes. Ainsi, les valeurs issues de tests de significativité et de multiplicité, ainsi que la valeur de l'Information Mutuelle (que nous avons introduit au chapitre précédent) servent à extraire les différentes relations afin de mettre en évidence celles dont la valeur de corrélation ρ_S tend vers $|1|$, de l'Information Mutuelle I tend vers $+\infty$ (dans le cas des relations composées de plusieurs segments) et dont la valeur de significativité et de multiplicité p et q (le *FDR*) tendent vers 0. Néanmoins afin de réduire le risque de fausse découverte, l'utilisation de tests statistiques n'est pas suffisante, car les données peuvent contenir des valeurs singulières qui ne sont pas forcément marginales.

6.2 PAMout : Détection de valeurs singulières via PAM

En plus des valeurs statistiques précédentes, l'intérêt d'une relation (*in extenso*, d'une corrélation) peut être défini en fonction des *valeurs singulières* qui peuvent influencer les valeurs descriptives de cette relation. Nous proposons dans cette partie une méthode permettant de mettre en évidence ces données de manière simple, rapide et objective.

6.2.1 De l'algorithme PAM à l'algorithme PAMout

Afin de mettre en évidence et de caractériser les valeurs singulières, qu'elles soient aberrantes, suspectes ou du bruit, nous proposons une approche basée sur l'algorithme d'apprentissage des *k-médoïdes*, aussi appelé *PAM* (Partitionning Around Medoids) [Kaufman et Rousseeuw 1990, van der Laan *et al.* 2003].

Le principe de l'algorithme *PAM* (donné à l'algorithme 6.1) consiste à partitionner l'espace des données autour de k médoïdes. Un médoïde est une donnée réelle de l'espace des données et qui va servir de centre à un des groupes. L'utilisateur doit fixer un nombre de groupes k que l'on souhaite obtenir. Au départ, l'algorithme choisit aléatoirement k données (ce sont les *médoïdes* initiaux). Ensuite, la distance entre chaque donnée et chaque médoïde est calculée et chaque donnée est affectée au groupe dont il est le plus proche du médoïde. La donnée la plus au centre de chaque groupe devient le nouveau médoïde. Tant que le partitionnement n'est pas stable, c'est-à-dire que les médoïdes changent, le processus est réitéré. Il est à noter que la stabilité du partitionnememnt dépend de son initialisation.

ALGORITHME **6.1** Algorithme PAM

Soit $D = d_1, d_2, \ldots, d_p$: l'ensemble des données
Soit k : le nombre de groupes
$R \leftarrow$ NULL {la liste des données organisées par groupe}
$M \leftarrow$ NULL {la liste des médoïdes}
$M \leftarrow$ choix aléatoire de k valeurs parmi D
Répéter
  Pour Tout $d_i \in D$ **Faire**
    Affecter d_i au groupe r_i dont le médoïde m_i est le plus proche
  Pour Tout $r_i \in R$ **Faire**
    $m_i \leftarrow$ calcul du nouveau médoïde de r_i
Jusqu'à convergence {Stabilisation des médoïdes}
Retourner R

Le principal inconvénient de l'algorithme PAM est la définition *a priori* de la valeur de k, le nombre de partitions qui sera extrait de nos données. Lorsqu'il s'agit d'étudier une base de données comprenant quelques dizaines d'enregistrements, l'utilisateur peut définir manuellement différentes valeurs de k et étudier les différentes partitions proposées afin de définir celle qui lui semble la plus pertinente. Le nombre de façons de partitionner un ensemble de n éléments en k sous-ensembles non vides est dénomé, *nombres de Stirling de deuxième espèce* [Adamchik 1997]. Le nombre définissant l'ensemble des k partitions possibles pour un ensemble de n éléments est appelé *nombre de Bell* [Rota 1964].

Or, dans le contexte de la Fouille de Données, c'est-à-dire l'exploration de grandes bases de données incluant de plusieurs centaines à plusieurs milliards d'enregistrements, il est nécessaire soit d'imposer une valeur fixe à k, soit de proposer une méthode permettant de trouver la valeur

optimale de k pour chaque ensemble de données étudié Ċette définition de valeurs optimales de k pose un problème, notamment dans le cadre de la détection de valeurs singulières. En effet, les nombres de valeurs de k possibles sont nombreuses et sont « observateur-dépendantes ». L'automatisation de la définition de k peut permettre de pallier à ce type de limitation liée aux valeurs élevées, respectives, du nombre de Stirling de Second espèce et du nombre de Bell, spécifiques à l'ensemble étudié.

Ray et Turi ont proposé une mesure pour définir le nombre de partitions optimales obtenues à l'aide de l'algorithme *k-moyennes* (ou en anglais, *k-means*) dans le contexte de la segmentation d'images colorées [Ray et Turi 1999]. L'objectif de cette mesure est d'optimiser le partitionnement d'un espace de données (une image). Elle consiste à minimiser la *distance intra-partition* (i) et à maximiser la *distance inter-partition* (I) [Frawley *et al.* 1992]. Ainsi, Ray et Turi calculent la somme des valeurs des distances intra-partitions (Σi) et la somme des valeurs des distances inter-partitions (ΣI) obtenues lors de l'exécution de l'algorithme des k-moyennes pour une valeur de k donnée. Ils en déduisent ensuite la mesure de validité v_k :

$$v_k = \frac{\Sigma i_k}{\Sigma I_k} \tag{6.1}$$

v_k reflète le rapport entre la *minimisation* des distances inter-individuelles et de la *maximisation* de distances inter-groupes. Plus v_k est petit, plus la *minimisation* (c'est-à-dire de la distance *au sein* des groupes) et la *maximisation* (c'est-à-dire de la distance *entre* les groupes) sont optimales. Ainsi, le premier minima de la valeur de v_k correspond au partitionnement le plus adapté aux données avec un nombre minimal de groupes. Nous reviendrons sur ce point dans le chapitre 8, lors des expérimentations.

Nous avons adapté la mesure de validité de Ray v_k à l'algorithme des k médoïdes comme le montre l'algorithme 6.2. Nous posons comme limite maximale à k, la valeur correspondant à la moitié du nombre d'individus N de l'ensemble à explorer (au-delà, les chances de découvrir des groupes à « 1 » individu étant élevé ce qui peut donc ne plus refléter une situation réaliste). Ainsi, plus le nombre k de médoïdes tend à être égal au nombre d'individus de la population étudiée, plus les chances que le nombre de médoïdes représentant chacun un seul et unique individu augmente. Si le partitionnement obtenu pour le k optimal, k_o, inclus une ou des partitions d'effectifs égale à « 1 », c'est-à-dire incluant un seul et unique individu, nous considérons cette (ou ces) partition(s) comme correspondant à la définition d'une (ou plusieurs) valeur(s) singulière(s). Si le partitionnement obtenu pour k_o ne contient pas de partitions d'effectifs « 1 », l'ensemble est considéré comme n'incluant pas de valeur singulière. Quand k est choisi de manière pertinente, un individu isolé peut être considéré comme une valeur singulière, car c'est un individu qui se différencie suffisament des autres pour être considéré comme divergent.

Une des limitations de PAM et de PAMout lors de leurs utilisations sur des grandes bases de données, est leur complexité qui est respectivement en $O(k(n-k)^2)$ et en $O(n^4)$ [Ng et Jiawei 1994, Park *et al.* 2006]. Néanmoins, malgré cette limitation, d'autres travaux, en Mécanique notamment, ont montré l'intérêt de PAM pour l'identification de valeurs singulières en analyse de données [Przybyklo 2005] avec la détermination manuelle de k.

6.2.2 Détection de valeurs aberrantes, suspectes et de bruit avec PAMout

Les valeurs singulières deviennent dans un contexte multivarié des données importantes. En effet, si une valeur d'un attribut pour un enregistrement (ou un exemple) donné est considérée comme « douteuse », elle peut conduire à la remise en cause de l'ensemble des résultats et l'exemple peut être totalement mis à l'écart pour tous les autres attributs. De manière plus

ALGORITHME **6.2** Algorithme PAMout

Soit $D = d_1, d_2, \ldots, d_n$: l'ensemble des données
Soit n : le nombre de données dans D
$nbMax \leftarrow (n/2) - 1$ {le nombre de groupes maximal}
$k \leftarrow 2$ {le nombre de groupes étudié}
$tabCluster \leftarrow$ NULL {tableau des groupes pour chaque k}
$tabVal \leftarrow$ NULL {tableau des validités pour chaque k}
$lOutliers \leftarrow$ NULL {liste des valeurs singulières}
{Recherche du k optimal}
Répéter
  $R \leftarrow$ PAM(D, k)
  Ajouter R à $tabCluster$
  $v \leftarrow$ validite(R)
  Ajouter v à $tabVal$
  $k \leftarrow k$+1
Jusqu'à ($k == nbMax$) OU ($k > 4$ ET $tabVal[k-3] < tabVal[k-2]$ ET $tabVal[k-2] > tabVal[k-1]$)
{Traitement du k optimal}
Si k != $nbMax$ **Alors**
  $k_o \leftarrow k$-2
  Pour $i = 1$ to k **Faire**
    $gp \leftarrow tabCluster[k_o + 1, i]$ {le ième groupe de la partition k}
    $n \leftarrow$ le nombre de données dans gp
    Si $n == 1$ **Alors**
      Ajouter gp[1] (la donnée) à $lOutliers$
  Retourner k_o et $lOutliers$

précise, il est important de connaître la nature d'une valeur singulière afin de pouvoir en déduire des connaissances pour le reste des analyses.

L'expert doit disposer de résultats facilement interprétables. Ainsi, nous proposons un guide d'interprétation des valeurs singulières en s'appuyant sur PAMout appliqué à un ensemble de données de 2 dimensions ($2D$). Ce guide peut être étendu à n dimensions (avec $n \geq 2$) en considérant que si au moins deux valeurs (c'est-à-dire deux dimensions) d'un exemple répondent aux critères définis dans la figure 6.1 alors cette valeur est « singulière ».

Cas	Dimension 1	Dimension 2	Ensemble	Type de données
A	0	0	0	« normal »
B	1	0	0	suspect
C	0	1	0	suspect
D	1	1	0	suspect/bruit
E	0	0	1	suspect/bruit
F	1	0	1	aberrant
G	0	1	1	aberrant
H	1	1	1	aberrant

FIGURE 6.1 – Guide d'interprétation des résultats de la mise en œuvre de PAMout sur des ensembles à *2* dimensions.

Afin d'expliquer ce guide d'interprétation (voir figure 6.1), nous considérons un ensemble de données à *2* dimensions, respectivement x et y :

- Si, au cours de l'exécution de PAMout, une donnée n'est pas considérée comme une valeur singulière que ce soit pour sa valeur en x, en y ou simultanément pour ses deux valeurs, alors cette donnée est considérée comme *normale* (cas A de la figure 6.1).
- Si une donnée sur l'une des 2 dimensions est détectée comme étant une valeur singulière, alors elle est considérée comme valeur *suspecte* (cas B et C de la figure 6.1).
- Si une donnée n'a pas été détectée lors de l'exécution de PAMout comme valeur singulière sur une dimension, mais uniquement sur l'étude des deux dimensions simultanément, alors cette donnée peut être soit une *valeur suspecte*, soit du *bruit* (cas D et E de la figure 6.1).
- Si une donnée dont au moins l'une des dimensions contient une valeur singulière et que cette donnée est aussi détectée comme valeur singulière lorsque les 2 dimensions sont étudiées simultanément, alors cette donnée est considérée comme *aberrante* (cas F, G et H de la figure 6.1).

Dans la partie expérimentations suivante, nous allons reprendre sur des exemples concrets chacun de ces cas afin de les étudier plus en détails.

6.2.3 Expérimentations sur des données artificielles

Les expérimentations que nous avons réalisées s'appuient sur l'implantation de PAMOUT sous la forme d'une librairie sous R [R Development Core Team 2006].

Les différentes exemples qui suivent permettent de mieux comprendre le fonctionnement de PAMOUT avec des ensembles de données dans un premier temps à *1 dimension* et dans un second temps à *2 dimensions*.

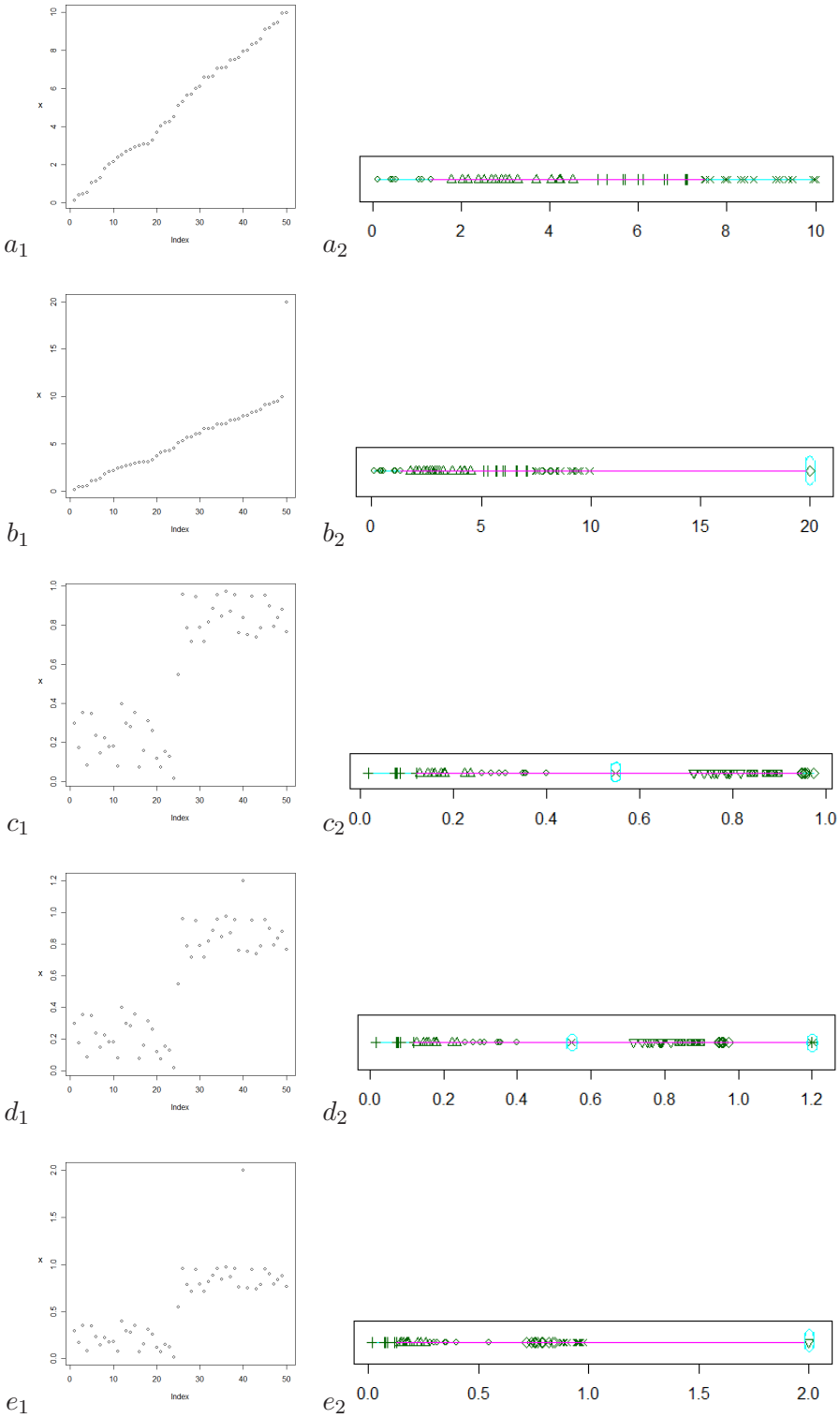


FIGURE 6.2 – Exemples d'ensembles de données univariées (indexées « 1 ») et de la représentation graphique de leur partitionnement (indexées « 2 »).

Valeurs singulières pour des données à une dimension

La figure 6.2 présente les ensembles des données univariés sur lesquelles nous avons testé PAMout. Les figures indexées « 1 » présentent des ensembles de données univariées (en ordonnée) identifiées par un index (en abscisse) qui n'est pas pris en compte pour le regroupement des données. Les figures indexées « 2 » présentent le résultat du regroupement des points en distinguant chaque groupe par un symbole différent. Les données détectées comme valeurs singulières sont entourées.

La figure 6.2a_1 présente un ensemble de données ne contenant pas *a priori* de valeur singulière. La mise en œuvre de PAMout permet de valider cette hypothèse. Sur la figure 6.2a_2, on ne distingue aucun point comme étant défini en tant que valeur singulière. En effet, les résultats de la figure 6.3 montrent que le regroupement *optimal* des données est obtenu pour $k = 4$ et que l'on obtient 4 groupes de données homogènes. Les approches statistiques basées sur les distributions ou sur l'étude des données marginales ne permettent pas de trouver ce type de résultats et conduisent à déclarer comme valeurs singulières des valeurs qui ne le sont pas. Les approches basées sur les distributions fournissent des résultats différents en fonction de la loi de distribution retenue. Or, dans le cas présent, comme les données ont été définies aléatoirement, elles ne répondent pas à une loi de distribution précise, les résultats seront donc erronés. Dans le cas du rognage, qu'il concerne la borne inférieure, supérieure ou les deux, des données sont inutilement supprimées par cette approche. Cependant, les approches basées sur les grilles [Wang *et al.* 1997] ou les densités [Breunig *et al.* 2000] obtiennent des résultats identiques à ceux de PAMout, si et seulement si elles sont paramétrées correctement.

#partition	v	#I/partition
2	8,123767	24/26
3	8,435848	13/17/20
4	**4,934696**	**7/12/14/17**
5	5,791000	7/9/9/11/14

FIGURE 6.3 – Résultats de PAMout sur les données univariées de la figure 6.8a_1.

La figure 6.2b_1 correspond aux mêmes données que celles présentées à la figure 6.2a_1, à différence près de la $50^{ème}$ donnée. Cette dernière a été modifiée afin d'être mise à l'écart des autres. La figure 6.2b_2 montre que cette donnée est détectée comme étant une valeur singulière. La figure 6.4 permet de préciser ce résultat. A $k_o = 5$, une valeur singulière est détectée. Cette donnée est définie comme une valeur singulière par des approches basées sur les distributions, car elle est très distante du reste des données. L'utilisation de la méthode de rognage sur la borne supérieure uniquement est pertinente, à condition de se limiter à une donnée, sinon des données « saines » seront perdues. Les approches issues de l'Intelligence Artificielle, telle que l'utilisation de grilles ou de regroupements basés sur la densité des données, ne sont efficaces que si elles sont paramétrées de manière optimale, c'est-à-dire ni trop stringente, ni trop laxiste.

La figure 6.2c_1 présente un ensemble de données constitué de 2 groupes de données séparés par un intervalle de même largeur que chacun des groupes. Au milieu de cet intervalle, se trouve une donnée isolée. Cette donnée est détectée comme valeur singulière comme le montre la figure 6.2c_2. La figure 6.5 montre que le k optimal est obtenu à la $7^{ème}$ itération ($k_o = 7$). La convergence est moins rapide que pour le cas précédent. La détection de cette donnée n'est pas possible avec les approches statistiques, car cette donnée n'est pas extrême. Les approches basées sur les grilles et la densité demandent chacune un paramétrage soigneux pour être efficace. Ces deux derniers

#partition	v	#I/partition
2	16,86179	24/26
3	17,67709	13/17/20
4	0,8161531	1/13/16/20
5	**0,7141015**	***1*/7/12/13/17**
6	0,7285032	1/7/8/9/11/14

FIGURE 6.4 – Résultats de PAMOUT sur les données univariées de la figure 6.8b_1.

algorithmes sont donc moins performants en terme de simplicité d'utilisabilisation par rapport à PAMOUT.

#partition	v	#I/partition
2	2,715794	24/26
3	3,920919	11/15/24
4	5,421616	11/11/13/15
5	6,796633	7/9/9/11/14
6	5,248556	6/7/8/9/9/11
7	**1,945677**	***1*/6/7/7/9/9/11**
8	1,987176	1/5/6/7/7/9/9

FIGURE 6.5 – Résultats de PAMOUT sur les données univariées de la figure 6.8c_1.

Les figures 6.2d_1 et 6.2e_1 présentent des données univariées dont la répartition est proche de celle présentée à la figure 6.2c_1.

La figure 6.2d_1 a la particularité de contenir 2 données isolées, distantes chacune d'une demi-largeur des groupes dominants de l'ensemble de données. Ces 2 données sont détectées comme des valeurs singulières comme le montre la figure 6.2d_2 et la figure 6.6 pour $k = 8$.

#partition	v	#I/partition
2	3,476398	24/26
3	5,054702	12/14/24
4	7,080064	11/12/13/14
5	8,918448	7/9/9/12/13
6	6,932822	6/7/8/9/9/11
7	1,955304	1/6/7/7/9/9/11
8	**1,071621**	***1*/*1*/6/6/7/9/9/11**
9	1,078611	1/1/5/6/6/6/7/9/9

FIGURE 6.6 – Résultats de PAMOUT sur les données univariées de la figure 6.8d_1.

La figure 6.2e_1 a la particularité de contenir 2 données isolées où la première est identique à celle présente entre les 2 groupes de données sur les figures 6.2c_1 et 6.2d_1. La seconde donnée isolée est extrême et placée à une distance beaucoup plus grande du reste des données. Dans ce contexte, seul ce second point est détécté par PAMOUT. La figure 6.7 indique une convergence rapide vers le partitionnement optimal avec $k_o = 5$. PAMOUT ne semble donc pas être efficace

dans ce contexte, car la première donnée isolée est trop proche des deux groupes principaux. Afin de détecter une telle donnée dans un tel contexte, il faudrait appliquer deux fois l'algorithme PAMout. Les approches statistiques et issues de l'Intelligence Artificielle fournissent les mêmes résultats que PAMout dans ce contexte.

#partition	v	#I/partition
2	6,163505	24/26
3	9,060200	12/14/24
4	0,638661	1/11/14/24
5	**0,636526**	***1*/11/11/13/14**
6	0,6401112	1/7/9/9/11/13

FIGURE 6.7 – Résultats de PAMout sur les données univariées (en ordonnée) de la figure 6.8e_1.

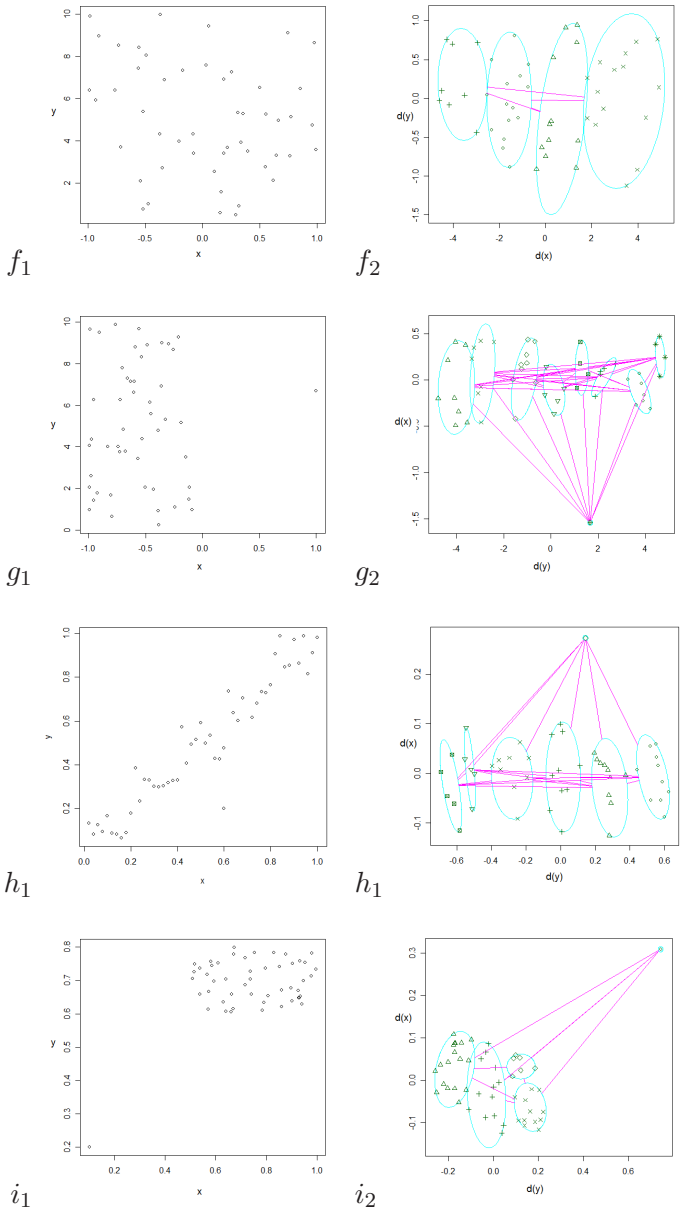
Valeurs singulières pour des données à deux dimensions

La figure 6.8 présente les ensembles des données bivariées sur lesquelles nous avons testé PAMout. Les figures indexées « 1 » représentent les valeurs des données bivariées étudiées. Les figures indexées « 2 » présentent le résultat du regroupement des points sur la base de la distance entre les sous-ensembles générés.

La figure 6.8f_1 présente une ensemble de données ne contenant pas *a priori* de données singulières. Cette situation correspond au cas A de notre guide d'interprétation. La mise en œuvre de PAMout valide cette hypothèse. La figure 6.8f_2, ne permettant pas de distinguer un point définissable comme une valeur singulière. Les résultats de la figure 6.9 indique que l'on obtient $k_o = 4$ soit 4 groupes de données homogènes. Les approches statistiques, qui nous ont servi de support dans la section précédente, ne trouve pas de validité ici car elles requièrent d'explorer chacune des deux dimensions et de tirer des conclusions sur la présence de valeurs singulières à partir uniquement de ces résultats . Les approches basées sur les grilles et la densité permettent une exploration globale des données (en 2 dimensions), mais les résultats obtenus sont critiquables car ils sont fortement dépendants des paramètres définis par l'utilisateur et ces paramètres sont très difficiles à définir en 2 dimensions. Ces commentaires sur les approches classiquement utilisées pour détecter les valeurs singulières sont aussi valables pour les expérimentations suivantes.

La figure 6.8g_1 présente des données identiques à celles de la figure 6.8f_1, à la nuance près, que toutes les valeurs ont été multipliées par -1 et que nous avons remplacé une valeur au hasard de l'ensemble défini en abscisse par la valeur 1. Cette dernière donnée est donc à l'écart des autres comme il est possible de le voir sur la figure 6.8g_1. PAMout a été appliqué à chacune des deux dimensions de l'ensemble de données. Comme le montre de la figure 6.10, il a détecté une valeur singulière pour $k = 4$ sur les valeurs en abscisse, ce qui correspond à la valeur que nous avons « prédéfinie ». La figure 6.11 montre que PAMout n'a pas détecté de valeurs singulières sur l'axe des ordonnées. Enfin, la figure 6.12 montre que la donnée définie comme une valeur singulière sur un des deux axes n'est pas détectée pour la valeur de k optimale ($k_o = 5$) mais plus tardivement pour $k = 9$. La figure 6.8g_2 permet de visualiser la répartition des sous-groupes de données pour $k = 9$. Notre valeur singulière est bien identifiée. Cette situation correspond aux cas B et C de notre guide d'interprétation : la valeur singulière que nous avons définie correspond à une valeur *suspecte*.

La figure 6.8h_1 illustre le cas E de notre guide d'interprétation : la valeur singulière correspond à du bruit dans les données. Cette figure présente un ensemble de données bivarié globalement



f_1 f_2 g_1 g_2 h_1 h_1 i_1 i_2

FIGURE 6.8 – Exemples d'ensembles de données bivariées (indexées « 1 ») et la représentation graphique de leur partitionnement (indexées « 2 »)

#partition	v	#I/partition
2	5,670684	24/26
3	13,74826	15/16/19
4	**4,12522**	**8/12/15/15**
5	4,153666	8/8/8/11/15

FIGURE 6.9 – Résultats de PAMOUT sur les données bivariées de la figure 6.8f_1.

#partition	v	#I/partition
2	43,699400	25/25
3	18,138420	13/16/21
4	0,6946494	1/13/15/21
5	0,6303497	1/9/10/15/15
6	0,5674212	1/8/10/10/10/11
7	0,5308647	1/4/5/9/10/10/11
8	0,5126224	1/4/4/5/7/9/10/10
9	**0,4726375**	***1*/4/4/5/5/6/7/8/10**
10	0,4897181	1/4/4/5/5/5/5/6/7/8

FIGURE 6.10 – Résultats de PAMOUT sur les données univariées (en abscisse) de la figure 6.8g_1.

croissant et incluant une donnée divergente. Cette donnée ne peut être détectée lors de l'analyse des deux dimensions prises séparément quelque soit la méthode utilisée (statistiques ou issues de l'Intelligence Artificielle), comme nous l'avons déjà noté dans la section précédente. PAMOUT, quant à lui, détecte cette donnée isolée uniquement lorsque l'on analyse simultanément les deux dimensions. La figure 6.8h_1 illustre les résultats présentés dans la figure 6.13.

La figure 6.8h_1 est un exemple du même type que celui présenté à la figure 3.1. Nous illustrons ici les cas *F*, *G* et *H* de notre guide d'interprétation. La donnée isolée est détectée par PAMOUT lorsque l'on considère les données univariées sur chacun des axes, mais aussi lorsque l'on considère ces données comme étant bivariées. Il s'agit donc d'une donnée *aberrante*. Cette donnée est détectée par les approches basées sur les distributions et le rognage dans l'analyse de chacune des deux dimensions. Les autres approches ne sont encore une fois efficace qu'en fonction de leur paramétrage.

PAMOUT montre ici son intérêt en ne faisant pas intervenir l'utilisateur et en fournissant un résultat pertinent.

Synthèse

Les expérimentations de l'algorithme PAMOUT que nous avons réalisées et celles relatives aux approches auxquelles nous l'avons comparé peuvent être synthétisées à l'aide de la figure 6.15. Les cas *a* à *e* correspondent aux situations univariées présentées plus haut et les cas *f* à *i* correspondent aux situations bivariées.

Les colonnes de cette figure concernent respectivement les grandes familles d'approches expérimentées. Les cellules restituent de manière qualitative les résultats obtenus pour chaque cas avec chaque type d'approches.

Il est intéressant de noter que les approches de définitions de valeurs singulières basées sur

#partition	v	#I/partition
2	11,87968	24/26
3	6,833955	15/16/19
4	4,271031	11/11/13/15/
5	2,785773	5/9/10/11/15
6	**2,722471**	**5/5/7/9/9/15**
7	3,112697	5/5/7/7/8/9/9

FIGURE 6.11 – Résultats de PAMOUT sur les données univariées (en ordonnée) de la figure 6.8g_1.

#partition	v	#I/partition
2	11,21456	24/26
3	5,403506	15/16/19
4	4,.421401	11/12/12/15
5	**3,49141**	**6/9/9/11/15**
6	3,926863	6/7/8/9/9/11

FIGURE 6.12 – Résultats de PAMOUT sur les données bivariées présentées de la figure 6.8g_1.

#partition	v	#I/partition
2	4,324343	19/31
3	12,18026	15/17/18
4	4,970390	10/12/14/14
5	2,926152	9/10/10/10/11
6	2,867896	5/5/9/10/10/11
7	**1,614080**	***1*/5/5/9/10/10/10**
8	1,637729	1/5/5/5/5/9/10/10

FIGURE 6.13 – Résultats de PAMOUT sur les données univariées (en ordonnée) de la figure 6.8h_1.

#partition	v	#I/partition
2	13,635400	24/26
3	13,645970	16/16/18
4	0,8885207	1/15/16/18
5	**0,7931279**	***1*/6/12/13/18**
6	0,8723061	1/6/9/10/12/12

FIGURE 6.14 – Résultats de PAMOUT sur les données univariées (en ordonnée) la figure 6.8i_1.

les distributions et le rognage, bien que fréquemment utilisées par les biologistes, sont des approches fortement voir strictement dépendantes des *desiderata* de l'utilisateur. Ces approches ne concernent que les valeurs extrêmes et ne sont pas efficaces dans toutes les situations que nous avons expérimentées. D'autres part, il est important de souligner que les approches basées sur le rognage des données ne sont applicables qu'aux ensembles univariés. Les approches basées sur les distributions présentent des résultats proches de ceux obtenus par le rognage, à la

	PAMout	**Distribution**	**Rognage**	**Densité**	**Grilles**
Paramétrage	Automatique	Manuel	Manuel	Manuel	Manuel
Cas a (0)	+	−	−	0	0
Cas b (1)	+	+	0	0	0
Cas c (1)	+	−	−	0	0
Cas d (2)	+	−	−	0	0
Cas e (2)	−	−	−	0	0
Cas f (0)	+	NA	NA	0	0
Cas g (1)	+	NA	NA	0	0
Cas h (1)	+	NA	NA	0	0
Cas i (1)	+	NA	NA	0	0

FIGURE 6.15 – Synthèse des résultats obtenues grâce à PAMout et les autres approches testées. + : résultats obtenus correspondant aux résultats attendus ; 0 : résultats obtenus partiellement correctes par rapport à ceux attendus ; − : résultats obtenus erronés ; NA : approches non adaptées au cas.

nuance près que ceux-ci sont dépendants non seulement d'un effectif à définir mais aussi du type de distribution statistique (généralement de type loi « Normale ») retenue par l'utilisateur. Ces mêmes approches ne sont, par ailleurs, pas utilisées par les biologistes en Génomique Fonctionnelle pour l'étude des ensembles bivariés car les modèles multidimensionnels sont trop complexes à interpréter. En effet, sans connaissance *a priori* du modèle de distributions des données sur chaque dimension, il n'est pas possible de proposer un traitement statistique adapté. Comparées à PAMout, ces deux approches ne sont pas efficaces et ne permettent pas, quelque soit le cas de figure retenu, d'obtenir des résultats pertinents.

Les approches permettant la définition de valeurs singulières basées sur la densité des données et sur les maillages (« grilles ») se sont avérées efficaces dans tous les cas que nous avons traités. Néanmoins, il est nécessaire de les paramétrer soigneusement pour obtenir des résultats pertinents pour un ensemble de données spécifique. Dans un cas, le e , où une donnée isolée non extrême qui semble visuellement distante de deux ensembles « denses » est détectée par ces deux approches et non par PAMout. Dans ce contexte, les deux approches restituent des résultats corrects, si et seulement si l'utilisateur prend le temps d'étudier la topologie des données pour définir la configuration optimale des algorithmes de recherche de valeurs singulières. Le cas f est un exemple pour lequel ces mêmes approches peuvent conduire à considérer un grand nombre de valeurs comme suspectes si les algorithmes de « détection » sont paramétrés de manière trop sensible. *A contrario* le cas h pose le problème de manière inverse où des valeurs singulières peuvent être ignorées si les paramétrages sont trop « laxistes ».

PAMout est une approche efficace qui permet la détection de valeurs singulières dans des ensembles de données multidimensionnelles d'effectifs modestes (quelques dizaines d'individus) et présentant un nombre très important de réplicats comme dans le contexte de la Génomique Fonctionnelle.

Il est très difficile de comparer le temps d'exécution de l'algorithme PAMout et des autres approches. En effet, les autres approches étant sujettes au paramétrage humain, elles restent très difficiles à évaluer quantitativement.

6.3 Conclusion

La mise en évidence de *valeurs singulières* au sein d'ensemble de données est très important. En effet, ce type de données peut biaiser les analyses qui sont faites sur les données et ainsi induire des erreurs d'interprétation des résultats [Knorr 2002, Undercoffer *et al.* 2003]. Les données singulières doivent donc être détectées en amont des analyses qui peuvent être faites sur les bases.

L'approche que nous avons proposons, PAMout, s'appuie sur un algorithme issu de travaux en apprentissage automatique non supervisé : l'algorithme des k-médoïdes. Le principe de PAMout est simple : l'algorithme regroupe de manière optimale les données, ce qui lui permet d'identifier de manière efficace les données aberrantes, suspectes et le bruit. Contrairement aux approches classiques de détection de valeurs singulières, PAMout ne nécessite pas de paramétrage particulier, ce qui le rend très facile d'utilisation pour tout utilisateur.

Les résultats de cet algorithme permettront d'informer l'expert du domaine de la qualité des données utilisées afin d'améliorer les résultats de la Fouille de Données comme nous le montrerons dans les expérimentations au chapitre 8.

Troisième partie

Expérimentations

Chapitre 7

DiscoClini : une interface interactive

Nous avons présenté dans les chapitres précédents la méthodologie que nous proposons pour découvrir des relations pouvant jouer le rôle de biomarqueurs d'états physio-pathologiques. Nous décrivons dans ce chapitre l'implantation du système : DiscoClini [Benis *et al.* 2003a, Benis *et al.* 2003c, Benis 2005, Benis 2007, Benis et Courtine 2009b, Benis et Courtine 2009a].

7.1 Le système DiscoClini

L'objectif du système DiscoClini est de permettre le calcul, la visualisation et l'exploration des relations existantes entre des ensembles de données issues des puces à ADNc et des données biocliniques. Il a été développé en R [R Development Core Team 2006] pour les calculs statistiques et l'identification des valeurs singulières et en PHP [3] pour l'interface graphique d'exploration des résultats. L'interface principale de DiscoClini permet d'accéder aux trois grandes étapes de notre approche :

1. **Définition des données** : l'expert définit les sources de données biocliniques et d'expression génique issues des puces à ADNc ;
2. **Mise en relation des données** : le système valide les données soumises, effectue les calculs nécessaires pour l'étude corrélationelle et la détection des valeurs singulières ;
3. **Exploration des résultats** : le système et l'expert « interagissent » afin de permettre à l'expert de sélectionner les relations qu'il souhaite valider biologiquement en laboratoire « humide ».

Nous allons dans les parties suivantes présenter plus en détails chacune de ces étapes dans le système DiscoClini.

7.2 Les sources de données

DiscoClini traite des données numériques issues de protocoles de recherche clinique. Les données à définir en entrée du système sont les suivantes (voir figure 7.1) :

1. la définition du protocole d'étude de mise en relation des deux types de données précédentes ;
2. des données d'expression génique issues de la mise en œuvre de puces à ADNc ;

3. `http://www.php.net/`

3. des données biocliniques collectées lors d'examens cliniques, de l'utilisation de questionnaires et d'analyses biologiques ;
4. des ressources externes liées à l'annotation des gènes mesurés notamment, mais pouvant aussi référer à la littérature.

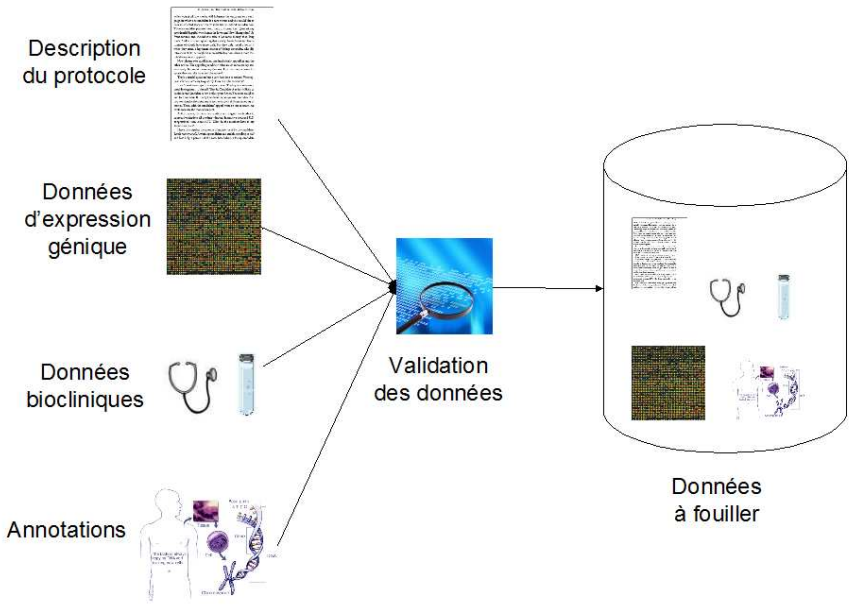


Figure 7.1 – Les sources de données en entrée du système DiscoClini.

La première étape de création d'une analyse sous DiscoClini consiste à la décrire. Pour cela, il faut remplir un formulaire donné à la figure 7.2. Un projet d'étude corrélationnelle est identifié *via* un « Nom court », qui peut être un acronyme (par exemple, *VLCD39*), un « Nom long » (par exemple, *Very Low Caloric Diet on 39 patients*), d'une adresse de messagerie électronique (« m-él de contact ») et d'un « mot de passe ». L'adresse électronique est utilisée pour que le système puisse envoyer un message à l'expert lorsque les traitements automatiques sont terminés. En combinaison avec le « mot de passe », elle permet de sécuriser l'accès aux résultats.

La description du protocole de recherche (*in extenso* de l'analyse à lancer) est définie en même temps que le chargement des données biomédicales par l'expert. L'expert a ainsi la possibilité de donner une description du contexte et de l'objectif de son analyse.

La seconde étape consiste à définir les données à traiter. Ces données sont téléchargées sur le serveur de calculs *via* l'interface graphique donnée à la figure 7.3. Cette interface permet à la fois de télécharger les données mais aussi de valider automatiquement le format de celles-ci.

Les données numériques se doivent de respecter un format particulier afin d'être exploitables par nos algorithmes (voir chapitre 5). Les données doivent ainsi être définies sous la forme d'un tableau composé de colonnes correspondant chacune aux valeurs relatives à un individu i_n et de lignes correspondant chacune aux valeurs relatives à un attribut (expression génique d'un gène ou mesure d'un paramètre bioclinique). La figure 7.4 montre un extrait d'un fichier contenant des valeurs d'expressions géniques issues de la mise en œuvre de puces à ADNc et la figure 7.5 montre un extrait de fichier contenant les valeurs de paramètres biocliniques.

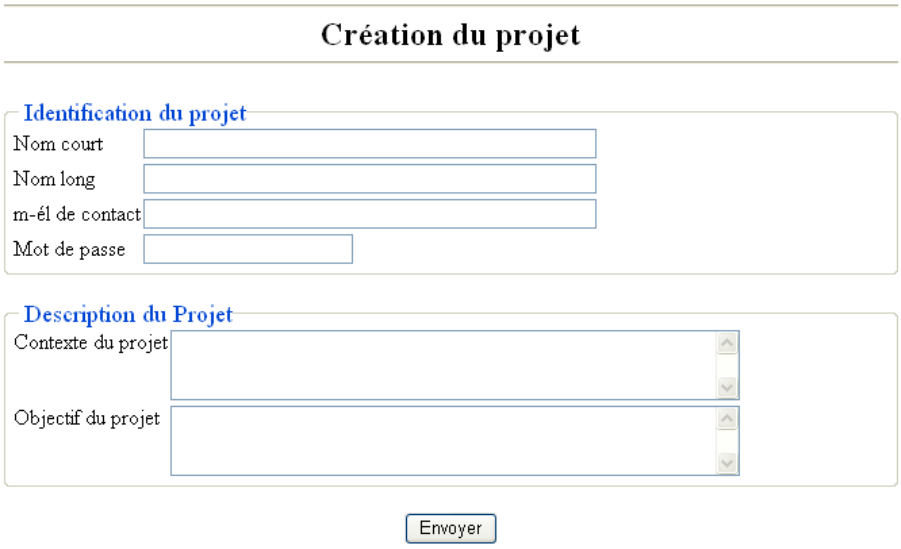


FIGURE 7.2 – Interface graphique de description de l'analyse dans le système DISCOCLINI.

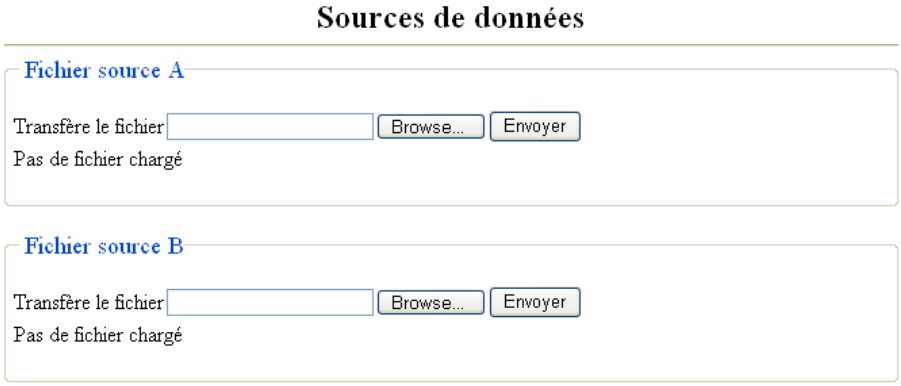


FIGURE 7.3 – Interface de téléchargement des fichiers de données sources dans le système DISCOCLINI.

Ainsi, le format du fichier soumis est de type *csv*[4]. Les valeurs sont séparées soit par des virgules ou des points-virgules. Le séparateur décimal est le point ou la virgule. Une fois, le format des deux fichiers validé, une mise en correspondance est faite. En effet, pour que l'analyse est lieu, il faut que les deux fichiers décrivent les mêmes individus, donc que chaque individu décrit par une puce à ADNc (ou les valeurs des différences entre deux puces à ADNc) soit aussi défini par des paramètres biocliniques. De plus, lors d'analyses où un même individu est représenté par plusieurs puces à ADNc et/ou plusieurs ensembles de mesures biocliniques, il est important que les identifiants de l'individu permettent la prise en compte de ces « répétitions ».

Les données numérico-symboliques relatives à l'annotation des données issues des puces à ADNc sont gérées par l'administrateur du système. Il a en charge de récupérer régulièrement ces données sur le site de « Stanford MicroArray Database » [5] afin de maintenir à jour les annotations des gènes et ce malgré l'évolution rapide de celles-ci. Les informations collectées sont celles liées

4. Comma-Separated Values
5. `http://genome-www5.stanford.edu/index.shtml`

	i_1	i_2	i_3	$\cdots$	i_{39}	i_{40}
$IMAGE$: 46936	0,17383	-0,87830	0,58277	$\cdots$	-0,99171	0,69618
$IMAGE$: 10512	0,80959	-0,51406	-0,21853	$\cdots$	-0,62747	0,33194
$IMAGE$: 74088	0,44535	-0,14982	0,85429	$\cdots$	0,26323	0,96770
$IMAGE$: 37664	NA	-0,78558	0,49005	$\cdots$	0,89899	0,76699
$IMAGE$: 17593	-0,88040	0,58487	0,28934	$\cdots$	NA	0,40275
$IMAGE$: 81169	0,51616	0,22063	-0,92510	$\cdots$	NA	-0,3851
$IMAGE$: 44745	0,15192	0,85639	-0,56086	$\cdots$	-0,96980	-0,67427
$IMAGE$: 8321	0,78768	-0,49215	-0,19662	$\cdots$	0,60556	0,31003
$IMAGE$: 71897	0,42344	-0,12791	NA	$\cdots$	0,24132	0,94579
$IMAGE$: 35473	0,5920	-0,76367	0,46814	$\cdots$	0,87708	0,58155
$IMAGE$: 15402	0,85849	0,56296	NA	$\cdots$	NA	NA
$IMAGE$: 78978	-0,49425	-0,19872	-0,90319	$\cdots$	-0,31213	0,1660

Figure 7.4 – Extrait d'un fichier contenant des valeurs d'expressions géniques issues de la mise en œuvre de puces à ADNc.

	i_1	i_2	i_3	$\cdots$	i_{39}	i_{40}
Âge	67,43	35,68	56,40	$\cdots$	45,37	66,62
IMC	20,84	27,9	34,81	$\cdots$	48,78	55,3
QUICKI	0,37	0,34	0,31	$\cdots$	0,35	0,32

Figure 7.5 – Extrait de fichier contenat les valeurs de paramétres biocliniques.

à l'ensemble des gènes qui apparaissent dans les analyses déjà faites sur DiscoClini. Si un utilisateur introduit un nouveau gène dans une analyse, une alerte est envoyée à l'administrateur afin qu'il procède à la mise à jour de la base d'annotations. Ces données incluent notamment des informations sur :

- les identifiants, les symboles et les noms des gènes présents sur les puces à ADNc ;
- leur localisation chromosomique ;
- leurs fonctions d'après l'ontologie *Gene Ontology* [Gene Ontology Consortium 2001, Gene Ontology Consortium 2004].

Une fois l'ensemble de ces données (description du contexte et des objectifs du protocole et valeurs d'expression génique et données biocliniques) chargées sur le serveur de calculs, les traitements statistiques et exploratoires de ces données vont être lancés automatiquement comme nous allons le voir dans le paragraphe suivant.

7.3 Traitements des données

Une fois les *données sources* enregistrées sur le serveur, deux analyses vont être lancées hors-ligne et en parallèle, comme le montre la figure 7.6 :

1. les analyses statistiques univariées et bivariées des relations entre les deux ensembles de données (voir chapitre 5) ;
2. la découverte de valeurs singulières grâce à l'algorithme PAMout (voir chapitre 6).

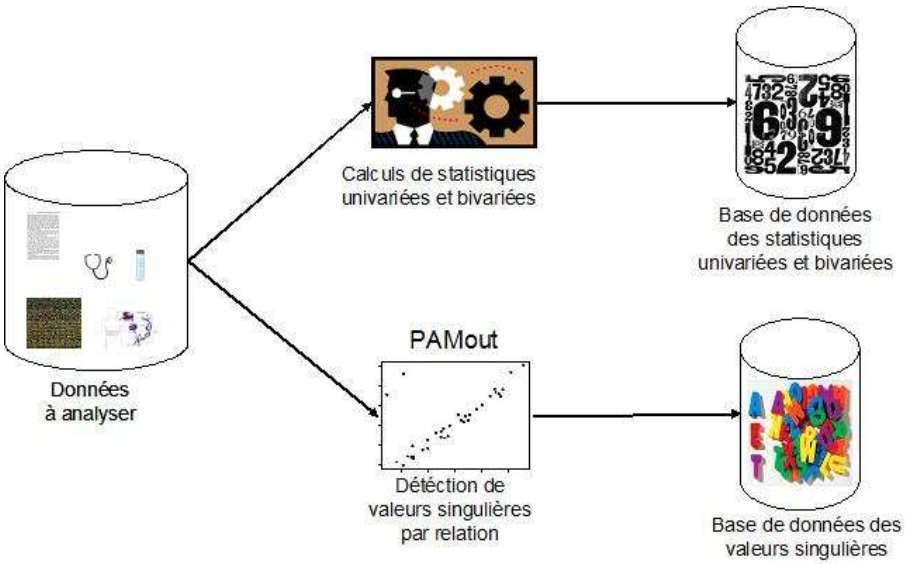


FIGURE 7.6 – Traitements hors-ligne des données par le système DISCOCLINI.

7.3.1 Calculs de statistiques univariées

Au cours d'un protocole expérimental incluant l'utilisation de puces à ADNc, il est possible que celles-ci ne soient pas issues d'un même lot. Cela peut poser différents problèmes au niveau des gènes (ou fragments) présents sur chaque puce, car les gènes présents sur chaque puce à ADNc peuvent être différents. D'autre part, au cours de la mise en œuvre de puces à ADNc en laboratoire « humide », c'est-à-dire lors par exemple des étapes d'hybridation ou d'acquisition du signal, il est possible que toute ou partie de la puce à ADNc devienne inexploitable (manipulations mal réalisées, dégâts physiques, contaminations biologiques ou chimiques,...). Afin de permettre une exploration cohérente et pertinente des données d'expression génique, il est nécessaire de réaliser des pré-traitements sur ces données afin d'avoir une vision précise des données exploitables, il faut donc commencer par définir pour chaque gène son effectif (nombre de puces où son expression est réellement mesurée).

DISCOCLINI réalise, en plus de ce comptage, des calculs de statistiques descriptives sur les les paramètres biocliniques, puis les valeurs d'expressions géniques. Ces différentes valeurs sont :

- le minimum, le maximum et l'étendue des données,
- la moyenne et l'écart-type à la moyenne,
- la médiane et l'écart-type à la médiane : elles donnent une meilleure vision des données car elles sont plus résistantes aux valeurs singulières,
- le coefficient d'asymétrie (en anglais, *Skewness*) et le coefficient d'aplatissement (en anglais, *Kurtosis*) : ils sont fournis à titre indicatif car ils sont basés sur la notion de loi Normale, qui comme on l'a vu précédemment n'est pas applicable à nos données. Cependant ces informations restent un point de repère important pour les biologistes même si d'un point de vue statistique « pure », elles n'ont pas de signification dans ce contexte.

7.3.2 Calculs de statistiques bivariées

Le début des analyses sur les données bivariées est le même que celui des données univariées. Les couples d'ensembles sont formés à partir des valeurs d'expression par gène et de celles d'un paramètre bioclinique. Seuls les couples bien formés sont gardés, c'est-à-dire ceux pour lesquelles

il y a à la fois une valeur d'expression génique et une valeur pour le paramètre bioclinique pour le même individu. Les autres sont écartés de l'analyse. De plus, pour que l'analyse puisse avoir lieu, il faut que l'effectif d'un couple inclus plus de 10 individus (voir partie II).

Une fois les couples bien identifiés, les calculs nécessaires à l'étude corrélationnelle sont réalisées sur le serveur avec le lancement en parallèle de l'approche *globale* et de l'approche *locale* présentées au chapitre 5.

7.3.3 Détection des valeurs singulières

En parallèle des calculs « statistiques » univariées et bivariées, DiscoClini lance aussi le détection des valeurs singulières présentent dans les données univariées et bivariées, en s'appuyant sur l'algorithne PAMout que nous avons présenté dans le chapitre 6. Cette fonctionnalité, comme les précédentes, est intégrée de manière automatique aux traitements des données.

7.3.4 Sauvegarde des résultats

Les résultats de ces différents traitements sont enregistrés sur le serveur sous la forme d'un *fichier texte*, par paramètre bioclinique et par type d'analyse (c'est-à-dire univarié, bivarié ou détection des valeurs singulières) dans lequel les individus sont représentés par colonne et leurs attributs par ligne. Dans le cas de la détection des valeurs suspectes, une information est ajoutée pour chaque attribut : la valeur optimale de k_o.

7.3.5 Informer l'expert

Une fois l'ensemble des calculs effectués hors-ligne, le serveur envoie un message électronique à l'utilisateur pour l'informer que les calculs sont finis et que les résultats peuvent être explorés *via* le serveur.

Dans la pratique, avec 40 individus décrits chacun par 1 puce à ADNc de 39927 gènes et de 22 paramètres biocliniques, un serveur de calculs quadri-processeurs Intel Xeon à 3,6 GHz et 4 Go de mémoire vive sous « Linux Debian Sarge » a besoin d'un peu moins de 24 heures de calculs s'il dispose de 4% des ressources du serveur. Le temps de calcul est partagé quasiment à part égale entre le calcul des relations et entre la détéction des valeurs singulières. Un serveur qui serait dédié à 100% permettrait de réduire les délais de calculs à quelques heures mais un serveur est rarement dédié à une seule tâche à la fois.

L'utilisateur peut alors se connecter au serveur en utilisant son adresse électronique, le mot de passe et l'intitulé court de son projet, qu'il a donné lors de la description de son analyse (et qui sont rappelés dans le message électronique de confirmation). Il peut alors explorer ses résultats. Au bout de 6 mois, une sauvegarde et une compression automatique des résultats est effectuée afin d'une part de réduire le taux d'occupation des ressources du serveur et d'autre part de pouvoir rapidement restituer les résultats à l'expert s'il souhait les étudier à nouveau sans avoir à réitérer le processus de soumission et de calculs. Pour une expérimentation, incluant 40 individus décrits chacun par 1 puce à ADNc de 39927 gènes et de 22 paramètres biocliniques le volume de données initialement soumis est d'environ 31 Mégaoctets et le volume de données numériques générés au cours de l'analyse est d'environ 200 Mégaoctets ().

7.4 Exploration des résultats

7.4.1 Reformulations symboliques et visuelles

L'exploration des résultats s'appuie sur des reformulations symboliques et visuelles des valeurs numériques obtenues dans l'étape précédente.

Les résultats obtenus par les calculs automatiques sont inexplorables par un expert-humain sans disposer d'outils lui permettant de les filtrer et de les synthétiser. La phase d'exploration du flux de DiscoClini (voir figure 7.7) va permettre à l'expert de définir dans un premier temps des contraintes sur les valeurs statistiques de corrélations, de significativité et d'effectifs (voir figure 7.8). Ces contraintes ne sont pas des *a priori* car elles ne sont pas liées à des Connaissances sur les gènes ou sur les paramètres cliniques, mais elles vont permettre uniquement de filtrer facilement les résultats, afin de limiter le nombre d'informations que l'utilisateur aura à explorer visuellement. Cependant, l'utilisateur est libre de ne mettre aucune contrainte et de visualiser ses 880000 résultats (40000 *gènes* $\times$ 22 *paramètres biocliniques*) mais cela lui permettrai un temps « infini ».

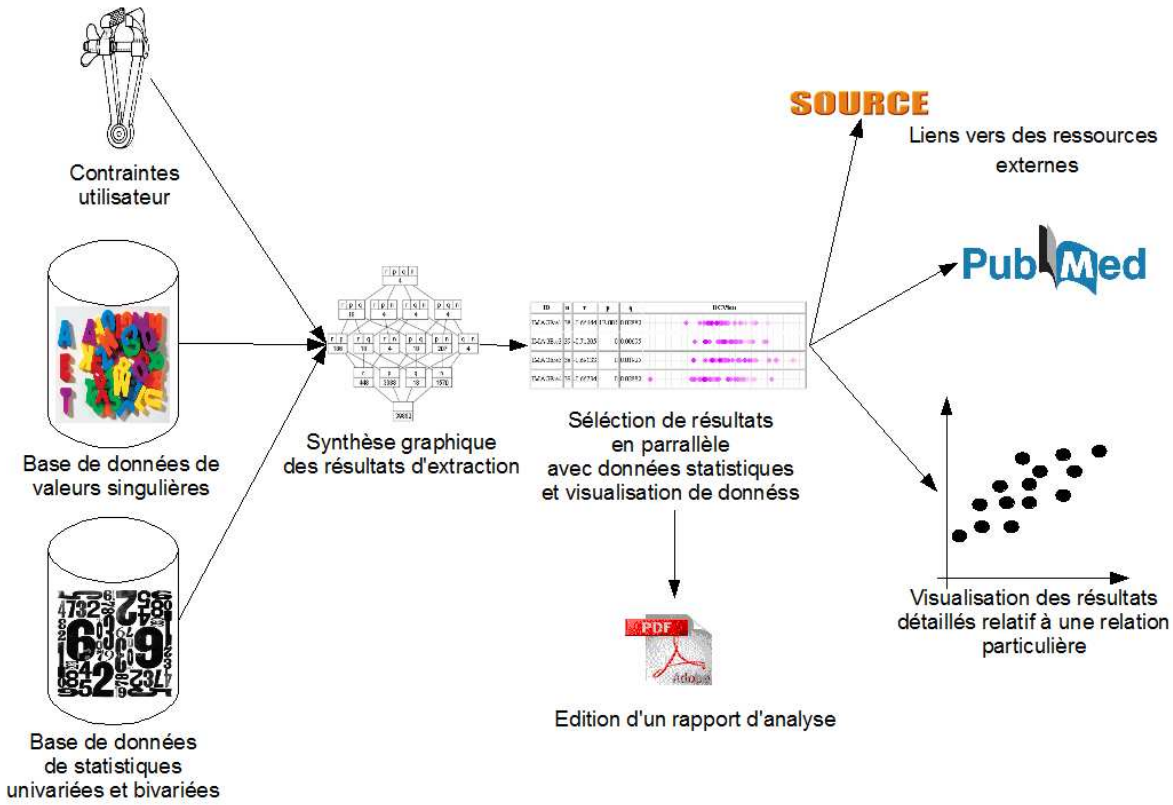


Figure 7.7 – Phase d'exploration des résultats et mise en relation avec des sources externes d'informations dans le système DiscoClini.

Les informations à sélectionner par l'utilisateur sont les suivantes :

– un des paramètres biocliniques ;

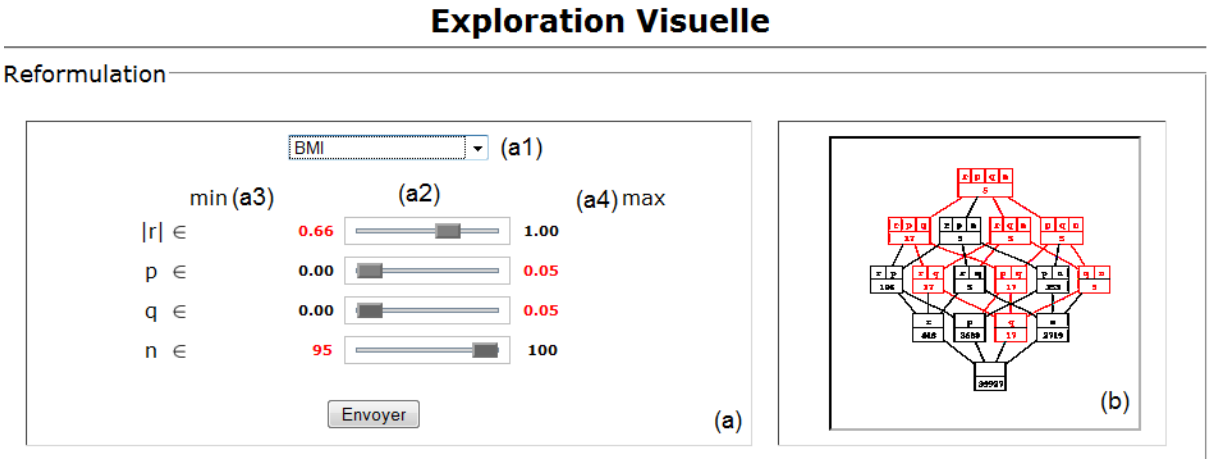


FIGURE 7.8 – Interface d'exploration des résultats de l'étude corrélationnelle dans DiscoClini.

– les valeurs minimales limites pour les valeurs de la corrélation $|\rho|$, de la significativité p, du test de multiplicité q et du nombre d'individus n.

Les minima (figure 7.8a3) et les maxima (figure 7.8a4) sont respectivement en noir quand ils sont statiques et en rouge quand ils sont ajustables grâce aux curseurs (figure 7.8a2).

Une fois ces valeurs modifiées, l'expert les valide et DiscoClini réalise une recherche dans les données des relations répondant à ces contraintes. Les résultats de cette exploration sont restitués sous la forme d'un diagramme de Hasse (figure 7.8b). Ces nœuds sont organisés hiérarchiquement dans le diagramme en fonction du nombre de contraintes que le nœud respecte. Chaque nœud de ce diagramme correspond au nombre de relations correspondant à une ou plusieurs des contraintes posées par l'expert sur les valeurs statistiques. Chaque nœud peut être « ouvert » (en cliquant dessus), l'ensemble des relations correspondantes à ce nœud apparaissent alors de manière détaillée comme présenté à la figure 7.11. Les informations fournies à l'expert sont les suivantes :

– la colonne *a* correspond au *nom court du gène* impliqué dans la relation avec le paramètre bioclinique défini *via* l'interface précédente (figure 7.8a1). Ce nom court est un lien dynamique vers le site de « Stanford MicroArray Database » [6]. Il permet d'afficher une page issue du site SOURCE (voir figure 7.9), qui inclut aussi bien des liens vers différentes sources d'informations sur le gène, comme par exemples sa localisation chromosomique, son expression dans différents tissus et chez différents organismes, son statut en terme d'évolution phylogénique (UniGene [7], Entrez Gene [8], OMIM [9], GenAtlas [10], GeneCard [11], Ensembl [12], MapView [13]... ; AceView [14], Genome Browser [15]) ;
– les colonnes *b* et *c* correspondent aux visualisations de la relation entre le paramètre bio-

6. http://genome-www5.stanford.edu/index.shtml
7. http://www.ncbi.nlm.nih.gov/sites/entrez?db=unigene
8. http://www.ncbi.nlm.nih.gov/sites/entrez/?db=gene
9. http://www.ncbi.nlm.nih.gov/sites/entrez?cmd=Search&db=omim
10. http://www.dsi.univ-paris5.fr/genatlas/
11. http://genome-www.stanford.edu/genecards/index.shtml
12. http://www.ensembl.org/Homo_sapiens/index.html
13. http://www.ncbi.nlm.nih.gov/mapview/maps.cgi?ORG=hum&MAPO=gene&MAP1=loc
14. http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/
15. http://genome.ucsc.edu/

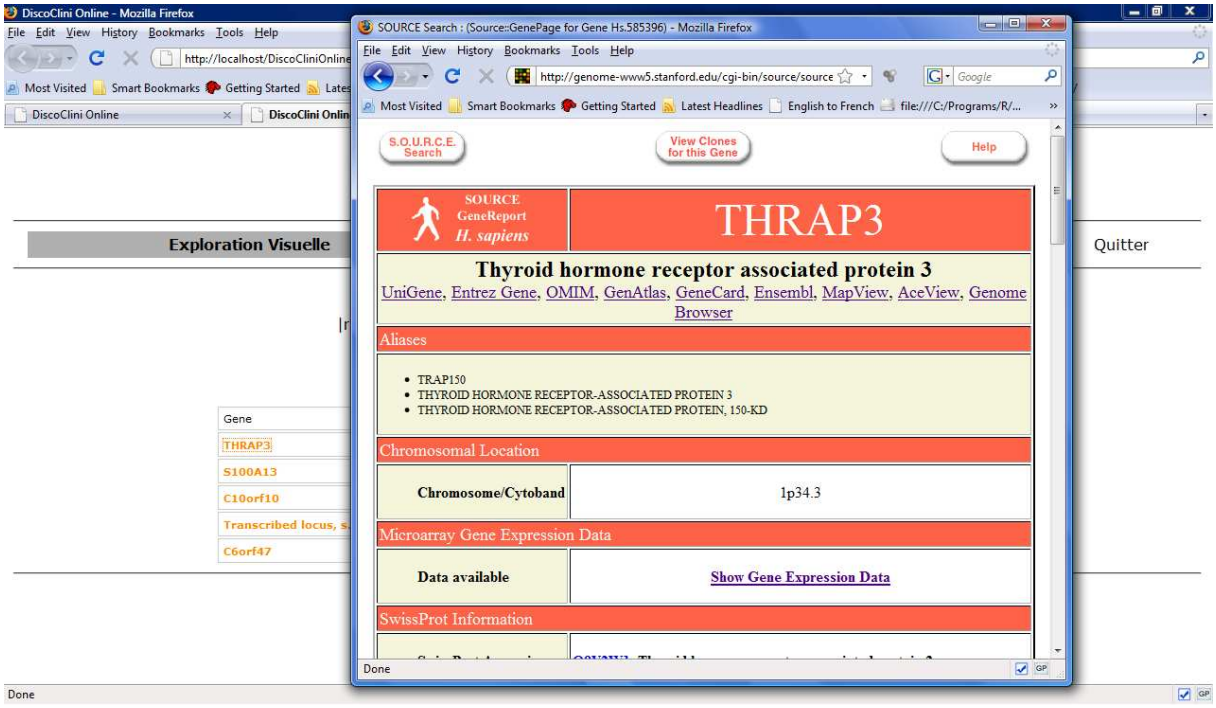


FIGURE 7.9 – Lien vers SOURCE à partir de DISCOCLINI.

clinique et les données d'expression du gène telle que nous les avons présentées au chapitre 5. La colonne b reprend le mode de visualisation défini par le langage L_V et la colonne c la représentation symbolique grâce au langage L_S étendu. Ces représentations sont un lien dynamique vers une représentation graphique en 2 dimensions de la relation, comme l'expert a traditionnellement l'habitude de la visualiser (voir figure 7.10) ;

– les colonnes d, e, f, g correspondent respectivement aux valeurs de l'effectif, du coefficient de corrélation de rang de Spearman, p sa significativité, q son taux de fausse découverte et p/q le rapport des deux dernières valeurs qui indique l'importance de l'ajustement de p par rapport à q.

Par exemple, la première ligne de la figure 7.11, concernent le gène « THRAP3 » (colonne a) qui présente une corrélation négative avec l'Indice de Masse Corporelle (visuellement perceptible *via* les colonnes b et c, et numériquement *via* la colonne e). Cette corrélation concerne une population de 38 individus (colonne d) et présentent un risque très réduit d'être une fausse découverte (colonnes f, g et h).

7.4.2 Prise en compte des valeurs singulières

Une fonctionnalité intéressante est celle qui permet d'informer l'expert sur le taux de « contamination » d'un nœud en terme de valeurs singulières (voir figure (i)). Il s'agit en fait du ratio de relations contenant au moins une valeur singulière par rapport au nombre de relations totales définies pour le nœud. Ce ratio est codé par une lettre (A,B,C,D,E,F), A correspondant à un nœud ne contenant pas de relations avec des valeurs singulières et F correspondant à une nœud fortement contaminé. Chaque lettre entre les deux extrêmes vont indiquer un degré de contamination plus ou moins important. Cette information est très importante lors de l'analyse

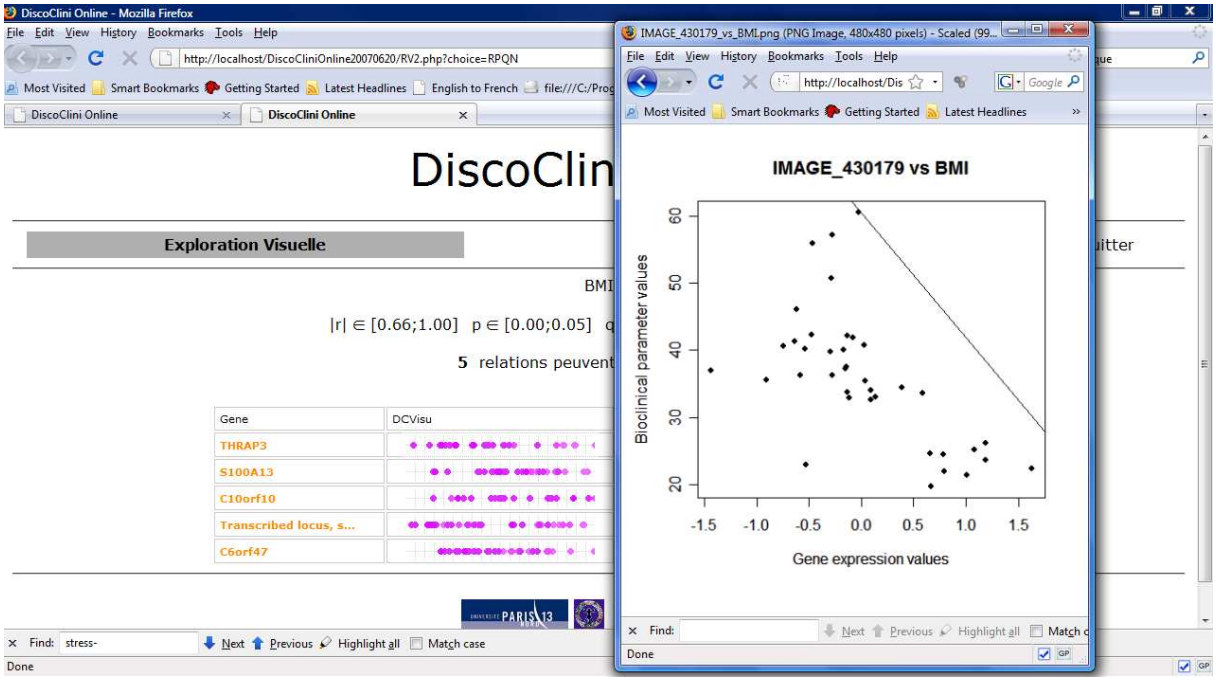


FIGURE 7.10 – Représentation graphique en 2 dimensions d'une relation dans DiscoClini.

(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
Gene	DCVisu	DCSymb	n	r	p	q	p/q	V.S.
THRAP3		\	38	-0.66444	0.00001	0.03069	0.00018	A
S100A13		\	39	-0.71205	<1E-005	0.00740	0.00005	A
C10orf10		\	38	-0.69133	<1E-005	0.02050	0.00008	A
Transcribed locus, s...		\	39	-0.66734	<1E-005	0.03069	0.00011	A
C6orf47		\	37	-0.66319	0.00001	0.03088	0.00025	A

FIGURE 7.11 – Interface graphique d'exploration des résultats d'un nœud de l'étude corrélationnelle dans le système DiscoClini et affichage de valeurs associées.

des résultats car plus un nœud est contaminé et moins l'expert peut avoir confiance dans les relations qui lui sont fournies. Les résultats ne sont néanmoins pas filtrés en fonction du nombre de valeurs singulières car cette information est fournie à titre indicatif et les valeurs singulières peuvent avoir une explication et une signification biologique.

7.4.3 Représentation chromosomique

Une autre fonctionnalité intéressante est la répartition chromosomique des gènes ayant des valeurs de corrélation intéressantes. Le principe est le même que le précédent sauf que les résultats sont donnés par chromosome et non plus par paramètre bioclinique, ainsi un diagramme de Hasse est associé à chaque chromosome. Ce mode de visualisation est très intéressant car certaines

pathologies sont liées à un chromosome donné. Par exemple, il a été montré que certaines obésités sont liées au gène *Ob* présent sur le chromosome 7 et qui influence la régulation de la Léptine qui est impliquée dans le métabolisme lipidique [16]. De même, dans le cadre du Diabète un lien avec le chromosome 6 a été montré par l'intermédiare du gène *Insulin-Dependent Diabetes Mellitus (IDDM1)* dont les mutations semblent influencer l'évolution d'un diabète de type 1 [17].

7.4.4 Édition d'un rapport

L'édition automatique d'un rapport d'exploration des résultats est une des fonctionnalités utile à l'expert dans la perspective de la diffusion et de la valorisation de ses travaux. Ainsi, l'expert lorsqu'il souhaite extraire des résultats du système, il peut demander la génération d'un rapport concernant les résultats relatifs à un paramètre bioclinique ou un chromosome donné. Ce rapport, généré en PDF, inclus les informations suivantes :

- le « nom court » de l'étude ;
- le « nom long » de l'étude ;
- l'adresse électronique de l'expert qui à soumis les données ;
- la date de soumission des données ;
- la date de fin des calculs ;
- le temps de calculs ;
- le nombre et la liste des individus inclus dans l'étude ;
- le nombre et la liste des gènes étudiés ;
- un histogramme de distribution du nombre de valeurs (biocliniques et gènes) présentes par individu ;
- un diagramme de Hasse correspondant au paramètre considéré avec les contraintes définies par l'expert ;
- un tableau semblable à celui de la figure 7.11 correspondant au contenu du nœud $\{r,\ p,\ q,\ n\}$ ou de celui qui lui est le plus proche (avec les valeurs de r et n les plus élevées pour une même relation) et qui contient des résultats répondant aux contraites fixées pour les valeurs de seuils de r et/ou p et/ou q et/ou n.

7.5 Conclusion

DiscoClini a été développé suivant un processus incrémental, afin d'impliquer l'expert tout au long de sa réalisation et de répondre le mieux possible à ses besoins. Il se décompose en deux phases de traitements : le premier « hors-ligne » pour la mise en relation des données issues des puces à ADNc avec les données biocliniques et le second « en-ligne » pour l'exploration et la visualisation des résultats. Ce système a été validé étape par étape dans le cadre de l'analyse de données réelles provenant de protocoles de recherche clinique sur les Obésités. Les résultats de ces expérimentations sont présentés dans le chapitre 8.

16. `http://www.ncbi.nlm.nih.gov/disease/Obesity.html`
17. `http://www.ncbi.nlm.nih.gov/disease/Diabetes.html`

Chapitre 8

Expérimentations étape par étape

Dans le cadre d'une collaboration avec le professeur Karine Clément et son équipe (Unité INSERM U 872 (Equipe 7) / Université Pierre et Marie Curie Paris VI - Centre de recherche des Cordeliers, AP/HP Pitié Salpêtrière, CRNH Ile de France), nous avons expérimenté et validé DiscoClini à chaque étape de son processus de développement en utilisant différentes bases de données. Parmi les ressources à notre disposition, nous nous sommes intéressés à des données issues de différents protocoles de recherche clinique sur les obésités et sur les pathologies qui leur sont associées [Basdevant *et al.* 1993, Wadden *et al.* 2002, Forga *et al.* 2002, Clément et Ferre 2003].

Tout d'abord, nous présentons la thématique de notre domaine d'application, l'Obésité, et les données desquelles nous disposons. Puis, dans un second temps, nous allons présenter les résultats d'expérimentations que nous avons réalisés avec DiscoClini à différentes étapes de son développement afin d'aider les biologistes dans leurs analyses. Nous ne présenterons pas tous les résultats que nous avons obtenus, mais nous nous concentrerons sur ceux qui nous paraissent les plus intéressants.

8.1 L'Obésité

L'Organisation Mondiale de la Santé (OMS) a reconnu l'Obésité comme une maladie chronique en 1997 [Organisation Mondiale de la Santé 2004, Organisation Mondiale de la Santé 2008]. L'OMS considéré l'Obésité comme une pandémie, bien qu'il ne s'agit pas d'une maladie infectieuse, car le nombre de personnes atteintes à travers le Monde est en constante augmentation depuis les dernières années [Organisation Mondiale de la Santé 2004, Organisation Mondiale de la Santé 2008].

L'Obésité est la conséquence de nombreux facteurs environnementaux, héréditaires et personnels ainsi que de facteurs socio-environnementaux comme le montre la figure 8.1. Les complications de l'Obésité sont elles aussi multiples et représentent des facteurs d'aggravation de l'état des sujets atteints de pathologies tels que les diabètes, les cancers et les accidents vasculaires.

L'Obésité est définie comme une maladie d'un organe, le tissu adipeux. L'OMS définit « *le surpoids et l'obésité comme une accumulation anormale ou excessive de graisse corporelle qui peut nuire à la santé* » [Organisation Mondiale de la Santé 2006]. Ainsi, les graisses (et autres lipides), tout comme les sucres (glucides), servent à stocker l'énergie dans le corps. Les sucres fournissent une énergie rapidement utilisable, alors que les graisses permettent de stocker beaucoup d'énergie dans peu d'espace. Quand l'organisme reçoit plus d'énergie qu'il n'en dépense, il met en œuvre des mécanismes de stockage des excès. La graisse est alors stockée dans les adipocytes. Ces

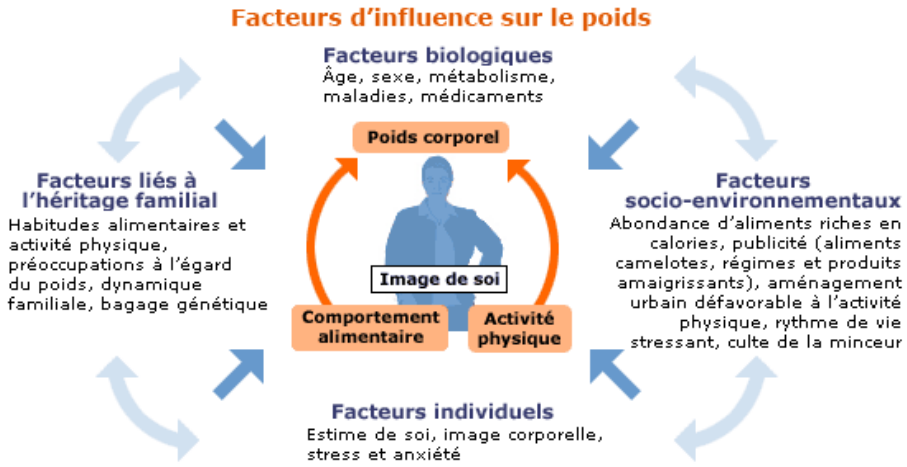


FIGURE 8.1 – Ensemble des facteurs influant sur le poids d'un individu (d'après `http://obesite.ulaval.ca`).

cellules constituent la majorité du tissu adipeux qui a pour fonction le stockage de la graisse et donc des réserves d'énergie. Si le stock de graisse croit anormalement, on distingue deux états physiologiques et morphologiques : le surpoids et l'Obésité. Dans le premier cas, les adipocytes stockent de plus en plus les graisses et grossissent ; dans le second cas, les adipocytes arrivent à saturation, ils se multiplient et on parle alors d'inflation adipocitaire. Cette dernière est variable d'un individu à l'autre en fonction de la quantité d'énergie en surplus. Dans tous les cas, l'Obésité peut être vue comme un déséquilibre de la balance énergétique (voir figure 8.2) : un déséquilibre entre les apports alimentaires et les dépenses énergétiques d'un individu.

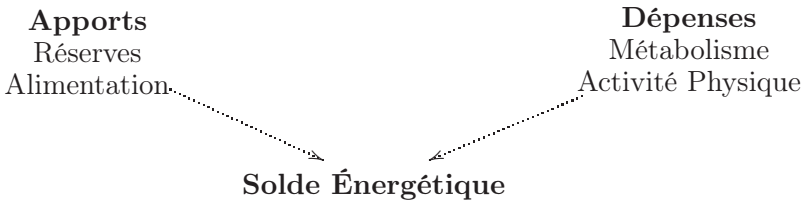


FIGURE 8.2 – Balance énergétique.

Les rôles de l'hérédité (de la génétique) sont de mieux en mieux connus pour cette pathologie. Des recherches ont notamment identifié des gènes responsables impliqués aussi bien dans la prédisposition à l'Obésité que dans son évolution [Clément *et al.* 2002, Perusse *et al.* 2005].

Des solutions thérapeutiques existent pour « traiter » l'Obésité. Elles peuvent s'appuyer sur une ou plusieurs des approches suivantes : la guidance diététique, l'activité physique, des traitements médicamenteux, des traitements chirurgicaux telles que la pose d'un anneau gastrique ou la réalisation d'un pontage gastrique (*gastric bypass*, en anglais), le soutien psychologique,...

Un des objectifs des chercheurs en Médecine des Obésités est de connaître davantage de gènes impliqués dans les mécanismes de la prise (*in extenso* de la perte) de poids afin d'être capable d'adopter la thérapeutique par rapport à l'origine de la pathologie et ce avec le maximum de chance de réussite. DISCOCLINI a pour objectif dans ce contexte de permettre l'identification de gènes qui peuvent présenter un intérêt comme biomarqueurs ou comme prédicteurs de l'état pondéral ou d'états associés à cette maladie.

Classification	Intervalle de définition
Manque de poids	$I.M.C. < 18,50$
Poids normale	$18,50 \geqslant I.M.C. \geqslant 24,99$
Surpoids	$25,00 \geqslant I.M.C. \geqslant 29,99$
Obésité de classe I	$30,00 \geqslant I.M.C. \geqslant 34,99$
Obésité de classe II	$35,00 \geqslant I.M.C. \geqslant 39,99$
Obésité de classe III (dite Obésité Morbide)	$I.M.C. \geqslant 40,00$

FIGURE 8.3 – Interprétation et classification des valeurs de l'I.M.C.

8.1.1 Sources de données issues des expérimentations biomédicales

Les données d'expression génique

L'ensemble des expérimentations que nous avons réalisées dans le cadre des travaux liés à la médecine de l'Obésité se sont appuyés sur des puces à ADNc pangénomiques de type *Stanford* [Sherlock *et al.* 2001] [18]. Ces *biopuces* permettent de mesurer l'expression de près de 40000 gènes (ou fragments de gènes) simultanément. Chacun de ces gènes est identifié par sa localisation sur la puce à ADNc, sa localisation chromosomique, ses fonctions lorsqu'elles ont été découvertes... Ces informations sont inclus dans une base d'annotations ; en général, lors du traitement des données, il est utile d'avoir une version, de l'une de ces bases d'annotations, la plus à jour possible, ces informations évoluant continuellement.

Les données biocliniques

Les données biocliniques auxquelles nous avons eu accès au cours de la collaboration avec l'équipe du professeur Clément sont nombreuses et hétérogènes de par leur nature et leur structuration. Cette non-homogénéité a eu plusieurs sources. La première était l'inexistence d'un Système d'Information regroupant l'ensemble des données disponibles dans les protocoles. La seconde est que l'équipe du professeur Clément travaille, aussi, sur des données issues de laboratoires partenaires français et européens qui ne récoltent pas forcément les mêmes informations sur les patients. Pour les différents protocoles, dans lesquels nous sommes intervenus, nous avons essentiellement travaillé avec les données suivantes (liste non exhaustive) :

- « Caractérisation de base » : Sexe, Age,... ;
- Anthropométrie : Poids, Taille, Indice de Masse Corporelle (voir tableau 8.3),... ;
- Bilan d'Insulino-Résistance : Glycémie, Insulinémie, QUICKI [Katz *et al.* 2000],... ;
- Bilan Lipidique : Cholestérolémie Totale (CT) , Cholestérolémie HDL (HDL), Triglycérides (TG), Cholestérolémie LDL (LDL, où $LDL = (CT - HDL - TG)/5$),... ;
- Bilan Léptinémique : Léptinémie ;

Il est important de noter que parmi ces données, il existe des indices calculés. L'Indice de Masse Corporelle, noté IMC, (*Body Mass Index ou BMI*, en anglais) est l'un d'entre eux. Son calcul consiste à diviser le poids P (exprimée en kilogrammes) par le carré de la taille T de la personne (en mètre) : $IMC = M/T^2$. A partir du résultat obtenu, il est possible de classer un individu dans une classe qui le positionne vis-à-vis de l'Obésité (voir figure 8.3).

Il existe également d'autres indices comme le rapport « Tour de Taille/Tour de Hanches », noté RTH (*Waist Hips Report (WHR)*, en Anglais) [Yusuf *et al.* 2005] qui reflète le niveau

18. http ://www.microarray.org/sfgf/

Expérimentation	Nombre d'individus	Nombre de paramètres biocliniques	Nombre de gènes	Origine des données
Muscle	9	2	1680	1 centre
Inflammation	29	≈ 10	≈ 40000	mixte
CTSS	39	1	≈ 400	mixte
Gastroplastie	10	>30	≈ 40000	1 centre
Basale	39	22	39739	mixte

FIGURE 8.4 – Récapitulatifs des données expérimentales disponibles.

d'adiposité viscérale, le QUICKI (*Quantitative Insulin Sensitivity Check Index*, en Anglais) [Katz *et al.* 2000] et le HOMA (*Homeostasis Model Assessments*, en Anglais) [Matthews *et al.* 1985] qui sont tous deux des indices de mesures de résistance à l'Insuline (généralement élevés chez l'Obèse). En générale, dans les données issues de protocole de recherche clinique, on dispose à la fois des données initiales, comme le poids, la taille, le tour de taille,... mais aussi des indices calculés. Cette remarque est importante car elle induit que les données biocliniques sont rarement indépendantes. Il faut donc le prendre en compte lors des analyses.

De manière synthétique, les différentes expérimentations que nous avons réalisées se sont appuyées sur des données venant de différents protocoles de recherche comme le montre la figure 8.4. Dans les parties suivantes, nous exposons une partie des résultats que nous avons obtenus pour ces différentes données.

8.2 Aide à la découverte de corrélation au sein de données basales

Dans le cadre de cette expérimentation, nous nous sommes concentrés sur un ensemble de données concernant 39 femmes (obèses et non-obèses) à l'état basal (c'est-à-dire avant toute intervention diététique, psychologique, pharmacologique et/ou chirurgicale). Pour chacun de ces sujets, nous disposons de données issues de puces à ADNc contenant entre 25 000 et 40 000 valeurs d'expressions géniques (le nombre de valeurs étant fonction de la qualité des données au terme des manipulations biologiques et des pré-filtrages réalisés). Nous disposons également de données biocliniques (de 3 à 22 valeurs en fonction de la complexité des protocoles d'origine des sujets). Nous avons essentiellement travailler dans cette expérimentation avec un paramètre clinique, l'IMC et nous allons montrer les résultats obtenus par les approches corrélationnelles sur ces données.

8.2.1 Traitements des données sources et exploration des corrélations obtenues

Nous avons calculé automatiquement les corrélations « globales » et « locales » entre toutes les valeurs d'expressions géniques et les valeurs biocliniques disponibles pour 39 927 gènes. Les calculs pour les 22 paramètres (avec le nombre de sujets variants) ont pris environ 24 heures. Ce temps est extrêmement court par rapport au temps que mettrait un biologiste en cherchant *a priori* les gènes et les paramètres d'intérêt avant de réaliser des calculs (plusieurs heures à plusieurs semaines).

Résultats par l'approche « globale » pour l'IMC

L'exploration des corrélations « globales » calculées entre l'Indice de Masse Corporel (I.M.C.) (39 valeurs disponibles) et les valeurs d'expressions géniques disponibles est donnée sous la forme du diagramme de Hasse présenté à la figure 8.5. Il prend en compte les seuils par défaut à savoir $|\rho_S| \geq 0,66$, $p \leq 0,05$, $q \leq 0,05$ et $n \geq 95\%$ (soit 38 individus). On peut observer qu'un nombre très restreint de relations répondent aux 4 contraintes simultanément. Ces relations sont les mêmes que celles des noeuds $\{r,p,n\}$, $\{r,q,n\}$ et $\{p,q,n\}$. Le noeud $\{r,p,q\}$ comprend 18 relations dont 14 « nouvelles » par rapport aux noeuds présentés précédemment. Cela signifie que ces relations sont potentiellement intéressantes, mais pour un effectif inférieur à 38 individus (puisque la condition sur n n'est pas vérifiée dans ce noeud).

Ce constat peut s'appliquer à q, qui est un indicateur du taux de fausse découverte. Ainsi, ce n'est pas parce que des relations sur les nœuds $\{r,p,n\}$ semblent intéressantes, mais non significatives pour q, qu'elles ne le sont pas. On retrouve cette situation pour d'autres paramètres biocliniques. Ainsi, le nombre de relations considérées comme significatives pour p est extrêmement important mais très réduit pour q et il n'est pas possible d'utiliser q comme paramètre discriminant. Il faut alors s'appuyer dans un premier temps sur r, p et n puis sur la littérature pour définir ces relations qui présentent un réel intérêt.

L'étape suivante de l'expérimentation a consisté à sélectionner un nœud du diagramme de Hasse afin de prendre connaissance de la liste des relations qui y sont définies et de visualiser ces relations. On obtient un tableau tel que celui de la figure 8.6 qui correspond au contenu du nœud $\{r,p,q,n\}$ du diagramme de Hasse précédent. Le paramétrage des seuils de pertinence a donc permis de mettre en évidence 5 gènes parmi les 39927 que l'on avait à traiter. Lorsque l'on clique sur le nom du gène, un lien est fait vers SOURCE [19] et lorsque l'on clique sur une relation représentée avec L_V (dans la colonne « DCVisu ») ou avec L_S (dans la colonne « DCSymb »), il est possible de visualiser la relation sous la forme d'un graphique, comme nous l'avons montré dans le chapitre 7.

Sur la figure 8.5, nous pouvons également observer le nœud $\{r,p,q\}$. Il est composé de 17 gènes et il permet de montrer que le $n \geq 95\%$ est une valeur de limitation des gènes potentiellement intéressants. En effet, dès que l'on fait abstraction de cette valeur de contrainte le nombre de relations augmente considérablement. De la même manière, les nœuds $\{r\}$, $\{p\}$ et $\{n\}$ analysés l'un après l'autre montrent qu'ils contiennent un très grand nombre de relations. Ces nœuds ne pourraient pas être efficacement étudiés par un biologiste.

D'autre part, les biologistes ont basé principalement leurs recherches de corrélation sur la valeur de significativité p de ces relations. Selon le diagramme 8.5 présenté, le nombre de relations ayant $p \leq 0,05$ est de 3688. Ceci soulève un problème en terme de multiplicité des tests. Pour définir des résultats les plus intéressants, il est nécessaire de mettre en œuvre comme nous le faisons, des tests de multiplicité (la valeur q), qui réduit le nombre des résultats considérés comme significatifs pour mettre en avant les plus « intéressants ». Ces résultats sont au nombre de 17 (nœud $\{p,q\}$) et correspondent aux 17 résultats associés aux nœuds $\{r,p\}$ et $\{r,p,q\}$. Cette relation entre r, p et q se retrouve quelque soit le paramètre bioclinique étudié.

L'histogramme de la figure 8.7 nous montre que le nombre d'individus sur lequel est calculé l'ensemble de nos relations varie en fonction des gènes. Il n'y a qu'environ 500 gènes pour lesquels on a tous les individus de définis. Nous pouvons aussi observer qu'en moyenne les relations sont calculées pour des ensembles incluant entre 20 et 30 individus. Ce dernier point est critique car il implique que les résultats seront difficilement généralisables à une grande population et cela est dû à de nombreuses données d'expression génique ou biocliniques manquantes.

19. http ://genome-www5.stanford.edu/cgi-bin/source/

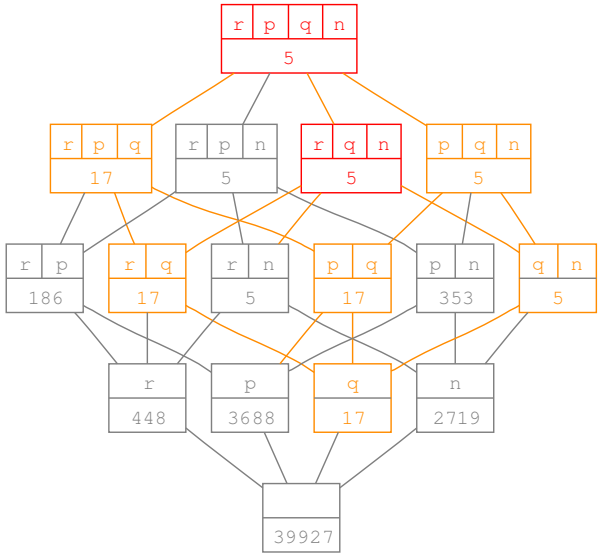


FIGURE 8.5 – Diagramme de Hasse généré pour les données basales avec l'IMC pour les valeurs de seuils suivantes : $r \geq 0.66$, $p \leq 0.05$, $q \leq 0.05$ et $n \geq 95\%$.

Gene	DCVisu	DCSymb	n	r	p	q	p/q
THRAP3		\	38	-0.66444	0.00001	0.03069	0.00018
S100A13		\	39	-0.71205	<1E-005	0.00740	0.00005
C10orf10		\	38	-0.69133	<1E-005	0.02050	0.00008
Transcribed locus, s...		\	39	-0.66734	<1E-005	0.03069	0.00011
C6orf47		\	37	-0.66319	0.00001	0.03088	0.00025

FIGURE 8.6 – Liste des gènes corrélés avec l'IMC issues du nœud $\{r, p, q, n\}$ du diagramme de Hasse de la figure 8.5.

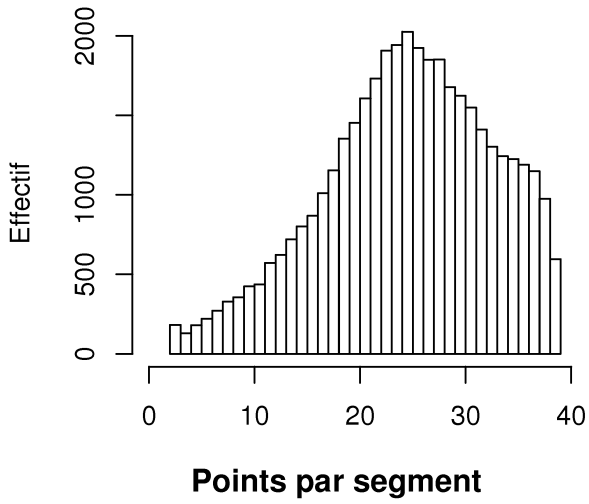


FIGURE 8.7 – Histogramme du nombre de données pour chaque relation « expression génique *vs.* IMC ».

La répartition des valeurs de corrélation varient entre -1 et $+1$ mais en fait très peu de gènes ont de très fortes ou de très faibles corrélations avec l'IMC (comme le montre la figure 8.8). Cela est logique d'un point de vue biologique car un petit nombre de gènes explique en général une pathologie.

Il est intéressant de remarquer à la figure 8.9 que les valeurs de p et de sa significativité sont distribuées telle que la littérature [Eisen 1999, Storey 2002] le recommande pour une bonne l'interprétation des valeurs du taux de fausse découverte. Plus précisément, le nombre de valeurs de p est décroissant sur l'intervalle $[0; 1]$. Nous avons donc un grand nombre des valeurs de p qui tendent vers 0 et un petit nombre qui tendent vers 1. Dans la pratique (sur d'autres paramètres biocliniques) nous avons rencontré le cas où q ne peut être exploité car les valeurs de p étaient homogénement distribuées sur l'intervalle $[0; 1]$.

La figure 8.10 permet de valider le fait que les meilleures valeurs de q sont liées aux meilleures valeurs de p, ainsi que l'ordre entre ces ensembles de valeurs statistiques ne change pas. Ainsi, on peut en conclure que les valeurs ρ_S, p et n semblent et peuvent être suffisantes pour évaluer l'intérêt d'une corrélation intéressante [Efron 2006c, Efron 2006a].

Pour conclure cette section, nous considérons qu'il est nécessaire de combiner quelques valeurs (indicateurs) statistiques comme ρ_S, p, q et n pour proposer à l'utilisateur une liste pertinente de relations entre l'expression génique et des valeurs biocliniques. Ceci permet de réduire le nombre de relations proposées et le risque de conduire l'expert dans une voie de recherche peu intéressante, voir erronée. Cependant, comme le nombre de relations *globales* peut être très réduit, il est alors nécessaire d'essayer de proposer à l'expert, des relations localement intéressantes. Nous présentons le résultat de ces analyses dans le paragraphe suivant.

Résultats issues de l'approche « locale »

Les résultats de l'approche « locale » présentés ici concernent les mêmes données que celles présentées pour l'approche « globale » (expression génique *versus* IMC). Le diagramme de Hasse de la figure 8.11 montre les résultats obtenus lors de la recherche de segments intéressants avec les

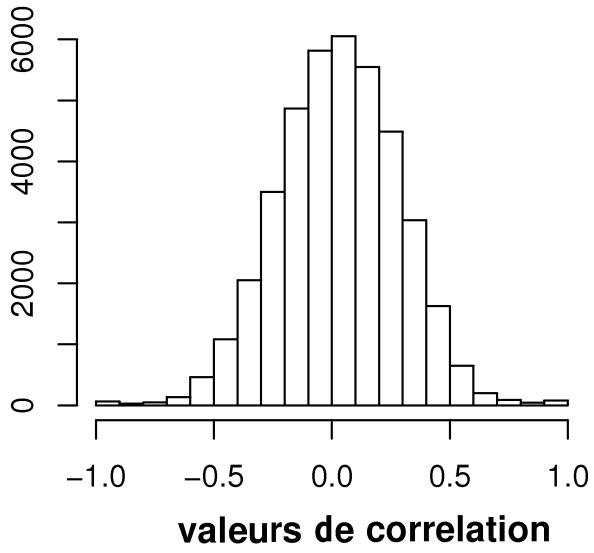


FIGURE 8.8 – Histogramme des valeurs de corrélation calculées par l'approche « globale ».

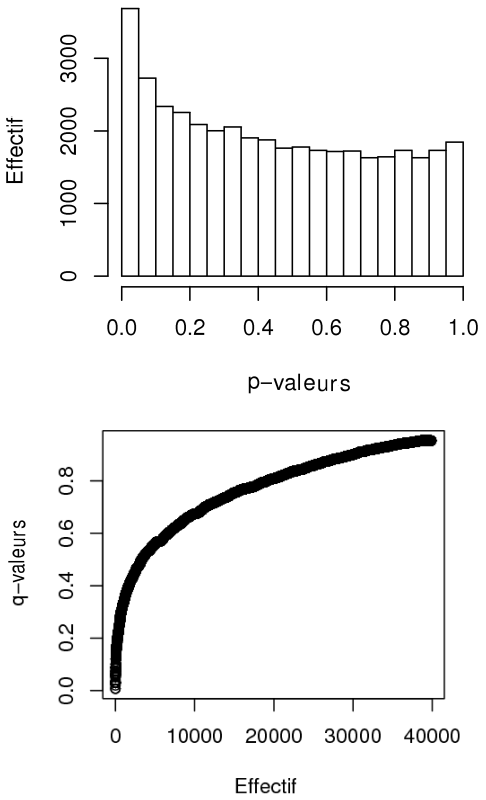


FIGURE 8.9 – Histogrammes de $p-valeur$ et de $q-valeurs$ associées dans le cadre des relations calculées par l'approche « globale ».

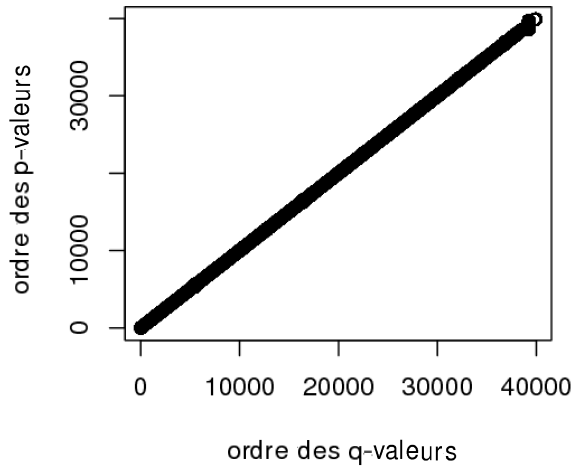


FIGURE 8.10 – Relation entre les p-valeurs et les q-valeurs pour chaque relation.

contraintes : $\rho_S \geq 0,66$, $p \leq 0,05$, $q \leq 0,05$ et $n \geq 20\%$ (soit incluant au minimum 8 individus par segment).

Nous pouvons observer que l'utilisation de q aide à réduire le nombre de relations considérées comme intéressantes. Bien que, dans le nœud $\{r, p, q, n\}$, le nombre de relations soit élevé (471), nous avons exploré les données définies par ce nœud en fonction du type de relations et de l'intérêt de ces relations pour le biologiste. DISCOCLINI a découvert 37 relations composées par 2 segments croissants (//), 44 composées par 2 décroissants (\\), 4 avec un segment croissant et un décroissant (/\) et 5 avec un segment décroissant et un croissant (\/). Nous pouvons conclure de ces résultats que le nombre de segments pouvant être étudiés avec profit est réduit (≈ 100).

Les situations particulières décrites par (/\) et (\/) sont exceptionnelles. Elles demandent une étude détaillée par l'expert afin d'être comprises et interprétées, car les connaissances actuelles en Génomique Fonctionnelle ne sont pas, d'après les biologistes, suffisantes pour expliquer ce type de variations.

L'histogramme de la figure 8.12 montre que le nombre de corrélations détectées inclus majoritairement 10 points par segment. Ceci peut trouver une explication dans la définition du nombre minimal de points utilisés pour le fenêtrage qui est égale à 10. Cependant, les segments qui contiennent plus de points sont plus intéressants car couvrant une population plus grande. Même en utilisant un fenêtrage, plus petit on retrouve des sous-populations d'environ 10 individus. Donc 10 semble être un bon compromis de taille de fenêtre dans le cadre des données que nous étudions.

La figure 8.13 permet d'observer que le nombre de corrélations tendant vers $|1|$ est beaucoup plus élevé que le nombre obtenu avec l'approche « globale ». Cela s'explique par le fait que le fenêtrage a pour but de détecter les meilleurs segments corrélés et *in extenso* permet de définir des corrélations meilleures et donc beaucoup plus homogènes. D'autre part, on observe la présence de deux pics pour les valeurs de corrélations, car les résultats ne sont pas restitués en valeur absolue ; ceci permet de montrer que l'approche de fenêtrage ne favorise pas plus un type de résultats (positif ou négatif) en particulier. De plus les valeurs p sont distribuées en respectant moins les règles éditées par la littérature. Néanmoins, toutes les valeurs de q sont inférieures à $0,5$ (voir la figure 8.14). Nous pouvons en déduire que tous les segments intéressants découverts sont « corrects » et les autres sont réellement non-intéressants. Le classement des ensembles des

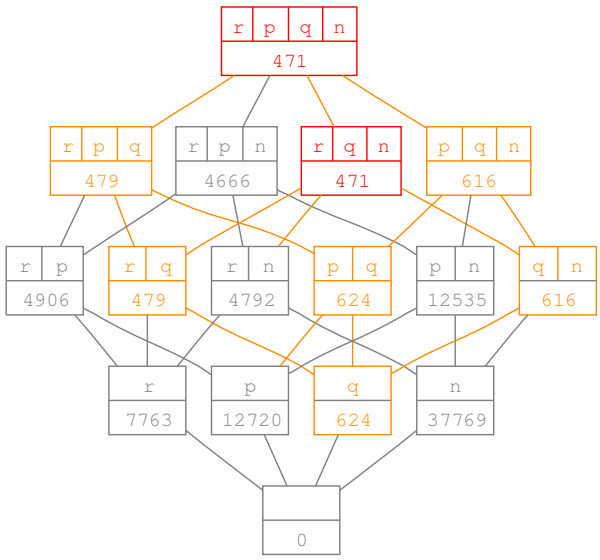


FIGURE 8.11 – Diagramme de Hasse généré pour les relations avec l'IMC avec les valeurs de seuils $\rho_S \geq 0,66$, $p \leq 0,05$, $q \leq 0,05$ and $n \geq 20\%$.

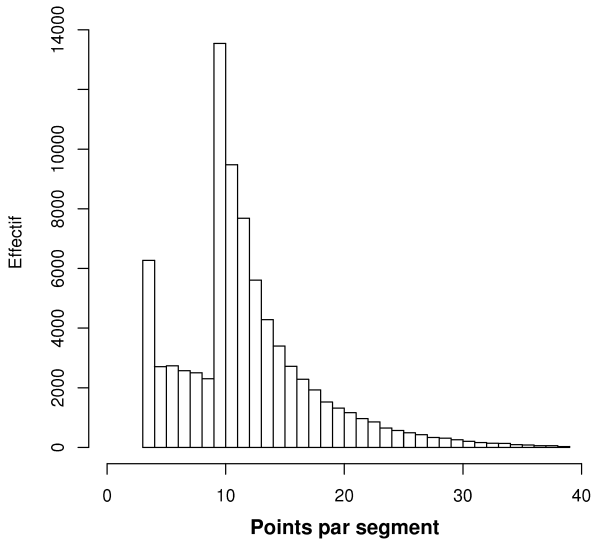


FIGURE 8.12 – Histogramme du nombre de valeurs par segments pour les relations « expression génique *vs.* IMC ».

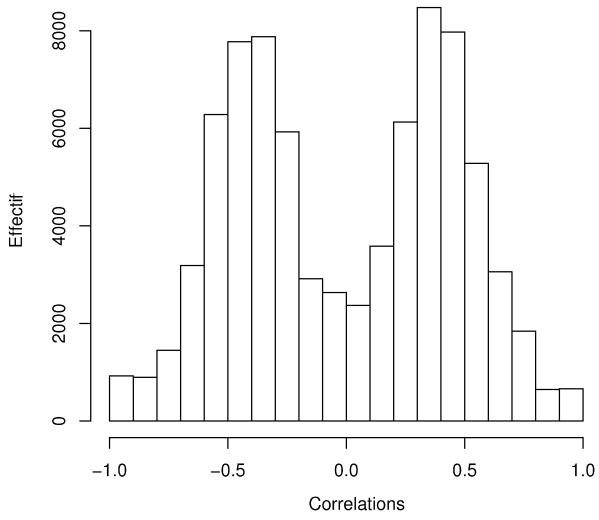


FIGURE 8.13 – Histogramme des valeurs de corrélations calculées par l'approche « locale ».

valeurs de p et q est presque le même, comme le montre la figure 8.15. Nous pouvons en conclure que les valeurs de q peuvent être utilisées comme un filtre pour réduire le nombre de corrélations présentées au biologiste, mais ne donnent pas plus d'informations sur la qualité d'une relation. Ainsi, elles vont permettre d'augmenter ses chances de faire de nouvelles « découvertes » mais elles ne seraient pas utilisées comme valeur de significativité des résultats.

L'approche « locale » permet de découvrir des relations intéressantes dans des données précédemment ignorées par l'approche « globale ». L'utilisation de résultats issus de cette seconde approche peut permettre de mettre en évidence des dysfonctionnements physiologiques pour des cas cliniques rares ou des sous-populations particulières.

Conclusion

Pour conclure sur cette partie de l'expérimentation sur les données basales, nous pouvons souligner que DISCOCLINI permet que ce soit avec l'approche « globale » ou « locale » de mettre en évidence les relations potentiellement existantes dans les données. Cette analyse est simple et demande un investissement réduit, en terme de temps à l'expert. L'association de l'automatisation du processus de calculs et d'outils simples de navigation pour les résultats et leurs visualisations permet d'accroître les chances de découvrir des biomarqueurs. Ces biomarqueurs ont besoin d'être par la suite validés par des expérimentations biologiques et médicales complémentaires avant d'aboutir à des applications diagnostiques ou prédictives réelles. Cependant, il est important de noter que ces résultats doivent être pondérés en fonction de la qualité des données. Nous allons nous concentrer sur cet aspect dans la partie suivante.

8.2.2 Détections de valeurs singulières

L'algorithme PAMOUT a été appliqué sur les données afin de détecter pour chaque gène (ensemble de valeurs d'expression génique), pour chaque paramètre bioclinique et pour chaque relation (couple « un gène *vs.* un paramètre bioclinique »), les individus pouvant impacter voir fausser les résultats. Nous avons étudié ces résultats à la fois dans une perspective univariée et

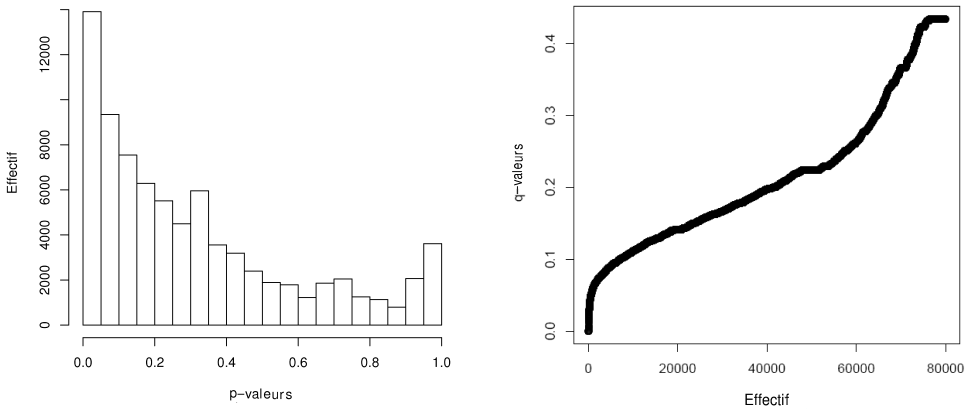


FIGURE 8.14 – Histogrammes des valeurs des p-valeurs et des q-valeurs associées dans le cadre des relations calculées par l'approche « locale ».

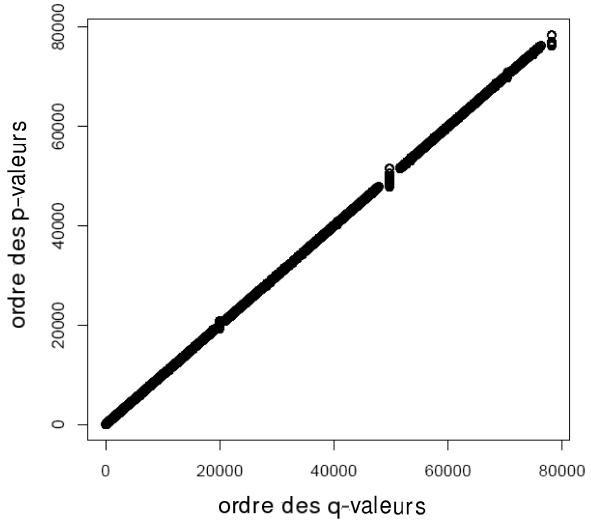


FIGURE 8.15 – Ordre des valeurs de q *versus* celles de p pour chaque relation « expression génique *vs.* IMC ».

bivariée.

L'étude corrélationnelle que nous avons réalisée et pour laquelle nous avons présenté les résultats auparavant a permis de mettre en avant des relations intéressantes pouvant aboutir à la définition de biomarqueurs. Néanmoins, il est pertinent de déterminer si ces relations sont ou non « faussées » par des valeurs singulières afin d'aider le biologiste à valider une découverte.

Dans les données biocliniques, la recherche de ce type d'individus peut être considérée comme simple, les données étant peu nombreuses. Néanmoins, cette tâche peut être définie comme partie intégrante d'un système d'aide à la décision [Degoulet et Fieschi 1998]. Notre objectif majeur dans cette thèse étant de réduire au maximum l'intervention humaine aux cours des études corrélationnelles, nous avons mis en œuvre PAMout sur ces données. En quelques secondes ($\approx$ 5 *sec.*), les 22 paramètres biocliniques ont été explorés automatiquement. Nous avons pu constater que les données collectées l'ont été de manière homogène (en terme de profil des individus au regard de ces valeurs), puisqu'aucune valeur singulière n'a été détectée. Cette analyse nous permet donc de valider les données biocliniques et de les considérer comme « fiables ».

Dans les données issues des puces à ADNc, la recherche de valeurs singulières a aussi été réalisée grâce à PAMout. Il est important de noter que certaines valeurs singulières ont déjà été enlevées par l'opérateur lors de la lecture des puces à ADNc. Environ 25 minutes sont nécessaires à l'exploration automatique des 40 000 gènes. Les résultats obtenus sont présentés dans la figure 8.16 (colonnes « VS/puce », « VS/puce (%) », « VS biv. » et « VS biv. (%) ») et le graphique de la figure 8.17.

La figure 8.16 inclut des informations relatives au statut pondéral (IMC) des individus étudiés (première colonne), ainsi que des données relatives aux valeurs collectées pour chacun d'eux (nombre de valeurs d'expression génique par puce à ADNc -*Exp./puce*-, *in extenso* par individu). Les résultats issus de la mise en œuvre de PAMout sur les données issues de puces à ADNc sont présentés :

- sous la forme de valeurs réelles $VS/puce$: le nombre de valeurs singulières détectées par puce à ADNc,
- sous la forme de valeur relative $VS/puce(\%)$: le pourcentage de valeurs singulières par puce à ADNc.

Les résultats issus de la mise en relation des données issues des puces à ADNc et des valeurs d'IMC sont présentées suivant les mêmes formulations (respectivement $VSbiv.$ et $VSbiv.(\%)$). Ces derniers résultats sont obtenus en environ 30 minutes pour l'exploration automatique d'environ 40 000 ensembles de valeurs mettant en relation des données d'expression génique et d'un paramètre bioclinique, par 39 individus. L'intégralité des valeurs d'IMC étant disponibles, les résultats de la figure 8.16 sont triés suivant ces données. Il est possible à l'aide de la colonne *Exp./puce* de constater que le nombre de valeurs collectées par puce à ADNc est très hétérogène entre 4 503 et 38 169 par puce à ADNc. Ainsi comme nous l'avons mentionné précédemment seul un nombre très restreint de relations inclus des valeurs relatives à tous les individus. Nous pouvons ainsi confirmer nos propos sur l'hétérogénéité de la qualité des données issues des puces à ADNc. Les pourcentages de valeurs singulières trouvées sur les puces renforcent ces propos car ils varient de $0,22\%$ (pour 38169 valeurs sur le support concerné) à $12,91\%$ (pour 15 399 valeur pour ce second support). La lecture de la figure 8.16 et du graphique 8.17 ne permettent pas de définir de lien entre l'hétérogénéicité du nombre de mesures par individu avec les valeurs du paramétre considéré.

Les données relatives aux valeurs singulières définies dans les mises en relation réalisées entre l'expression génique et l'IMC sont porteuses d'informations complémentaires à celles que nous venons d'exposer. Le nombre de valeurs définies comme singulières au cours de la mise en œuvre de PAMout sur les données bivariées indique un accroissement notable de ces valeurs par rap-

Individu	Exp./puce	VS/puce	VS/puce (%)	IMC	VS biv.	VS biv. (%)
BUR	36525	1961	5,37	19,72	4906	**13,43**
omega1	16513	743	4,50	21,40	1653	**10,01**
H	37198	2047	5,50	21,99	2689	7,23
omega8	7391	311	4,21	22,40	401	5,43
B	34946	3991	**11,42**	22,89	4660	**13,33**
omega3	23984	759	3,16	23,60	978	4,08
omega6	4503	113	2,51	23,90	159	3,53
A	12512	319	2,55	24,56	330	2,64
D	36546	885	2,42	24,62	1098	3,00
omega9	16442	706	4,29	25,20	915	5,57
F	22294	768	3,44	26,17	830	3,72
DEP	33256	285	0,86	32,60	1061	3,19
ZEAU	31553	1168	3,7	32,90	1462	4,63
DGDP	34956	126	0,36	32,97	1686	4,82
DIDT	15399	1988	**12,91**	33,64	696	4,52
PC	18541	243	1,31	33,79	1110	5,99
KAM	37090	69	0,19	34,00	2174	5,86
NGO	26544	716	2,7	34,50	1754	6,61
DHDR	16848	657	3,9	35,36	1077	6,39
PN	31366	115	0,37	35,51	1075	3,43
PE	29882	386	1,29	36,25	1080	3,61
PO	35754	533	1,49	36,31	2010	5,62
DADJ	31393	378	1,20	36,90	1635	5,21
GAN	36894	130	0,35	37,20	2276	6,17
DA	27032	1317	4,87	37,56	1991	7,37
ZEMB	19963	624	3,13	39,80	986	4,94
VAR	32206	997	3,1	40,00	1268	3,94
PX	37123	400	1,08	40,12	2438	6,57
BR	24498	583	2,38	40,66	1342	5,48
DAR	40,80	29711	263	0,89	1471	4,95
V	13812	419	3,03	41,36	742	5,37
DCHDS	38169	84	0,22	41,90	2693	7,06
CAR	42,20	28770	695	2,42	1670	5,80
II	7293	278	3,81	42,28	373	5,11
CAL	12660	539	4,26	46,00	959	7,58
I	5671	243	4,28	50,74	375	6,61
DOR	13883	1297	9,34	55,87	2013	14,5
DEC	30812	2103	6,83	57,19	2920	9,48
LEF	35144	195	0,55	60,50	5751	**16,36**

FIGURE 8.16 – Récapitulatif des nombres de valeurs singulières détectées par individu (identifié par un identifiant interne) lors de l'analyse des puces à ADNc (colonnes « VS/puce » et « VS/puce (%) ») mais aussi des relations « expression génique *vs.* IMC » (colonnes « VS biv. » et « VS biv. (%) »). La valeur de l'IMC de chaque individu est donnée à la colonne « IMC ».

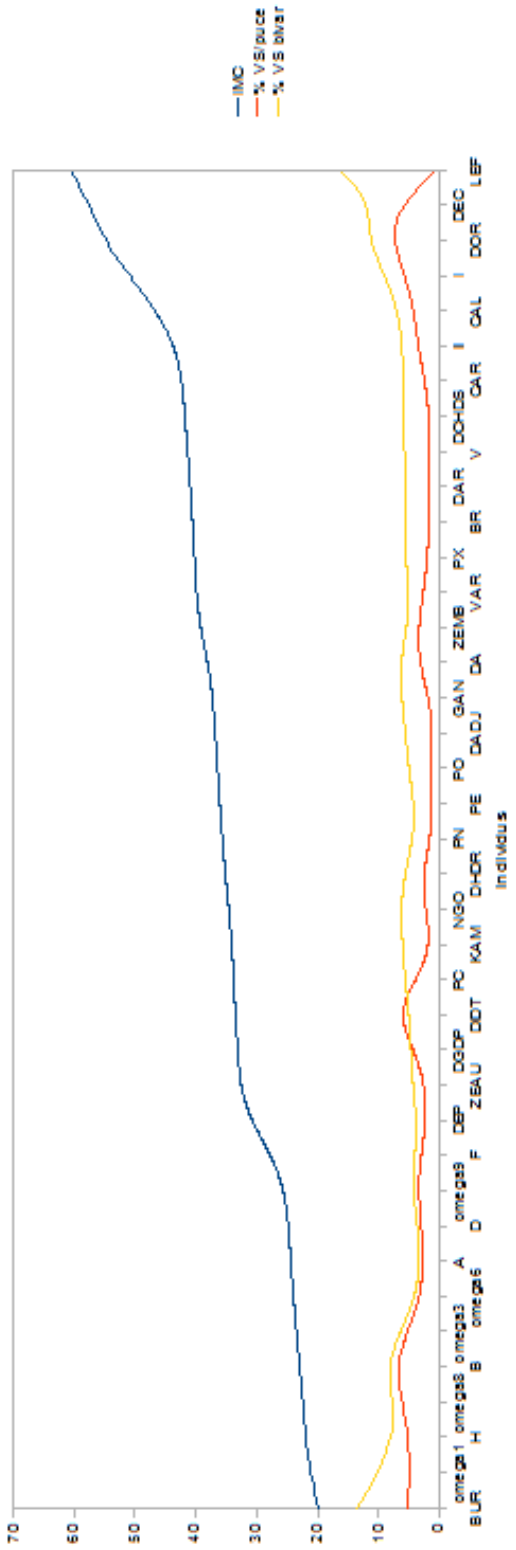


FIGURE 8.17 – Synthèse graphique des résultats présentés dans le tableau 8.16, c'est-à-dire les pourcentages de valeurs singulières pour les valeurs d'expression génique par puce à ADNc et les données « expression génique *vs.* IMC ».

port aux valeurs univariées (issues uniquement des puces à ADNc). Ces valeurs varient de $2,64\%$ (12515 valeurs d'expression génique et un $IMC = 24,56$ pour l'individu A) à $16,36\%$ (pour 35144 valeurs d'expression génique et un $IMC = 60,5$ pour l'individu LEF). Il est important de souligner que l'individu LEF qui présente le plus grand nombre de valeurs singulières pour les données bivariées (e.g. *expression génique vs. IMC*) est celui qui présente un des plus petits nombres de valeurs singulières pour les données univariées. Cela s'explique par le fait que LEF à l'IMC le plus grand de la population étudiée. Ainsi, cette valeur marginale fait que malgré des valeurs issues de la puce à ADNc (que l'on peut qualifier de bonne qualité au vue des résultats obtenus), cet individu se retrouve fréquemment isolé du reste de la population pour les relations impliquant des gènes pour lesquels les valeurs manquantes sont relativement importantes. Le même comportement est observable sur l'extrême opposé des valeurs d'IMC, où la valeur la plus faible est égale à $19,72$ (individu BUR). Le nombre de valeurs singulières qui lui est associé est multiplié par trois entre l'exploration univariée et l'exploration bivariée. De manière synthétique, plus les valeurs sont marginales dans l'un des deux ensembles de données (*a fortiori*, le plus complet) plus ces valeurs lorsqu'elles sont liées à d'autres (ensembles moins complets) ont un risque d'être définis comme aberrantes (au sens de la figure 6.1 correspondant au guide d'interprétation des résultats issus de PAMout). Graphiquement (voir la figure 8.17) on note ainsi que les extrémités de la courbe $\%VSbivar$ sont respectivement décroissante et croissante.

Des cas particuliers peuvent exister. Le nombre de valeurs singulières peut décroître entre les résultats issus de l'exploration des données univariées et les résultats issus de l'exploration des données bivariées (par exemple, pour l'individu $DIDT$, on passe de $12,91\%$ à $4,52\%$). Cela s'explique par l'existence d'ensembles de données incomplets. Les distances inter-individus augmentent et s'homogénéisent et il n'est alors plus possible de distinguer un individu particulier défini *via* les données univariées, d'un autre « normal ». Graphiquement (voir la figure 8.17), on note une inversion de tendances, pour ce type d'individus, pour les courbes $\%VS/puce$ et $\%VSbivar$.

Ces différentes expérimentations de PAMout dans le cadre de l'exploration des données « basales » nous ont permis de valider l'intérêt de notre approche, en plus des expérimentations que nous avons réalisées sur des données artificielles. PAMout permet en combinant les différents résultats obtenus de distinguer les individus suspects des individus aberrants pour une relation donnée. Néanmoins ces résultats ne sont pas « sûrs » pour une expérimentation dans son ensemble (un ensemble de relations) quand les données utilisées sont lacunaires de manière hétérogène, c'est-à-dire que l'on ne peut pas comparer les résultats d'individus qui ne sont pas présents dans les mêmes relations. PAMout permet ainsi d'informer l'expert sur la qualité globale des données et des résultats obtenus. PAMout n'a pas pour vocation d'encourager l'expert à supprimer des valeurs pour améliorer des résultats, d'une part car le nombre d'individus-sources de données est restreint et d'autre part car toute omission ne ferait que rendre les résultats peu généralisables.

8.3 DiscoClini et la Recherche Médicale

Le système DiscoClini a participé à un certain nombre de « découverte médicale » comme nous allons le voir dans cette partie.

8.3.1 Étude de l'impact de l'adrénaline sur l'expression des gènes dans le muscle

Les premières expérimentations de DiscoClini sur des données réelles se sont inscrites dans le cadre d'une étude d'impact de l'adrénaline sur l'expression des gènes dans le muscle squelet-

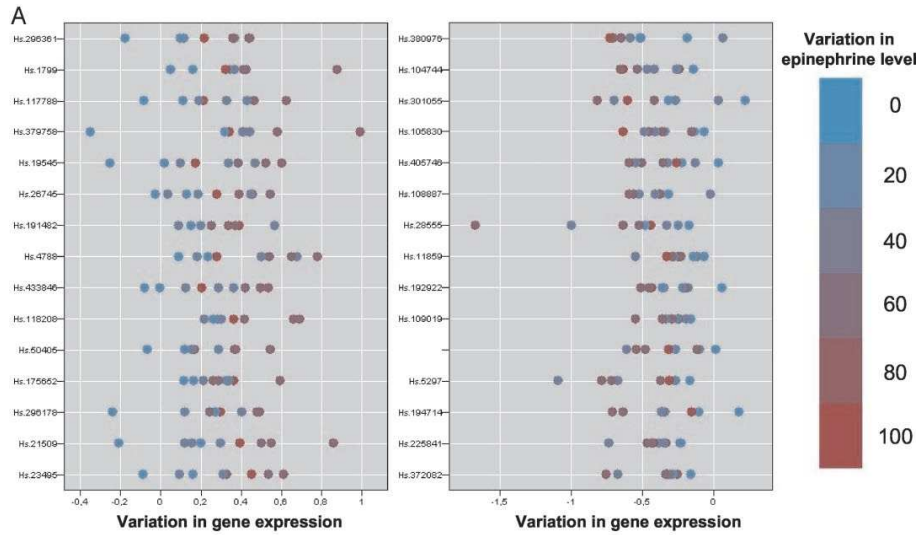


FIGURE 8.18 – Représentation graphique de la variation de l'expression génique par rapport aux taux d'adrénaline [Viguerie *et al.* 2004].

tique chez 9 hommes jeunes sains [Viguerie *et al.* 2004]. Pour chacun des sujets, nous disposions de données biologiques et d'expressions géniques obtenues à 2 dates. Les premières avant que les sujets soient traités et les secondes après qu'ils aient été perfusés pendant 6 heures avec une solution d'adrénaline. Les données d'expression génique correspondaient à celles de 1680 gènes (1206 étaient sur-exprimés et 474 sous-exprimés) que les biologistes ont sélectionnés grâce notamment à une analyse différentielle sur environ 43 000 présents sur les puces à ADNc. Nous avons calculé de manière automatique les différences d'expression avant et après la perfusion d'adrénaline ainsi que celles des données relatives au métabolismes. Nous avons réalisé de calculs des corrélations globales sur ces données. Sur les meilleurs résultats (arbitrairement les 15 premiers résultats), nous avons appliqué le langage L_V pour permettre une visualisation synthétique des résultats. Les figures 8.18 et 8.19 [Viguerie *et al.* 2004] montre la première version du langage L_V qui a été développée. La figure 8.18 présente l'expression génique en fonction du taux d'adrénaline pour les 15 meilleurs gènes sur-exprimés (à droite) et les 15 meilleurs gènes sous-exprimés (à gauche). La figure 8.19 correspond à la variation de l'expression génique *vs.* la réponse métabolique pour les 20 meilleures valeurs de corrélation de Spearman parmi les gènes sur-exprimés. Le gradient des couleurs correspond aux variations de l'adrénaline circulante pour la figure 8.18 et de la réponse métabolique pour la figure 8.19 entre les deux dates extrêmes de l'expérience. La ligne encadrée sur la figure 8.19 est celle correspondant à la meilleure valeur de corrélation ($\rho_s = 0,93$), c'est-à-dire à un gène dont l'expression est très fortement corrélée aux variations de la réponse métabolique.

Ces deux expérimentations (calculs automatiques des corrélations et visualisation) ont contribué, d'un point de vue biomédicale, à définir les liens entre l'adrénaline et les mécanismes inflammatoires. De plus, nous avons avec cette étude validée, d'un point de vue informatique, l'intérêt des biologistes pour l'approche de calcul systématique de l'ensemble des données génomiques et biocliniques disponibles dans un protocole de recherche clinique. Enfin, cette expérimentation nous a ainsi permis de valider notre mode de visualisation : cette représentation a pour eux le même intérêt que les dendrogrammes de Eisen ([Eisen 1999]) dans des contextes différents.

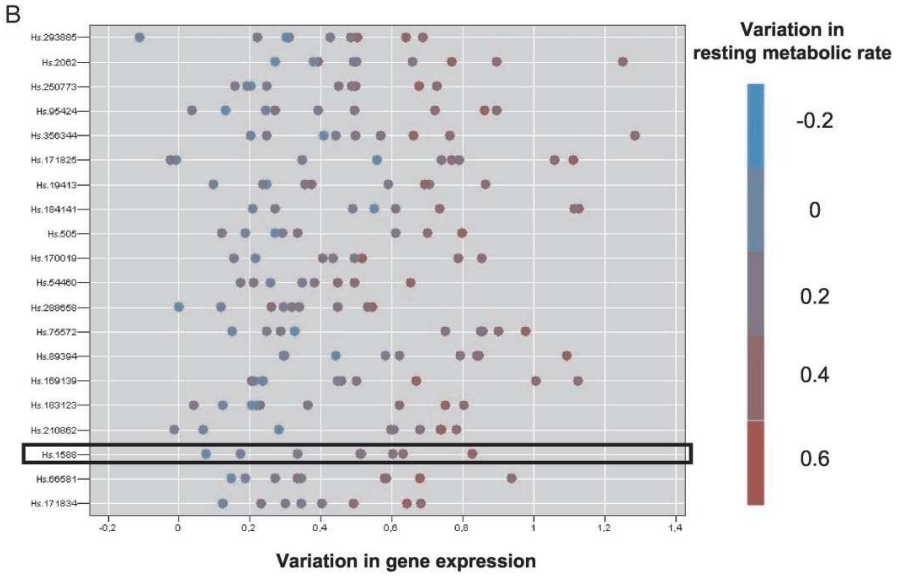


FIGURE 8.19 – Représentation graphique de la variation de l'expression génique par rapport à la réponse métabolique [Viguerie *et al.* 2004].

Étude des variations de l'expression génique dans le tissu adipeux

Une autre expérimentation a été réalisée dans le but de valider une intuition biologique : l'Obésité est une pathologie ayant des facteurs inflammatoires et ceux-ci sont perceptibles à l'échelle génomique.

Nous avons appliqué notre approche dans le cadre de l'analyse des données d'un protocole visant à étudier les variations de l'expression génique dans le tissu adipeux sous-cutané chez 29 femmes obèses au cours de deux protocoles de (forte) restriction calorique.

La mise en œuvre de DISCOCLINI a contribué à définir qu'une perte de poids même faible a une influence sur le profil inflammatoire en induisant une baisse des facteurs pro-inflammatoires et une hausse des anti-inflammatoires [Clément *et al.* 2004]. Nous avons recherché de manière systématique toutes les corrélations globales entre les variations de poids chez ces femmes. Parmi les relations les plus intéressantes, nous avons mis en évidence avec les biologites, des cytokines comme les interleukines IL-6, IL-12 et des facteurs de croissance qui sont impliqués dans la cicatrisation et le contrôle négatif de l'inflammation comme *Transforming growth factors (TGF)*.

Ainsi, l'utilisation de DISCOCLINI dans cette recherche a permis de valider l'hypothèse fondamentale qui consiste à rechercher sans *a priori* des gènes pouvant être de potentiels biomarqueurs.

Aide à la découverte de la Cathepsine S

Une contribution directe de DISCOCLINI a été l'aide à la découverte de la Cathepsine S (noté CTSS), comme nouveau marqueur de l'adiposité [Taleb *et al.* 2005]. Les biologistes se sont plus particulièrement intéressés aux corrélations pouvant exister entre l'expression génique et l'Indice de Masse Corporelle (IMC) de 28 sujets *obèses* et de 11 sujets *sains*. DISCOCLINI a été mis en œuvre afin d'optimiser le temps de travail de l'expert dans le but de calculer les valeurs de corrélations globales. Cette expérimentation riche en *a priori* a permis :

– d'un point de vue biologique de caractériser la CTSS comme biomarqueur de l'Obésité ;

– d'un point de vue informatique de montrer l'intérêt double de l'automatisation des calculs de mises en relation des données.

Les biologistes nous ont fournis une liste de quelques centaines de gènes pré-sélectionnés à l'aide de tests statistiques de significativité en fonction de leurs valeurs d'expressions géniques ainsi que de leurs connaissances. La mise en œuvre de DISCOCLINI dans ces conditions a permis d'obtenir des résultats détaillés en quelques minutes mais fortement biaisés par les *a priori* initiaux. Le gène de la Cathépsine S a été défini comme le gène ayant la huitième meilleure corrélation avec l'IMC avec $\rho_S = 0,58$. Sur la base de leurs *a priori*, les biologistes n'ont pas à notre connaissance été intéressé par les sept relations précédentes. Il est important de souligner que ces huit premières « meilleures » relations présentées toutes une valeur tel que $|\rho_S| < 0,66$ et un effectif tel que $n < 95\%$ (où n est l'effectif total) soit 37 individus. Ainsi, au regard de notre guide d'interprétation, CTSS n'aurait pas été étudié en « priorité » dans le cadre d'une exploration systématique sans *a priori* des résultats. Avec notre approche sans *a priori* CTSS se positionne en $447^{ème}$ position avec $\rho_S = 0,58$, $p \approx 0,00$, $q = 0,14$ et $n = 32$. CTSS est ainsi précédé de résultats qui semblent être plus intéressants d'un point de vue statistique.

Cette expérimentation nous a donné la possibilité de constater que notre approche automatique permet d'extraire des résultats qui peuvent être pertinents pour le biologiste lorsqu'il filtre en amont et qu'il ignore des découvertes qui peuvent être potentiellement plus intéressantes du point de vue biomédicale. Bien que ceci ne soit pas dans les objectifs initiaux de DISCOCLINI, celui-ci peut ainsi être utilisé avec profit quand l'expert a une idée plus ou moins précise de ce qu'il cherche à découvrir ou à valider.

8.4 Conclusions

Les différentes expérimentations que nous avons réalisées de DISCOCLINI sur des données issues de la Recherche Clinique nous ont permis d'assister les experts dans leurs découvertes publiées [Viguerie *et al.* 2004, Clément *et al.* 2004, Taleb *et al.* 2005, Benis et Courtine 2009b, Benis et Courtine 2009a]. Ces même expérimentations nous ont permis de valider les différents éléments composants de notre méthodologie. Ainsi, nous avons montré l'intérêt de réaliser le calcul sans *a priori* de toutes les valeurs de corrélations possibles entre deux ensembles de données d'expression génique et de données biocliniques. Puis nous avons empiriquement monter l'intérêt pour l'expert d'une représentation graphique lui permettant de naviguer dans les résultats tout en mettant en évidence les plus « intéressants ». Cette même restitution des résultats d'une étude corrélationnelle s'appuyant sur les guides d'interprétation de valeurs de corrélation, que nous proposons pour les applications en Génomique Médicale Fonctionnelle. Enfin les expérimentations que nous avons réalisées de PAMOUT nous ont permis de valider cette approche intégrée à DISCOCLINI.

Chapitre 9

Utilisabilié et usages de DiscoClini

Évaluer l'utilisabilité et l'usage aussi bien d'une méthode que d'un système est un processus important qui a pour objectif de montrer les avantages et les inconvénients aussi bien aux niveaux conceptuelles que de l'ingénierie applicative. Cette étape importante peut être réalisée à la fois tout au long mais aussi à la fin du processus de développement d'un produit. Dans ce chapitre, nous traitons des évaluations de l'utilisabilité et de l'usage de DiscoClini qui ont été réalisées aussi bien en cours de développement, mais aussi que sur une version finalisée du produit.

9.1 Utilisabilité d'un système

L'utilisabilité telle qu'elle est définie par la norme ISO 9241 [International Organization for Standardization 2000] est « *le degré selon lequel un produit peut être utilisé, par des utilisateurs définis et identifiés, afin d'atteindre des buts définis avec efficacité, fiabilité et satisfaction, dans un contexte d'utilisation spécifié* ».

9.1.1 Critères d'utilisablité

Cette notion d'utilisabilité intègre des concepts plus larges tels que l'ergonomie des Interfaces Homme-Machine [Shneiderman et Plaisant 2004] et le Facteur Humain [Dejours 2005]. L'utilisabilité est défini, selon Nielsen [Nielsen 1994, Costabile 2001] suivant cinq caractéristiques majeures d'un système utilisable, qui sont les suivants :

- l'efficacité signifie que l'application permet d'atteindre le résultat prévu ;
- la fiabilité signifie que l'application permet d'atteindre le résultat avec un effort moindre ou requiert un temps minimal ;
- la satisfaction, qui est en soit une notion subjective, signifie que l'application offre un confort (évaluation subjective de l'interaction pour l'utilisateur) à l'utilisation ;
- la facilité d'apprentissage signifie que l'utilisateur acquiert par la pratique et/ou un enseignement les connaissances nécessaires à la mise en œuvre de l'application ;
- la facilité d'appropriation signifie que l'utilisateur considère l'application comme un élément à part entière de son environnement.

Une méthode, un système ou un processus peuvent respecter une majorité voir tous les critères d'utilisabilité mais être inutile. L'objectif majeur du développement de l'une de ces approches est l'adéquation entre l'activité et l'outil (la méthode, le système ou le processus) qui permettra de dire que ce dernier est utile. Afin de répondre à cette contrainte, il est nécessaire que des

« utilisateurs identifiés » existent, qu'ils aient un « but défini » dans un « contexte d'utilisation spécifié ».

9.1.2 Formulaire d'utilisabilité

Afin de réaliser l'évaluation aussi bien de l'utilisabilité que des usages possibles de DiscoClini, nous nous sommes appuyés sur l'*USE Questionnaire* proposé par [Lund 2002] (voir figure 9.1). Ce questionnaire permet initialement de collecter, *via* 30 questions, l'avis des utilisateurs concernant les 5 caractéristiques définies par Nielsen [Nielsen 1994]. Les réponses peuvent être rendues sous la forme d'un score (c'est-à-dire une « note ») soit d'une réponse en langage naturel (qui peut, par exemple, avoir été obtenue au cours d'un entretien entre un *utilisateur* et un *chargé d'évaluation*. Ce test se décompose en 4 rubriques :

1. La première est relative à l'utilité, c'est-à-dire si la méthode, le système ou le processus trouve une application dans un contexte donné pour un utilisateur défini.
2. La seconde concerne la facilité d'utilisation de la méthode, du système ou du processus évalué. Il s'agit, ici, de considérer si ce qui est évalué est utilisable par l'utilisateur dans son environnement et avec son Savoir.
3. La troisième partie de l'*USE Questionnaire* (et *in extenso* de son adaptation) s'intéresse aux facilités d'apprentissage, d'adaptation et d'appropriation de l'objet du test, par ceux qui ont été, sont ou seront amenés à l'utiliser. Ainsi, cette partie permet de savoir si pour mettre en œuvre la méthode, le système ou le processus concerné, l'apprentissage et/ou l'intégration au processus de travail (plus généralement, la réalisation d'une activité complexe) ont été plus ou moins faciles et plus ou moins rapides.
4. La dernière partie du questionnaire permet d'évaluer la satisfaction des utilisateurs, c'est-à-dire qu'il est question de savoir si la méthode, le système ou le processus qu'ils ont eu à mettre en œuvre, pour ce test, comble un besoin de manière suffisante au point de vouloir l'intégrer pleinement à leur activité et/ou de le recommander à leur communauté d'intérêt (professionnelle ou non).

De manière générale, on peut observer que ce questionnaire ne s'intéresse pas aux aspects techniques de la méthode, du système ou du processus évalué. Il est orienté dans son ensemble vers les besoins de l'utilisateur et ses demandes.

9.2 Évaluation de DiscoClini

L'évaluation de DiscoClini a été réalisée en continu au cours de son développement (théorique et pratique) auprès de 6 « génomiciens ». D'autre part, une évaluation ponctuelle de l'Interface Homme-Machine et du flux de données qui lui est associé a été réalisée auprès d'un public hétérogène (c'est-à-dire 23 personnes dont des biologistes génomiciens et non génomiciens et des non biologistes), comme le montre la figure 9.2. La répartition des évaluateurs en fonction de leur âge permet de voir qu'ils ont entre 20 et 40 ans (voir figure 9.3). Il est important de noter que tous sont familiers des outils d'analyse de données. Ces évaluations ont permis de collecter des commentaires pertinents sur notre système.

9.3 Résultats des évaluations

Nous synthétisons, en nous appuyant sur la structuration de l'*USE questionnaire*, les résultats obtenus (voir figure 9.4) par une analyse *a posteriori* des commentaires faits par les évaluateurs :

DiscoClini:

Questionnaire for User Interface Satisfaction

Questionnaire built on Nielson recommendation [Nielson 1994, Lund 2002].

USEFULNESS	1	2	3	4	5	6	7	NA
1. It helps me be more effective.	Disagree						Agree	
2. It helps me be more productive.	Disagree						Agree	
3. It is useful.	Disagree						Agree	
4. It gives me more control over the activities in my life.	Disagree						Agree	
5. It makes the things I want to accomplish easier to get done.	Disagree						Agree	
6. It saves me time when I use it.	Disagree						Agree	
7. It meets my needs.	Disagree						Agree	
8. It does everything I would expect it to do.	Disagree						Agree	

EASE OF USE	1	2	3	4	5	6	7	NA
9. It is easy to use.	Disagree						Agree	
10. It is simple to use.	Disagree						Agree	
11. It is user friendly.	Disagree						Agree	
12. It requires the fewest steps possible to accomplish what I want to do with it.	Disagree						Agree	
13. It is flexible.	Disagree						Agree	
14. Using it is effortless.	Disagree						Agree	
15. I can use it without written instructions.	Disagree						Agree	
16. I don't notice any inconsistencies as I use it.	Disagree						Agree	
17. Both occasional and regular users would like it.	Disagree						Agree	
18. I can recover from mistakes quickly and easily.	Disagree						Agree	
19. I can use it successfully every time.	Disagree						Agree	

EASE OF LEARNING	1	2	3	4	5	6	7	NA
20. I learned to use it quickly.	Disagree						Agree	
21. I easily remember how to use it.	Disagree						Agree	
22. It is easy to learn to use it.	Disagree						Agree	
23. I quickly became skillful with it.	Disagree						Agree	

SATISFACTION	1	2	3	4	5	6	7	NA
24. I am satisfied with it.	Disagree						Agree	
25. I would recommend it to a friend.	Disagree						Agree	
26. It is fun to use.	Disagree						Agree	
27. It works the way I want it to work.	Disagree						Agree	
28. It is wonderful.	Disagree						Agree	
29. I feel I need to have it.	Disagree						Agree	
30. It is pleasant to use.	Disagree						Agree	

	1	2	3	4	5	6	7	NA

Enter the main **negative** aspects:

1.

2.

3.

Enter the main **positive** aspects:

1.

2.

3.

Comments on questions:

FIGURE 9.1 – USE questionnaire [Lund 2002].

Domaine principal de compétence	Nombre d'évaluations
Genomique/Biotechnologies	6 (+6)
Médecine/Pharmacie	8
Finance/Mercatique/Droit	5
Informatique/Multimédia	2
Statistiques/Mathématiques	2
	23 (+6)

Figure 9.2 – Répartition du nombre d'utilisateurs par domaine de compétence.

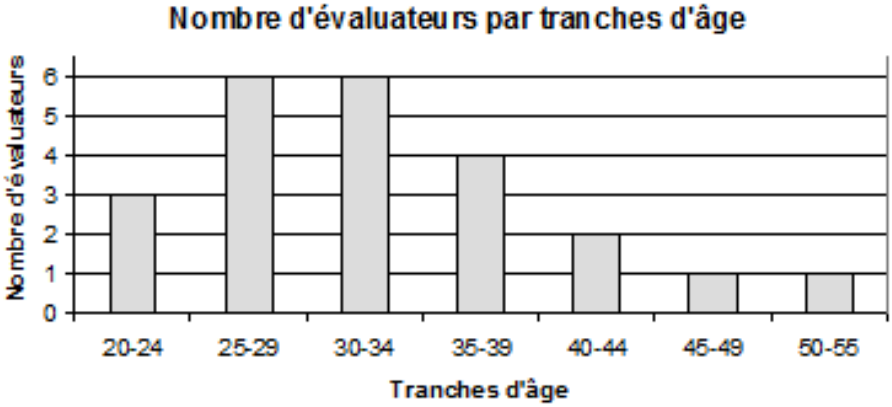


Figure 9.3 – Répartion des évaluateurs par tranches d'âge.

– des utilisateurs et des récipiendaires des résultats de la mise en œuvre de toute ou partie de DiscoClini (la majorité de ces résultats ont été obtenus au cours de discussions informelles et les questionnaires complétés par la suite),
– des biologistes et des non-biologistes qui ont eu une démonstration et des explications détaillées relatives à DiscoClini.

Nous traitons simultanément les aspects relatifs à ces deux aspects qui sont globalement indissociables.

	Nombre d'évaluations
Questionnaires complets	18
Questionnaires incomplets ou faits oralement au cours de la construction de l'outil	5 (+6)

Figure 9.4 – Répartion du nombre de formulaires complets/incomplets.

9.3.1 Utilité

L'*utilité* de DiscoClini a été prouvé de manière indirecte par la publication de résultats de validations biologiques et médicales issues de son utilisation [Viguerie *et al.* 2004, Clément *et al.* 2004, Taleb *et al.* 2005]. Ainsi, cette rubrique de l'*USE questionnaire* et *de facto* sujette à des opinions positives de la part de l'expert car ce dernier a été demandeur de résultats issus de la mise en œuvre de DiscoClini et a découvert aussi bien de nouvelles voies de recherche que des

biomarqueurs relatives à un état pathologique, comme nous l'avons développé dans le chapitre 8.

Aux éléments de l'*USE questionnaire* : « Il m'aide à être plus efficace », « Il m'aide à être plus productif », « Il est utile », « Il fait plus facilement les choses que je veux accomplir », « Il m'assiste quand je l'utiliser », « Il répond à mes besoins », les retours de l'expert (au sens générale du terme et tel que nous l'avons défini plus tôt dans ce manuscrit) sont globalement positifs, voir très positifs. En effet, en ne fournissant que deux fichiers de données, l'un comportant des données biocliniques et un second un ensemble de données d'expression génique, le système valide la structure de ces sources de données et la présence (à l'aide de l'identifiant de chaque individu) de données biocliniques et génomiques pour chacun. Les analyses réalisées par la suite de manière automatique ne souffrent (ou tout du moins le risque en est fortement réduit) pas de la mise en relation de données issues de deux populations différentes.

D'autre part, après la mise en relation automatique (l'éxecution des calculs « statistiques » et la recherche de valeurs suspectes), l'utilisateur de DiscoClini est informé par courrier électronique.

La présentation des résultats *via* une interface graphique simple (figure 7.8), avec des valeurs pré-sélectionnées pour les « filtres » statistiques présente un intérêt pour l'expert du domaine. La présentation des résultats potentiellement pertinents regroupés au sein d'un diagramme de Hasse permet un aperçu rapide du nombre de relations pouvant être explorées avec intérêt. La seule limitation de ce diagramme a été noté par l'ensemble des personnes auxquelles DiscoClini a été présenté, comme n'étant pas facilement compréhensible sans explications à un utilisateur non initié pour cette représentation graphique. Cette difficulté réside essentiellement dans la compréhension des liens entre les nœuds du diagramme.

Enfin, l'utilité de la visualisation en parallèle de relations (figure 7.11) a présenté pour les experts et les non experts un grand intérêt. En effet, chacun de ces types d'utilisateurs a souligné le gain de temps apportée par la visualisation simultanée de plusieurs relations dans un même référentiel visuel (tableau) et graphique (gradient de couleurs). Les liens, associés aux identifiants des gènes, permettant d'obtenir plus d'informations biologiques et médicales sur ces gènes ont été d'autre part considérés comme un plus qui réduit le temps de recherche d'informations complémentaires *a posteriori*. Ils permettent d'avoir pour les experts une idée précise du rôle des gènes découverts.

PAMout est un composant de la méthode DiscoClini à part entière dans le système DiscoClini ; néanmoins il est important de souligner que PAMout permet aux experts d'exclure de leurs expérimentations des données d'expression génique issues de puces à ADNc « défectueuses ». La seule limitation que nous avons pu observer s'inscrit dans le cadre de la résitution des résultats issus de la mise en œuvre de PAMout. En effet, l'utilisation d'une notation scolaire anglosaxone (valeurs de A à F) pour définir la présence plus ou moins importante de relations incluant des valeurs singulières n'est pas immédiatement comprise par les évaluateurs évoluant dans un environnement francophone.

Pour résumé, l'utilité de DiscoClini a été démontré aussi bien par des publications de résultats issus de sa mise en œuvre pour tout ou partie dans le domaine de la recherche sur les Obésités et d'autre part par les commentaires généralement positifs obtenus concernant ces différentes fonctionnalités (voir figure 9.5) par un public à la fois de biologistes/génomiciens et de non-experts du domaine utilisant déjà des méthodes d'analyses de données.

9.3.2 Facilité d'utilisation

La *facilité d'utilisation* de DiscoClini n'a pas été évaluée par les experts du domaine au sens strict du terme. En effet, ils n'ont pas bénéficié de l'implémentation complète du système au cours de leurs travaux. Néanmoins, les commentaires relatifs à l'*utilité* de DiscoClini peuvent être étendus à son utilisabilité en tant que méthode qui a permis les productions et l'utilisation de résultats scientifiques valorisées et valorisables. Ainsi, avec les versions antérieures, les utilisateurs ont eu à fournir les fichiers sources pour par la suite recevoir des fichiers de résultats classés en fonction de valeurs de seuil d'intérêts définies dans les guides d'interprétations que nous avons proposés. La différence avec la version actuelle réside dans le fait que tout le flux de données est « en ligne ».

Les commentaires sur la facilité d'utilisation de DiscoClini dans sa version finale ont essentiellement été liés à la prise en main du diagramme de Hasse. Cependant, l'utilisateur (expert du domaine d'application ou non) apprécie, une fois qu'il a compris le mode de lecture de cette représentation graphique, la manière selon laquelle l'information est représentée et synthétisée. D'autre part, les non-experts du domaine d'application, c'est-à-dire, les « utilisateurs-testeurs » non biologistes y ont vu de potentielles applications dans leurs domaines respectifs (la finance et la mercatique, principalement) (voir figure 9.5).

9.3.3 Facilité d'apprentissage et d'appropriation

Les résultats de l'évaluation des *facilités d'apprentissage, d'adaptation et d'appropriation* de DiscoClini (voir figure 9.5) découlent des étapes précédentes de l'évaluation (c'est-à-dire de celles relatives à l'*utilité* et à la *facilité d'utilisation*). Les deux principales parties du système qui demandent un apprentissage sont le diagramme de Hasse et la visualisation des relations *via* les langages L_V et L_S. Comme nous l'avons expliqué précédemment avec des explications « minimales », tout utilisateur de « DiscoClini » l'utilise de manière autonome mais tout utilisateur a besoin d'une formation minimale pour pouvoir le faire.

Les experts du domaine se sont indirectement appropriés DiscoClini *via* l'utilisation au sein de publications de ce système, même si les articles ne font pas toujours référence au système en tant que tel. La seule limite que nous avons constaté dans l'adaptation à DiscoClini est le besoin constant des « génomiciens » avec lesquels nous avons travaillés d'appliquer des filtres en amont sur les données, ce qui conduit à une réduction significative du nombre de découvertes potentiellement « surprenantes » qui peuvent être faites.

9.3.4 Satisfaction

Les utilisateurs de DiscoClini, qu'ils soient des biologistes experts du domaine d'application (des génomiciens) ou qu'ils soient des chercheurs et/ou des professionnels issus d'autres domaines, ont été dans l'ensemble satisfait avec des résultats favorables compris entre 80 et 100% pour chaque section du questionnaire USE. Seul, un des deux mathématiciens/statisticiens n'a pas été pleinement satisfait de DiscoClini car le système ne fournit pas de résultats sous la forme d'équations complexes. Ceci implique le taux de satisfaction moyen pour le groupe « mathématiciens/Statisticiens » (voir figure 9.5).

Globalement, la publication de résultats, la modification d'ensembles de données étudiées et la demande d'une ouverture à d'autres domaines montrent la satisfaction des utilisateurs pour DiscoClini en tant que système automatique d'aide à la découverte de relations plus ou moins complexes dans de très grandes ensembles de données.

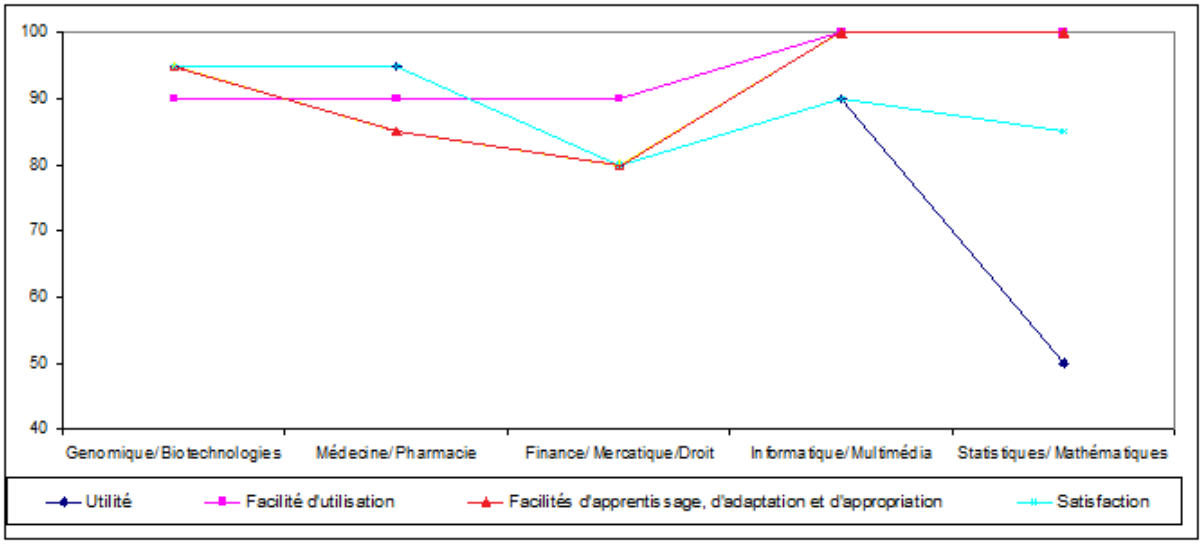


FIGURE 9.5 – Synthèse graphique des résultats de l'évaluation.

9.4 Conclusion

L'évaluation de l'usage et de l'utilisabilité de DISCOCLINI nous a permis de constater que le système présente un intérêt important en Génomique Médicale Fonctionnelle. Bien que le diagramme de Hasse comme modalité de représentation visuelle des groupes de relations potentiellement intéressantes nécessite quelques explications auprès des utilisateurs, ce diagramme une fois compris est apprécié comme outil de navigation dans les nombreux résultats disponibles. La visualisation en parallèle des relations définies comme « intéressantes » a été au cours de l'évaluation « en continu » comme un des atouts majeurs de la méthode et a été utilisée avec intérêt par les biologistes [Viguerie *et al.* 2004].

Conclusion et perspectives

Bilan

L'objectif principal de cette thèse était de proposer un flux de données et un système de visualisation permettant de faciliter la découverte de relations dans des données de Génomique Médicale Fonctionnelle. Ces données ont la particularité de contenir un nombre réduit de réplicats (d'individus) et beaucoup de descripteurs (attributs). Les données que nous avons traitées sont des données génomiques issues de l'expression génique à un instant donné grâce à des puces à ADNc et des données biocliniques issues d'analyses de laboratoire et de la pratique clinique Notre but a été de mettre en relation ces deux types de données afin d'aider à la découverte de biomarqueurs d'états physiologiques particuliers pour une meilleure compréhension de pathologies complexes, comme l'Obésité. Ainsi, nous avons abordé cette mise en relation grâce à une approche statistique et nous avons proposé dans ce cadre deux méthodologies. La première, globale, permet d'avoir une idée générale des relations pouvant exister entre l'expression génique et des paramètres biocliniques. Elle contribue à l'aide à la découverte de biomarqueurs pouvant être utilisés pour une population « large ». La seconde est locale et permet au sein des mêmes données de rechercher à l'aide d'un algorithme de fenêtrage des biomarqueurs utilisables auprès de sous-populations présentant des comportements définissables dans un intervalle restreint de valeurs.

Nous avons présenté notre système DiscoClini. Dans un premier temps, il calcule sans *a priori* toutes les corrélations possibles entre les données d'expression et les données biocliniques, et il détecte les éventuelles valeurs singulières apparaissant dans les données. Les résultats sont ensuite explorés de manière *guidée* avec des *a priori* réduits au minimum et n'incluant pas de Connaissance du domaine d'application. L'étape de fouille des résultats est structurée de manière à faire *interagir* l'expert avec le système en lui proposant des représentations graphiques simples, que ce soit pour visualiser ou pour comparer les relations. De plus, l'utilisateur est alerté de la présence de valeurs singulières dans les relations, ce qui lui permet de les prendre en compte dans son analyse et ainsi de réduire le risque de fausses découvertes. De plus, DiscoClini offre des liens directs vers des données externes, ce qui permet à l'expert de répondre rapidement à ces questions concernant un gène. Enfin, la génération d'un rapport d'exploration des données synthétisant les résultats les plus intéressants, permet à l'expert d'approfondir son travail hors ligne.

De manière globale, notre approche dans son ensemble a permis aux biologistes de réaliser (ou de valider) des découvertes en Génomique Fonctionnelle des Obésités. DiscoClini [Benis et Courtine 2009a, Benis et Courtine 2009b] a contribuer à la compréhension de mécanismes physiologiques [Viguerie *et al.* 2004, Clément *et al.* 2004] et à la définition de biomarqueurs [Taleb *et al.* 2005].

L'évaluation de l'utilisabilité de DiscoClini par des utilisateurs experts et non-experts a permis de valider le système aussi bien d'un point de vue de son usage applicatif que d'un point

de vue cognitif (compréhensibilité du système).

Perspectives

Les perspectives de nos travaux peuvent s'inscrire dans trois domaines : l'adaptation du système DiscoClini à l'utilisateur, la définition de relations complexes et la gestion des ressources de recherche clinique.

Le système DiscoClini

Une des perspectives de recherche que nous proposons est celui de l'amélioration de DiscoClini en terme d'interactivité avec l'utilisateur [Cohn *et al.* 1996, Moskovitch *et al.* 2008]. Ainsi, ces améliorations peuvent s'inscrivent dans trois directions.

La première consiste à permettre au système de s'adapter automatiquement aux paramétrages effectués par chaque utilisateur. Plus un utilisateur travaillera avec le système DiscoClini, plus le système s'adaptera aux valeurs de paramétrages effectués. De cette manière lorsqu'un utilisateur se connectera au système, il serait identifié et les valeurs des paramètres de l'étape d'exploration et de visualisation seront chargées de manière spécifique ainsi que régulièrement ajustées en fonction de ses comportements face au système.

La seconde direction a pour objectif de permettre la prise en compte des valeurs singulières détectées par PAMout dès l'étape d'exploration des résultats. Les résultats les plus intéressants même incluant des valeurs singulières seraient mis en avant par rapport à des résultats intéressants mais n'incluant pas de valeurs singulières. L'objectif, ici, est de s'appuyer sur un « coefficient de confiance » pour chaque valeur singulière détectée.

La troisième perspective d'amélioration du système DiscoClini correspond à une personnalisation du rapport d'exploration des résultats qui permettra à l'expert de sélectionner aussi bien les informations qu'il souhaite y voir apparaître, leur ordonnancement et leur format d'export. Cela donnera la possibilité d'une réutilisation des résultats issus de DiscoClini dans d'autres systèmes ou dans le cadre de publications.

Enfin, la quatrième perspective de notre système serait de le rendre suffisamment générique pour pouvoir le rendre applicable à d'autres domaines. En effet, lors des tests d'usages et d'utilisabilités, nous nous sommes rendus compte que les experts non-biologistes trouvaient l'outil pertinent dans leur cas d'études.

Par exemple, dans le cadre de la Mercatique, DiscoClini pourrait être utilisé pour étudier les comportements de groupes d'utilisateurs en fonction des produits qu'ils auraient déjà consommés et afin de proposer des produits qu'ils sont susceptibles de consommer par la suite. Dans le cadre de l'accidentologie, DiscoClini pourrait être utilisé pour faciliter la découverte de relations entre les causes et/ou les conséquences de différents accidents sur la base d'un très grands nombre de paramètres tels que le type d'accident, la description des protagonistes (victimes ou non), leurs états physiques et psychologiques, ... Dans les Sciences de l'Environnement, DiscoClini pourrait permettre de faire facilement de découvrir des relations entre les différents paramètres permettant de caractériser les écosystèmes. Le flux de fouille de données utilisé dans DiscoClini doit être adapté à chacun de ces cas avant de pouvoir être utilisé, car dans chacun de ces cas, les données utilisées ont des caractéristiques qui leur sont propres et les méthodes à appliquer ne sont donc pas les mêmes. Cependant, il est intéressant de noter que cette adaptation est simple et offre de nombreuses nouvelles voies d'applications à notre système et ce en ne changeant pas les outils de visualisation mis en œuvre.

Définition des relations complexes

L'utilisation du coefficient de corrélation permet de rechercher et de définir des relations intéressantes de manière approximative mais efficacement et rapidement. Une des perspectives de ce travail est liée à l'amélioration des résultats de nos analyses par la prise en compte de relations plus complexes entre des données d'expression génique et les données biocliniques. Comme nous l'avons noté, la relation entre deux ensembles de données peut être définie par une fonction mathématique. Ce type d'approche va permettre de décrire des relations plus complexes [Lullman et Mohr 2003, Simon 2006], mais elle requiert un nombre important de données, donc au moins plusieurs dizaines, voir plusieurs centaines d'individus pour que le résultat soit le plus précis possible. Hors, aujourd'hui le nombre d'individus inclus dans un protocole de recherche clinique reste encore insuffisant pour que cette approche soit viable. Mais le nombre d'individus inclus dans les protocoles étant de plus en plus nombreux, nous pouvons espérer que cela soit possible d'ici quelques années.

De plus, la définition de fonctions précises pour décrire les données impliquent la mise en place de méthodes de visualisation adaptées (par exemple, la schématisation de la courbe) afin de faciliter l'exploration des résultats qui seront obtenus. Avoir des fonctions plus complexes implique l'utilisation de méthodes de représentation et/ou de visualisation simplificatrices, car comme nous l'avons déjà mentionné, il est difficile de se représenter mentalement une telle fonction arithmétique complexe.

Gestion des ressources de Recherche Clinique

Les problèmes soulevés par l'hétérogénéité et la multiplicité des sources de données est un problème essentiel. En effet, les connaissances requises pour faire d'une relation intéressante une découverte sont nombreuses et complexes. Cela suggère le besoin de proposer un modèle standard et uniformisé pour la structuration des données sources. Un tel modèle doit permettre la prise en compte de l'accroissement du volume et de l'évolution constante des connaissances dans les différentes sources de données.

Ce problème soulève lui-même un autre problème relatif aux sources de données elles-même. Il existe des modèles de Systèmes d'Information pour des données issues d'expérimentations mettant en œuvre des biopuces, pour des données d'annotations géniques et génomiques, pour des données biologiques, cliniques, sociologiques, économiques, épidémiologiques... mais l'ensemble de ces systèmes sont aujourd'hui distincts. Ainsi, il n'existe pas de méthodologie proposant un modèle de Système d'Information Informatisé permettant la gestion totale (et unifiée) de protocoles de recherche clinique. Nous avons réalisé parallèlement à cette thèse des travaux (non présentés ici) et qui traitent de cette problématique [Benis 2003, Benis *et al.* 2003e, Benis *et al.* 2003f]. Ces travaux nous semblent essentiels dans le monde actuel où toute donnée est source d'information et vice versa. L'implantation et la mise en œuvre de telles approches permettraient des collaborations interdisciplinaires plus efficaces [Mondada 2005] et d'envisager l'automatisation de bout en bout du processus d'aide à la *Découverte* en se basant sur des quantités d'informations beaucoup plus importantes et ayant une meilleur qualitée.

Néanmoins, la mise en œuvre technique de ce type d'environnement nécessite le respect d'un cadre juridique particulier et strict [Code de la Santé Publique 2008, Code de la propriété intellectuelle 2008]. La réglementation est telle que les données collectées et traitées au cours d'un protocole de recherche clinique doivent être gérées suivants des règles de *Qualité Totale* [International Organization for Standardization 2007a, International Organization for Standardization 2007b], ce qui implique *in extenso* une gestion centralisée des ressources, et qui inclut la traça-

bilité de toutes les opérations effectuées dans le système.

Une autre perspective de nos travaux est liée à la définition d'un système comme celui que nous venons de décrire. Ce type de système devrait nous permettre d'étendre notre approche à la recherche de corrélations entre des données d'expression génique et d'autres biologiques/non biologiques (SNP, QTL, données épidémiologiques...). Cela permettrait, d'un point de vue biomédical, par exemple, de savoir s'il existe un lien entre le niveau d'expression d'un gène et une mutation génique et ainsi de pouvoir généraliser les résultats à une sous-population beaucoup mieux décrites en terme générique.

Bibliographie

[Abdi 2007] H. Abdi. *Encyclopedia of Measurement and Statistics*, chapter Bonferroni and Sidak corrections for multiple comparisons. Thousand Oaks, CA : Sage, 2007.

[Abou 2002] C. Abou. Les biopuces. Technical report, Dossier du CEA, 2002.

[Adamchik 1997] V. Adamchik. On stirling numbers and euler sums. *Journal of Computational and Applied Mathematics*, (79) : 119 ?130, 1997.

[Afifi et Azen 1979] AA. Afifi et SP. Azen. *Statistical analysis : A computer oriented approach.* Academic Press, New York, 1979.

[Agrawal *et al.* 1998] R. Agrawal, J. Gehrke, D. Gunopulos et P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In LM. Haas et A. Tiwary, editors, *SIGMOD 1998, Proceedings ACM SIGMOD International Conference on Management of Data, June 2-4, 1998, Seattle, Washington, USA*, pages 94–105. ACM Press, 1998.

[Al-Shahrour *et al.* 2004] F. Al-Shahrour, R. Diaz-Uriarte et J. Dopazo. Fatigo : a web tool for finding significant associations of gene ontology terms with groups of genes. *Bioinformatics*, 20(4) : 578–580, March 2004.

[Ancelle 2002] T. Ancelle. *Statistique épdidémiologie.* Maloine, 2002.

[Ankerst *et al.* 1999] M. Ankerst, MM. Breunig, HP. Kriegel et J. Sander. Optics : ordering points to identify the clustering structure. In *Proc. ACM SIGMOD 99 Int. Conf. on Management of Data*, 1999.

[Arredondo-Vega *et al.* 1998] F.X. Arredondo-Vega, I. Santisteban, S. Daniels, S. Toutain et M.S. Hershfield. Adenosine deaminase deficiency : genotype-phenotype correlations based on expressed activity of 29 mutant alleles. *Am J Hum Genet*, 63(4) : 1049–1059, 1998.

[Ashrafi *et al.* 2006] GH. Ashrafi, DR. Brown, KH. Fife et MS. Campo. Down-regulation of mhc class i is a property common to papillomavirus e5 proteins. *Virus Res.*, 120 : 208–211, September 2006.

[Azaïs et Bardet 2006] J.M. Azaïs et J.M. Bardet. *Le modèle linéaire par l'exemple - Régressions, analyse de la variance et plans d'expériences illustrées avec R, SAS, et Splus.* Dunod, 2006.

[Bacon 1620] F. Bacon. *Novum Organum, Livre I, 95, Chapitre La fourmi, l'araignée, l'abeille.* 1620.

[Bagaria et Bagaria 2007] V. Bagaria et S. Bagaria. A geographic information system to study trauma epidemiology in india. *J Trauma Manag Outcomes.*, 26 : 1–3, 2007.

[Balakrishnama et Ganapathiraju 1998] S. Balakrishnama et A. Ganapathiraju. Linear discriminant analysis - a brief tutorial. Technical report, Mississippi State University, Department of Electrical and Computer Engineering, Institute for Signal and Information Processing, 1998.

[Barnett et Lewis 1994] V. Barnett et T. Lewis. *Outliers in statistical data.* John Wiley, New York, 1994.

[Barrett et Edgar 2006] T. Barrett et R. Edgar. Mining microarray data at NCBI's gene expression omnibus (GEO)*. *Methods Mol Biol*, 338 : 175–190, 2006.

[Basdevant *et al.* 1993] A. Basdevant, M. le Barzic et B. Guy-Grand. *Les obésités*. Ardix Médical, 1993.

[Beddo et Kreuter 2004] V. Beddo et F. Kreuter. A handbook of statistical analyses using spss. *Journal of Statistical Software, Book Reviews*, 11(2) : 1–4, 6 2004.

[Ben-Gal 2005] I. Ben-Gal. *Data Mining and Knowledge Discovery Handbook : A Complete Guide for Practitioners and Researchers*, chapter Outlier detection. Kluwer Academic Publishers, 2005.

[Benis *et al.* 2003a] A. Benis, R. Cancello, C. Carette, M. Courtine, B. Hanczar, V. Pelloux, K. Clément et JD. Zucker. Combining gene expression and clinical data to predict response in obesity treatment. In *6th International Meeting of the Microarray Gene Expression Data Society (MGED 6)*, Aix-en-Provence, France. 2003.

[Benis *et al.* 2003b] A. Benis, R. Cancello, C. Carette, M. Courtine, B. Hanczar, V. Pelloux, K. Clément et JD. Zucker. Expression génique et données cliniques : prédire la réponse dans un traitement de l'obésité. In *Journée Jeunes Chercheurs IFR 58 - (Régulations et Communications Cellulaires)*, Paris, France. 2003.

[Benis *et al.* 2003c] A. Benis, R. Cancello, C. Carette, M. Courtine, B. Hanczar, V. Pelloux, K. Clément et JD. Zucker. Gene expression meets clinical practice. In *European Conference on Computational Biology 2003 (ECCB 2003)*, Paris, France. 2003.

[Benis *et al.* 2003d] A. Benis, C. Carette, R. Cancello, B. Hanczar, V. Pelloux, K. Clément et JD. Zucker. Introduction de données cliniques dans l'analyse des pucesà cDNA. In *21ème Réunion Scientifique de l' Association Française d'Etudes et de Recherches sur l'Obésité (AFERO)*, Paris, France. 2003.

[Benis *et al.* 2003e] A. Benis, A. Michaut, JD. Zucker, P. Barbe, A. Basdevant, O. Ziegler, M. Laville et K. Clément. ObMinder : Un entrepôt de données pour la recherche clinique sur les obésités. In *21ème Réunion Scientifique de l' Association Française d'Etudes et de Recherches sur l'Obésité (AFERO)*, Paris, France. 2003.

[Benis *et al.* 2003f] A. Benis, A. Michaut, JD. Zucker, M. Laville et K. Clément. Recherche clinique dans l'obésité : l'entrepôt de données " ObMinder ". In *Reunion 2003 des Centres de Ercherches en Nutrition Humaine*, France. 2003.

[Benis et Courtine 2009a] A. Benis et M. Courtine. Biomarker discovery in medical genomics data. In *International Conference on Bioinformatics & Computational Biology, BIOCOMP*, pages 265–272, Las Vegas, USA. CSREA Press, 2009.

[Benis et Courtine 2009b] A. Benis et M. Courtine. Un système pour l'extraction de corrélations linéaires dans des données de génomique médicale. In *Extraction et Gestion des Connaissances, EGC*, pages 467–468, Strasbourg, France. Cépaduès-Éditions, 2009.

[Benis 2003] A. Benis. Intégration de données cliniques pour la classification de données issues de puces à cDNA : Application à la génomique fonctionnelle de l'obésité. Dea en informatique médicale et technologies de la communication / master of sciences in medical informatics and communication technologies, Université Pierre et Marie Curie - Paris VI, 2003.

[Benis 2005] A. Benis. Categorizing gene expression correlations with bioclinical data : an abstraction based approach. In *Symposium of Abstraction, Reformulation and Approximation, SARA*, number 3607, pages 352–353. Lectures Notes of Computer Sciences, Springer Verlag, 2005.

[Benis 2007] A. Benis. Discoclini : Un environnement pour l'aide à la découverte de corrélations entre des données d'expression génique et des données biocliniques. In *Journée Francophone d'Informatique Médicale*, number CD, Mali. 2007.

[Benjamini *et al.* 2001] Y. Benjamini, D. Drai, G. Elmer, N. Kafkafi et I. Golani. Controlling the false discovery rate in behavior genetics research. *Behav Brain Res*, 125(1–2) : 279–84, 2001. 0166-4328 Comment Journal Article.

[Benjamini et Hochberg 1995] Y. Benjamini et Y. Hochberg. Controlling the false discovery rate : A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57 : 289–300, 1995.

[Benjamini et Yekuteli 2001] Y. Benjamini et D. Yekuteli. The control of the false discovery rate in multiple testing un dependency. *The Annals of Statistic*, 29(4) : 1165–1188, 2001.

[Berrar 2003] D.P. Berrar. Integration of microarray data for a comparative study of classifiers and identification of marker genes. In *The 4th International Conference on Critical Assessment of Microarray Data Analysis 2003 (CAMDA03)*, 2003.

[Bertrand et Garnier 2005] A. Bertrand et PH. Garnier. *Psychologie Cognitive*. Studyrama, 2005.

[Besson 2005] J. Besson. *Découverte de motifs pertinents pour l'analyse du transcriptome : applications à l'insulino-résistance*. Thèse de doctorat, National des Sciences Appliqués de Lyon, 2005.

[Beuscart *et al.* 2009] R. Beuscart, J. Bénichou, P. Roy et C. Quantin. *Biostatistique*. Omniscience, 2009.

[Bhardwaj et Lu 2005] N. Bhardwaj et H. Lu. Correlation between gene expression profiles and protein-protein interactions within and across genomes. *Bioinformatics*, 21(11) : 2730–2738, 2005.

[Brazma *et al.* 2001] A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. A. Ball, H. C. Causton, T. Gaasterland, P. Glenisson, F. C. Holstege, I. F. Kim, V. Markowitz, J. C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo et M. Vingron. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet*, 29(4) : 365–371, 2001.

[Brazma *et al.* 2003] A. Brazma, H. Parkinson, U. Sarkans, M. Shojatalab, J. Vilo, N. Abeygunawardena, E. Holloway, M. Kapushesky, P. Kemmeren, G.G. Lara, A. Oezcimen, P. Rocca-Serra et S.A. Sansone. Arrayexpress a public repository for microarray gene expression data at the ebi. *Nucleic Acids Research*, 31(1) : 68–71, 2003.

[Breton 2002] L. Breton. *GranuLab : un système d'aide à la découverte scientifique appliqué à la physique des milieux granulaires*. Thèse de doctorat, Université Paris 6, 24 Janvier 2002.

[Breunig *et al.* 2000] MM. Breunig, HP. Kriegel, RT. Ng et J. Sander. Lof : Identifying density-based local outliers. In *Proc. ACM SIGMOD 2000 Int. Conf. On Management of Data*, 2000.

[Brodley et Friedl 1996] C. Brodley et M Friedl. Identifying and eliminating mislabeled training instances. In *AAAI/IAAI, Vol. 1*, pages 799–805, 1996.

[Brunet 2002] O. Brunet. *Etude de la connaissance dans le cadre d'observations partielles : la logique de l'observation*. Thèse de doctorat, Université Joseph Fourrier, Grenoble, EXMO-HELIX, Octobre 2002.

[Burges 1998] C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2 : 121–167, 1998.

[Butte 2002] A. Butte. The use and analysis of microarray data. *Nat Rev Drug Discov*, 1(12) : 951–960, December 2002.

[Carletti 1989] G. Carletti. *Comparaison empirique de méthodes statistiques de détection de valeurs anormales à une et à plusieurs dimensions*. Thèse de doctorat, Fac. Univ. Sci. Agron., Gembloux, Belgique, 1989.

[Caruana et Niculescu-Mizil 2006] R. Caruana et A. Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23 rd International Conference on Machine Learning*, Pittsburgh, PA. 2006.

[Caussinus *et al.* 2003] H. Caussinus, M. Fekri, S. Hakam et Ruiz-Gazen A. A monitoring display of multivariate outliers. *Computational statistics & data analysis*, 44 : 237–252, 2003.

[Causton *et al.* 2003] H. Causton, A. Brazma et J. Quackenbush. *Microarray Gene Expression Data Analysis*. Wiley, John & Sons, 2003.

[Cecil *et al.* 2002] C.E. Cecil, R.H. Chiang et E. Lim. An intelligent middleware for linear correlation discovery. *Decis. Support Syst.*, 32(4) : 313–326, 2002.

[Chakravarti *et al.* 1967] I. Chakravarti, R. Laha et J. Roy. *Handbook of Methods of Applied Statistics*. John Wiley and Sons, 1967.

[Chiang *et al.* 2005] R.H. Chiang, C.E. Cecil et E. Lim. Linear correlation discovery in databases : a data mining approach. *Data and Knowledge Engineering*, 53(3) : 311–337, 2005.

[Christopher 2004] P. Christopher. Their U.S. military intervention decision-making process : who participates, and how ? *Journal of Political and Military Sociology*, 32(1) : 19–43, 2004.

[Cios *et al.* 2007] Krzysztof J. Cios, Witold Pedrycz, Roman W. Swiniarski et Lukasz A. Kurgan. *Data Mining : A Knowledge Discovery Approach*. Springer, 1 édition, February 2007.

[Clément *et al.* 2002] K. Clément, P. Boutin et P. Froguel. Genetics of obesity. *Am J Pharmacogenomics*, 2(3) : 177–187, 2002.

[Clément *et al.* 2004] K. Clément, N. Viguerie, C. Poitou, C. Carette, V. Pelloux, C.A. Curat, A. Sicard, S. Rome, A. Benis, J.D. Zucker, H. Vidal, M. Laville, G.S. Barsh, A. Basdevant, V. Stich, R. Cancello et D. Langin. Weight loss regulates inflammation-related genes in white adipose tissue of obese subjects. *FASEB J*, 18(14) : 1657–1669, November 2004.

[Clément et Ferre 2003] K. Clément et P. Ferre. Genetics and the pathophysiology of obesity. *Pediatr Res*, 53(5) : 721–725, 2003.

[Cleveland 1979] W. S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.*, 74 : 829–836, 1979.

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

[Code de la propriété intellectuelle 2008] Code de la propriété intellectuelle. *Code de la propriété intellectuelle*. Journal officiel de la République française, 2008.

[Code de la Santé Publique 2008] Code de la Santé Publique. *Code de la Santé Publique*. Journal officiel de la République française, 2008.

[Cohen 1988] J. Cohen. *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates Inc, US, 1988.

[Cohn *et al.* 1996] D. Cohn, Z. Ghahramani et M. Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4 : 129, 1996.

[Cook et Weisberg 1980] RD. Cook et S. Weisberg. Characterisations of an empirical influence function for detecting influential cases in regression. *Technometrics*, 22 : 495–508, 1980.

[Cornuéjols *et al.* 2002] A. Cornuéjols, L. Miclet et Y. Kodratoff. *Apprentissage artificiel : Concepts et algorithmes*. Editions Eyrolles, Paris, France, 2002.

[Costabile 2001] M. Costabile. Usability in the software life cycle. In *Handbook of Software Engineering and Knowledge Engineering*, pages 179–192. World Scientific Publishing, 2001.

[Cour des comptes 2009] de la République Française Cour des comptes . Rapport public annuel - la gestion du GIP dossier médical personnel. Technical report, Cour des comptes de la République Française, 2009.

[Courtine 2002] M. Courtine. *Regroupement conceptuel de données structurées pour la fouille de données*. Thèse de doctorat, Université Paris 6, 13 décembre 2002.

[Croux 2000] C. Croux. Outlier resistant estimators for canonical correlation analysis. In *Compstat 2000 - Proceedings in Computational Statistics : 14th Symposium Held in Utrecht*, pages 301–306, New York, LLC, USA. Springer-Verlag, 2000.

[Datta 2003] S. Datta. Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics*, 19 : 459–466, 2003.

[de Lichtenberg *et al.* 2005] U. de Lichtenberg, L.J. Jensen, S. Brunak et P. Bork. Dynamic complex formation during the yeast cell cycle. *Science*, 307(5710) : 724–727, 2005.

[de Rosnay 1975] J. de Rosnay. *Le macroscope*. Editions du Seuil, 1975.

[Debouk et Goodfellow 1999] C. Debouk et P. Goodfellow. Dna microarrays in drug discovery and development. *Nature Genetics*, 1 : 48–50, 1999.

[Degoulet et Fieschi 1998] P. Degoulet et M. Fieschi. *Informatique Médicale*. Masson, 3 édition, 02 1998.

[Dejours 2005] C. Dejours. *Le facteur humain*. Presses Universitaires de France, 2005.

[Diehn *et al.* 2003] M. Diehn, G. Sherlock, G. Binkley, H. Jin, J.C. Matese, T. Hernandez-Boussard, Rees C.A., J.M. Cherry, D. Botstein, P.O. Brown et A.A. Alizadeh. SOURCE : a unified genomic resource of functional annotations, ontologies, and gene expression data. *Nucleic Acids Research*, 31 : 219–223, 2003.

[Dodge et Rousson 1999] Y. Dodge et V. Rousson. *Analyse de régression appliquée*. Dunod, 1999.

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

[Dudoit et Shaffer 2003] S. Dudoit et C.B.J. Shaffer. Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18,1 : 71–103, 2003.

[Dudoit et van der Laan 2007] S. Dudoit et MJ. van der Laan. *Multiple Testing Procedures and Applications to Genomics*. Series in Statistics. Springer, 2007.

[Efron 2006a] B. Efron. Correlation and large-scale simultaneous significance testing. http ://www-stat.stanford.edu/ ckirby/brad/papers/, 2006.

[Efron 2006b] B. Efron. Local false discovery rates. http ://www-stat.stanford.edu/ ckirby/brad/papers/, 2006.

[Efron 2006c] B. Efron. Size, power, and false discovery rates. http ://www-stat.stanford.edu/ ckirby/brad/papers/, 2006.

[Eisen *et al.* 1995] M.B. Eisen, P.T. Spellman, P.O. Brown et D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, 95 : 14863–14868, 1995.

[Eisen 1999] M.B. Eisen. Cluster and treeview - manual, 1999.

[Eisensmith *et al.* 1996] R.C. Eisensmith, D.R. Martinez, A.I. Kuzmin, A.A. Goltsov, A. Brown, R. Singh, L.J.II Elsas et S.L. Woo. Molecular basis of phenylketonuria and a correlation between genotype and phenotype in a heterogeneous southeastern US population. *Pediatrics*, 97(4) : 521–516, 1996.

[Engeli *et al.* 2003] S. Engeli, M. Feldpausch, K. Gorzelniak, F. Hartwig, U. Heintze, J. Janke, M Möhlig, A.F.M. Pfeiffer, F.C. Luft et A.M. Sharma. Association between adiponectin and mediators of inflammation in obese women. *Diabetes*, 52 : 942–947, 2003.

[Escalante 2005] H. Escalante. A comparison of outlier detection algorithms for machine learning. In *Proceedings of CIC-2005*, 2005.

[Ester *et al.* 1996] M. Ester, HP. Kriegel, J. Sander et X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*, 1996.

[Everitt 2002] BS. Everitt. *The Cambridge dictionary of statistics*. University Press, Cambridge, UK, 2nd édition, 2002.

[Fayyad *et al.* 1996] U.M. Fayyad, D. Haussler et P.E. Stolorz. KDD for science data analysis : Issues and examples. In *Knowledge Discovery and Data Mining*, pages 50–56, 1996.

[Forga *et al.* 2002] L. Forga, E. Petrina et J. Barberia. Complicaciones de la obesidad. *Anales del sistema sanitario de Navarra*, 25(Suppl 1) : 117–126, 2002.

[Frawley *et al.* 1992] W. Frawley, G. Piatetsky-Shapiro et C. Matheus. Knowledge discovery in databases : An overview. *AI Magazine*, 13(3) : 57–70, 1992.

[Friedman et Bitterer 2009] T. Friedman et A. Bitterer. De la qualité des données à la pertinence des décisions. World Wide Web electronic publication, 2009.

[Fu *et al.* 2005] R. Fu, D.K. Dey et K.E. Holsinger. Bayesian models for the analysis of genetic structure when populations are correlated. *Bioinformatics*, 21(8) : 1516–1529, 2005.

[Gene Ontology Consortium 2000] Gene Ontology Consortium. Gene Ontology : tool for the unification of biology. *Nature Genet.*, 25 : 25–29, 2000.

[Gene Ontology Consortium 2001] Gene Ontology Consortium. Creating the Gene Ontology resource : design and implementation. *Genome Research*, 11 : 1425–1433, 2001.

[Gene Ontology Consortium 2004] Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research*, 32 (Database issue) : D258–D261, 2004.

[Georgiev 2007] I. Georgiev. A mixture-distribution factor model for multivariate outliers. *Econometrics Journal*, 10(3) : 605–636, November 2007.

[Gibson et Muse 2003] F.G. Gibson et S. Muse. *Précis de génomique*. De Boeck, 2003.

[Giudici et Passerone 2002] P. Giudici et G. Passerone. Data mining of association structures to model consumer behaviour. *Computational statistics and data analysis*, 38(2) : 533–541, 2002.

[Gnanadesikan et Kettering 1972] R. Gnanadesikan et JR. Kettering. Robust estimates, residuals and outlier detection with multiresponse data. *Biometrics*, 28 : 81–124, 1972.

[Goldner et Messier 2002] M. Goldner et S. Messier. Introduction note on the correlation between the molecular and clinical aspects of infection. *The Canadian Journal of Infectious diseases*, 13(1) : 28–30, January/February 2002.

[Golub *et al.* 1999] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield et E.S. Lander. Molecular classification of cancer : class discovery and class prediction by gene expression monitoring. *Science*, 286(5439) : 531–537, 1999.

[Grawe et Mulligan 2002] N. Grawe et C. Mulligan. Economic interpretations of intergenerational correlations. *Journal of Economic Perspectives*, 16 : 45–58, 2002.

[Griffiths 1980] D. Griffiths. *A Pragmatic Approach to Spearman's Rank Correlation Coefficient*, chapter Teaching Statistics 2, pages 10–13. 1980.

[Grubbs 1969] FE. Grubbs. Procedures for detecting outlying observations. *Technometrics*, 11 : 1–21, 1969.

[Guthke *et al.* 1997] R. Guthke, W. Schmidt-Heck et F. Meyer. Data analysis and knowledge acquisition in biotechnology. In *1st Int. Data Analysis Symposium, Aachen*, September 1997.

[Hampel *et al.* 1986] FR. Hampel, EM. Ronchetti, PJ. Rousseeuw et WA. Stahel. *Robust Statistics : The Approach Based on Influence Functions*. Wiley, 1986.

[Han et Kamber 2006] Jiawei Han et Micheline Kamber. *Data Mining, Second Edition, Second Edition : Concepts and Techniques (The Morgan Kaufmann Series in Data Management Systems) (The Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann, January 2006.

[Hanczar *et al.* 2004] B. Hanczar, M. Courtine, A. Benis, C. Hennegar, K. Clément et JD. Zucker. A prototype-gene based dimension reduction approach for microarray data classification. *SIGKDD Explorations Special Issue on Microarray (Special Interest Group on Knowledge Discovery in Data and Data Mining)*, 5 : 21–28, 2004.

[Hardoon *et al.* 2006] D. Hardoon, C. Saunders, S. Szedmak et J. Shawe-Taylor. A correlation approach for automatic image annotation. pages 681–692. 2006.

[Hargitai *et al.* 2005] J. Hargitai, J. Zernant, G.M. Somfai, R. Vamos, A. Farkas, G. Salacz et R. Allikmets. Correlation of clinical and genetic findings in hungarian patients with stargardt disease. *Invest Ophthalmol Vis Sci*, 46(12) : 4402–4408, 2005.

[Hercberg 2009] S. Hercberg. Etude NUTRINET-SANTE. Technical report, Ministère de la Santé et des Sports, Unité de Recherche en Epidémiologie Nutritionnelle, U557 Inserm/U1125 Inra/Cnam/Paris 13,, 2009.

[Higami *et al.* 2004] Y. Higami, T.D. Pugh, G.P. Page, D.B. Allison, T.A. Prolla et R. Weindruch. Adipose tissue energy metabolism : altered gene expression profile of mice subjected to long-term caloric restriction. *FASEB Journal*, 18 : 415–417, 2004.

[Higgins 2004] AJ. Higgins. Beyond biochemical profiling for biomarker and target discovery. *Current Drug Discovery*, pages 21–24, February 2004.

[Hinneburg et Keim 1998] A. Hinneburg et DA. Keim. An efficient approach to clustering in large multimedia databases with noise. In *Proc. 4th Int. Conference on Knowledge Discovery in Databases (KDD'98)*, 1998.

[Hopkins 2004] W.G. Hopkins. A new view of statistics : Sportsci. Online available : http ://www.sportsci.org/resource/stats/index.html, 2004.

[Huber *et al.* 2006] W. Huber, J. Toedling et L.M. Steinmetz. Transcript mapping with high-density oligonucleotide tiling arrays. *Bioinformatics*, 22(16) : 1963–1970, 2006.

[Hugueney 2003] B. Hugueney. *Représentation symbolique de courbes numériques.* Thèse de doctorat, Université Paris 6, 10 Janvier 2003.

[Ihaka 1996] R. Ihaka, R.and Gentleman. R : a language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5 : 299–314, 1996.

[International Organization for Standardization 2000] International Organization for Standardization. *ISO 9241-11 : Ergonomic requirements for office work with visual display terminals (VDTs) – Part 9 : Requirements for non-keyboard input devices.* 2000.

[International Organization for Standardization 2007a] International Organization for Standardization. ISO Directives, Part 1 : Procedures for the Technical Work. Technical report, 2007.

[International Organization for Standardization 2007b] International Organization for Standardization. ISO Directives, Part 2 : Rules for the structure and drafting of International Standards. Technical report, 2007.

[Jensen *et al.* 2006] L.J. Jensen, J. Saric et P. Bork. Literature mining for the biologist : from information retrieval to biological discovery. *Nature Reviews Genetics*, 7 : 119–129, 2006.

[Jiang *et al.* 2004] H. Jiang, Y. Deng, HS. Chen, L. Tao, Q. Sha, J. Chen, CJ. Tsai et S. Zhang. Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes. *BMC Bioinformatics*, 5, 2004.

[Jiang et Omer 2007] B. Jiang et I. Omer. Spatial topology and its structural analysis based on the concept of simplicial complex. *Transactions in GIS*, 11 : 943, 2007.

[Katz *et al.* 2000] A. Katz, S. S. Nambi, K. Mather, A. D. Baron, D. A. Follmann, G. Sullivan et M. J. Quon. Quantitative insulin sensitivity check index (QUICKI) : a simple, accurate method for assessing insulin sensitivity in humans. *Journal of Clinical Endocrinology and Metabolism*, 85(7) : 2402–2410, 2000.

[Kaufman et Rousseeuw 1990] L. Kaufman et PJ. Rousseeuw. *Finding Groups in Data : An Introduction to Cluster Analysis.* Wiley, New York, 1990.

[Knorr *et al.* 2000] E. Knorr, R. Ng et V. Tucakov. Distance-based outliers : Algorithms and applications. *The VLDB Journal*, 8(3) : pp. 237–253, February 2000.

[Knorr 2002] E. Knorr. *Outliers and data mining : finding exceptions in data.* Thèse de doctorat, University of British Colombia, 2002.

[Kohonen 1997] T. Kohonen. *Self-Organizing Maps.* Springer, 1997.

[Kotsiantis et Pintelas 2004] S. Kotsiantis et P. Pintelas. Recent advances in clustering : A brief survey. *Transactions on Information Science and Applications*, 1(1) : 73–81, 2004.

[Kotsiantis 2007] S.B. Kotsiantis. Supervised Machine Learning : A review of classification techniques. *Informatica*, 31 : 249–268, 2007.

[Kovalerchuk 2001a] B. Kovalerchuk. Review of visual correlation methods. Online available : http ://www.cwu.edu/ borisk/visualization/review2b.pdf, 2001.

[Kovalerchuk 2001b] B. Kovalerchuk. Visualization and decision-making using structural information. In *Proceedings of International Conference on Imaging Science, Systems, and Technology (CISST'2001)*, pages 478–484, 2001.

[Kulkarni et Simon 1988] D. Kulkarni et H.A. Simon. The processes of scientific discovery : The strategy of experimentation. *Cognitive Science*, 12 : 139–175, 1988.

[Labart *et al.* 2000] L. Labart, A. Morineau et M. Piron. *Statistique exploratoire multidimensionnelle*. 3ème édition, 2000.

[Labrecque *et al.* 1999] L.G. Labrecque, S.A. Xue, P. Kazembe, J. Phillips, I. Lampert, N. Wedderburn et B.E. Griffin. Expression of epstein-barr virus lytically related genes in african burkitt's lymphoma : correlation with patient response to therapy. *Int J Cancer*, 81(1) : 6–11, 1999.

[Lander 1999] ES. Lander. Array of hope. *Nature Genetics*, 21 : 3–4, 1999.

[Langley et Nordhausen 1986] P. Langley et B. Nordhausen. A framework for empirical discovery. In *Proceedings of the Internatinal Meeting on Advances in Learning*, Les Arcs, France. 1986.

[Langley 1999] P. Langley. The computer-aided discovery of scientific knowledge. In *Proceedings of the First International Conference on Discovery Science*, Fukuoka, Japan. Springer, 1999.

[Lebart *et al.* 2000] L. Lebart, A. Morineau et M. Piron. *Statistique exploratoire multidimensionnelle*. Dunod, Paris, France, 2000.

[Lebbah 2003] M. Lebbah. *Carte topologique pour données qualitatives : application à la reconnaissance automatique de la densité du trafic routier, Mémoire de Thèse de Doctorat.* Thèse de doctorat, Université de Versailles Saint-Quentin-en-Yvelines, 2003.

[Levy *et al.* 2007] S. Levy, G. Sutton, P.C. Ng, L. Feuk et A.L. and Halpern. The diploid genome sequence of an individual human. *PLoS Biology*, 5(10), 2007.

[Lonning *et al.* 2007] P.E. Lonning, R. Chrisanthar, V. Staalesen, S. Knappskog et J. Lillehaug. Adjuvant treatment : the contribution of expression microarrays. *Breast Cancer Research*, 9(Suppl 2) : S14–S20, 2007.

[Lullman et Mohr 2003] M. Lullman et K. Mohr. *Atlas de poche de pharmacologie*. Atlas de poche Flammarion, 3ème édition, 2003.

[Lund 2002] A. Lund. Measuring usability with the use questionnaire. *Usability and User Experience*, 8-2, 2002.

[MacQueen 1967] J.B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Symposium on Math, Statistics, and Probability*, volume 1, pages 281–297, 1967.

[Madeira et Oliveira 2004] SC. Madeira et AL. Oliveira. Biclustering algorithms for biological data analysis : A survey. *IEEE Transcations On Computational Biology and Bioinformatics*, 1(1) : 24–45, 2004.

[Maglott *et al.* 2005] D. Maglott, J. Ostell, K. D. Pruitt et T. Tatusova. Entrez Gene : gene-centered information at NCBI. *Nucleic Acids Res*, 33, January 2005.

[Mahajan *et al.* 2000] R. Mahajan, K. Brown et V. Atluri. The evolution of microprocessor packaging. *Intel Technology Journal*, Q3 : 1–10, 2000.

[Malandain 2006] G. Malandain. *Les mesures de similarité pour le recalage des images médicales.* Habilitation à diriger des recherches, Université de Nice Sophia-Antipolis, Ecole Doctorale STIC, 2006.

[Marton *et al.* 1999] M. Marton, J. derisi, H. Bennett, V. Iyer, M. Meyer, C. Roberts, R. Stoughton, J. Burchard, D. Slade, H. Dai, D. Bassett, L. Hartwell, P. Brown et S. Friend. Drug traget validation and identification of secondary drug target effects using DNA microarrays. *Nature Medicime*, 4 : 1293–1301, 1999.

[Matthews *et al.* 1985] D.R. Matthews, J.P. Hosker, A.S. Rudenski, B.A. Naylor, D.F. Treacher et R.C. Turner. Homeostasis model assessment : insulin resistance and beta-cell function from fasting plasma glucose and insulin concentrations in man. *Diabetologia*, 28(7) : 412–419, 1985.

[McClure et Wit 2004] J. McClure et E. Wit. *Statistics for Microarrays*. Wiley, 2004.

[Meur *et al.* 2004] N. Meur, G. Lamirault, A. Bihouee, M. Steenman, H. Bedrine-Ferran, R. Teusan, G. Ramstein et JJ. Leger. A dynamic, web-accessible resource to process raw microarray scan data into consolidated gene expression values : importance of replication. *Nucleic Acids Res.*, 32 : 5349–5358, 2004.

[Meyn 2007] S. Meyn. *Control Techniques for Complex Networks*. Cambridge University Press, December 2007.

[Michaut 2006] C. Michaut. Les puces à adn améliorent leur fiabilité. *La Reccherche*, 402, 2006.

[Mirnics *et al.* 2004] K. Mirnics, P. Levitt et D.A. Lewis. DNA microarray analysis of postmortem brain tissue. *Int Rev Neurobiol*, 60 : 153–181, 2004.

[Missal *et al.* 2006] K. Missal, M.A. Cross et D. Drasdo. Gene network inference from incomplete expression data : transcriptional control of hematopoietic commitment. *Bioinformatics*, 22(6) : 731–738, 2006.

[Mondada 2005] L. Mondada. *Chercheurs en interaction Comment émergent les savoirs*. Lausanne, 2005.

[Morris et Truskowski 2003] RJT Morris et BJ. Truskowski. The evolution of storage systems. *IBM Systems Journal*, 42(2) : 205–217, 2003.

[Moskovitch *et al.* 2008] R. Moskovitch, N. Nir Nissim et Y. Elovici. Acquisition of malicious code using active learning. *KDD08 Workshop on Privacy, Security and Trust in KDD (PinKDD08)*, 2008.

[Motulsky 1999] HJ. Motulsky. *Analyzing Data with GraphPad Prism*. GraphPad Software Inc., San Diego CA, 1999.

[Motulsky 2002] HJ. Motulsky. *Biostatistique : Une approche intuitive*. DeBoeck University, 2002.

[Mukherjee *et al.* 2005] G. Mukherjee, N. Abeygunawardena, H ; Parkinson, S. Contrino, S. Durinck, A. Farne, E. Holloway, P. Lilja, Y. Moreau, A. Oezcimen, T. Rayner, A. Sharma, A. Brazma, U. Sarkans et M. Shojatalab. Plant-based microarray data at the european bioinformatics institute. introducing at miamexpress, a submission tool for arabidopsis gene expression data to arrayexpress. *Plant Physiology*, 139 : 632–636, 2005.

[Mullins *et al.* 2006] I. M. Mullins, M. S. Siadaty, J. Lyman, K. Scully, C. T. Garrett, W. G. Miller, R. Muller, B. Robson, C. Apte, S. Weiss, I. Rigoutsos, D. Platt, S. Cohen et W. A. Knaus. Data mining and clinical data repositories : Insights from a 667,000 patient data set. *Comput Biol Med*, 36(12) : 1351–1377, December 2006.

[Munoz-Garcia *et al.* 1990] J. Munoz-Garcia, JL. Moreno-Rebollo et A. Pascual-Acosta. Outliers : a formal approach. *Int. Statist. Rev.*, 58 : 215–226, 1990.

[Myatt 2006] G.J. Myatt. *Making Sense of Data : A Practical Guide to Exploratory Data Analysis and Data Mining*. John Wiley, 2006.

[Natarajakumar *et al.* 2004] B. Natarajakumar, V. Kurisunkal, R.K. Moore et D. Braaten. Rain Heights Over the Oceans : Relation to Rain Rates. pages 126–132, Cairns, Great Barrier Reef, Australia. URSI Commission F Triennium Open Symposium, June 2004.

[National Institutes of Health 2000] National Institutes of Health. The practical guide identification, evaluation, and treatment of overweight and obesity in adults. Technical report, National Institutes of Health, october 2000.

[Nelson 2004] D. Nelson. *The Penguin Dictionary of Statistics*. Penguin, 2004.

[Ng et Jiawei 1994] RT. Ng et H. Jiawei. Efficient and effective clustering methods for spatial data mining. pages 144–155, 1994.

[Nielsen 1994] J. Nielsen. John Wiley & Sons, 1994.

[Nisbet *et al.* 2009] R. Nisbet, J. Elder et G. Miner. *Handbook of Statistical Analysis and Data Mining Applications*. Elsevier, 2009.

[O'Carroll *et al.* 2003] PW. O'Carroll, WA. Yasnoff, ME. Ward, LH. Ripp et EL. Martin. *Public Health Informatics and Information Systems*. Springer, 2003.

[Ohsawa et Yada 2009] Y. Ohsawa et K. Yada. *Data Mining for Design and Marketing*. Chapman and Hall CRC, 2009.

[Organisation Mondiale de la Santé 2004] Organisation Mondiale de la Santé. Obesity : preventing and managing the global epidemic. Technical report, Organisation Mondiale de la Santé, Geneva, 2004.

[Organisation Mondiale de la Santé 2008] Organisation Mondiale de la Santé. *World Health Statistics 2008*. WHO Press, 2008.

[Organisation Mondiale de la Santé 1995] Organisation Mondiale de la Santé. Rapport sur la santé dans le monde. Technical report, OMS, 1995.

[Organisation Mondiale de la Santé 2006] Organisation Mondiale de la Santé. Obésité et surpoids. *Aide-mémoire*, 311, Septembre 2006.

[Pakhira 2008] MK. Pakhira. Fast image segmentation using modified CLARA algorithm. In *ICIT '08 : Proceedings of the 2008 International Conference on Information Technology*, pages 14–18, Washington, DC, USA. IEEE Computer Society, 2008.

[Papadimitriou *et al.* 2006] S. Papadimitriou, J. Sun et PS. Yu. Local correlation tracking in time series. In *ICDM '06 : Proceedings of the Sixth International Conference on Data Mining*, pages 456–465, Washington, DC, USA. IEEE Computer Society, 2006.

[Park *et al.* 2006] H-S. Park, J-S Lee et C-H. Jun. A k-means-like algorithm for k-medoids clustering and its performance. In *Proceedings of the 36th CIE Conference on Computers & Industrial Engineering*, pages 1222–1231, 2006.

[Peirce 1852] B. Peirce. Criterion for the rejection of doubtful observations. *Astronomical Journal*, 45 : 161–163, 1852.

[Perusse *et al.* 2005] L. Perusse, T. Rankinen, A. Zuberi, Y.C. Chagnon, S.J. Weisnagel, G. Argyropoulos, B. Walts, E.E. Snyder et C. Bouchard. The human obesity gene map : the 2004 update. *Obesity Research*, 13(3) : 381–490, 2005.

[Pevsner 2005] J. Pevsner. *Bioinformatics and Functional Genomics*. John Wiley & Sons Inc, 2005.

[Piatetsky-Shapiro et Frawley 1991] G. Piatetsky-Shapiro et W. Frawley. *Knowledge Discovery in Databases*. Cambridge, Mass. : AAAI/MIT Press, 1991.

[Pinto *et al.* 2005] F.R. Pinto, L. Ashley Cowart, Y.A. Hannun, B. Rohrer et J.S. Almeida. Local correlation of expression profiles with gene annotations-proof of concept for a general conciliatory method. *Bioinformatics*, 21(7) : 1037–1045, 2005.

[Planchon 2005] V. Planchon. Traitement des valeurs aberrantes : concepts actuels et tendances générales. *Biotechnol. Agron. Soc. Environ.*, 9(1) : 19–34, 2005.

[Porter *et al.* 1994] M.J. Porter, J.K. Field, S.F. Leung, D. Lo, J.C. Lee, D.A. Spandidos et C.A. van Hasselt. The detection of the C-MYC and RAS oncogenes in nasopharyngeal carcinoma by immunohistochemistry. *Acta Otolaryngol*, 114(1) : 105–109, 1994.

[Przybyklo 2005] A. Przybyklo. Optimisation par algorithme de groupement de la construction automatique de bases de connaissance floues. Master's thesis, Ecole Polytechnique de Montréal - Département de Génie Mécanique., 2005.

[Qiao *et al.* 2004] J.G. Qiao, Y.Q. Zhang, Y.C. Yin et Z. Tan. Expression of Survivin in pancreatic cancer and its correlation to expression of Bcl-2. *World J Gastroenterol*, 10(18) : 2759–2761, 2004.

[Quinlan 1993] JR. Quinlan. *C4.5 : programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[Quinlan 1996] JR. Quinlan. Boosting, Bagging, and C4.5. pages 725–730, 1996.

[R Development Core Team 2006] R Development Core Team. *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2006.

[Rajman et Lebart 2008] M. Rajman et L. Lebart. Similarités pour données textueles. In Presses universitaires de Lyon, editor, *JADT 2008 : actes des 9es Journées internationales d'Analyse statistique des Données Textuelles*, 2008.

[Ramaswamy et Golub 2002] S. Ramaswamy et T.R. Golub. DNA Microarrays in Clinical Oncology. *J Clin Oncol*, 20(7) : 1932–1941, 2002.

[Ray et Turi 1999] S. Ray et R.H. Turi. Determination of number of clusters in k-means clustering and application in colour image segmentation. In N. R. Pal, A. K. De et J. Das, editors, *Proceedings in the 4th International Conference on Advances in Pattern Recognition and Digital Techniques (ICAPRDT'99)*, pages 137–143, Calcutta, India. Narosa Publishing House, 1999.

[Rayner *et al.* 2006] T.F. Rayner, P. Rocca-Serra, Spellman P.T., H.C. Causton, A. Farne, E. Holloway, R.A. Irizarry, J. Liu, D.S. Maier, M. Miller, K. Petersen, J. Quackenbush, G. Sherlock, C.J. Stoeckert, J. White, P.L. Whetzel, F. Wymore, H. Parkinson, C.A. Sarkans, U.and Ball et A. Brazma. A simple spreadsheet-based, miame-supportive format for microarray data : Mage-tab. *BMC Bioinformatics*, 7 : 489–502, 2006.

[Rosen *et al.* 2005] J.E. Rosen, N.G. Costouros, D. Lorang, A.L. Burns, H.R. Alexander, M.C. Skarulis, C. Cochran, J.F. Pingpank, S.J. Marx, A.M. Spiegel et S.K. Libutti. Gland size is associated with changes in gene expression profiles in sporadic parathyroid adenomas. *Ann Surg Oncol*, 12(5) : 412–416, 2005.

[Rota 1964] G-C. Rota. The number of partitions of a set. *American Mathematical Monthly*, 5(71) : 498–504, 1964.

[Rousseeuw et Leroy 2003] PJ. Rousseeuw et AM. Leroy. *Robust Regression and Outlier Detection*. Wiley, 2003.

[Schena *et al.* 1995] M. Schena, D. Shalon, R.W. Davis et P.O. Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235) : 467–470, 1995.

[Scherf *et al.* 2000] U. Scherf, DT. Ross, M. Waltham, JK. Smith JH.and Lee, L. Tanabe, KW. Kohn, Reinhold WC., TG. Myers, DT. Andrews, DA. Scudiero, Eisen MB., EA. Sausville,

Y. Pommier, D. Botstein, PO. Brown et JN. John N. Weinstein. A gene expression database for the molecular pharmacology of cancer. *Nature Genetics*, 24, March 2000.

[Schervish 1996] MJ. Schervish. P values : What they are and what they are not. *The American Statistician*, 50(3) : 203–206, August 1996.

[Sengur 2008] A. Sengur. An expert system based on linear discriminant analysis and adaptive neuro-fuzzy inference system to diagnosis heart valve diseases. *Expert Syst. Appl.*, 35(1-2) : 214–222, 2008.

[Shannon *et al.* 2002] W. Shannon, M. Watson, A. Perry et K. Rich. Mantel statistics to correlate gene expression levels from microarrays with clinical covariates. *Genetic Epidemiology*, 23 : 87–96, 2002.

[Sherlock *et al.* 2001] G. Sherlock, T. Hernandez-Boussard, A. Kasarskis, G. Binkley, J.C. Matese, S.S. Dwight, M. Kaloper, S. Weng, H. Jin, C.A. Ball, M.B. Eisen, P.T. Spellman, P.O. Brown, D. Botstein et J. M. Cherry. The Stanford Microarray Database. *Nucleic Acids Res*, 29(1) : 152–155, 2001.

[Shi *et al.* 2005] L. Shi, F. Goodsaid, U. Scherf, R. Puri, J. Warrington, J. Collins, R. Setterquist, L. Zhang, L. Lamarcq, M. Elliot, C. VanHuffel, R. Shippy, Y. Luo, S. Baker, G. Fischer, D. Dix, M. Salit, Z. Szallasi, R. Jensen et W. Tong. The maqc project : Establishing qc metrics and thresholds for microarray quality control. Technical report, Food and Drug Administration - National Center for Toxicological Research - USA, 2005.

[Shneiderman et Plaisant 2004] B. Shneiderman et C. Plaisant. *Designing the User Interface : Strategies for Effective Human-Computer Interaction.* Addison Wesley, 4th édition, 2004.

[Siegel et Castellan 1988] S. Siegel et N. J. Castellan. *Nonparametric statistics for the behavioral sciences.* McGraw-Hill Book Company, New York, USA, 1988.

[Simon 2006] N. Simon. *Pharmacocinétique de population. Introduction à Nonmem.* 2006.

[Slimani *et al.* 2007] T. Slimani, B. Ben Yaghlane et K. Mellouli. Une extension de mesure de similarité entre les concepts d'une ontologie. In *4th International Conference : Sciences of Electronic, Technologies of Information and Telecommunications, Tunisia*, 2007.

[Slonim 2002] DK. Slonim. From patterns to pathways : gene expression data analysis comes of age. *Nature Genetics*, pages 502–508, 2002.

[Spellman *et al.* 2002] P.T. Spellman, M. Miller, J. Stewart, C. Troup, U. Sarkans, S. Chervitz, D. Bernhart, G. Sherlock, C. Ball, M. Lepage, M. Swiatek, W.L. Marks, J. Goncalves, S. Markel, D. Iordan, M. Shojatalab, A. Pizarro, J. White, R. Hubley, E. Deutsch, M. Senger, B. J. Aronow, A. Robinson, D. Bassett, C.J. Jr. Stoeckert et A. Brazma. Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol*, 3(9), 2002.

[Storey et Tibshirani 2003] J.D. Storey et R. Tibshirani. Statistical significance for genomewide studies. *Proc Natl Acad Sci USA*, 100(16) : 9440–9445, August 2003.

[Storey 2002] J.D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 64(3) : 479–498, 2002.

[Storey 2005] J.D. Storey. The optimal discovery procedure : A new approach to simultaneous significance testing. *UW Biostatistics Working Paper Series*, Working Paper : 259, September 2005.

[Stratowa *et al.* 1999] C. Stratowa, G. Loffler, P. Haberl, N. Schweifer, P. Lichter, H. Dohner et KK. Wilgenbus. Correlation of clinical data with expression profiles in B-cell chronic lymphocytic leukaemia as determined by cDNA microarray analysis. *Nature Genetics*, 23 : 76, 1999.

[Struyf *et al.* 1997] A. Struyf, M. Hubert et P. Rousseeuw. Clustering in an object-oriented environment. *Journal of Statistical Software*, 1, February 1997.

[Symmans *et al.* 2007] W. Symmans, F. Peintinger, C. Hatzis, R. Rajan, H. Kuerer, V. Valero, L. Assad, A. Poniecka, B. Hennessy, M. Green, A.U. Buzdar, S.E. Singletary, G.N. Hortobagyi et L. Pusztai. Measurement of Residual Breast Cancer Burden to Predict Survival After Neoadjuvant Chemotherapy. *J Clin Oncol*, 25(28) : 4414–4422, 2007.

[Takase *et al.* 2000] S. Takase, K. Suruga et T. Goda. Regulation of vitamin a metabolism-related gene expression. *Br J Nutr.*, 84 Suppl 2 : 217–221, December 2000.

[Taleb *et al.* 2005] S. Taleb, D. Lacasa, J.P. Bastard, C. Poitou, R. Cancello, V. Pelloux, N. Viguerie, A. Benis, J.D. Zucker, J.L. Bouillot, C. Coussieu, A. Basdevant, D. Langin et K. Clément. Cathepsin S, a novel biomarker of adiposity : relevance to atherogenesis. *FASEB J*, 19(11) : 1540–1542, 2005.

[Tan *et al.* 2003] PK. Tan, TJ. Downey, ELJr. Spitznagel, P. Xu, D. Fu, DS. Dimitrov, RA. Lempicki, BM. Raaka et MC. Cam. Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res.*, 31 : 5676–5684, 2003.

[Tan et Gilbert 2003] A.C. Tan et D. Gilbert. Ensemble machine learning on gene expression data for cancer classification. *Appl Bioinformatics*, 2(3 Suppl) : 75–83, 2003.

[Tanay *et al.* 2002] A. Tanay, R. Sharan et R. Shamir. Discovering statistically significant biclusters in gene expression data. In *In Proceedings of ISMB 2002*, pages 136–144, 2002.

[Temanni *et al.* 2005] M.R. Temanni, B. Hanczar et J.D. Zucker. Combinaison des données d'expressions géniques et des données cliniques pour améliorer la qualité de la prédiction de la survie à 5 ans de patients atteints de cancer. In *11ème Journées Francophones Informatique Médicale*, 2005.

[The Royal Swedish Academy of Sciences 2006] The Royal Swedish Academy of Sciences. Molecular basis of eukaryotic transcription. *Advanced information on the Nobel Prize in Chemistry*, 2006.

[Tusher *et al.* 2001] VG. Tusher, R. Tibshirani et G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *PNAS*, 98 : 5116–5121, 2001.

[Uguen *et al.* 2007] G. Uguen, G. Coppin, C. Lassudrie et P. Lenca. A social approach of the individual decision making process by the search for a dominance structure : risk management by a business process manager. In *Computational Methods for Modelling and learning in Social and Human Sciences - Bilingual Conference*, 2007.

[Undercoffer *et al.* 2003] JL. Undercoffer, A. Joshi et H. Shah. Fuzzy Clustering for Intrusion Detection. In *Proceedings of the 12th IEEE International Conference on Fuzzy Systems*, pages 1274–1278, April 2003.

[Unnthorsson *et al.* 2003] R. Unnthorsson, TP. Runarsson et MT. Jonsson. Model selection in one-class v-SVMs using RBF kernels. In *COMADEM - Proceedings of the 16th international congress*. Vaxj University, August 2003.

[van der Laan *et al.* 2003] M. van der Laan, K. Pollard et J. Bryan. A new partitioning around medoids algorithm. *Journal of Statistical Computation and Simulation*, 73(8) : 575–584, 2003.

[van Rossum *et al.* 2005] E.F.C. van Rossum, H. Russcher et S.W.J. Lamberts. Genetic polymorphisms and multifactorial diseases : facts and fallacies revealed by the glucocorticoid receptor gene. *Trends in Endocrinology & Metabolism*, 16(10) : 445–450, December 2005.

[Vapnik 1998] V. Vapnik. *Statistical learning theory*. Wiley, 1998.

[Varela 1997] F. Varela. *Invitation aux sciences cognitives*. Points Sciences, 1997.

[Vazquez Martinez *et al.* 1998] C. Vazquez Martinez, P. Galan, P. Preziosi, L. Ribas, LL. Serra et Hercberg S. The SUVIMAX study : the role of antioxidants in the prevention of cancer and cardiovascular disorders. *Rev Esp Salud Publica*, 72(3) : 173–183, 1998.

[Venables et Ripley 2002] WN. Venables et BD. Ripley. *Modern Applied Statistics with {S}*. Fourth édition, 2002.

[Viguerie *et al.* 2004] N. Viguerie, K. Clément, P. Barbe, M. Courtine, A. Benis, D. Larrouy, B. Hanczar, V. Pelloux, C. Poitou, Y. Khalfallah, G. S. Barsh, C. Thalamas, J. D. Zucker et D. Langin. In vivo epinephrine-mediated regulation of gene expression in human skeletal muscle. *J Clin Endocrinol Metab*, 89(5) : 2000–2014, 2004.

[Wadden *et al.* 2002] T.A. Wadden, K.D. Brownell et G.D. Foster. Obesity : responding to the global epidemic. *J Consult Clin Psychol*, 70(3) : 510–525, 2002. 0022-006x Journal Article Review Review, Tutorial.

[Waltz 2001] C. Waltz. Les puces sous toutes les coutures. . . Technical report, Dossier "Biopuces" CEA, 2001.

[Wang *et al.* 1997] W. Wang, J. Yang et R. Muntz. Sting : A statistical information grid approach to spatial data mining. In *VLDB '97 : Proceedings of the 23rd International Conference on Very Large Data Bases*, pages 186–195, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc., 1997.

[Watada et Yamashiro 2006] Junzo Watada et Kozo Yamashiro. A data mining approach to consumer behavior. In *ICICIC '06 : Proceedings of the First International Conference on Innovative Computing, Information and Control*, pages 652–655, Washington, DC, USA. IEEE Computer Society, 2006.

[Watson et Crick 1953] J.D. Watson et F.H. Crick. Molecular structure of nucleic acids ; a structure for deoxyribose nucleic acid. *Nature*, 171(4356) : 737–738, 1953.

[Werner 2003] M. Werner. *Identification of Multivariate Outliers in Large-Scale Data Sets*. Thèse de doctorat, University of Colorado at Denver, 2003.

[Whetzel *et al.* 2006] P.L. Whetzel, H. Parkinson, H.C. Causton, L. Fan, J. Fostel, G. Fragoso, L. Game, M. Heiskanen, N. Morrison, P. Rocca-Serra, S.A. Sansone, Taylor C., J. White et C.J. Stoeckert. The MGED Ontology : a resource for semantics-based description of microarray experiments. *Bioinformatics*, 22(7) : 866–882, 2006.

[Whistler *et al.* 2003] T. Whistler, E. Unger, R. Nisenbaum et S. Vernon. Integration of gene expression, clinical, and epidemiologic data to characterize chronic fatigue syndrome. *Journal of Translational Medicine*, 1(1) : 10, 2003.

[Wolfe 2000] M. Wolfe. Metadata, knowledge management, and communications. *Canadian Journal of Communication*, 25(4), 2000.

[Wu 2008] F. X. Wu. Genetic weighted k-means algorithm for clustering large-scale gene expression data. *BMC bioinformatics*, 9 Suppl 6, 2008.

[Yamanishi *et al.* 2003] Y. Yamanishi, J.P. Vert, A. Nakaya et N. Kanehisa. Extraction of correlated gene clusters from mutliple genomic data by generalized kernel canonical correlation analysis. *Bioinformatics*, 19 : 323–330, 2003.

[Yusuf *et al.* 2005] S. Yusuf, S. Hawken, S. ?unpuu, L. Bautista, M. Grazia Franzosi, P. Commerford, CC. Lang, Z. Rumboldt, CL. Onen, L. Lisheng, S. Tanomsup, PJr. Wangai, F. Razak, AM. Sharma et SS. Anand. Obesity and the risk of myocardial infarction in 27 000 participants from 52 countries : a case-control study. *Lancet*, 366 : 1640–1649, 2005.

[Zou *et al.* 2003] K.H. Zou, K. Tuncali et S.K. Silverman. Correlation and simple linear regression. *Radiology*, 227(3) : 617–622, June 2003.

[Zuo 2006] Y. Zuo. Multidimensional trimming based on projection depth. *The Annals of Statistics*, 34 : 2211–2251, 2006.

Résumé

La Fouille de Données est un domaine de recherche émergent en Informatique Médicale. Elle consiste à extraire de la Connaissance de grandes bases de données, pour aider, par exemple, à la prédiction et au diagnostic des maladies multifactorielles. Aujourd'hui, les protocoles de recherche clinique ne se contentent plus de collecter des données uniquement médicales, mais ils s'intéressent aussi aux données génétiques, issues notamment des puces à ADNc. Les analyses doivent donc prendre en compte l'ensemble de ces données et ce en dépit de leur nature différente et de leur qualité relative. Actuellement, les approches « classiques » couramment utilisées par les biologistes dans ce contexte se contentent d'étudier qu'une infime partie des données en se fondant sur des a priori importants. Nos travaux reposent sur l'automatisation de ce processus d'analyse dans un système : DiscoClini. Dans un premier temps, cette thèse s'intéressera à la mise en place d'un flux de données adapté aux données que nous souhaitons traiter (données biocliniques versus données issues de puces à ADNc) et fondé sur une mesure statistique, la corrélation de Spearman. Dans un second temps, les valeurs singulières, dues à la qualité relative des données et sources d'erreurs lors des analyses, seront identifiées automatiquement grâce à une méthode de classification, PAMout. Enfin, l'ensemble de ces résultats sera présenté de manière accessible aux experts-biologistes afin de les aider dans leur analyse finale des données. Des expérimentations dans le domaine de l'étude des Obésités ont été menées. Elles nous ont permis de valider pas à pas notre processus de Fouille de Données et de découvrir de « potentiels » biomarqueurs. Une analyse plus globale d'usage et d'utilisabilité a montré l'intérêt du système dans son ensemble.

Mots-clés: Fouille de Données, Découverte de Connaissance, Corrélation, Valeurs singulières, Visualisation

Abstract

Data Mining is an emerging area in Medical Informatics research field. It consists in extracting knowledge from large databases in order to assist, for example prediction and diagnosis of multifactorial diseases. Nowadays, clinical research protocols are no longer limited to collect only medical data, but they are also regarding to other kinds of data such as genomic data from cDNA microarrays. The analysis has to take into account all these data in despite of their different nature and their relative quality. Currently, the « classical » approaches commonly used by biologists in this context simply explore a tiny part of the data based on major *a priori*. Our work is based on automating the analysis process in a system : DiscoClini. Firstly, this PhD dissertation focuses on the definition of a data workflow adapted to data that we want to deal with (bioclinical data *versus* data from cDNA microarrays) and mainly based on a statistic, the Spearman's rank correlation coefficient. In a second step, outliers, due to the relative quality of data and sources of errors in analysis are automatically identified thanks to a classification method, PAMout. Finally, all these results will be presented in an easy way to biologist experts in order to help to analyze them. Experiments related to researches in obesity medicine have been done. They allowed us to validate our Data Mining process, step by step, and to discover « potential » biomarkers. Evaluation of use and usability has shown the benefits of the system as a whole.

Keywords: Data Mining, Knowledge Discovery, Correlation, Outliers, Visualization