De l'analyse de données d'expression à la reconstruction de réseau de gènes

Marie Agier

N° d'Ordre : D.U. 1707
EDSPIC : 363

Université Blaise Pascal Clermont-Ferrand II

École doctorale
sciences pour l'ingénieur de Clermont-Ferrand

THÈSE

présentée par

Marie AGIER

pour obtenir le grade de

Docteur d'université

Spécialité : Informatique

De l'analyse de données d'expression à la reconstruction de réseaux de gènes

Soutenue publiquement le 7 décembre 2006 devant le jury :

Pr. Alain Quilliot (LIMOS Clermont-Fd) Président du jury
Pr. Christine Froidevaux (LRI Paris) Rapporteur
Pr. Michel Scholl (CNAM Paris) Rapporteur
Pr. Einoshin Suzuki (Kyushu Univ. Japan) Examinateur
Pr. Michel Schneider (LIMOS Clermont-Fd) Examinateur
Pr. Jean-Marc Petit (LIRIS Lyon) Directeur de thèse
Dr. Véronique Vidal (Société Diagnogène) Co-encadrant
Dr. Christian Pradeyrol (Société Diagnogène) Membre invité

Résumé

Ce travail de thèse s'inscrit dans le contexte émergent de la fouille de données d'expression de gènes. Elle s'est déroulée, dans le cadre d'un contrat CIFRE au sein de trois structures : la société Diagnogène, le LIMOS et le Centre de Lutte contre le Cancer de la Région Auvergne.
Ce travail a consisté tout d'abord à pré-traiter et à analyser l'ensemble des données d'expression dans le cadre de deux principaux projets dont l'objectif global est d'améliorer le diagnostic et le pronostic du cancer du sein et de mieux cibler les traitements, grâce à la technologie des biopuces. Un processus de pré-traitement des données d'expression optimisé et spécialement adapté aux biopuces de la société Diagnogène a été mis en place. De plus, plusieurs analyses ont été réalisées et ont permis d'identifier des gènes permettant de mettre en évidence un profil d'expression caractéristique du statut ganglionnaire des patientes et de la réponse thérapeutique à un traitement chimiothérapeutique particulier, le docétaxel.
Ensuite, une nouvelle technique de reconstruction de réseaux de gènes basée sur la notion de règles entre gènes a été proposée, l'idée étant d'offrir aux biologistes la possibilité de choisir parmi plusieurs sémantiques, le sens des règles qu'ils souhaitent générer. L'originalité de ce travail est de proposer un cadre global pouvant inclure un grand nombre de sémantiques pour les règles et d'utiliser des méthodes identiques de génération, de post-traitement et de visualisation des règles pour toutes les sémantiques proposées. La notion de sémantiques bien-formées i.e. pour lesquelles les axiomes d'Armstrong sont justes et complets, est introduite. Un résultat est également donné permettant de savoir simplement si une sémantique est ou non bien-formée. Une visualisation des règles sous forme de réseaux globaux i.e. incluant plusieurs sémantiques est proposée. Finalement, cette approche a été développée sous forme d'un module appelé RG intégré à un logiciel d'analyse de données d'expression de gènes, le logiciel MeV de TIGR.

Mots-clés :
Data mining ; Découverte de règles entre attributs ; Système d'inférence d'Armstrong ; Données d'expression ; Réseaux de gènes.

Remerciements

Mes premiers remerciements vont tout naturellement à mon directeur de thèse, le Professeur Jean-Marc PETIT, pour m'avoir guidée, encouragée, conseillée pendant ces trois années tout en me laissant une grande liberté. Je ne le remercierai jamais assez pour son soutien toujours présent fasse aux difficultés que j'ai pu rencontrer.

Cette thèse a été réalisée en grande partie au LIMOS dont je remercie vivement le directeur le Professeur Alain QUILLIOT. L'accueil chaleureux qui m'a été donné m'a permis de travailler dans les meilleures conditions. Je remercie également Mr QUILLIOT pour avoir accepté de présider mon jury de thèse.

Je tiens ensuite à remercier très chaleureusement le Docteur Christian PRADEYROL, directeur de la société Diagnogène, qui m'a permis d'obtenir un financement CIFRE. Je le remercie pour son soutien tout au long de cette thèse, pour ces quelques discussions très intéressantes que nous avons pu partager et pour avoir accepté de participer à mon jury de thèse.

Je tiens également à apporter mes remerciements au Professeur Yves-Jean BIGNON, directeur du Laboratoire d'Oncologie Moléculaire du Centre Jean Perrin. Ses remarques scientifiques étaient toujours très constructives pour mon travail. Je le remercie également pour les responsabilités qui m'ont été confiées dans le projet international Breast Med Consortium.

Je remercie également le Professeur Christine FROIDEVAUX et le Professeur Michel SCHOLL d'avoir accepté de rapporter mon travail de thèse, le Professeur Michel SCHNEIDER et le Professeur Einoshin SUZUKI pour l'intérêt qu'ils ont porté à mon travail et pour avoir accepté de participer à mon jury de thèse.

Je remercie mon encadrant Véronique Vidal ainsi que toutes les personnes avec qui j'ai travaillé au sein de la société Diagnogène, notamment les 2 Valérie, Chantal, Laurence et Jean-Baptiste. Je remercie également les personnes du LOM avec qui j'ai travaillé et tout particulièrement Marie-Laure, Laurence, Marie, Jérôme, Thomas et Fabrice.

Je remercie également mes « collègues de bureau » avec qui j'ai partagé des moments de joie mais aussi de doute : Fréd, Leïla et Fawzy, tous mes collègues de l'ISIMA et tout particulièrement Hélène, Julien, Yoan et Pierre pour les moments de détente ainsi que toutes les personnes que j'ai pu oublier.

Enfin, je tiens à remercier ma famille pour son encouragement et surtout la personne qui m'a toujours soutenue pendant ces années parfois difficiles : mon amour Julien.

Table des matières

Liste des figures

Liste des tableaux

Chapitre 1

Introduction

Sommaire

1.1 Contexte du travail

L'achèvement du séquençage du génome de plusieurs espèces et en particulier de l'humain, marque le début de la **génomique fonctionnelle** (ou post-génomique). Découvrir la fonction précise des gènes dans la cellule et étudier leurs interactions sont les grands enjeux de cette nouvelle étape sur laquelle reposent beaucoup d'espoirs.

Etudier la fonction des gènes et leurs interactions requiert une analyse de leur expression i.e. mesurer l'activité des gènes dans une cellule à un instant donné. Cette analyse de l'**expression des gènes** n'est pas nouvelle, toutefois son caractère global est une avancée très importante pour les biologistes. En effet, jusqu'à présent la génétique classique permettait d'analyser un petit nombre de gènes à la fois, aujourd'hui diverses techniques

permettant d'étudier simultanément l'expression de milliers de gènes sont disponibles comme par exemple les puces à ADN ou les puces à oligonucléotides.

La technologie des **puces à ADN** ou **biopuces** [87] permet par exemple l'étude simultanée de plus de 50 000 gènes et offre ainsi la possibilité d'étudier des génomes entiers comme le génome humain qui compte environ 30 000 gènes.

Les champs d'application des outils de la post-génomique sont très nombreux et concernent toutes les sciences du vivant (médecine, pharmacologie, agronomie). De grands espoirs reposent dans la génomique fonctionnelle notamment pour améliorer la compréhension des maladies génétiques, comme le cancer et développer de nouvelles molécules thérapeutiques.

L'application de ces nouvelles approches expérimentales génère des quantités très importantes de données. L'exploitation efficace de ces données repose sur la **bioinformatique**, discipline regroupant toutes les applications informatiques appliquées à la biologie. Les méthodes de **fouille de données** (ou data mining) [44] ont pour objectif d'extraire de la connaissance à partir de toutes ces données et sont naturellement applicables aux données d'expression.

Cette thèse s'inscrit dans le contexte émergent de la **fouille de données d'expression de gènes**. Elle s'est déroulée, dans le cadre d'un contrat CIFRE, au sein de trois structures, la société **Diagnogène**, le **LIMOS** Laboratoire d'Informatique, de Modélisation et d'Optimisation des Systèmes de l'Université Blaise Pascal et le Laboratoire d'Oncologie Moléculaire du **Centre Jean Perrin**. Ces trois entités sont décrites ci-après.

1.1.1 La société Diagnogène

La société **Diagnogène** [5] est spécialisée dans le développement et l'analyse de puces à ADN et de puces CGH. Partenaire privilégié d'Agilent Technologies [1], la société dispose de l'intégralité de la plate-forme technologique (cf figure 1.1) permettant l'analyse du génome grâce aux puces CGH (Comparative Genomic Hybridization) et l'analyse du transcriptome grâce aux puces à ADN.

La société Diagnogène oriente ses projets de recherche vers la cancérologie, et plus particulièrement vers les cancers qui représentent des problèmes de santé publique, le cancer du sein, de l'ovaire et de la prostate.

Les tumeurs résultant de l'accumulation et de combinaisons de multiples altérations moléculaires sont très hétérogènes. Cependant, les indications thérapeutiques et diagnostiques sont fondées sur des facteurs pronostics qui ne révèlent pas cette hétérogénéité. La technique des puces à ADN est donc très intéressante car elle permet une caractérisation moléculaire globale des tumeurs. Elles permettent aussi d'identifier de nouvelles classes pronostiques dans des groupes de tumeurs d'apparence clinique et histologique homogènes

mais hétérogènes sur le plan évolutif.

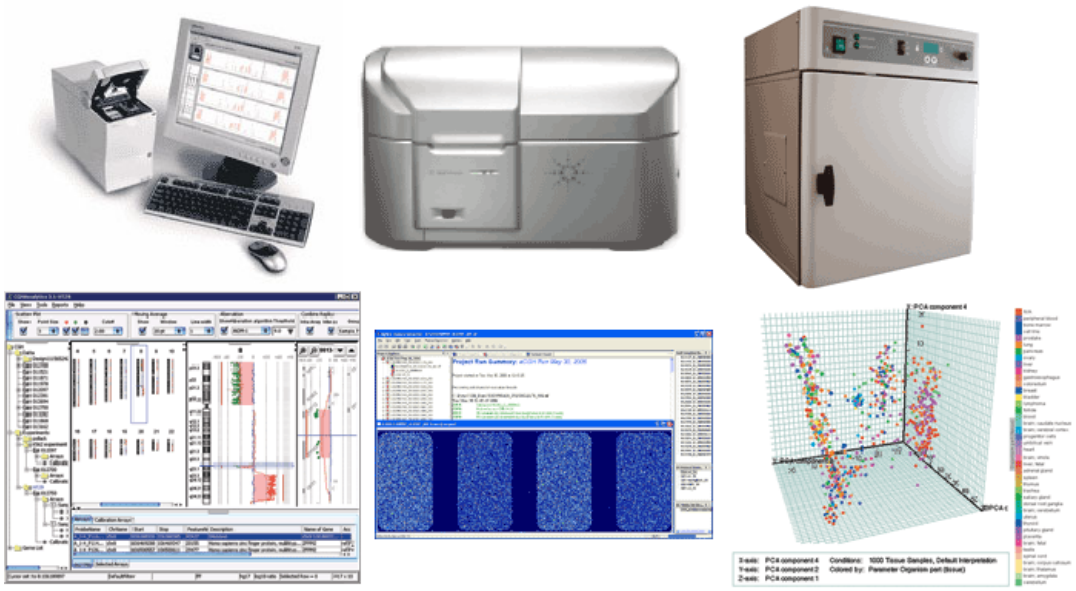


FIGURE 1.1 – Plate-forme Agilent Technologies

La société Diagnogène a développé deux biopuces permettant l'étude de tumeurs du sein dont l'objectif est de développer un outil d'aide au diagnostic et au pronostic (biopuces Tumeurs Mammaires nommées TM) ainsi qu'à l'orientation thérapeutique (biopuces Chimio-Sensibilité nommées CS).

1.1.2 Le LIMOS

Le Laboratoire d'Informatique, de Modélisation et d'Optimisation des Systèmes [6], unité mixte de recherche Université Blaise Pascal/CNRS UMR 6158, centre son activité sur l'informatique des systèmes organisationnels, et décompose celle-ci selon 3 orientations : La Recherche Opérationnelle et l'Aide à la Décision, les Systèmes d'Information et de Communication, et la Productique.

Ce travail de thèse s'est déroulé au sein de l'équipe **Bases de Données** de l'axe **Systèmes d'Information et de Communication** et concerne plus spécifiquement la partie Fouille de Données.

La **Fouille de Données** ou Extraction de Connaissances à partir de Données (ECD) [44] est un processus complexe qui se déroule suivant une suite d'opérations, partant des données pour arriver à la connaissance (cf figure 1.2). Il s'agit à partir de données brutes, de produire de la connaissance utile afin que les experts du domaine puissent prendre les bonnes décisions.

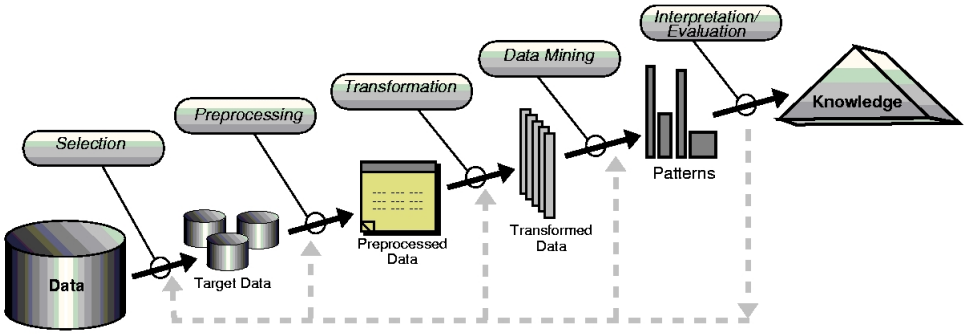


FIGURE 1.2 – Extraction de connaissances à partir des données

Les étapes de pré-traitement concernent la mise en forme des données selon leur type (numériques, ordinales, symboliques, images, ...) ainsi que le nettoyage des données (suppression du bruit, normalisation, traitement des données manquantes, des données extrêmes, etc.). Cette première phase est cruciale car de cette étape dépend la qualité des données analysées agissant directement sur la qualité des résultats et donc de la connaissance extraite.

L'objectif de l'étape de "data mining" est de donner du sens aux données explorées. Plusieurs méthodes et algorithmes ont été développés, émanant de plusieurs disciplines : statistiques, intelligence artificielle, bases de données, algorithmique, visualisation, etc.

Dans ce travail de thèse, nous appliquons le processus global décrit dans la figure 1.2 aux données d'expression de gènes.

1.1.3 Le Centre Jean Perrin

Les missions essentielles du **Centre Jean Perrin** [3], le Centre de Lutte contre le Cancer de la région Auvergne, concernent la prévention, le dépistage, le diagnostic et le traitement des cancers. Il participe également à la surveillance prolongée des résultats thérapeutiques, à l'éducation sanitaire, à l'action sociale, à la recherche et à l'éducation universitaire et post-universitaire.

Le plateau technique des Centres regroupe des équipements permettant de diagnostiquer et de mettre en œuvre les différentes thérapeutiques : chirurgie, radiothérapie et traitements médicaux et biologiques. Les médecins des centres sont spécialisés dans ces différentes disciplines et ont l'habitude de travailler ensemble pour assurer les traitements les mieux adaptés à chaque cas particulier.

Comme les 19 autres Centres Régionaux de Lutte contre le Cancer de France, le Centre Jean Perrin a pour mission les soins et également la recherche. Aujourd'hui, la recherche a

une dimension nationale, voire internationale qui s'organise le plus souvent par l'intermédiaire des essais cliniques et thérapeutiques. Pour cela, les médecins du Centre disposent d'attachés de recherche, de techniciens, de biostatisticiens et d'infirmières spécialisées. Cette équipe collabore au sein de tous les services et laboratoires de l'établissement et bénéficie du soutien de correspondants régionaux.

Le **Laboratoire d'Oncologie Moléculaire** appartient à l'unité d'oncogénétique qui comprend :

- Une structure de consultation d'oncogénétique où sont reçues les familles chez lesquelles une prédisposition héréditaire aux cancers est suspectée. Ceci permet un conseil génétique et une prise en charge médicale adaptée à tous les membres d'une famille considérée à risque.
- Le Laboratoire de Diagnostic Génétique et Moléculaire qui assure le diagnostic d'altération germinale des gènes impliqués dans le risque héréditaire de cancer du sein.
- Le Laboratoire d'Oncologie Moléculaire qui met en œuvre les techniques de biologie moléculaire permettant d'analyser les fonctions des gènes pouvant être à l'origine du développement d'un cancer.

Dans ce laboratoire, une équipe de chercheurs essaie de comprendre le mécanisme d'action des gènes prédisposant aux cancers, notamment du sein et de l'ovaire.

1.2 Domaine d'application

1.2.1 Le cancer du sein

Le cancer est une maladie liée à la prolifération incontrôlée de cellules de l'organisme devenues anormales, dites cellules malignes. Le cancer touche environ 10 millions de personnes dans le monde et constitue l'une des premières causes de mortalité dans les pays développés.

Les travaux de cette thèse concernent plus particulièrement le **cancer du sein** qui est le cancer le plus fréquent chez la femme. On compte 43000 nouveaux cas et 11000 décès par an en France.

Plusieurs **facteurs de risque** sont susceptibles de jouer un rôle dans l'apparition du cancer du sein, notamment les facteurs hormonaux : il est actuellement bien établi que l'âge des premières règles, de la ménopause et l'âge à la première grossesse ont une influence sur l'apparition d'un cancer du sein. Un antécédent de cancer du sein dans une famille augmente également le risque de chaque apparenté de développer un cancer du sein au cours de sa vie. Une alimentation riche en graisses et la consommation d'alcool favoriserait enfin l'apparition d'un cancer du sein.

Plusieurs **facteurs pronostics** classiques sont aujourd'hui pris en compte comme

l'envahissement ganglionnaire axillaire, le grade histopronostique, la taille tumorale histologique ou les récepteurs hormonaux. Ainsi, la mise en place de moyens de dépistage et de prévention permet un diagnostic des tumeurs mammaires à un stade précoce, qui permet de définir ensuite une orientation thérapeutique.

Quatre principaux **moyens thérapeutiques** sont employés pour traiter le cancer du sein :

- La chirurgie : Si le diagnostic est suffisamment précoce, l'acte chirurgical est de plus en plus souvent limité à l'ablation du morceau de sein qui porte la tumeur, on parle de tumorectomie. Lorsque la tumeur est plus volumineuse, une mastectomie, c'est-à-dire l'ablation de la totalité du sein, est nécessaire. Ce traitement chirurgical est complété localement par une radiothérapie externe.
- La chimiothérapie : Les médicaments chimiothérapeutiques sont des médicaments interférant dans le métabolisme et la vie cellulaire et qui, de ce fait, sont cytotoxiques, c'est par ce mécanisme que la chimiothérapie permet d'inhiber la croissance tumorale. Les produits utilisés les plus fréquemment sont les anthracyclines, le cyclophosphamide, le 5-fluoro-uracile et, plus récemment, la vinorelbine et les taxanes.
- La radiothérapie : La radiothérapie tient une place importante dans le traitement loco-régional des cancers infiltrants du sein, soit associée à la chirurgie à titre pré ou post-opératoire, soit seule. Elle est indispensable après traitement chirurgical conservateur. L'irradiation de base englobe la totalité du sein.
- L'hormonothérapie : Longtemps représentée par la castration chirurgicale ou radiothérapique, l'hormonothérapie soustractive est actuellement le plus souvent réalisée. L'hormonothérapie additive est également possible avec le tamoxifène, qui est la molécule le plus couramment utilisé que ce soit en phase adjuvante ou métastatique.

Les progrès des connaissances réalisés ces dernières années sur le cancer du sein ont permis d'étendre les indications de traitement conservateur, de diminuer les indications d'irradiation, de développer une hormonothérapie avec peu de séquelles et de limiter les effets secondaires de la chimiothérapie. Néanmoins, le cancer du sein reste un cancer très imprévisible et hétérogène d'un malade à l'autre. Afin d'adapter les traitements au pronostic de la tumeur, des indicateurs biologiques sensibles sont nécessaires.

L'**analyse du transcriptome des tumeurs** par puces à ADN est une piste actuellement explorée par de nombreuses équipes, l'objectif global étant de mieux comprendre le mécanisme d'action des gènes prédisposant au cancer du sein, d'améliorer le diagnostic et le pronostic de la pathologie (c'est-à-dire prévoir le développement et la progression des tumeurs) et étudier la sensibilité des patientes aux différentes drogues afin de mieux adapter les traitements thérapeutiques.

Ce travail de thèse s'est essentiellement articulé autour de deux principaux projets dont l'objectif global est d'améliorer le diagnostic et le pronostic du cancer du sein et de mieux cibler les traitements grâce à la technologie des puces à ADN.

La première étude concerne un facteur pronostic important qu'est l'envahissement

ganglionnaire. Le second projet vise à analyser la sensibilité des patientes à un traitement chimiothérapeutique particulier.

1.2.2 Etude de l'envahissement ganglionnaire

Nous avons vu précédemment qu'un facteur de pronostic important du cancer du sein, consiste à déterminer si des métastases sont présentes dans les ganglions lymphatiques, on parle alors d'envahissement ganglionnaire.

L'une des méthodes jusqu'ici couramment employées pour caractériser l'envahissement ganglionnaire consiste à effectuer un curage axillaire. Il s'agit alors de prélever des ganglions axillaires (i.e. des aisselles) et de réaliser des coupes de ces ganglions en coloration hématoxyline et éosine. Toutefois, cette méthode est responsable de la plupart des symptômes fonctionnels post-chirurgicaux du cancer du sein. En effet, les conséquences sont d'une part, l'atteinte du drainage lymphatique du sein et du bras du côté opéré et d'autre part, l'augmentation de la sensibilité aux infections au niveau du bras. De plus, la proportion des tumeurs dépourvues d'envahissement ganglionnaire (N-) approche les 80% dans les pays occidentaux, ce qui veut dire que près de 4 femmes sur 5 subissent un curage axillaire inutile. Enfin, La classification des tumeurs mammaires en stade N- (pas d'envahissement ganglionnaire) ou N+ (envahissement ganglionnaire) est d'autant plus exact que le nombre de ganglions prélevés est élevé, ainsi les résultats présentent un certain risque d'erreur.

Une alternative au curage axillaire est l'étude du ganglion sentinelle. Ce dernier est défini comme le premier ganglion qui reçoit le flux lymphatique de la tumeur primaire. Cette technique est donc une méthode efficace pour évaluer le statut ganglionnaire axillaire et éviter le curage chez les patientes ayant un cancer du sein de petite taille sans envahissement ganglionnaire clinique. En effet, un ganglion sentinelle non atteint exclut en théorie un envahissement ganglionnaire métastatique du reste de l'aisselle et rend donc le curage inutile. La technique du ganglion sentinelle n'est toutefois pas encore standardisée et les taux de faux négatifs ne sont pas négligeables.

L'objectif de cette étude est d'étudier la corrélation entre l'expression des gènes des tumeurs mammaires et le statut ganglionnaire des patientes. L'idée à terme serait de proposer un nouvel outil de diagnostic simple à mettre en œuvre et peu coûteux permettant ainsi d'éviter les curages axillaires.

1.2.3 Etude de la sensibilité au docétaxel

Nous avons vu précédemment que les taxanes sont des médicaments chimiothérapeutiques utilisés dans le traitement du cancer du sein. Les taxanes sont des dérivés de plantes

très actifs sur les tumeurs du sein mais ils sont relativement toxiques : neuropathies, aplasies parfois profondes et risque de survenue d'œdèmes de membres voire des séreuses.

Les taxanes comprennent le **docétaxel**, qui a d'abord été indiqué dans le traitement de cancers du sein en monothérapie, puis en association avec les anthracyclines. Son efficacité démontrée, le docétaxel a été ensuite incorporé dans le traitement des stades précoces en adjuvant (chimiothérapie réalisée après chirurgie) et en néoadjuvant (chimiothérapie réalisée avant chirurgie). Le docétaxel semble être le seul taxane à démontrer un bénéfice en terme de survie sans rechute, quel que soit le statut hormonal de la patiente.

Dans les protocoles de chimiothérapie néoadjuvante, l'administration du docétaxel après des anthracyclines permet une amélioration de la survie, une meilleure réponse clinique entraînant un plus grand nombre de traitements conservateurs et une augmentation du taux de réponse complète histologique. L'efficacité du docétaxel est donc démontrée dans les formes avancées et précoces du cancer du sein.

Le plus souvent, la chimiothérapie est administrée par voie systémique, ce qui explique que la toxicité qu'elle induit soit générale. Une conséquence importante liée à la propriété principale de ce type de produits sur la division cellulaire, est d'affecter de la même façon les cellules normales et plus particulièrement les tissus présentant un taux de renouvellement important. Compte tenu de la cytotoxicité liée à l'utilisation d'un anticancéreux tel que le docétaxel, il est nécessaire d'identifier les patientes qui peuvent bénéficier de ce type de traitement. En effet, les cellules tumorales peuvent être résistantes à un traitement de chimiothérapie, lequel traitement s'avère alors inutile et toxique pour la patiente.

Comme moyen de prédiction, on peut citer par exemple l'étude in vitro des modifications cellulaires causées par un agent anticancéreux, qui se fait sur des cellules isolées qui, après incubation, peuvent être altérées par la présence d'un anticancéreux, ou bien au contraire résistantes à cet anticancéreux. Toutefois, ces conditions sont loin de celles des traitements in vivo.

L'objectif de cette étude est d'étudier la corrélation entre l'expression des gènes des tumeurs mammaires et la résistance au docétaxel. L'idée à terme serait de proposer un nouvel outil qui permette de déterminer au préalable si une patiente sera sensible ou bien au contraire résistante au docétaxel.

1.3 Contributions

La première partie de ce travail de thèse a consisté à pré-traiter et à analyser l'ensemble des données d'expression pour les différents projets présentés, en assurant une veille technologique. La seconde partie plus prospective, a permis la mise en place d'une nouvelle technique de reconstruction de réseaux de gènes basée sur la notion de règles entre gènes.

1.3.1 Pré-traitement et analyse des données d'expression

Le processus d'extraction de connaissances à partir des données d'expression est décrit dans la figure 1.3. Il consiste d'une part à analyser les images et à normaliser les données et d'autre part à appliquer des méthodes statistiques et informatiques pour obtenir de la connaissance utile pour les experts.

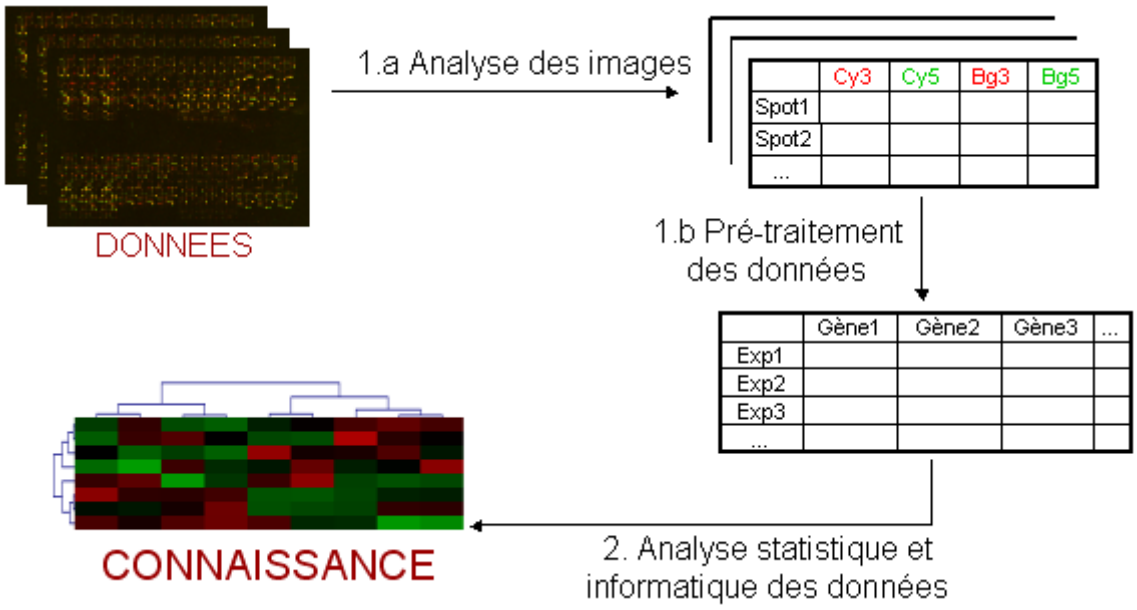


FIGURE 1.3 – Traitement des données d'expression

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

Pour les biopuces TM/CS DIAGNOGENE utilisées pour l'étude du statut ganglionnaire, différents outils et méthodes ont été comparés aux différentes étapes pour arriver à un processus optimisé dont les points clés sont les suivants :

- Scanning des images à plusieurs réglages du scanner.
- Pour chaque spot, choix du "meilleur" réglage.
- Normalisation selon la méthode du LOWESS [35, 81].
- Test de Student pour détecter des outliers parmi les réplicats d'un gène.

L'analyse des images a été réalisée sous Genepix Pro 6.0. (Molecular Devices [7]) et la méthode du LOWESS a été exécutée sous MIDAS [83] de TIGR [11]. Les autres points ont été implémentés en Visual Basic pour en faciliter l'accès aux utilisateurs. Ce nouvel outil offre donc l'originalité de considérer plusieurs réglages de scanner et la possibilité de

détecter des outliers (le détail de ce processus est donné dans la section 5.1).

Une fois les données pré-traitées, la seconde étape consiste à **analyser les données** pour en extraire de la connaissance. Différentes méthodes peuvent être appliquées aux données d'expression de gènes, les plus couramment utilisées dans le cadre de cette thèse sont :

- La recherche de gènes différentiellement exprimés entre différentes conditions expérimentales grâce notamment aux tests d'hypothèse.
- Le regroupement d'échantillons ou de gènes ayant des profils d'expression similaires grâce aux méthodes de clustering.
- La visualisation de données en dimension réduite grâce par exemple à l'Analyse en Composantes Principales.
- La prédiction de classe à partir d'exemples d'apprentissage.

Le choix des méthodes dépend toujours de l'objectif global de l'étude. Divers logiciels dédiés ou non aux données d'expression permettent de réaliser ces analyses, les plus couramment utilisés au cours de cette thèse sont MultiExperimentViewer (MeV) [83] de TIGR [11], GeneSpring GX (Agilent [1]), SAS [10] et WEKA [99].

Dans le cadre des projets présentés précédemment, plusieurs analyses ont été réalisées et ont permis d'identifier des gènes permettant de mettre en évidence un profil d'expression caractéristique du statut ganglionnaire et de la réponse thérapeutique au docétaxel (cf section 5.1 et 5.2).

Pour la première étude, ce travail a contribué au dépôt du brevet français n°FR2875512 et du brevet international n°WO2006032769 : "Biopuce de diagnostic du caractère métastatique ou localisé d'une tumeur mammaire, procédé et kit d'utilisation". Dans le cadre du second projet, ce travail a contribué au dépôt du brevet français n°FR2874622 et du brevet international n°WO2006027463 : "Biopuce de détermination d'une sensibilité à un médicament anticancéreux et son procédé d'utilisation".

1.3.2 Des règles aux réseaux de gènes

Nous avons vu qu'un des défis majeurs fixé à présent et rendu possible notamment grâce à la technologie des puces à ADN, est de découvrir les **interactions** entre les différents gènes.

Dans le cadre de ce travail de thèse, nous nous sommes intéressés plus particulièrement à la découverte de règles entre gènes, à partir de données d'expression. La notion de **règles** entre attributs est très générale, allant des règles d'association en fouille de données aux dépendances fonctionnelles en bases de données. Malgré cette diversité, la **syntaxe** des

règles est toujours la même : $X \rightarrow Y$ avec X et Y deux ensembles d'attributs, seule leur **sémantique** i.e. leur signification diffère.

Compte tenu des différents objectifs biologiques possibles, il peut être utile de découvrir différents types de règles entre gènes et ne pas se restreindre à une seule sémantique. L'idée de notre approche est alors d'offrir aux biologistes la possibilité de choisir parmi plusieurs sémantiques, le sens des règles qu'ils souhaitent générer.

L'originalité de ce travail est de proposer un cadre global pouvant inclure un grand nombre de sémantiques pour les règles et d'utiliser des méthodes identiques de génération, de post-traitement et de visualisation des règles pour toutes les sémantiques proposées.

Nous présentons tout d'abord les principaux éléments caractérisant une sémantique puis nous donnons une définition de la satisfaction des règles pour les sémantiques ainsi spécifiées [18, 19].

Nous introduisons ensuite la notion de "sémantiques bien-formées" [16]. Il s'agit des sémantiques pour lesquelles les axiomes d'Armstrong sont justes et complets. Les intérêts pratiques d'une sémantique bien-formée se situent à la fois sur le raisonnement qui est rendu possible sur les règles mais aussi sur la définition des couvertures des règles et des possibilités algorithmiques qui en découlent lorsque l'on s'intéresse à leur découverte à partir des données.

Nous donnons également un résultat important permettant de savoir simplement si une sémantique est ou non bien-formée, et ceci grâce à des restrictions syntaxiques [15, 14].

A partir de données d'expression, nous proposons de découvrir différents types de règles entre gènes [13, 17]. Cette approche constitue une nouvelle méthode de reconstruction de réseaux de gènes à partir de données d'expression, les règles peuvent en effet être visualisées sous forme de graphes [12]. L'originalité de notre approche est d'une part de superposer des règles avec des sémantiques différentes au sein d'un même support visuel et d'autre part de ne générer que les règles qui impliquent des gènes dits "centraux". Ceux-ci sont spécifiés en amont par les experts et permettent de limiter la génération des règles aux seuls gènes qui intéressent les spécialistes.

L'approche proposée a été développée spécifiquement pour les données d'expression dans un nouveau module appelé RG (Rule Generation), intégré à un logiciel gratuit et open-source consacré à l'analyse de données d'expression de gènes, le logiciel MeV (MultiExperimentViewer) [83].

Finalement, des expérimentations ont été menées dans le cadre des projets présentés dans la section précédente. Ce travail s'inscrit dans le cadre du projet GeneRules, débuté en 2004, dont le site web est le suivant : http ://www.isima.fr/agier/GeneRules.

1.4 Organisation du mémoire

Dans la partie I, nous introduisons tout d'abord les notions biologiques nécessaires à la compréhension du mémoire puis nous présentons les données d'expression de gènes. Nous décrivons ensuite le processus d'extraction de connaissances à partir de ces données d'expression, qui comporte deux principales étapes : le pré-traitement et l'analyse des données. Enfin, nous donnons quelques résultats obtenus dans le cadre des étude sur l'envahissement ganglionnaire et la sensibilité au docétaxel.

Nous présentons dans la partie II notre approche basée sur la découverte de différents types de règles entre gènes. Nous présentons tout d'abord le cadre général de l'approche proposée permettant d'inclure un grand nombre de sémantiques pour les règles. Nous présentons ensuite différentes sémantiques spécialement adaptées aux données d'expression de gènes. La dernier chapitre introduit la méthode utilisée de génération des règles.

Dans la partie III, nous faisons le lien entre l'approche proposée basée sur la notion de règles et la reconstruction de réseaux de gènes à partir de données d'expression. Nous présentons ensuite l'outil réalisé et enfin les réseaux obtenus à partir des données de tumeurs mammaires.

Finalement, nous concluons ce travail et donnons quelques perspectives dans la partie IV.

Première partie

Pré-traitement et analyse des données d'expression

Chapitre 2

Du génome au transcriptome

Sommaire

Nous présentons tout d'abord le passage du génome (regroupant l'ensemble des gènes d'une cellule) au transcriptome (regroupant l'ensemble des gènes **actifs** dans cette cellule). Nous expliquons ensuite le principe des biopuces ou puces à ADN, permettant d'étudier l'activité des gènes. Enfin nous donnons quelques plans expérimentaux possibles pour ce type d'études.

2.1 L'Information Génétique

Un être vivant est composé de milliard de cellules, qui migrent et se spécialisent au cours du développement embryonnaire, c'est-à-dire qu'elles acquièrent une forme et une structure particulières en relation avec une fonction donnée. La spécialisation des cellules dépend du programme génétique dont une partie seulement s'exprimera selon la spécialisation de la cellule. Le programme génétique dirige l'ensemble de la construction fonctionnelle de l'organisme et détermine l'identité de chaque individu. Il est situé dans le noyau de chaque cellule et est porté par les chromosomes (cf figure 2.1).

L'ADN (Acide DésoxyriboNucléique) est une macromolécule (la plus grosse du vivant) constituée de deux chaînes de nucléotides torsadées en double hélice. Chaque nucléotide résulte de l'association d'un sucre, d'un groupement phosphate et d'une base azotée. Seules quatre bases sont utilisées, ne donnant ainsi naissance qu'à quatre nucléotides : l'Adénine (A), la Thymine (T), la Guanine (G) et la Cytosine (C). Les bases sont complémentaires

deux à deux, l'Adénine ne peut se lier qu'à la Thymine et la Guanine qu'à la Cytosine (cf figure 2.2).

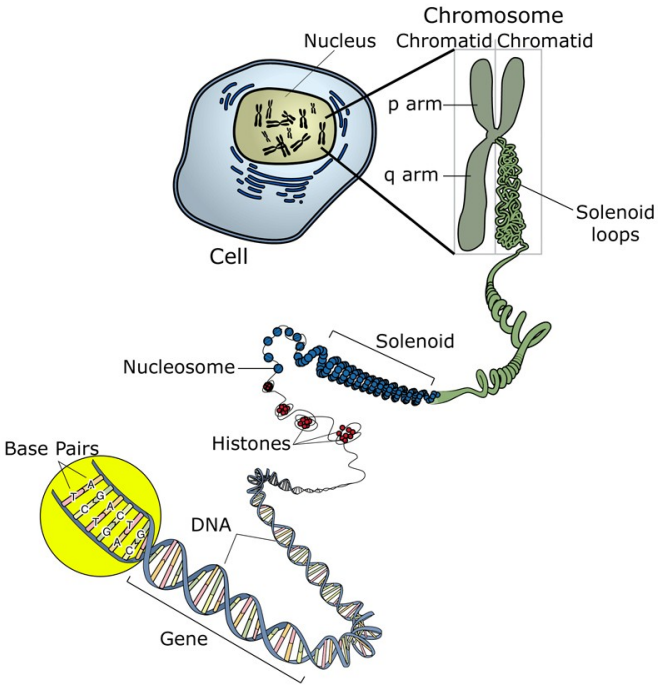


FIGURE 2.1 – L'information génétique au sein d'une cellule [4]

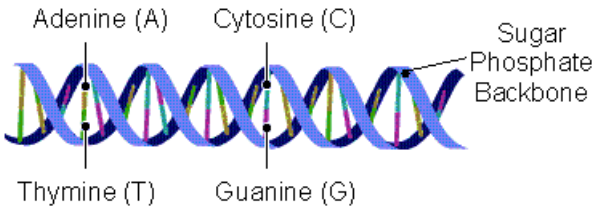


FIGURE 2.2 – Composition de l'ADN

La spécification de chaque molécule d'ADN réside dans l'ordre d'enchaînement des différentes bases qui constitue la séquence des bases. L'information génétique réside dans cette séquence, elle est écrite dans un langage codé à l'aide d'un alphabet à 4 lettres $\{A, T, G, C\}$. Le génome humain compte environ 3,3 milliards de bases. Pour une espèce donnée, la séquence de la molécule d'ADN est globalement semblable. Il existe cependant des variations entre les séquences d'ADN de chaque individu, il s'agit des polymorphismes au nombre de 3 millions environ dans l'espèce humaine. Ceux-ci peuvent être silencieux ou bien se manifester par la modification d'un caractère observable ou dans certains cas être à l'origine de dysfonctionnement grave de l'organisme.

Le gène est l'unité de base de l'information génétique, il code pour une protéine res-

ponsable d'un caractère. Les gènes (environ 30 000 chez l'homme) sont disséminés le long du double brin d'ADN, ils sont donc caractérisés par une séquence nucléotidique. Les gènes ont la particularité d'être morcelés en de nombreux fragments informationnels appelés exons, séparés par des morceaux de séquences appelés introns, dont le rôle est encore inconnu (cf figure 2.3).

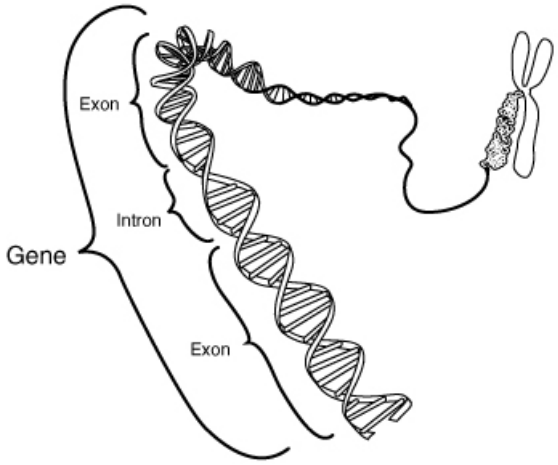


FIGURE 2.3 – Composition d'un gène [8]

L'ensemble des séquences des gènes mises bout à bout représente moins de 2% de la totalité de la molécule d'ADN. L'ensemble de ces gènes constitue le **génome** des êtres vivants.

Toutes les cellules d'un organisme possèdent la même information génétique i.e. le même génome, mais selon leurs besoins, elles n'expriment qu'une partie de cette information pour réaliser la synthèse des protéines indispensables aux grandes fonctions cellulaires.

Le passage du gène à la protéine, du noyau vers le cytoplasme de la cellule, se déroule en 2 étapes : la transcription et la traduction (cf figure 2.4).

1. **La transcription** : Selon les besoins de la cellule, **plusieurs copies** des gènes nécessaires à son bon fonctionnement, vont être produites sous forme d'ARNm (Acide RiboNucléique messager). L'ARNm est aussi constitué de quatre bases, trois sont similaires à celles de l'ADN : l'Adénine (A), la Guanine (G) et la Cytosine (C), la quatrième étant l'Uracile (U). Cette séquence d'ARNm se forme selon la complémentarité A/U et C/G. L'ensemble des séquences des gènes **exprimés** représente le **transcriptome**, qui est le reflet de l'activité cellulaire à un instant donné.
2. **La traduction** : La séquence des nucléotides de l'ARNm est ensuite convertie grâce en particulier aux ribosomes, en une séquence d'acides aminés pour constituer une protéine. Un même gène peut produire plusieurs protéines, on suppose ainsi que probablement un million de protéines peuvent être produites dans les cellules humaines. L'ensemble des protéines à un instant donné correspond au **protéome**,

qui assure le développement, la croissance et le fonctionnement de la cellule (donc de l'organisme).

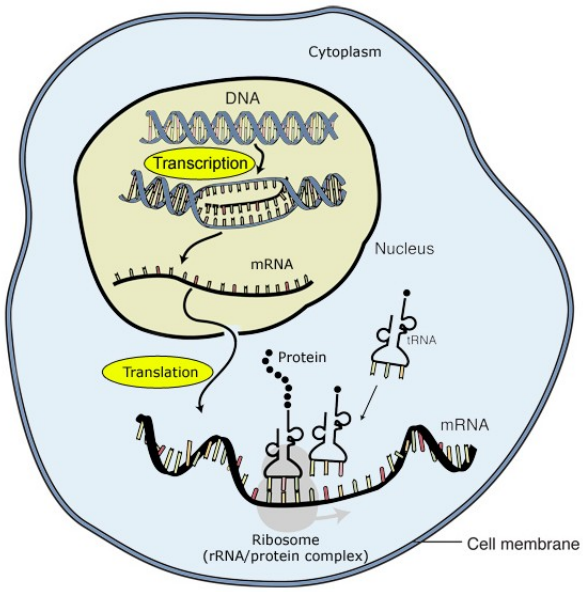


FIGURE 2.4 – Etapes de transcription et de traduction [4]

2.2 Les Biopuces ou Puces à ADN

Etudier l'ensemble du transcriptome d'une cellule est d'un enjeu considérable afin de mieux comprendre les mécanismes fonctionnels de l'organisme. Dans le passé, les biologistes étaient capables de mesurer le niveau d'expression (i.e. le nombre de transcrits) d'un petit nombre de gènes à la fois. La technologie des biopuces [87] leur permet maintenant d'étudier des milliers de gènes simultanément, une avancée qui leur permettra de déterminer les relations complexes entre les gènes.

Les applications des biopuces sont très nombreuses. En effet, il est possible aujourd'hui de connaître la dynamique du transcriptome et des réseaux génétiques impliqués, on peut aussi classifier par exemple des tumeurs selon leur signature moléculaire. Inversement, il est possible d'explorer les gènes pour connaître la fonction d'un gène inconnu. Ces connaissances peuvent servir ensuite à une meilleure compréhension des maladies et à l'élaboration de nouveaux médicaments.

Le principe général des biopuces est illustré dans la figure 2.5 et détaillé dans la suite.

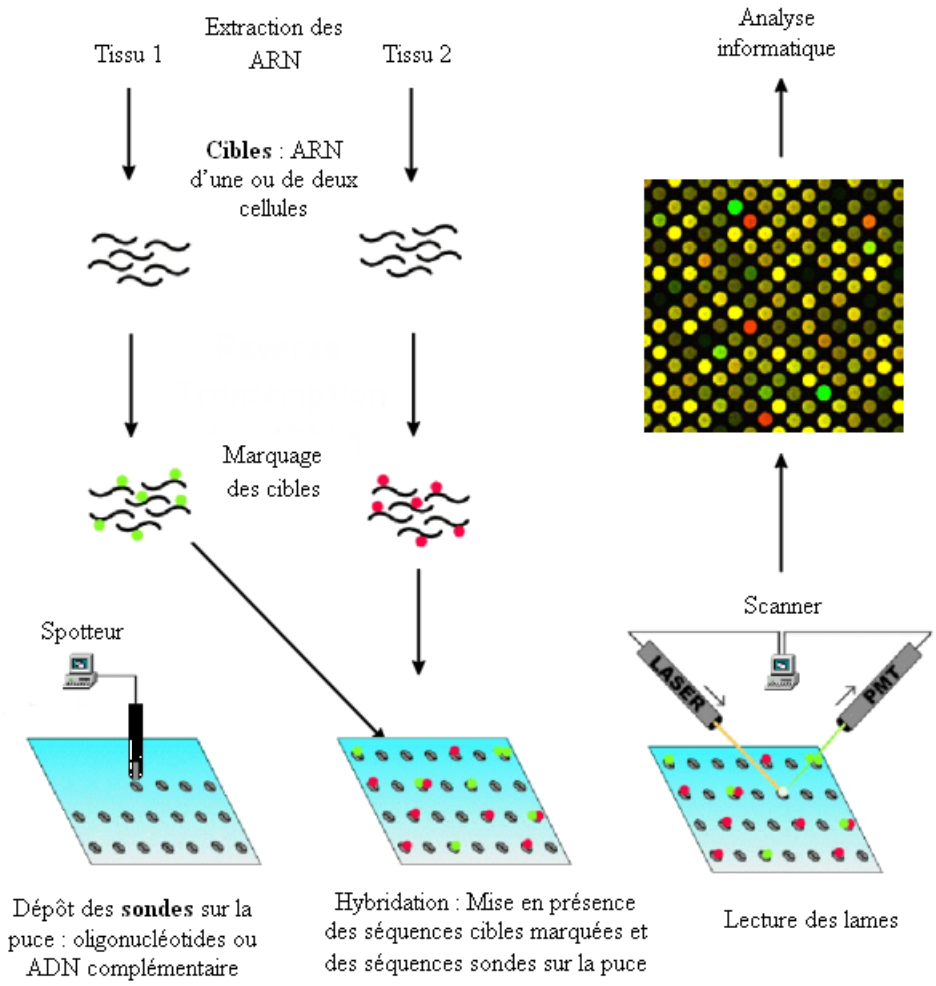


FIGURE 2.5 – Principe des biopuces [9]

2.2.1 Mode de Fabrication des Biopuces

– **Choix et dépôt des sondes moléculaires** La première étape de fabrication des puces consiste à choisir les gènes que l'on souhaite étudier, ce choix se faisant en fonction notamment de l'espèce étudiée et des différents objectifs de l'étude. Ensuite, on ne dépose pas la séquence entière des gènes sur la lame mais simplement des séquences de nucléotides représentatifs de ces gènes, appelées sondes moléculaires. Différentes méthodes algorithmiques sont développées pour assurer la spécificité de la séquence déposée. Ces séquences peuvent être de différentes tailles, les oligonucléotides sont de courtes séquences de quelques dizaines de bases, des séquences plus importantes (500 à 1000 bases) peuvent également être déposées, il s'agit alors d'ADN complémentaire.
Deux principaux types de puces sont proposées actuellement, les puces dites *pangénomiques*, regroupant des dizaines de milliers de dépôts sur la lame (par exemple l'ensemble du génome humain), ou bien des puces dites *puces à façon* qui regroupent quelques centaines de gènes caractéristiques par exemple de la pathologie étudiée.
Une fois les sondes choisies, l'appareil de dépôt, le "spotteur", les dépose sur le support, par exemple une lame de verre (type lame de microscope) recouvert d'une surface réactive. Pour chaque gène, un nombre suffisant de sondes sera déposé de façon à "supporter" l'ensemble des transcrits du gène. De plus, certaines gènes peuvent être déposés à différents endroits sur la lame, il s'agit alors de réplicats techniques qui permettent de s'assurer de la qualité des données.

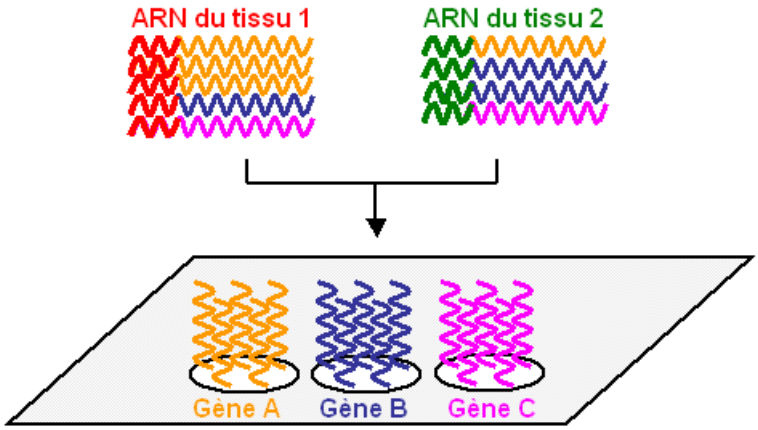


FIGURE 2.6 – Préparation des sondes et des cibles

– **Extraction et marquage des ARN cibles** Parallèlement, des ARN dont la qualité a été préalablement vérifiée, sont extraits d'un ou de deux tissus que l'on souhaite étudier (les cibles) puis marqués à l'aide d'un fluorochrome, en général Cyanine 5 (Cy5) et Cyanine 3 (Cy3) qui génèrent des signaux fluorescents respectivement rouges et verts (cf figure 2.6). Sur une même puce, il est en effet possible d'étudier un ou deux échantillons, on parle alors respectivement de *simple couleur* et de *double couleur*. Cette dernière permet de comparer les profils d'expression de deux tissus, par exemple un tissu sain et un tissu pathologique.

– **Mise en contact des sondes et des cibles** Une fois les sondes déposées, la lame est mise en contact avec le transcriptome du ou des tissus choisis, il s'agit de l'étape d'hybridation. Pendant cette phase, les gènes exprimés dans le tissu vont s'hybrider avec les séquences correspondantes déposées sur la lame, par le principe de complémentarité des bases.
Lorsque deux cellules sont analysées simultanément, cette technique permet de comparer le nombre de copies d'un gène dans les deux échantillons. Par exemple dans la figure 2.7, nous pouvons constater que le gène A est 3 fois plus exprimé dans le tissu 1 que dans le tissu 2, le gène B est 2 fois plus exprimé dans le tissu 2 que dans le tissu 1, tandis que le gène C est autant exprimé dans les deux tissus.

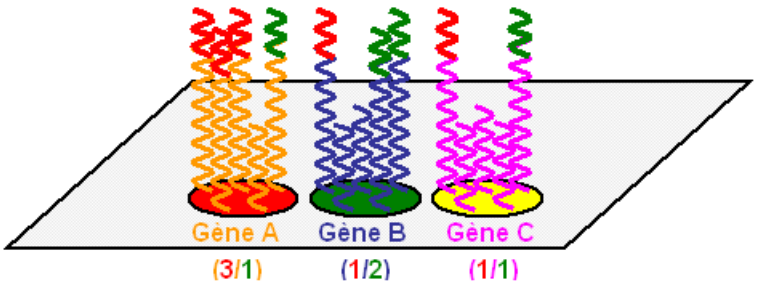


FIGURE 2.7 – Mise en contact des sondes et des cibles

– **Lecture des lames par le scanner** Après un lavage minutieux, les lames sont lues par un scanner qui permet de visualiser les différents niveaux d'expression. En effet, plus le nombre de copies d'un gène est important dans un tissu, plus le signal fluorescent émis par le marqueur sera important. Par exemple dans la figure 2.7, pour le gène A, le signal émis par le marqueur du tissu 1 (i.e. le marqueur rouge) sera trois fois plus important que le signal émis par le marqueur du tissu 2 (i.e. le marqueur vert), ainsi le **spot** correspondant sur la lame sera plutôt de couleur rouge. De la même manière, le spot correspondant au gène B sera plutôt vert et le spot correspondant au gène C sera jaune. L'image produite par le scanner, avec des spots à différents niveaux de rouge et de vert, est ensuite analysée par un logiciel d'analyse d'images, permettant à partir des intensités des spots de quantifier l'expression des gènes. Cette étape est développée dans le chapitre 3.

2.2.2 Plans expérimentaux

Les biopuces permettent donc d'identifier les gènes qui sont actifs dans une ou plusieurs cellules. Cette technique permet par exemple d'observer le transcriptome d'une cellule à différents instants, pour identifier les délais d'activation des gènes. Il est également possible d'étudier le transcriptome de différents types de tumeurs pour déterminer les gènes qui ne s'expriment pas de la même façon dans toutes les tumeurs, etc. En fonction du problème traité, différents plans expérimentaux peuvent alors être envisagés.

Prenons un cas classique où l'étude consiste à analyser le transcriptome de deux types d'échantillons (par exemple cellules saines vs cellules pathologiques, régime alimentaire A vs régime alimentaire B...). Notons que dans la suite, nous emploierons le terme général d'**échantillons** plutôt que de cellules ou tissus. L'objectif est d'identifier les gènes différentiellement exprimés entre ces deux conditions c_1 et c_2 i.e. découvrir les gènes dont le niveau d'expression est différent entre les échantillons de la condition c_1 et les échantillons de la condition c_2. Notons n_1 (resp. n_2) le nombre d'échantillons disponibles pour la condition c_1 (resp. c_2), plusieurs plans expérimentaux sont alors envisageables :

- **Plan expérimental n°1** : Si $n_1 = n_2 = 1$ i.e. un seul échantillon est disponible pour chaque condition alors une biopuce sera réalisée avec le premier échantillon marqué en Cy5 (rouge) et le second échantillon marqué en Cy3 (vert).

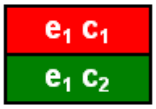


FIGURE 2.8 – Plan expérimental n°1

- **Plan expérimental n°2** : Si $n_1 = n_2 = n$ alors il est possible de réaliser n biopuces comme précédemment avec en Cy5 un échantillon de la condition c_1 et en Cy3 un échantillon de la condition c_2. Dans ce cas, le problème est de savoir quels échantillons coupler. Si chaque échantillon de la condition c_1 est "lié" à un échantillon de la condition c_2, par exemple cellule saine vs cellule malade d'un même individu, avec n individus, alors ce peut être un plan intéressant. Toutefois, la spécificité des individus entrera en compte dans les résultats.

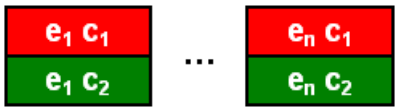


FIGURE 2.9 – Plan expérimental n°2

- **Plan expérimental n°3** : Une autre possibilité est de réaliser $n_1 + n_2$ biopuces où chaque biopuce est réalisée avec un des $n_1 + n_2$ échantillons en Cy5 et une référence commune en Cy3. Ce plan présente l'avantage d'avoir une référence identique à toutes les biopuces, ce qui assure d'une part d'avoir des données comparables et d'autre part de "gommer" certains artefacts dus à la technologie (cf section 3.1.2). Toutefois, se pose le problème du choix de la référence.
  Deux types de références peuvent être envisagées : Une référence commerciale universelle, qui garantit que l'expression des gènes dans la référence sera toujours identique au cours du temps et pour toutes les expérimentations (et éventuellement entre plusieurs plates-formes) ou bien une référence fabriquée spécialement pour l'étude, composée d'un pool d'ARN. Le choix de la référence va dépendre notamment de la

durée de l'étude car s'il s'agit d'une étude à long terme (par exemple à but diagnostic) où des échantillons seront régulièrement ajoutés, la deuxième solution ne sera pas pertinente.

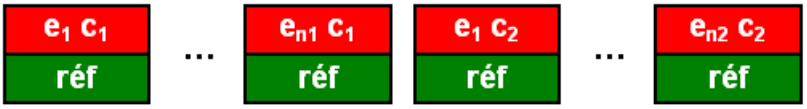


FIGURE 2.10 – Plan expérimental n°3

– **Plan expérimental n°4** : Enfin, il est possible de réaliser n_1+n_2 biopuces en simple couleur, où chaque biopuce est réalisée avec un des $n_1 + n_2$ échantillons en Cy5. La simple couleur apporte tout d'abord l'avantage de ne pas avoir à faire ce choix de la référence, qui est toujours très fastidieux. Il permet également de travailler directement sur des intensités et non sur des ratios d'intensités (cf section 3.1.2). Toutefois, comme il n'y a pas de référence commune à toutes les biopuces (comme dans le cas précédent), il conviendra de renforcer les contrôles qualité.

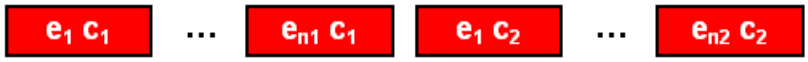


FIGURE 2.11 – Plan expérimental n°4

Sans être exhaustif, cet exemple montre l'importance et la difficulté de définir correctement le plan expérimental [24, 96]. Il faut bien voir que chaque plan expérimental a ses avantages et ses inconvénients et qu'il n'existe pas toujours de plan optimal. Des études préconisent également un nombre minimal de biopuces en fonction de l'étude réalisée [105, 65], toutefois il est évident que plus ce nombre sera important, plus les résultats seront significatifs. D'autre part, des paramètres supplémentaires sont à prendre en compte : Par exemple, pour les plans expérimentaux n°3 ou n°4, il conviendra de ne pas réaliser les n_1 biopuces correspondant à la condition c_1 un jour et les n_2 autres biopuces un autre jour, ou bien les n_1 premières par un premier manipulateur et les n_2 autres par un second manipulateur car il sera difficile de déterminer de quel facteur résulte la variabilité de l'expression des gènes. Cette étape est réellement à établir avec attention car il sera difficile de faire marche arrière une fois les hybridations lancées.

Notons également que d'autres expérimentations peuvent être réalisées afin d'ajouter un contrôle qualité supplémentaire. Par exemple, les réplicats de biopuces i.e. avec les mêmes échantillons réalisés à des jours différents, permettent de tester la reproductibilité des expérimentations. Les dye-swap (i.e. l'échantillon 1 en Cy3 vs l'échantillon 2 en Cy5 et inversement) permettent de contrôler les effets des deux marqueurs (cf section 3.1.2).

Chapitre 3

Pré-traitement des données d'expression

Sommaire

Le processus d'extraction de connaissances à partir des données d'expression se décompose en deux principales étapes, le pré-traitement des données permettant d'assurer la fiabilité des résultats fournis puis l'analyse statistique et informatique des données permettant de donner du sens à toutes ces informations. Diverses méthodes d'analyse de données sont particulièrement utilisées pour les données d'expression [80], nous les développons plus en détail dans le chapitre 4.

Auparavant, nous décrivons le processus de pré-traitement des données d'expression. Ce processus commence par l'analyse des images générées par le scanner, puis regroupe diverses méthodes permettant de corriger les biais qui ont pu apparaître au cours des expérimentations.

3.1 Des images aux données numériques

Nous avons vu dans la section 2.2.1, le principe de fabrication des biopuces. Au terme des différentes étapes, le scanner produit une image avec des spots de différentes intensités en fonction du niveau d'expression des gènes correspondants. L'intensité des spots est alors quantifiée à l'aide d'un logiciel d'analyse d'images.

Ensuite il faut corriger les données, l'objectif étant de s'assurer de la qualité des données pour chaque biopuce, et d'obtenir des données comparables entre les différentes biopuces. Les variations peuvent se situer à l'intérieur d'une même puce (suivant le marqueur ou l'aiguille utilisés, les conditions d'hybridation...) et entre plusieurs puces (suivant les lots de lame, le jour de dépôt des sondes...).

Comme logiciels de pré-traitement, on peut citer Feature Extraction (Agilent [1]), Spotfinder et MIcroarray Data Analysis System (MIDAS) [83] de TIGR [11], MADSCAN [78], GenePix Pro (Molecular Devices [7]), ImaGene (BioDiscovery [2]), VARAN [49] ou limmaGUI [98].

3.1.1 Analyse des images

Le processus d'analyse des images se décompose en trois étapes, tout d'abord l'**adressage des spots**. Une grille est positionnée sur l'image permettant d'indiquer où se situent les spots (cf figure 3.1). A cette grille est généralement lié l'ensemble des annotations correspondant aux différents gènes i.e. l'identifiant des gènes, leur nom, le n° de l'oligonucléotide...

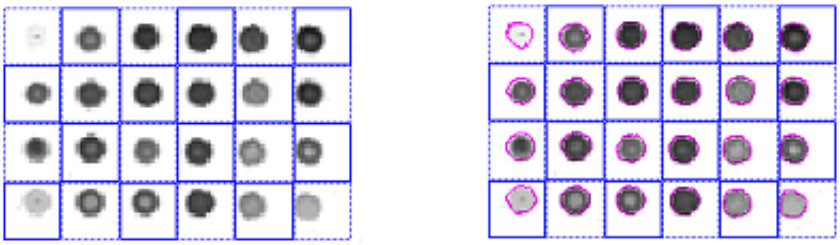


FIGURE 3.1 – Adressage puis segmentation des spots [83]

Ensuite, l'étape de **segmentation des spots** consiste à classifier les pixels comme "signal" ou comme "bruit de fond", c'est-à-dire engendrés par des défauts apparaissant sur la lame. Le calcul du bruit de fond se fait grâce à différents algorithmes, la plupart permettent de le calculer localement puisque le bruit peut varier d'une région à une autre de la lame. Les principaux facteurs de bruit de fond sont les changements physiques et chimiques selon les régions de la lame, la température, l'hybridation non spécifique...

Enfin, le logiciel **quantifie le signal**, c'est-à-dire qu'il détermine l'intensité du signal et du bruit de fond de chaque spot pour le ou les marqueurs fluorescents. Ces intensités vont de 0 à 65525. Pour chaque spot, sont également indiqués d'autres paramètres permettant de déterminer la qualité globale du spot, par exemple, les intensités médiane et moyenne des pixels et du bruit de fond, le diamètre du spot, le nombre de pixels saturés... Les spots peuvent par exemple être entourés de différentes couleurs en fonction de leur qualité (cf figure 3.2).

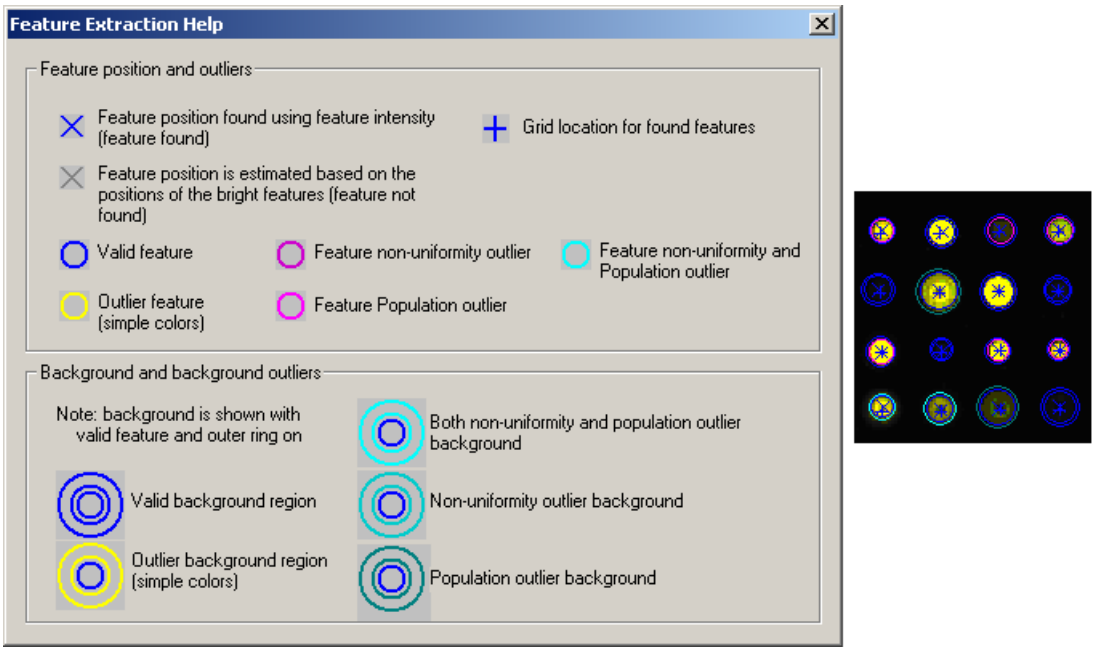


FIGURE 3.2 – Indication de la qualité des spots [1]

Les spots dont la qualité pose problème, i.e. les spots dont l'intensité est trop faible, les spots saturés, les spots non détectés à cause de leur forme... pourront ensuite être supprimés pour les analyses ou bien pris en compte mais avec un certain degré de confiance.

Notons que les images peuvent être scannées à différents réglages du scanner, influant alors sur la qualité des données extraites. L'objectif est d'obtenir le meilleur réglage afin que le minimum de spots soient saturés mais aussi que le maximum de spots soient supérieurs au bruit de fond. Généralement, l'image analysée correspond au meilleur compromis toutefois, il est possible de prendre en compte plusieurs réglages du scanner (cf section 5.1).

Une fois que toutes les informations ont été collectées pour chaque spot, il est souhaitable d'appliquer un pré-traitement aux données générées, l'objectif étant de corriger certains effets néfastes qui apparaissent régulièrement.

3.1.2 Normalisation des données

Une première correction doit être appliquée intra-lames pour corriger les effets dus par exemple aux marqueurs fluorescents, puis une seconde correction inter-puces permet d'obtenir des données comparables [81, 102, 103]. Auparavant, une petite remarque est donnée sur la prise en compte des logarithmes.

Ratios et logarithmes

Dans les approches "double couleur", la valeur considérée pour un gène est le ratio de l'intensité en Cy5 sur l'intensité en Cy3 (ou inversement). Toutefois, comme la distribution des ratios n'est pas une distribution normale, il convient de considérer plutôt les **logarithmes** des ratios qui ont une distribution normale. Par exemple, le passage en log_2 permet d'avoir une symétrie entre les gènes sous- et sur-exprimés.

Dans la table 3.1, sont présentés les intensités de différents gènes. La symétrie apportée par le passage en log est bien visible dans ce tableau. De plus, celui-ci permet de réduire l'influence de grandes valeurs. Par exemple, les gènes g_1 et g_5 auront moins d'influence sur la distribution des logs que sur la distribution des ratios. Les logarithmes sont donc usuellement considérés plutôt que les ratios.

Gènes	g_1	g_2	g_3	g_4	g_5
$Cy5$	100	20000	5000	20000	60000
$Cy3$	60000	40000	5000	10000	100
$Ratio$	0.002	0.5	1	2	600
Log_2	-9.23	-1	0	1	9.23

TABLE 3.1 – Passage en log

Normalisation intra-puces

La plupart des méthodes de normalisation intra-puces consistent à limiter les différences de luminosité des deux marqueurs Cy3 et Cy5. En effet, la Cyanine 3 a tendance à être toujours plus intense que la Cyanine 5, mais cela ne signifie pas que le gène est plus exprimé dans la première cellule. Il faut donc corriger cet effet.

Une méthode simple, nommée **Total Intensity Normalization** consiste à calculer pour chaque biopuce, la somme des intensités pour Cy5 et pour Cy3. Puis le ratio de ces sommes est calculé :

$$r = \sum_i (Cy5_i) / \sum_i (Cy3_i).$$

Ensuite chaque intensité en Cy3 est multiplié par ce ratio :

$$new(Cy3_i) = old(Cy3_i) * r.$$

Ainsi, après cette correction, la somme des intensités en Cy3 est égale à la somme des intensités en Cy5.

Cette méthode permet donc d'effacer le biais introduit par les marqueurs mais calcule un seul facteur multiplicateur r pour tous les spots de la lame. Or diverses études ont montré qu'il y avait une dépendance systématique entre la valeur du log2(ratio) et l'intensité des spots. Cet effet peut se voir notamment dans les RI-plot qui tracent les

valeurs des ratios : $log_2(Cy5/Cy3)$ en fonction des valeurs d'intensité : $log_{10}(Cy5 * Cy3)$ (cf figure 3.3).

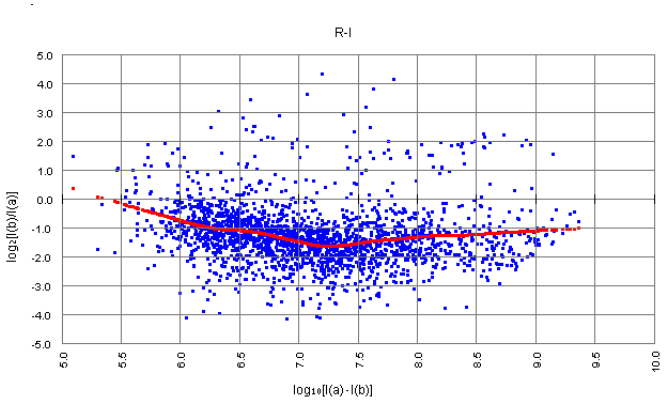


FIGURE 3.3 – RI-Plot sous MIDAS [83]

Nous pouvons voir dans la figure 3.4, la correction apportée par la méthode Total Intensity Normalization. Bien que les intensités soient maintenant centrées sur 0, nous pouvons constater que cet effet de dépendance n'est pas corrigé.

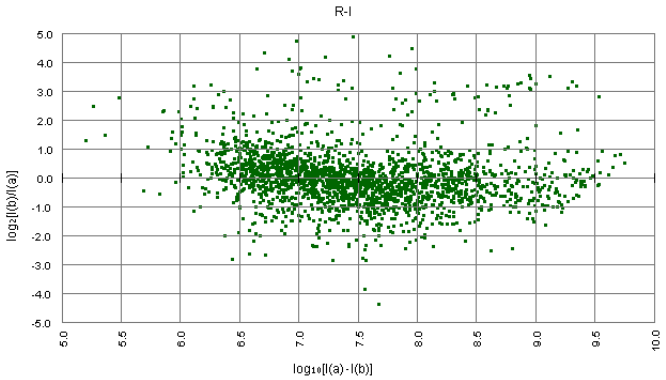


FIGURE 3.4 – Correction apportée par la méthode Total Intensity Normalization

Afin de prendre en compte ces effets, une méthode est largement utilisée aujourd'hui, il s'agit de la méthode du **Lowess** [35, 81]. Le principe est de déterminer la dépendance entre les ratios et les intensités en calculant l'équation $y(x)$ de la courbe dessinée dans la figure 3.3, avec $x_i = log_{10}(Cy5_i * Cy3_i)$ et $y_i = log_2(Cy5_i/Cy3_i)$. Cette fonction corrige ensuite la valeur du log2(ratio) pour chaque point de façon à obtenir :

$$new(log_2(Cy5_i/Cy3_i)) = old(log_2(Cy5_i/Cy3_i)) - y(x_i).$$

Cela revient à multiplier chaque intensité en Cy3 par un ratio différent pour chaque point :

$$new(Cy3_i) = old(Cy5_i) * 2^{y(x_i)}.$$

Finalement, cette méthode permet d'obtenir une nouvelle courbe de dépendance droite centrée sur 0 (cf figure 3.5).

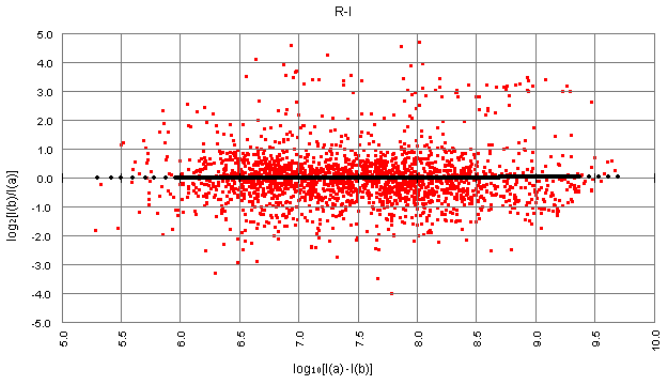


FIGURE 3.5 – Correction apportée par la méthode du Lowess

Normalisation inter-puces

Pour faire en sorte que les données soient comparables, une normalisation inter-puces peut être appliquée. En simple couleur, des contrôles internes peuvent être déposés en quantités connues (spike genes), permettant de vérifier que les données des différentes biopuces sont bien à la même échelle.

De manière générale, les données peuvent être centrées. En effet, l'information liée à la moyenne peut être utile en soi mais est rarement très informative : cela concerne l'expression moyenne d'un gène pour toutes les puces ou celle d'une puce pour tous les gènes. Il peut être intéressant de procéder à un double centrage à la fois en lignes et en colonnes du tableau des données. Par contre, la réduction ses données d'expression n'est pas souhaitable car en ramenant à un les variances des gènes, les effets de sur ou sous-expression de certains d'entre eux peuvent être éliminés.

3.2 Exemples expérimentaux

Une fois les données pré-traitées, celles-ci sont prêtes à être analysées pour apporter de la connaissance aux experts. Les données sont généralement représentées sous forme d'une **relation** où les lignes (appelées aussi **tuples**) représentent les échantillons et les **attributs** en colonne représentent les gènes. Voici trois exemples simplifiés de données d'expression pré traitées, représentant les protocoles les plus couramment rencontrés au cours de cette thèse. Le premier exemple permet de comparer des données obtenues pour deux conditions expérimentales. Le deuxième exemple présente des données d'expression temporelles et le troisième exemple regroupe des données temporelles obtenues pour différentes cellules.

Exemple 1:
Les données de la relation r_1 (cf table 3.2) représentent les niveaux d'expression de 5 gènes pour 7 échantillons. Les échantillons t_1 à t_4 appartiennent à la classe c_1 et les échantillons t_5 à t_7 appartiennent à la classe c_2.

r_1	a_1	a_2	a_3	a_4	a_5	*classe*
t_1	1.7	1.4	1.7	0.9	−1.9	c_1
t_2	1.8	1.6	1.9	0.8	2.0	c_1
t_3	−0.5	0.4	−1.4	1.0	0.4	c_1
t_4	−0.4	0.1	−0.4	0.8	−1.0	c_1
t_5	−1.4	1.8	1.3	1.9	0.1	c_2
t_6	−1.3	1.7	1.4	1.8	1.7	c_2
t_7	−1.4	−1.4	1.9	1.8	−2.0	c_2

TABLE 3.2 – Relation r_1

Remarquons que les données sont généralement représentées par une échelle de couleur allant du vert (gènes exprimés dans la cellule 1) au rouge (gènes exprimés dans la cellule 2) comme décrit dans la figure 3.6 pour les gènes a_1, a_2, a_3 et a_4. Elles peuvent également être représentées sous forme graphique (cf figure 3.7).

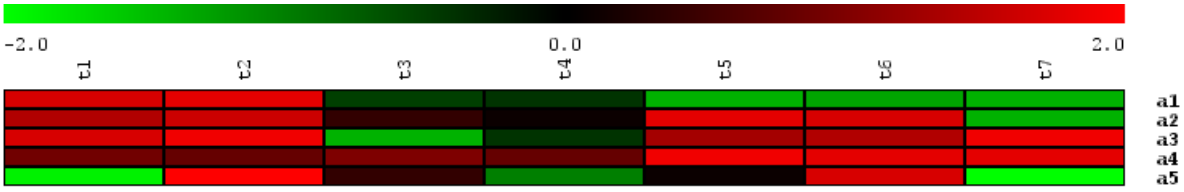


FIGURE 3.6 – Représentation classique des données d'expression de gènes

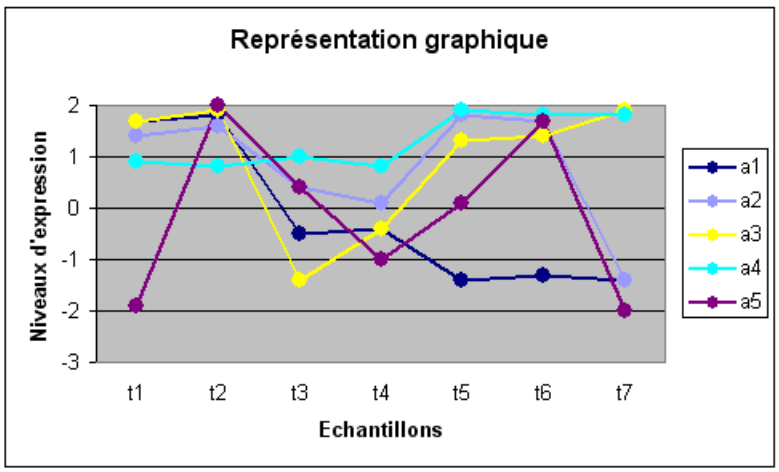


FIGURE 3.7 – Représentation graphique des niveaux d'expression des gènes

Exemple 2:
Les données de la relation r_2 (cf table 3.3) représentent les niveaux d'expression de 3 *gènes pour* 6 *échantillons.*

r_2	a_1	a_2	a_3	*temps*
t_1	0.4	−1.5	0.4	0
t_2	1.5	−0.3	−2.0	1
t_3	−0.7	0.8	1.6	2
t_4	0.4	1.8	−0.9	3
t_5	−1.4	0.5	−0.1	4
t_6	1.9	1.7	2.0	5

TABLE 3.3 – Relation r_2

Il s'agit ici de données temporelles, les échantillons correspondent aux profils d'une même cellule à 6 instants consécutifs. Par exemple, l'échantillon t_1 correspond au profil de la cellule au temps $t = 0$, l'échantillon t_2 correspond à son profil au temps $t = 1$ min ... et l'échantillon t_6 correspond à son profil au temps $t = 60$ min.

Exemple 3:
Les données de la relation r_3 (cf table 3.4) représentent les niveaux d'expression de 5 *gènes pour* 9 *échantillons. Les échantillons t_1, t_4 et t_7 correspondent au profil d'une première cellule aux temps $t = 0$, $t = 1$ et $t = 2$ (par exemple avant traitement, pendant traitement et après traitement). De même, les échantillons t_2, t_5 et t_8 correspondent au profil d'une deuxième cellule et les échantillons t_3, t_6 et t_9 correspondent au profil d'une troisième cellule.*

r_3	a_1	a_2	a_3	a_4	a_5	*cellule*	*temps*
t_1	−0.4	0.3	−2.0	−1.5	1.4	c_1	0
t_2	0.3	−1.9	−0.5	1.4	−0.4	c_2	0
t_3	1.1	−0.5	−1.9	1.2	0.7	c_3	0
t_4	0.9	1.5	−0.6	−0.3	1.7	c_1	1
t_5	1.8	−0.3	0.6	−0.5	−1.1	c_2	1
t_6	−0.5	1.8	1.2	1.1	−0.5	c_3	1
t_7	0.7	1.3	2.0	1.8	0.4	c_1	2
t_8	0.2	1.5	1.8	1.9	−1.1	c_2	2
t_9	−0.2	0.4	−0.5	−1.5	1.9	c_3	2

TABLE 3.4 – Relation r_3

Chapitre 4

Analyse des données d'expression de gènes

Sommaire

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

Ensuite, des méthodes non supervisées permettent d'explorer les données, permettant ainsi de savoir quels gènes ou quels échantillons ont des profils d'expression similaires. Si des informations supplémentaires sont connues à priori sur les données et que nous souhaitons utiliser ces connaissances, des méthodes supervisées peuvent alors être appliquées.

Le choix des méthodes dépend toujours de l'objectif global de l'étude. Divers logiciels dédiés aux données d'expression permettent de réaliser ces analyses, on peut citer GeneSpring GX (Agilent [1]), Bioconductor [48], MultiExperimentViewer (MeV) [83] de TIGR [11], Acuity (Molecular Devices [7]), GeneSight (BioDiscovery [2]) ou GEPAS [59]. Les logiciels de data mining généraux peuvent également être utilisés comme SAS [10] ou WEKA [99].

4.1 Gènes différentiellement exprimés

Plusieurs méthodes existent pour déterminer les gènes différentiellement exprimés [41, 89, 70, 77]. Nous allons voir par la suite la méthode du Fold Change ainsi que différents tests statistiques.

Pour ces différentes méthodes, nous nous plaçons dans le cas général où nous avons l classes $c_1, c_2, ...c_l$ à tester comportant respectivement $n_1, n_2, ..., n_l$ échantillons (comme dans la relation r_1 décrite dans la table 3.2 où $l = 2$, $n_1 = 4$ et $n_2 = 3$).

Notons moy_k^j la moyenne des valeurs du gène a_j pour la classe c_k et std_k^j l'écart-type de ces valeurs. Nous cherchons donc les gènes qui s'expriment différemment entre les différentes classes.

4.1.1 Fold Change

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

On parle alors de ϵ-Fold Change et une valeur $\epsilon = 2$ est couramment utilisée comme dans la figure 4.1 où en rouge sont donnés les gènes 2 fois plus exprimés dans la classe c_1 et en vert les gènes 2 fois plus exprimés dans la classe c_2.

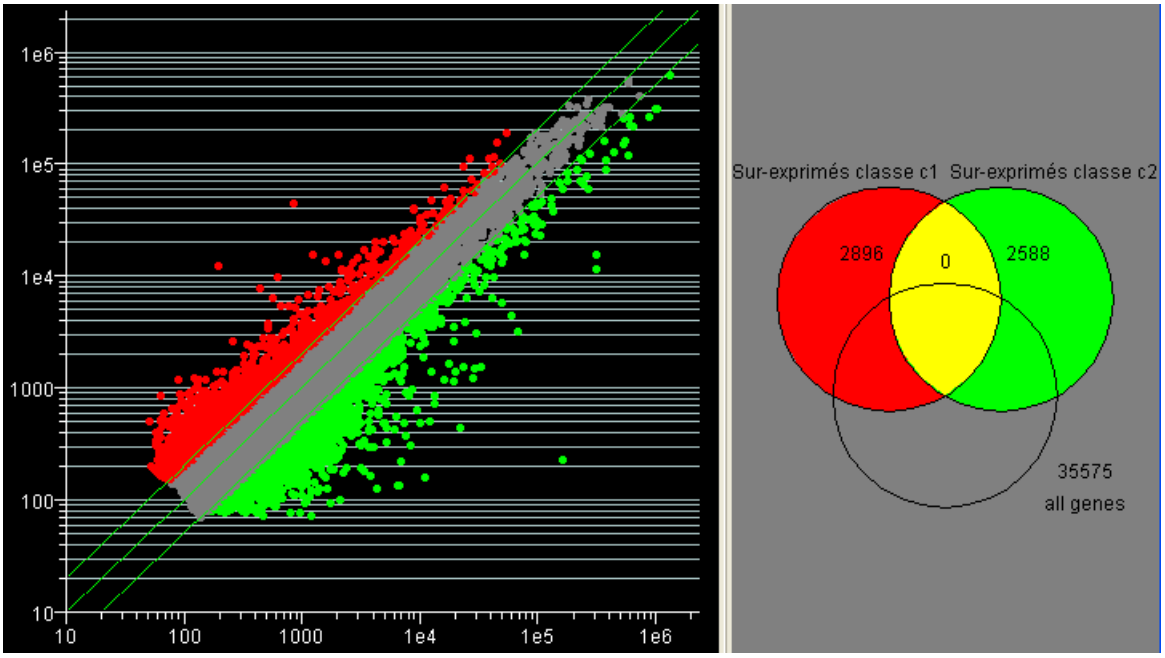


FIGURE 4.1 – Méthode du Fold Change

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

4.1.2 Tests d'hypothèse

Un test d'hypothèse consiste à tester une hypothèse statistique H_0, appelée hypothèse nulle contre une hypothèse alternative H_1. L'objectif est de décider à partir des données étudiées quelle est l'hypothèse à retenir.

La première étape consiste à définir avec précision les deux hypothèses H_0 et H_1, l'hypothèse H_1 regroupant toutes les hypothèses hormis H_0.

La seconde étape consiste à calculer une statistique de test, notée par exemple T (fonction des données), dont la distribution est connue si H_0 est vraie. Puis il faut comparer la valeur de la statistique obtenue à partir des données notée T_{obs} à la distribution attendue de T.

La troisième étape consiste à fixer un seuil α et de calculer la région de rejet correspondant i.e. l'ensemble des valeurs de la statistique de test pour lesquelles l'hypothèse nulle H_0 est rejetée.

En fonction de cette région de rejet et de la valeur de T_{obs}, la décision est prise de rejeter ou non H_0 :

– Si T_{obs} appartient à la région critique alors H_0 est rejetée avec le risque α.

– Si T_{obs} n'appartient pas à la région critique alors l'hypothèse H_0 ne peut être rejetée.

Si on note Γ la région de rejet alors $\alpha = Proba(T \in \Gamma | H_0)$. La probabilité α, dit **risque de première espèce**, correspond à la probabilité de rejeter H_0 à tort. La probabilité d'accepter H_0 à tort, notée β, est le risque de seconde espèce (cf tableau 4.1).

Réalité	H_0 non rejetée	H_0 rejetée
H_0 vraie	$1-\alpha$	α
H_1 vraie	β	$1-\beta$

TABLE 4.1 – Risques de première et de seconde espèces

Attention à l'interprétation du résultat, accepter H_0 ne signifie pas que cette hypothèse est vraie mais seulement que les observations disponibles ne sont pas incompatibles avec cette hypothèse et que l'on n'a pas de raison suffisante de lui préférer l'hypothèse H_1 compte tenu des résultats expérimentaux.

Le degré de signification (ou **p-value**) est la probabilité de commettre une erreur de première espèce si l'on décide de rejeter l'hypothèse nulle dès que la valeur de la statistique est supérieure à la valeur observée : $p\text{-}value = Proba(|T| > T_{obs} | H_0)$. Lors d'un test au

niveau α, l'hypothèse nulle est rejetée si $p\text{-}value < \alpha$.

Ici, nous souhaitons identifier les gènes dont la différence des moyennes d'expression entre deux ou plusieurs classes est statistiquement significative. Pour chaque gène a_j, l'hypothèse nulle testée est la suivante : $H_0 : moy_1^j = moy_2^j = ... = moy_l^j$.

Pour simplifier, dans la suite nous considérons un gène a_j quelconque et nous noterons $moy_k = moy_k^j$ et $std_k = std_k^j$. De plus, notons moy la moyenne totale des valeurs prises par le gène a_j étudié.

En général, on choisit $\alpha = 0.05$ ou $\alpha = 0.01$. Si la p-value est inférieure à 0.05, on dit que la différence entre les moyennes est **statistiquement significative**. Si la p-value est inférieure à 0.01, on dit que la différence entre les moyennes est **hautement significative**.

Plusieurs tests sont alors possibles : des tests paramétriques qui requièrent ou non des hypothèses sur la distribution des données et des tests non paramétriques.

Test de Student et ANOVA

Il s'agit ici de tests paramétriques avec hypothèse d'égalité des variances. Le test de Student est utilisé lorsque deux classes sont comparées, l'ANOVA est utilisée lorsque le nombre de classes est supérieur à 2.

Le principe de l'ANOVA est le suivant :

- Calculer la Somme des Carrés inter-classes correspondant à la variation inter-classes :

$$SC_{inter} = \sum_{k=1..l} n_k * (moy_k - moy)^2$$

- Calculer les Carrés Moyens inter-classes :

$$CM_{inter} = SC_{inter}/(k-1)$$

- Calculer la Somme des Carrés intra-classes correspondant à la variation intra-classes :

$$SC_{intra} = \sum_{k=1..l} \sum_{i=1..n_k} (x_{i,k} - moy_k)^2$$

- Calculer les Carrés Moyens intra-classes :

$$CM_{intra} = SC_{intra} / \sum_{k=1..l} n_k - 1$$

- Calculer la statistique de test :

$$F_{obs} = CM_{inter}/CM_{intra}$$

Sous l'hypothèse H_0, F suit une loi de Fisher avec $ddl_1 = k - 1$ et $ddl_2 = \sum_{k=1..l} n_k - 1$.

Si 2 classes sont comparées, un test de Student est utilisé avec la statistique T suivante :

$T_{obs} = [\sqrt{n_1 + n_2 - 2} * (moy_1 - moy_2)] / [\sqrt{1/n_1 + 1/n_2} * \sqrt{(n_1 - 1) * {std_1}^2 + (n_2 - 1) * {std_2}^2}]$

Sous l'hypothèse H_0, T suit une loi de Student à $n_1 + n_2 - 2$ degrés de liberté.

Test de Welch et Welch ANOVA

Ces tests paramétriques ont l'avantage de ne pas requérir d'hypothèse d'égalité des variances mais n'assure qu'un contrôle approximatif du risque de première espère contrairement aux précédents tests. Le test de Welch est utilisé lorsque deux classes sont comparées, la Welch ANOVA est utilisée lorsque le nombre de classes est supérieur à 2.

Le principe de la Welch ANOVA est le suivant :

- Calculer le poids des groupes : $w_k = n_k((n_k - 1)/SS_k)$
- Calculer la somme des poids : $w = \sum_{k=1..l} w_k$
- Calculer la moyenne pondérée : $\widetilde{moy} = \sum_{k=1..l} moy_k / w$
- Calculer la Somme des Carrés inter-classes correspondant à la variation inter-classes :

$$SC_{inter} = \sum_{k=1..l} w_k * (moy_k - \widetilde{moy})^2$$

- Calculer les Carrés Moyens inter-classes :

$$CM_{inter} = SC_{inter}/(k - 1)$$

- Calculer Z :

$$Z = (1/(l^2 - 1)) \sum_{k=1..l} (1 - w_k/w)^2/(n_k - 1)$$

- Calculer les Carrés Moyens intra-classes :

$$CM_{intra} = 1 + 2(l - 2)Z$$

- Calculer la statistique de test :

$$W_{obs} = CM_{inter}/CM_{intra}$$

Sous l'hypothèse H_0, W suit une loi de Fisher avec $ddl_1 = k - 1$ et ddl_2 l'entier le plus proche de $1/(3Z)$.

Si 2 classes sont comparées, un test de Welch est utilisé avec la statistique T suivante :

$$T_{obs} = |moy_1 - moy_2| / \sqrt{{S_1}^2/n_1 + {S_2}^2/n_2}$$

Sous l'hypothèse H_0, T suit une loi de Student à ddl degrés de liberté où ddl est l'entier le plus proche de $({S_1}^2/n_1 + {S_2}^2/n_2)^2 / ({S_1}^2/((n_1 - 1) * n_1) + {S_2}^2/((n_2 - 1) * n_2))$.

Test de Wilcoxon-Mann-Whitney et de Kruskal-Wallis

Ces tests sont équivalents aux tests de Student et à l'ANOVA mais au lieu de prendre en compte les valeurs $x_{i,k}$ du gène considéré, ces tests considèrent les rangs $r_{i,k}$ des valeurs. Ceci permet en outre de corriger la présence de valeurs extrêmes dans les données.

Les tests que nous avons présentés permettent de tester si un gène est ou non différentiellement exprimé entre les différentes classes avec un risque α. Or sur une biopuce, plusieurs milliers de gènes peuvent être étudiés, réalisant ainsi autant de tests d'hypothèse.

Si $m = 20000$ tests sont réalisés à un risque $\alpha = 0.05$, alors environ 1000 gènes seront détectés comme différentiellement exprimés à tort. Le tableau suivant résume les différents cas possibles :

Réalité	H_0 non rejetée	H_0 rejetée	Total
H_0 vraie	U	V	m_0
H_1 vraie	T	S	m_1
	W	R	m

TABLE 4.2 – Répartition des faux positifs

L'objectif est donc de minimiser V i.e. le nombre de gènes détectés comme différentiellement exprimés alors qu'en réalité ils ne le sont pas, on les appelle les **faux positifs** et T i.e. ne pas oublier des gènes qui sont réellement différentiellement exprimés i.e. minimiser le nombre de **faux négatifs**.

Plusieurs procédures permettant de contrôler ce risque d'erreur ont alors été proposées [97, 47], celles basées sur le Family Wise Error Rate et celles basées sur le False Discovery Rate :

- Le **Family Wise Error Rate** (FWER) est la probabilité d'avoir au moins un faux positif : $FWER = P(V > 0)$. Contrôler le FWER au seuil α, par exemple égal à 0.05 permet d'être confiant à 95% de n'avoir aucun faux positifs.

  La **méthode de Bonferroni** [60] consiste alors à réaliser les tests pour chaque gène à un seuil de α/m, ainsi le FWER sera forcément inférieur ou égal à α. Malheureusement en pratique, lorsque le nombre de gènes est grand, cette méthode fournit souvent une liste de gènes vide.

- Le **False Discovery Rate** (FDR) contrôle l'espérance du taux de faux positifs : $FDR = E(Q)$ où $Q = V/R$ si $R \neq 0$ et $Q = 0$ si $R = 0$. Contrôler le FDR au seuil α, par exemple égal à 0.05 permet d'affirmer qu'en moyenne le taux de faux positifs est inférieur à 5%. Ce critère est donc moins restrictif que le FWER ($FDR \leq FWER$).

La **méthode de Benjamini et Hochberg** [28] consiste à ordonner les m p-values correspondant aux m tests et à rechercher le plus haut rang s des p-values tel que $p\text{-}value(s) \geq \alpha * s/m$. Cette méthode est assez intéressante pour les données d'expression de gènes car elle permet de bien contrôler le taux de faux positifs toutefois, il est également possible d'obtenir un tout petit nombre de gènes.

4.2 Analyses non supervisées

L'objectif des méthodes non supervisées est d'explorer les données "à l'aveugle" pour découvrir des connaissances intéressantes. Les méthodes de clustering permettent par exemple de regrouper les échantillons ou les gènes ayant des profils d'expression similaires. Les méthodes statistiques comme l'Analyse en Composantes Principales permettent quant à elles de réduire le nombre de dimensions pour visualiser les données.

4.2.1 Méthodes de clustering

L'objectif des méthodes de clustering est de regrouper les échantillons ou les gènes ayant des profils d'expression similaires. Deux types d'approches existent : les classifications hiérarchiques [42] ou les classifications non hiérarchiques [67, 91].

Classification hiérarchique

Il s'agit ici d'une méthode itérative, en effet, à chaque étape les échantillons ou les gènes les plus proches sont regroupés ensemble jusqu'à obtenir une structure d'arbre hiérarchique reliant tous les échantillons et/ou tous les gènes (cf figure 4.2).

Le principe de la classification hiérarchique est le suivant [42] :

1. Calculer les distances 2 à 2 entre tous les individus (échantillons ou gènes) puis regrouper les deux individus les plus proches.
2. Calculer les distances 2 à 2 entre tous les individus restants et les groupes formés puis regrouper les deux individus ou groupes les plus proches.
3. Répéter l'étape 2 jusqu'à ce que tous les individus soient regroupés dans un seul cluster.

Cet algorithme nécessite alors un critère de ressemblance entre les individus et un critère d'agrégation entre les groupes d'individus. Plusieurs distances peuvent être utilisées pour mesurer la ressemblance entre les individus comme la distance euclidienne, le coefficient de corrélation de Pearson ou la distance de Manhattan.

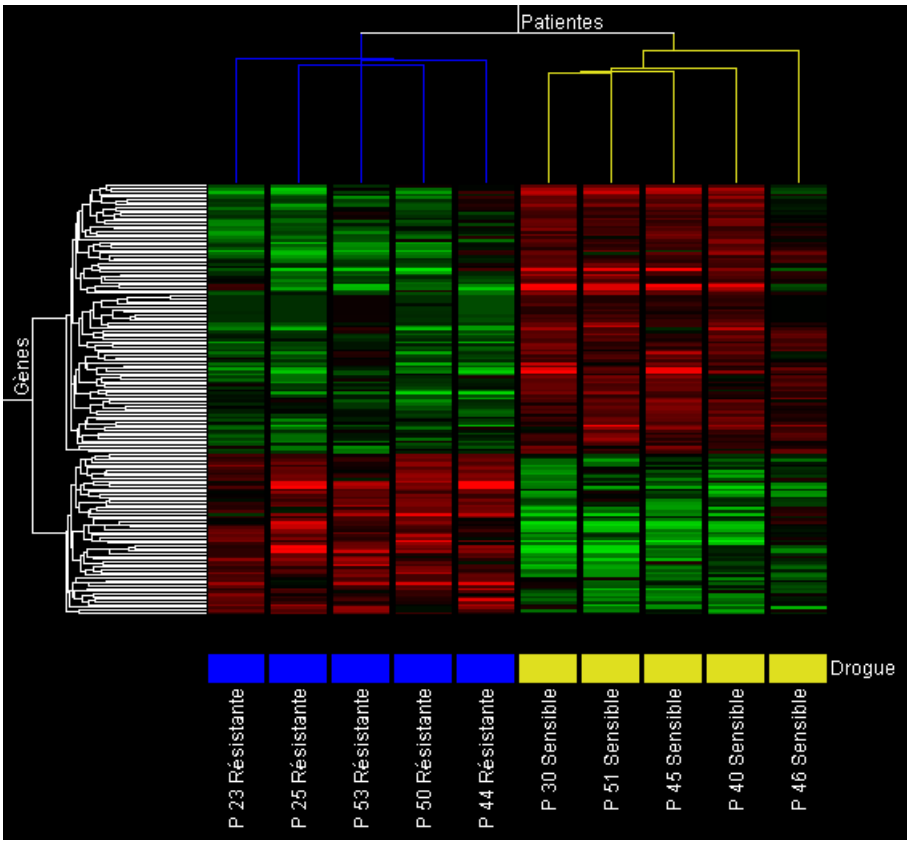


FIGURE 4.2 – Classification hiérarchique sur les gènes et les échantillons

Pour calculer la distance entre deux clusters, trois principales méthodes existent :
- La distance minimum ou **Single Linkage** : la distance entre 2 clusters c1 et c2 est définie par la plus courte distance séparant un individu de c1 et un individu de c2.
- La distance moyenne ou **Average Linkage** : Ce critère consiste à calculer la distance moyenne entre tous les individus de c1 et tous les individus de c2.
- La distance maximum ou **Complete Linkage** : la distance entre les 2 clusters c1 et c2 est ici définie par la plus grande distance séparant un individu de c1 et un individu de c2.

La difficulté du choix du critère d'agrégation réside dans le fait que ces critères peuvent déboucher sur des résultats différents (cf figure 4.3). Pour les données d'expression de gènes, les deux dernières méthodes sont les plus utilisées.

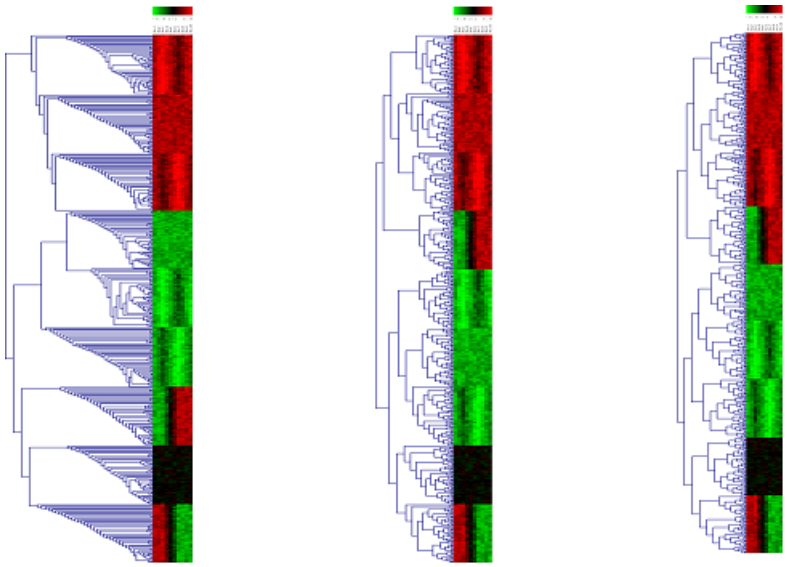


FIGURE 4.3 – Classifications hiérarchiques avec différents critères d'agrégation

Classification non hiérarchique

Les méthodes non hiérarchiques regroupent les individus dans K clusters distincts et non plus sous forme d'arbre hiérarchique comme précédemment. Deux principales méthodes sont couramment utilisées, la méthode du K-means et la méthode SOM.

Méthode du K-means La méthode du K-means est une méthode non hiérarchique qui regroupe les individus dans K clusters, où K est déterminé par l'utilisateur. L'objectif est de minimiser la variabilité intra-clusters et de maximiser la variabilité inter-clusters.

Le principe de la méthode du K-means est le suivant :

1. Les individus sont affectés à un des K clusters de façon aléatoire de façon à ce que les K clusters aient (à peu près) le même nombre d'individus.
2. Calculer le centre de gravité de chaque cluster.
3. Calculer la distance entre chaque individu et chaque centre de gravité.
4. Si le gène est le plus proche du centre de gravité de son cluster, alors il reste affecté à ce cluster sinon il est réaffecté au cluster le plus proche.
5. Répéter les étapes 2 à 4 jusqu'à ce que tous les gènes soient affectés au cluster le plus proche.

Cet algorithme est relativement simple à comprendre et à implémenter, il est également assez rapide en temps de calcul mais le nombre de clusters à former reste un choix difficile. De plus, le regroupement final dépendra de l'affectation aléatoire de la première étape et peut donc varier.

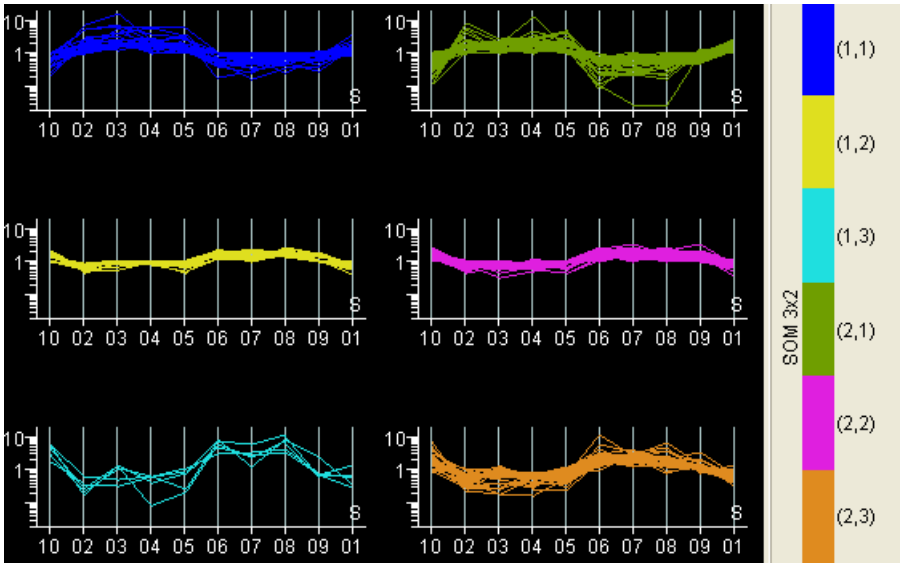


FIGURE 4.4 – Méthode SOM avec une grille 3*2

Self Organizing Maps Comme le K-means, la méthode SOM [67, 91] regroupe les individus en différents clusters de façon à minimiser la variabilité intra-clusters. En plus, la méthode SOM renseigne sur la similarité entre les différents clusters. L'utilisateur spécifie le nombre de lignes et de colonnes d'une grille à 2 dimensions, ainsi une grille 3*2 aura 6 cases correspondant aux 6 clusters (cf figure 4.4).

Il se peut contrairement à la méthode du K-means que la méthode SOM génère des clusters ne comportant aucun individu. L'avantage de cette méthode est qu'elle apporte une idée de la proximité des clusters puisque deux clusters proches sur la grille auront des profils d'expression assez similaires.

Toutefois, comme précédemment, il est difficile de déterminer la dimension de la grille auparavant, le mieux reste de lancer l'algorithme pour différentes dimensions.

4.2.2 Analyse en Composantes Principales

Considérons une relation $n \times m$ où n est le nombre d'échantillons et m le nombre de gènes. L'ensemble des échantillons peut se voir comme un nuage de points dans $\mathbb{R}^m$ et l'ensemble des gènes peut se voir comme un nuage de points dans $\mathbb{R}^n$. Or, il est impossible de représenter graphiquement ces nuages de points dès lors que m ou n sont supérieurs à 3.

L'objectif de l'Analyse en Composantes Principales (ou Décomposition en Valeurs Singulières [22]) est de donner une vue la moins détériorée possible de ces nuages de points dans un espace de dimension réduite.

Cette méthode se base essentiellement sur le calcul de l'**inertie** d'un nuage de points, qui mesure la dispersion des individus autour de son centre de gravité (cette valeur généralise la variance au cas multidimensionnel).

Supposons que nous souhaitons explorer le nuage des échantillons. Le principe est de chercher un nouveau plan de dimension m, dans lequel le nuage de points est toujours le même et son inertie totale ne change pas. Mais où par contre la répartition de cette dispersion globale est modifiée de telle sorte que les premiers facteurs restituent à eux seuls une grande partie de l'information, permettant ainsi de négliger les derniers facteurs et donc de proposer un plan de dimension inférieure mais résumant bien la dispersion du nuage. Cela revient finalement à un changement de base dans un espace vectoriel.

Soit I_g l'inertie du nuage des points dans le plan de dimension m. Le plan P_k de dimension $k < m$ sur lequel le nuage des points projeté aura une inertie I'_g maximale, correspond au plan engendré par les k vecteurs propres M-normés associés aux k plus grandes valeurs propres de VM, où V est la matrice des variances-covariances et M est la matrice des distances. Il suffit alors de diagonaliser VM pour obtenir les m valeurs propres $(\lambda_1, ..., \lambda_m)$ et les m vecteurs propres associés $(\vec{u_1}, ..., \vec{u_m})$. Parmi les m vecteurs propres obtenus, on conserve k vecteurs propres correspondant alors aux axes principaux. Les k composantes principales correspondent aux nouvelles coordonnées de l'ensemble des points sur ces axes principaux.

Le nombre k de composantes pertinentes peut être choisi de sorte que la part d'inertie expliquée par l'ensemble de ces composantes soit supérieure à une valeur seuil fixée à priori par l'utilisateur. D'autres méthodes permettent également de choisir un nombre pertinent de composantes, toutefois pour les données d'expression, la première composante explique généralement à elle seule une grande partie de l'inertie du nuage, comme l'illustre par exemple la figure 4.5.

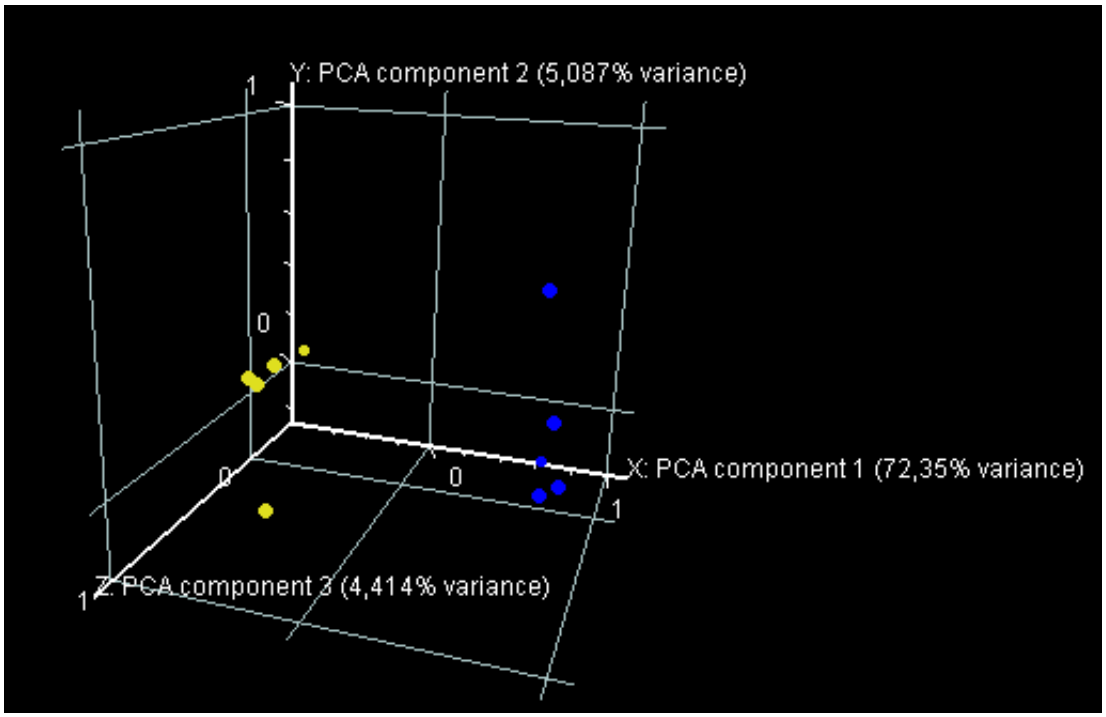


FIGURE 4.5 – ACP sur dix échantillons

Il faut rester prudent quant aux résultats d'une ACP. Il faut tout d'abord noter que lorsque des projections de points sont éloignées sur un axe (ou sur un plan), on peut assurer que les points sont éloignés dans l'espace d'origine. En revanche, deux individus dont les projections sont proches peuvent ne pas être proches dans l'espace.

La qualité de représentations des individus peut par exemple être vérifiée en étudiant l'angle entre un point et un nouvel axe (ou un plan). Plus le point est proche de l'axe, plus l'individu correspondant à ce point est bien représenté par sa projection sur l'axe.

De plus, un individu contribuera d'autant plus à la confection d'un axe que sa projection sur cet axe sera éloignée du centre de gravité du nuage. Inversement, un individu dont la projection sur un axe sera proche du centre de gravité contribuera faiblement à l'inertie portée par cet axe. L'étude de ces contributions permet d'interpréter les nouveaux axes en fonction des individus. L'étude des corrélations entre les variables et les nouveaux facteurs nous permettent de la même façon de donner un sens aux nouveaux axes.

4.3 Analyses supervisées

Les analyses sont dites supervisées lorsqu'elles reposent sur l'utilisation d'exemples d'apprentissage. Le principe des méthodes de prédiction consiste ainsi à étudier le comportement d'individus dont la classe à prédire est connue, puis de construire un modèle à

partir de cette connaissance pour enfin appliquer ce modèle à de nouveaux individus dont la classe est inconnue.

4.3.1 Méthodes de prédiction

Différentes méthodes de prédiction ont été proposées comme la méthode des k plus proches voisins, les support vector machines, les arbres de décision, les réseaux bayésiens ou les réseaux de neurones.

La méthode des **k plus proches voisins** consiste par exemple à donner pour un nouvel individu la classe la plus fréquente chez ces k plus proches voisins ou bien la classe de son voisin dans le cas où $k = 1$. Cette méthode nécessite le choix d'une distance, par exemple la distance Euclidienne pour mesurer le voisinage de chaque individu. La méthode PAM (Prediction Analysis for Microarrays) [93] est dérivée de cette méthode des k plus proches voisins.

Le principe des **support vector machines** [33, 68] consiste à représenter l'ensemble des individus dont la classe est connue dans un plan de dimension réduite comme pour l'Analyse en Composantes Principales (cf section 4.2.2). L'objectif cependant ici est de trouver le plan le plus discriminant possible i.e. tel que les individus d'une même classe soient les plus proches possibles et les individus de classes différentes soient les plus éloignés (cf figure 4.6).

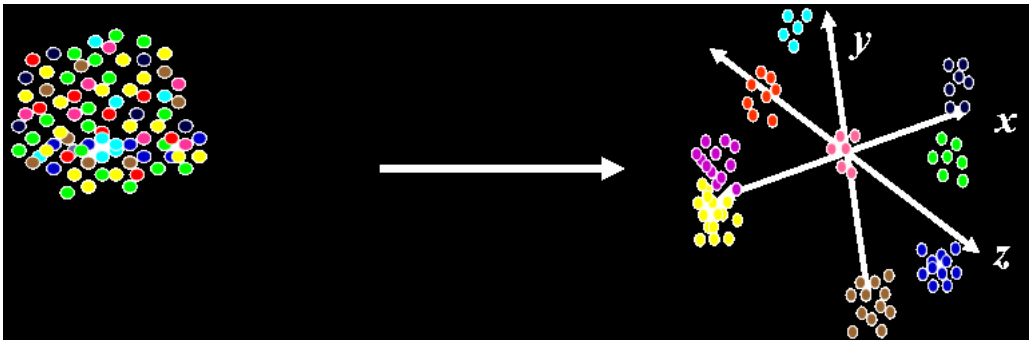


FIGURE 4.6 – Support Vector Machines

Ensuite, un nouvel individu est placé dans ce plan et selon sa position, il sera possible de prédire sa classe.

4.3.2 Validation du modèle prédictif

Un modèle prédictif est d'autant plus intéressant qu'il présente un bon taux de prédiction. L'objectif est donc de comparer plusieurs modèles pour déterminer celui qui semble le plus satisfaisant.

Pour mesurer le taux de prédiction, l'ensemble des individus dont la classe est connue est partitionnée en un **jeu d'entraînement** à partir duquel le modèle sera construit (celui-ci peut par exemple contenir 75% des individus) et un **jeu de validation** sur lequel sera testé le modèle (cf figure 4.7).

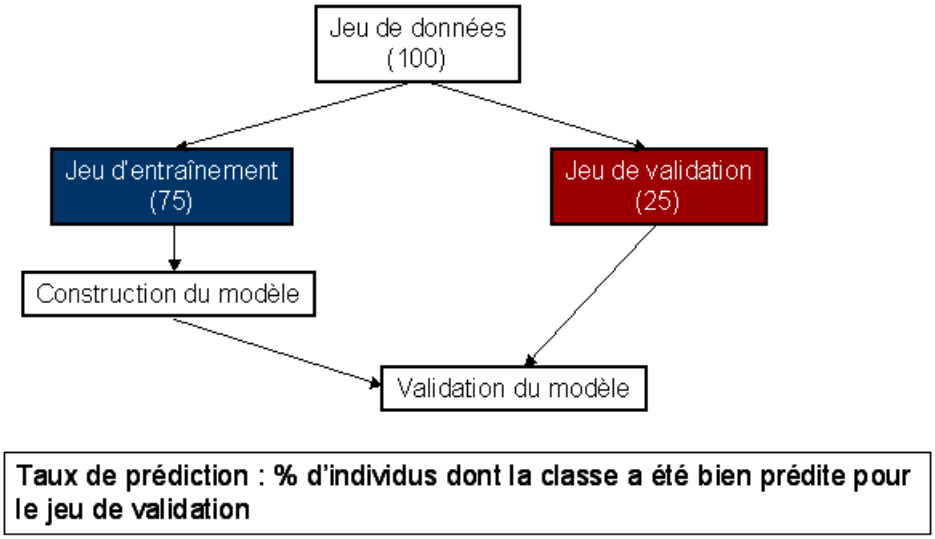


FIGURE 4.7 – Construction et validation d'un modèle de prédiction

Le taux de prédiction correspond alors au nombre d'individus dans le jeu de validation dont la classe a été bien prédite.

Cette méthode a cependant l'inconvénient de dépendre des individus inclus dans le jeu d'entraînement. En effet, si on modifie le jeu d'entraînement, le taux de prédiction peut être meilleur ou moins bon. Une autre méthode plus élaborée permet alors de prendre en compte ces remarques, il s'agit de la **k-fold cross-validation**.

Le principe de la k-fold cross validation consiste à séparer le jeu de données en k groupes et d'utiliser à chaque fois un des groupes comme jeu de validation (cf figure 4.8).

Dans ce cas, k modèles prédictifs sont construits et pour chacun d'entre eux un taux de prédiction est obtenu. Le modèle final pris en compte est alors le modèle construit avec tous les individus et le taux de prédiction est la moyenne des taux obtenus avec les k modèles.

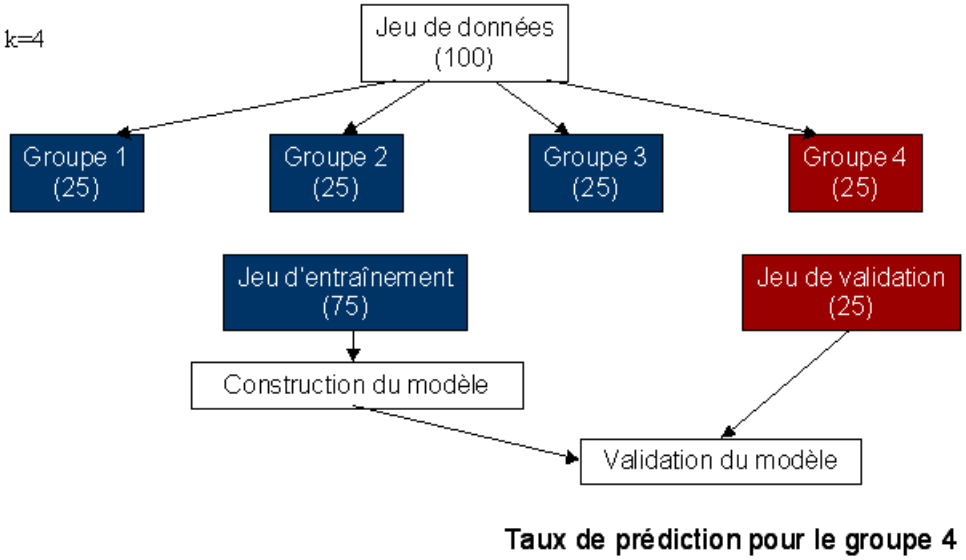


FIGURE 4.8 – k-Fold Cross-Validation (k=4)

Si k est égal au nombre total d'individus, on parle alors de **Leave-One-Out Cross-Validation** (LOOCV). Dans ce cas, à chaque itération, un individu est mis de côté pour la construction du modèle et on teste si la classe prédite est correcte pour cet individu.

Notons que les méthodes de prédiction peuvent être appliquées à partir de listes de gènes découvertes par exemple par des tests d'hypothèses (cf section 4.1), mais des méthodes supplémentaires [50, 55] peuvent être appliquées afin de chercher parmi les gènes donnés en entrée de l'algorithme quels sont ceux permettant d'obtenir le meilleur modèle prédictif.

Chapitre 5

Mise en œuvre

Sommaire

Dans ce chapitre, nous indiquons tout d'abord les protocoles expérimentaux des deux applications présentées dans la section 1.2, ainsi que le processus de pré-traitement appliqué sur les données. Nous présentons ensuite quelques résultats obtenus à partir des méthodes détaillées précédemment.

5.1 Etude de l'envahissement ganglionnaire

5.1.1 Protocole expérimental

Pour cette étude, des fragments chirurgicaux de tumeurs mammaires chez des patientes opérées au Centre Jean Perrin ont été prélevés, puis la qualité des ARN extraits a ensuite été vérifiée.

Les biopuces utilisées sont des biopuces TM/CS DIAGNOGENE, il s'agit de puces dites à façon (cf section 2.2.1), elles permettent d'analyser le niveau d'expression d'environ 800 gènes connus pour être impliqués dans le cancer du sein. Ces gènes sont tous déposés trois fois sur la puce pour un meilleure contrôle de la reproductibilité. Un pool d'ARN de

lignées cellulaires tumorales est utilisé comme ARN référence (cf plan expérimental n° 3 de la section 2.2.2).

Pour l'étude du statut ganglionnaire, il est important de connaître également la réceptivité hormonale (RH) des tumeurs qui est un des paramètres les plus influant sur l'expression des gènes.

Pour ce projet, les profils transcriptomiques des tumeurs mammaires de 39 patientes sont disponibles, celles-ci se répartissent de la façon suivante (cf table 5.1).

Caractéristiques	RH-	RH+	Total
N-	9	8	17
N+	11	11	22
Total	20	19	39

TABLE 5.1 – Répartition des patientes selon les paramètres N et RH

5.1.2 Pré-traitement des données

Les biopuces ont tout d'abord été scannées à **plusieurs réglages** du scanner. Pour chaque réglage, les images ont été analysées avec le logiciel Genepix Pro 6.0. (Molecular Devices [7]). Celui-ci permet d'obtenir différentes données pour chaque spot de la puce, notamment la médiane des intensités des pixels du spot ainsi que la médiane des intensités des pixels du bruit de fond local pour Cy5 et Cy3. Le pourcentage de pixels saturés pour Cy5 et Cy3 est également donné, un signal est alors considéré comme saturé si le pourcentage de pixels saturés est supérieur ou égal à 50%. Pour éviter de perdre de l'information, la valeur des spots saturés a cependant été conservée. De plus, une information est donnée sur la qualité de chaque spot, notamment s'il y a eu une poussière ou une trace sur ce spot, il est considéré comme "bad" ou bien si le signal est trop faible pour être détecté. Ces spots ainsi que les spots dont l'intensité (- bruit de fond) est inférieure à un seuil donné (fixé à 1000) ont alors été considérés comme "faibles".

Un outil a ensuite été créé en Visual Basic, prenant en compte les trois réglages pour chaque biopuce et permettant de choisir pour chaque spot **le meilleur réglage**. Le réglage conservé est celui pour lequel la qualité globale du spot est la meilleure. Par exemple, si pour un réglage le spot était saturé est pour un autre réglage il ne l'était pas alors ce deuxième réglage a été préféré au premier. Si pour plusieurs réglages, un spot avait la même qualité globale, le plus fort (resp. le moins fort) réglage a été choisi si l'intensité du spot était plutôt faible (resp. plutôt élevée). En fonction de la qualité des spots, a été prise en compte soit la valeur de l'intensité moins le bruit de fond ou bien la valeur seuil fixé à 1000 pour les signaux faibles.

Pour chaque réglage, les données ont ensuite été normalisées avec le logiciel MIDAS [83] de TIGR [11], par la méthode du **Lowess** sur l'ensembe des spots. Le log2 du ratio

Cy5/Cy3 a ensuite été calculé pour chaque spot avec les valeurs normalisées obtenues pour le meilleure réglage.

Comme chaque gène est déposé trois fois sur la puce, nous devons nous assurer que les trois valeurs sont proches, plusieurs méthodes ont été proposées à cet effet [71, 29]. Après plusieurs comparaisons, nous avons opté tout d'abord pour une méthode permettant de **détecter un outlier** parmi les 3 valeurs, puis un contrôle de la distance est effectué entre les valeurs restantes à l'issue du premier filtre. Un test de Student à 95% a tout d'abord été appliqué à chaque gène afin de détecter un outlier parmi ses 3 valeurs. Par exemple, si la valeur du spot n°3 n'appartenait pas à l'intervalle de confiance des deux autres valeurs, alors ce spot a été considéré comme outlier. L'intervalle de confiance est calculé de la façon suivante :

$$IC = [moy(s1, s2) - (12.706 * std(s1, s2)/\sqrt{2}) \ ; \ moy(s1, s2) + (12.706 * std(s1, s2)/\sqrt{2})].$$

Une vérification a ensuite été réalisée afin de s'assurer que la distance entre les valeurs non outliers n'était pas trop importante. Pour cela, un nouveau test de Student à 95% a été appliqué sur la distribution de ces distances et une distance a été considérée comme trop importante si elle n'appartenait pas à l'intervalle suivant :

$$[moy(distances) - 1.96 * std(distances) \ ; \ moy(distances) + 1.96 * std(distances)].$$

L'ensemble de ces calculs ont été implémentés dans l'outil créé en Visual Basic.

Finalement, la moyenne des spots restants a été calculée pour chaque gène, et si tous les spots étaient faibles alors une valeur du log2(ratio) de 0 a été attribuée au gène correspondant et une valeur manquante a été attribuée aux gènes dont tous les spots étaient "Bad".

5.1.3 Analyse des données

Tout d'abord, une Analyse en Composantes Principales a été réalisée sur l'ensemble des gènes de la biopuce (cf figure 5.1).

Sur cette ACP, les points jaunes correspondent aux tumeurs RH- tandis que les points bleus correspondent aux tumeurs RH+. Nous voyons alors que ce paramètre a une grande influence sur l'expression de l'ensemble des gènes.

Notre objectif est de déterminer les gènes différentiellement exprimés entre les tumeurs N+ et les tumeurs N-. Or lorsque nous appliquons un test d'hypothèse (par exemple une Welch-ANOVA) sur l'ensemble des tumeurs, les gènes différentiellement exprimés entre les tumeurs RH+ et RH- ont tendance à systématiquement ressortir, gommant ainsi l'influence des gènes spécifiques à l'envahissement ganglionnaire.

Les tumeurs RH+ et RH- ont alors été étudiées séparément. Tout d'abord, un test de Welch avec une p-value à 0.05 a été appliqué sur les 20 tumeurs RH-. Cette méthode a

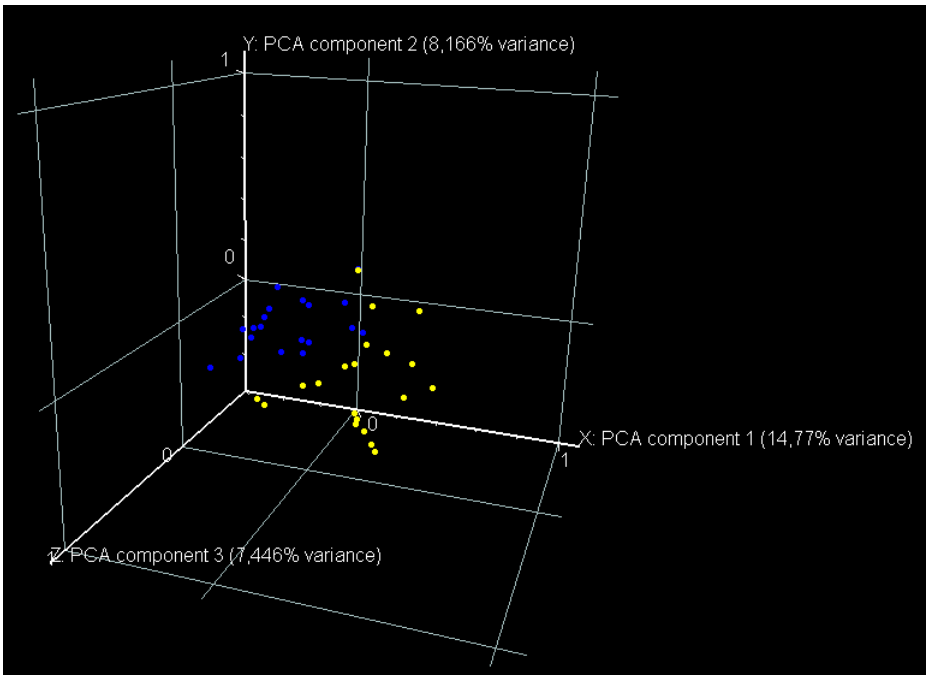


FIGURE 5.1 – Analyse en Composantes Principales à partir de l'ensemble des gènes

permis d'identifier 73 gènes comme étant différentiellement exprimés entre les tumeurs N- et les tumeurs N+. Aucune correction n'a été appliquée car avec une correction, aucun gène n'était significatif, il faut donc rester prudent car sur ces 73 gènes, environ 35 peuvent avoir été obtenus par chance.

Une classification hiérarchique a alors été appliquée sur ces 73 gènes (cf figure 5.2).

Cette arbre montre bien que ces gènes permettent de distinguer les deux types de tumeurs N+ en bleu et N- en jaune, situées sur 2 branches différentes.

De la même façon, un test de Welch avec une p-value à 0.05 a été appliqué sur les 19 tumeurs RH+. Cette méthode a permis d'identifier 33 gènes comme étant différentiellement exprimés entre les tumeurs N- et les tumeurs N+. Une classification hiérarchique a ensuite été appliquée sur ces 33 gènes, l'arbre obtenu est donné dans la figure 5.3.

Cet arbre montre qu'une tumeur N- se retrouve parmi les tumeurs N+. Les gènes obtenus permettent à peu près de distinguer les deux types de tumeurs N+ et N- mais le nombre de gènes est relativement faible et peuvent avoir été obtenus par chance. Un seul gène a d'ailleurs été obtenu différentiellement exprimé pour les tumeurs RH- et pour les tumeurs RH+.

Ces résultats montrent qu'il semble plus facile de distinguer les tumeurs dont le statut ganglionnaire est positif des tumeurs dont le statut ganglionnaire est négatif parmi les tumeurs RH- que parmi les tumeurs RH+.

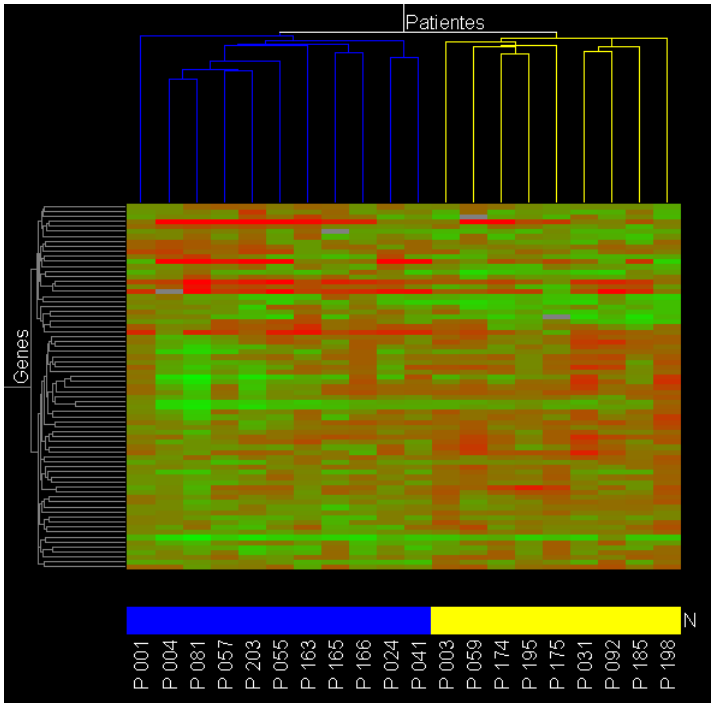


FIGURE 5.2 – Classification hiérarchique sur les tumeurs RH-

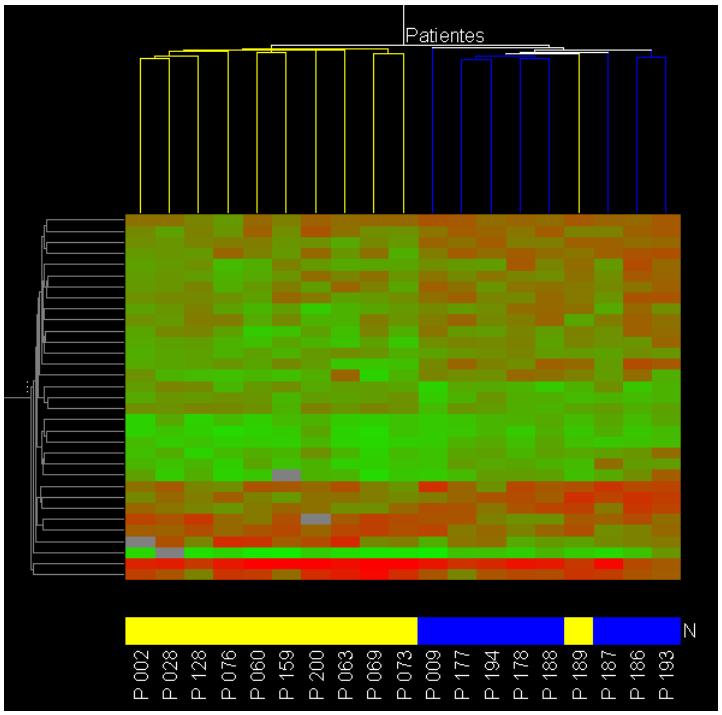


FIGURE 5.3 – Classification hiérarchique sur les tumeurs RH+

Pour tester la pertinence des gènes obtenus, des modèles de prédiction ont ensuite été construits à partir de ces différents profils et comparés entre eux.

Tout d'abord, des modèles permettant de prédire le statut RH des tumeurs ont été construits à partir d'une liste de 126 gènes obtenus par un test de Welch, comme étant différentiellement exprimés entre les deux types de tumeurs. Ces modèles ont été générés sous GeneSpring [1] en utilisant la méthode des k plus proches voisins et des support vector machines. Finalement, le modèle retenu a été obtenu avec la méthode des suport vector machines sur l'ensemble des 126 gènes. Ce modèle présente un taux de prédiction de 100% en Leave-One-Out Cross-Validation.

Ensuite, de la même façon des modèles ont été construits pour prédire le statut ganglionnaire pour les tumeurs RH- puis pour les tumeurs RH+. Dans les deux cas, le modèle retenu et présentant les meilleurs résultats, a été obtenu avec la méthode des support vector machines sur l'ensemble des 73 gènes discriminant pour les tumeurs RH- et les 33 gènes discriminants pour les tumeurs RH+. Le taux de prédiction obtenu pour les tumeurs RH- est de 100% tandis qu'il n'est que de 79% pour les tumeurs RH+.

Cela confirme les résultats obtenus avec les analyses non supervisées à savoir que les gènes obtenus pour les tumeurs RH+ sont moins discriminants que ceux obtenus pour les tumeurs RH-.

5.2 Etude de la sensibilité au docétaxel

5.2.1 Protocole expérimental

Les patientes du Centre Jean Perrin prises en compte dans cette étude, ont reçu une chimiothérapie néo-adjuvante composée de 6 cures de docétaxel. Les ARN sont extraits à partir des biopsies avant traitement et leur qualité a ensuite été vérifiée.

Les biopuces utilisées ici sont des biopuces humaines 44K AGILENT [1], il s'agit de puces pangénomiques (cf section 2.2.1), elles permettent d'analyser le niveau d'expression de plus de 40 000 gènes humains. Un pool d'ARN de lignées cellulaires tumorales est utilisé comme ARN référence (cf plan expérimental n° 3 de la section 2.2.2).

La réponse thérapeutique a ensuite été évaluée après l'intervention chirurgicale. Cette étude regroupe 10 patientes réparties de la façon suivante (cf tableau 5.2).

L'objectif de ce projet est d'identifier les gènes différentiellement exprimés entre les patientes ayant bien répondu au docétaxel (sensible) et les patientes ayant résisté au docétaxel.

Patientes	Réponse
P001	Résistante
P002	Résistante
P003	Résistante
P004	Résistante
P005	Résistante
P006	Sensible
P007	Sensible
P008	Sensible
P009	Sensible
P010	Sensible

TABLE 5.2 – Descriptif des patientes

5.2.2 Pré-traitement des données

Les images (obtenues à partir d'un seul réglage du scanner) ont été analysées avec le logiciel Genepix Pro 6.0. (Molecular Devices [7]). Les données ont ensuite été normalisées avec le logiciel GeneSpring GX [1], par la méthode du **Lowess**. Pour chaque spot, le log2 du ratio Cy5/Cy3 a été calculé avec les valeurs normalisées. Une normalisation inter-puces a ensuite été appliquée permettant de centrer l'ensemble des gènes autour de la médiane.

5.2.3 Analyse des données

L'objectif ici est d'identifier les gènes différentiellement exprimés entre les patientes "sensibles" et les patientes "résistantes". Un premier filtre a été appliqué permettant de supprimer environ 10 000 gènes dont l'expression ne variait pas entre les différentes patientes. Ensuite, les 3 214 gènes présentant un Fold Change supérieur à 2 (i.e. les gènes au moins deux fois plus exprimés chez les "sensibles" ou chez les "résistantes") ont été conservés.

Sur ces gènes, un Welch test avec une p-value à 0.05 a alors été appliqué permettant d'identifier 101 gènes différentiellement exprimés. Une Analyse en Composantes Principales (cf figure 5.4) ainsi qu'une classification hiérarchique (cf figure 5.5) ont alors été réalisées sur ces 101 gènes.

Le premier axe de l'ACP dont l'inertie est supérieure à 50% sépare bien les deux types de patientes (sensibles en bleu et résistantes en jaune). De même, ces patientes se retrouvent bien sur deux branches distinctes dans l'arbre obtenu. Ces résultats montrent que les gènes obtenus permettent bien de distinguer les deux types de patientes.

Des modèles de prédiction ont également été construits à partir de ces différents profils et comparés entre eux. Ces modèles ont été générés sous GeneSpring [1] en utilisant la

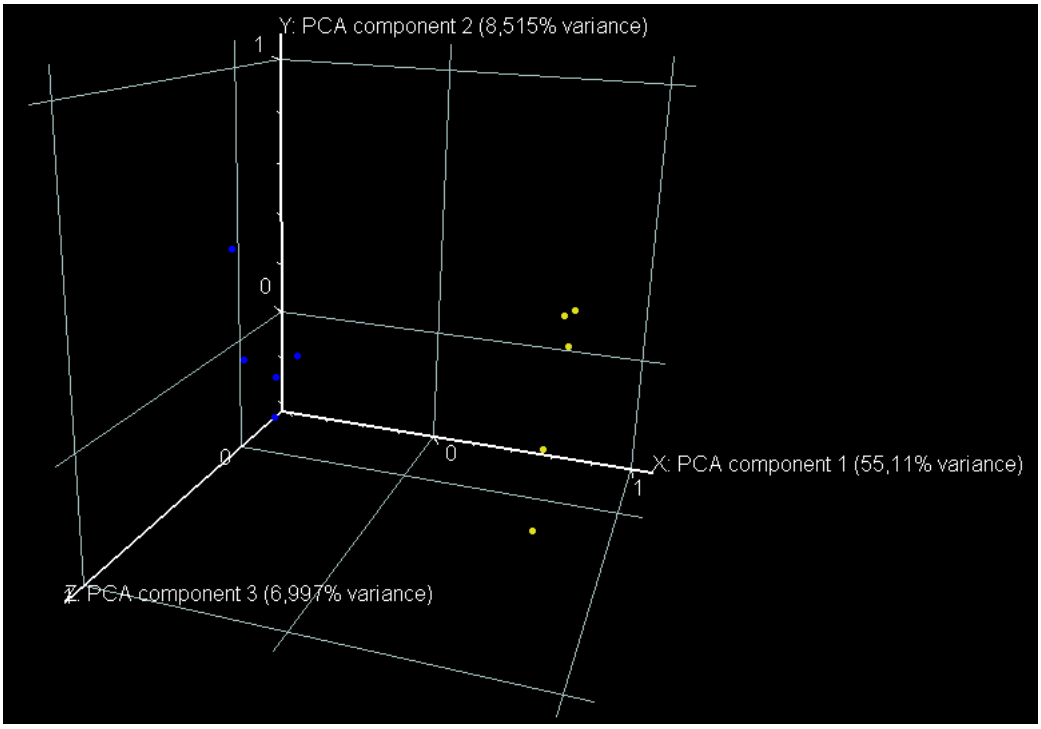


FIGURE 5.4 – Analyse en Composantes Principales

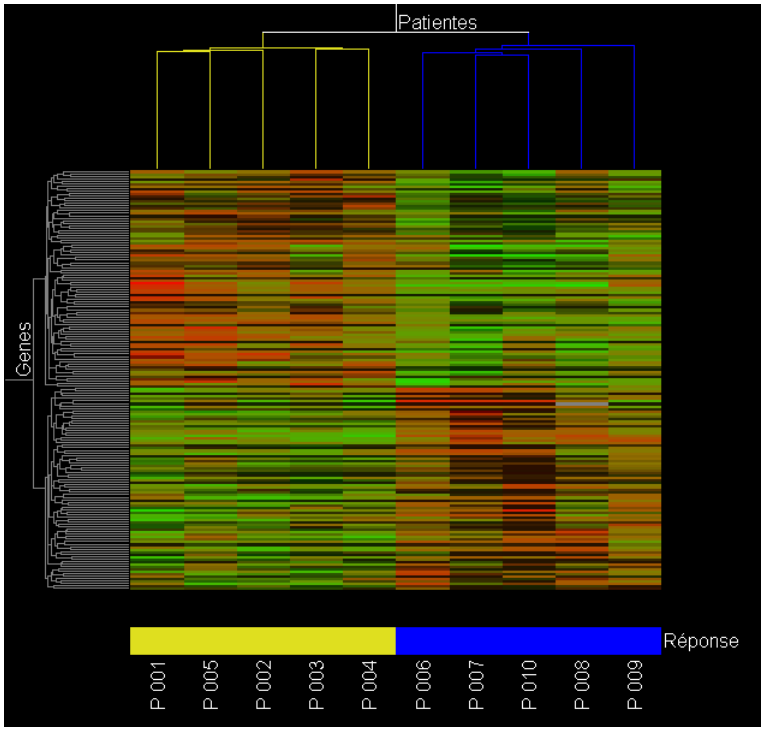


FIGURE 5.5 – Classification hiérarchique

méthode des k plus proches voisins et des support vector machines. Finalement, le modèle retenu a été obtenu avec la méthode des suport vector machines sur l'ensemble des 101 gènes et présente un taux de prédiction de 100% en Leave-One-Out Cross-Validation.

Ces gènes permettent ainsi de mettre en évidence un profil d'expression caractéristique de la réponse thérapeutique au docétaxel.

Deuxième partie

Différents types de règles entre gènes

Chapitre 6

Présentation de l'approche

Sommaire

Nous avons vu dans la partie précédente, plusieurs méthodes d'analyses de données permettant d'extraire de la connaissance à partir des données d'expression. Dans cette partie, une nouvelle technique est présentée, elle consiste à découvrir différents types de règles entre gènes.

Ce chapitre présente tout d'abord l'approche générale que nous avons suivie. Le chapitre 7 introduit ensuite différents types de règles pouvant être découvertes à partir de données d'expression de gènes et le principe de génération de ces règles est détaillé dans le chapitre 8.

6.1 Préliminaires

Pour motiver l'approche proposée, considérons la relation quelconque r donnée dans la table 6.1. Cette relation est composée de 8 tuples (t_1 à t_8) et de 9 attributs : a_1, a_2, a_3 et a_4 sont des attributs numériques, a_5, a_6 et a_7 sont des attributs binaires et a_8 et a_9 sont des attributs catégoriels.

Nous noterons par $t_i[a_j]$ la valeur prise par l'attribut a_j pour le tuple t_i, par exemple $t_1[a_2] = 2.1$.

La méthode que nous proposons consiste à découvrir différents types de règles entre

r	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9
t_1	2.7	2.1	−2.7	0.1	$faux$	$faux$	$faux$	c_1	T
t_2	−0.8	−4.0	1.4	5.4	$vrai$	$faux$	$vrai$	c_1	T
t_3	−0.1	2.4	4.5	−0.3	$faux$	$faux$	$faux$	c_1	T
t_4	−2.1	−0.7	−1.7	−1.4	$faux$	$vrai$	$faux$	c_1	T
t_5	−1.2	−2.4	−0.1	−0.2	$vrai$	$faux$	$vrai$	c_2	S
t_6	1.2	−3.1	2.4	3.1	$faux$	$faux$	$vrai$	c_2	S
t_7	2.7	5.4	−2.4	2.1	$faux$	$vrai$	$vrai$	c_3	T
t_8	−2.1	3.4	3.4	−1.4	$faux$	$faux$	$faux$	c_3	T

TABLE 6.1 – Relation r composée de 8 tuples et de 9 attributs

attributs. Une **règle** est une expression de la forme $X \rightarrow Y$ qui se lit "X implique Y" avec X et Y deux ensembles d'attributs, par exemple : $\{a_6, a_7\} \rightarrow \{a_5\}$ qui sera notée par la suite $a_6a_7 \rightarrow a_5$ par commodité.

Maintenant il faut se poser la question de savoir que signifie une règle. Par exemple, la règle $a_6a_7 \rightarrow a_5$ peut signifier qu'à chaque fois que a_6 et a_7 sont $faux$ alors a_5 est aussi $faux$, mais la règle $a_6a_7 \rightarrow a_5$ peut tout aussi bien signifier qu'à chaque fois que a_6 et a_7 sont $faux$ alors a_5 est $vrai$. C'est ce que nous appelons la **sémantique** de la règle i.e. la *signification*, le *sens* donné à la règle.

Or dans la relation r, nous pouvons remarquer qu'à chaque fois que a_6 et a_7 sont $faux$ alors a_5 est aussi $faux$, on dit alors que la règle $a_6a_7 \rightarrow a_5$ est **satisfaite** dans la relation r avec cette sémantique, nommée par commodité s, qui sera notée $r \models_s a_6a_7 \rightarrow a_5$. De la même façon, nous pouvons donner d'autres règles satisfaites dans la relation r, chacune avec une sémantique particulière comme décrit dans l'exemple 4.

Exemple 4:

- *On remarque que pour tout tuple t de la relation r, si $t[a_1] \geq 2.0$ alors $t[a_2] \geq 2.0$.*
- *On remarque que pour tout tuple t, si $t[a_2]$ et $t[a_3]$ sont strictement positifs alors $t[a_1]$ est strictement négatif.*
- *On remarque que pour tout couple de tuples distincts $\{t_i, t_j\}$ tel que $t_i[a_8] = t_j[a_8] =$ 'c_1', si $|t_j[a_2] - t_i[a_2]| > 3.0$ alors $|t_j[a_3] - t_i[a_3]| > 3.0$.*
- *On remarque que pour tout couple de tuples distincts $\{t_i, t_j\}$ tel que $t_i[a_8] = t_j[a_8] \in \{$'$c_2$', '$c_3$'$\}$, si $d(t_i[a_1a_2], t_j[a_1a_2]) \geq 5.0$ alors $d(t_i[a_3], t_j[a_3]) \geq 5.0$, où d est la distance euclidienne.*

Cet exemple nous montre qu'il est possible de définir un très grand nombre de sémantiques pour les règles.

Compte tenu des différents objectifs biologiques possibles, il peut être utile de découvrir différents types de règles entre gènes et ne pas se restreindre à une seule sémantique. L'idée de notre approche est alors d'offrir aux biologistes la possibilité de choisir parmi plusieurs sémantiques, le sens des règles qu'ils souhaitent générer.

L'originalité de ce travail est de proposer un cadre global pouvant inclure un grand nombre de sémantiques pour les règles et d'utiliser des méthodes identiques de génération, de post-traitement et de visualisation des règles pour toutes les sémantiques proposées.

Nous donnons tout d'abord une définition générique des sémantiques puis nous nous intéressons plus particulièrement à un certain type de sémantiques appelées sémantiques bien-formées.

6.2 Définition générique des sémantiques

A partir des exemples donnés précédemment, nous pouvons dégager deux caractéristiques inhérentes aux sémantiques qui nous seront utiles pour définir proprement cette notion :

- *Les sous-ensembles de la relation sur lesquels la règle s'applique*. Il est possible par exemple d'étudier tous les tuples un à un comme pour les deux premières sémantiques de l'exemple 4, de considérer tous les couples de tuples ou bien uniquement certains tuples ou certains couples de tuples comme pour les deux dernières sémantiques de l'exemple.
- *Les prédicats qui donnent réellement le sens de la règle* : "**si** $Pred_1$ est vrai pour X **alors** $Pred_2$ est vrai pour Y". Notons que ces deux prédicats peuvent être les mêmes. Par exemple, pour la première sémantique, les deux prédicats (identiques) peuvent être formulés de la façon suivante : $[\ \forall A \in X, t[A] \geq 2.0\ ]$. Pour la deuxième sémantique, les deux prédicats sont différents et peuvent être formulés ainsi : $Pred_1 = [\ \forall A \in X, t[A] > 0.0\ ]$ et $Pred_2 = [\ \forall A \in X, t[A] < 0.0\ ]$.

Les sémantiques données précédemment peuvent alors se caractériser à partir d'un ensemble, qui sera noté c, contenant les sous-ensembles r' de r devant être considérés (par exemple, $c(r) = \{\{t\} \mid t \in r\}$), et de deux prédicats $Pred_1$ et $Pred_2$. Les prédicats sont des **conditions** devant être définies sur X (ou Y) et r' **seulement**. Aucun autre attribut ou sous-ensemble de r n'est autorisé dans leur définition. En d'autres termes, ils doivent être définis sur $\pi_X(r')$ (ou $\pi_Y(r')$).

Afin de préciser les conditions admissibles, nous en donnons une définition inductive comme suit :

Une **condition simple** sur un ensemble d'attributs X et une relation r', est une expression de la forme : `<terme>` θ `<terme>`, où :

- θ est un opérateur de comparaison : $=, <, >, \leq, \geq, \neq$.
- Un `<terme>` est un des éléments suivants (avec $A, B \in X$, $Y \subseteq X$ et $t \in r'$) :
  - Une valeur dans $\pi_X(r')$: $t[A], t[B], ...$
  - Une constante : $a, b, 8, \varepsilon, null, ...$
  - Une fonction : $fct(r', X), fct(r', A), fct(t, Y)$, ... e.g. $d(t[A], t[B])$ ou $|r'|$.

Une **condition** sur X et r', est une expression composée d'une ou plusieurs **conditions simples** sur X et r', liées grâce aux connecteurs logiques : AND, OR, NOT, (). De plus, les variables $A, B, Y, ...$ (resp. t) sont introduites en utilisant les quantifieurs $\forall$ et $\exists$ sur X (resp. r') dans la partie déclarative de la **condition**.

Ainsi formulée, la sémantique classique des dépendances fonctionnelles peut par exemple se caractériser par l'ensemble c et les prédicats suivants :

- $c(r) = \{\{t_i, t_j\} \mid t_i, t_j \in r \text{ AND } t_i \neq t_j \}$.
- $Pred_1(X, \{t_i, t_j\}) = Pred_2(X, \{t_i, t_j\}) = [\ \forall A \in X, t_i[A] = t_j[A]\ ]$.

Nous avons ainsi dissocié les caractéristiques propres des sémantiques, de la définition de la satisfaction des règles (cf définition 1). L'exemple 5 ci-dessous reprend les illustrations données dans l'exemple 4 et précise les sémantiques sous jacentes.

Exemple 5:
Les quatre sémantiques notées s_1, s_2, s_3 et s_4, sont caractérisées par les ensembles et prédicats suivants.

- *Sémantique s_1*
  - $c(r) = \{\{t\} \mid t \in r\}$.
  - $Pred_1(X, \{t\}) = Pred_2(X, \{t\}) = [\ \forall A \in X, t[A] \geq 2.0\ ]$.

- *Sémantique s_2*
  - $c(r) = \{\{t\} \mid t \in r\}$.
  - $Pred_1(X, \{t\}) = [\ \forall A \in X, t[A] > 0.0\ ]$.
  - $Pred_2(X, \{t\}) = [\ \forall A \in X, t[A] < 0.0\ ]$.

- *Sémantique s_3*
  - $c(r) = \{\{t_i, t_j\} \mid t_i, t_j \in r \text{ AND } t_i[a_8] = t_j[a_8] = \text{`}c_1\text{`} \text{ AND } t_i \neq t_j\}$.
  - $Pred_1(X, \{t_i, t_j\}) = Pred_2(X, \{t_i, t_j\}) = [\ \forall A \in X, |t_j[A] - t_i[A]| > 3.0\ ]$.

- *Sémantique s_4*
  - $c(r) = \{\{t_i, t_j\} \mid t_i, t_j \in r \text{ AND } t_i[a_8] = t_j[a_8] \in \{\text{`}c_2\text{`}, \text{`}c_3\text{`}\} \text{ AND } t_i \neq t_j\}$.
  - $Pred_1(X, \{t_i, t_j\}) = Pred_2(X, \{t_i, t_j\}) = [\ d(t_i[X], t_j[X]) \geq 5.0\ ]$, *où d est la distance euclidienne.*

Notation 1:
Dans la suite, nous dénotons par C la classe de l'ensemble des sémantiques pouvant être définies par un ensemble c et deux prédicats $Pred_1$ et $Pred_2$ définis pour un ensemble d'attributs.

6.2 Définition générique des sémantiques

Nous définissons à présent la satisfaction des règles pour les sémantiques appartenant à la classe C :

Définition 1:
Soient r une relation définie sur U, $X, Y \subseteq U$ deux ensembles d'attributs et $s \in C$ une sémantique caractérisée par c, $Pred_1$ et $Pred_2$.
La règle $X \rightarrow Y$ est ***satisfaite*** *dans r pour la sémantique s, notée $r \models_s X \rightarrow Y$ si et seulement si :*

$$\forall r' \in c(r),$$

$$si\ Pred_1(X, r')\ est\ vrai\ alors\ Pred_2(Y, r')\ est\ vrai.$$

Le cas des sémantiques pathologiques

Certaines sémantiques sont définies de telle sorte que quelle que soit la relation r et quels que soient les ensembles d'attributs X et Y, toutes les règles $X \rightarrow Y$ sont fausses (resp. vraies). En pratique, ces sémantiques que nous appellerons sémantiques **pathologiques**, ne présentent aucun intérêt et ne seront pas prises en compte. Leur définition est donnée dans la suite :

Définition 2:
Soit $s \in C$ une sémantique caractérisée par un ensemble c et deux prédicats $Pred_1$ et $Pred_2$. La sémantique s est dite ***pathologique*** *si pour toute relation r sur U, une des conditions suivantes est vraie :*
– *$c(r)$ est égal à l'ensemble vide.*
– *$\forall X \subseteq U$ et $\forall r' \in c(r)$, $Pred_1(X, r')$ est vrai et $Pred_2(X, r')$ est faux.*
– *$\forall X \subseteq U$ et $\forall r' \in c(r)$, $Pred_1(X, r')$ est faux.*
– *$\forall X \subseteq U$ et $\forall r' \in c(r)$, $Pred_2(X, r')$ est vrai.*

Des sémantiques pathologiques sont données dans les exemples 6 et 7 :

Exemple 6:
Soit s la sémantique partiellement définie comme suit :

1. $c(r) = \{\{t\} \mid t \in r \text{ AND } \forall A \in U, t[A] < 0\}$.
2. $Pred_1(X, \{t\}) = [\ \forall A \in X, t[A] > 0\ ]$.

La sémantique s est clairement une sémantique pathologique puisque $Pred_1$ est toujours faux quelle que soit la définition de $Pred_2$. Pour cette sémantique, toutes les règles sont satisfaites pour n'importe quelle relation.

Exemple 7:
Soit s la sémantique caractérisée par l'ensemble c et les prédicats $Pred_1$ et $Pred_2$ suivants :

1. $c(r) = \{\{t_i, t_j\} \mid t_i, t_j \in r \text{ AND } \forall A \in U, t_i[A] < t_j(A)\}$.
2. $Pred_1(X, \{t_i, t_j\}) = [\ \exists A \in X, t_i[A] < t_j(A)\ ]$.
3. $Pred_2(X, \{t_i, t_j\}) = [\ \exists A \in X, t_i[A] > t_j[A]\ ]$.

La sémantique s est aussi clairement une sémantique pathologique puisque $Pred_1$ est toujours vrai et $Pred_2$ est toujours faux. Ainsi, pour cette sémantique, aucune règle n'est satisfaite quelle que soit la relation.

Notation 2:
Dans la suite, les notations suivantes sont utilisées :
– *C_P dénote la classe des sémantiques pathologiques de C, i.e. $C_P \subset C$.*
– *C_X dénote la classe des sémantiques non pathologiques de C, i.e. $C_X = C \setminus C_P$.*

6.3 Sémantiques bien-formées

Nous nous intéressons maintenant à un groupe particulier de sémantiques que nous appellerons "**sémantiques bien-formées**". Ces sémantiques ont la particularité de satisfaire des propriétés intéressantes venant de la théorie des dépendances fonctionnelles et plus particulièrement des axiomes d'Armstrong [76, 39]. Ce cadre offre plusieurs avantages que nous développons dans la suite, auparavant nous donnons la définition des sémantiques bien-formées :

Définition 3:
*Une sémantique s est dite **bien-formée** si les axiomes d'Armstrong sont justes et complets pour s.*

Rappelons le système d'axiomes d'Armstrong [23] pour un ensemble de règles F défini sur U un ensemble d'attributs :

1. (réflexivité) si $X \subseteq Y \subseteq U$ alors $F \vdash Y \rightarrow X$
2. (augmentation) si $F \vdash X \rightarrow Y$ et $W \subseteq U$, alors $F \vdash XW \rightarrow YW$
3. (transitivité) si $F \vdash X \rightarrow Y$ et $F \vdash Y \rightarrow Z$ alors $F \vdash X \rightarrow Z$

La notation $F \vdash X \rightarrow Y$ signifie qu'une preuve de $X \rightarrow Y$ peut être obtenue à partir de F en utilisant les axiomes d'Armstrong. De plus, pour une sémantique donnée s, la notation $F \models_s X \rightarrow Y$ signifie que pour toute relation r définie sur tout ensemble d'attributs U, si $r \models_s F$ alors $r \models_s X \rightarrow Y$.

Le système d'axiomes d'Armstrong est juste et complet si ces trois axiomes ne génèrent pas de règles incorrectes (la justesse) et s'ils génèrent bien toutes les règles possibles

pouvant être déduites de F (la complétude). Montrer que le système d'axiomes d'Armstrong est juste et complet pour une sémantique donnée s revient donc à montrer que si $F \vdash X \rightarrow Y$ alors $F \models_s X \rightarrow Y$ (la justesse) et que si $F \models_s X \rightarrow Y$ alors $F \vdash X \rightarrow Y$ (la complétude). Finalement, pour toute sémantique bien-formée s, $\vdash$ et $\models_s$ coïncident.

Les axiomes d'Armstrong ont été prouvés justes et complets pour les dépendances fonctionnelles [94]. Avec notre terminologie, la sémantique des dépendances fonctionnelles est donc bien-formée.

Dans un contexte de découverte de connaissances, la découverte de règles satisfaites dans une relation pour une sémantique bien-formée offre les avantages suivants :

- Il est tout d'abord possible de **raisonner** sur les règles à partir des axiomes d'Armstrong sans accéder aux données. A partir d'un ensemble de règles F, il est possible de savoir si une règle est **impliquée** par cet ensemble de règles en temps linéaire [26]. Ainsi, si on dispose d'une relation r qui satisfait F alors on sait que toutes les règles pouvant être déduites de F par les axiomes d'Armstrong seront satisfaites dans cette relation.
- Nous pouvons également travailler sur des "petites" **couvertures** des règles [75, 54] et proposer un processus de découverte spécifique à la couverture considérée mais applicable à *toute* sémantique bien-formée, ce qui laisse entrevoir une grande généricité dans le traitement opérationnel.
- De plus, il est aussi possible de proposer des couvertures pour les règles approximatives [51] (cf chapitre 8).

6.4 Nouvelles restrictions syntaxiques

Nous avons défini la classe C_X des sémantiques non pathologiques de C permettant de capturer un large choix de sémantiques.

Comme nous nous intéressons plus particulièrement aux sémantiques bien-formées, nous pouvons alors nous poser la question suivante : "Est-ce qu'il y a une équivalence entre la classe des sémantiques bien-formées et la classe C_X ?"

Malheureusement mais de façon non surprenante, la réponse est négative comme le montre le contre-exemple suivant.

Exemple 8:
La sémantique $s_2 \in C_X$ n'est pas bien-formée puisque l'axiome de réflexivité n'est pas juste pour cette sémantique. En effet, il est possible de donner une relation r telle que $r \not\models_{s_2} AB \rightarrow A$.

Ainsi pour chaque sémantique de la classe C_X, nous devons vérifier si le système d'axiomes d'Armstrong est juste et complet, la preuve étant relativement longue et fastidieuse.

Pour pallier ce problème, nous définissons une nouvelle classe de sémantiques $C_A \subseteq C_X$ en espérant que celle-ci coïncide avec les sémantiques bien-formées :

Notation 3:
Dans la suite, nous notons C_A la classe des sémantiques pouvant être caractérisées par un ensemble c et un prédicat $Pred$ défini pour un seul attribut.

La sémantique des dépendances fonctionnelles appartient par exemple à la classe C_A puisqu'elle peut se caractériser par l'ensemble c et le prédicat suivants :
- $c(r) = \{\{t_i, t_j\} \mid t_i, t_j \in r \text{ AND } t_i \neq t_j \}$.
- $Pred(A, \{t_i, t_j\}) = [\ t_i[A] = t_j[A]\ ]$.

Pour les sémantiques appartenant à la classe C_A, la définition de la satisfaction des règles devient :

Définition 4:
Soient r une relation définie sur U, $X, Y \subseteq U$ deux ensembles d'attributs et $s \in C_A$ une sémantique caractérisée par c et $Pred$.
La règle $X \rightarrow Y$ est ***satisfaite*** *dans r pour la sémantique s, notée $r \models_s X \rightarrow Y$ si et seulement si :*

$$\forall r' \in c(r),$$

$$si\ \forall A \in X, Pred(A, r')\ est\ vrai\ alors\ \forall A \in Y, Pred(A, r')\ est\ vrai.$$

La différence est double par rapport à la définition 1 :

- Premièrement, il n'y a plus qu'un seul prédicat et non plus deux comme précédemment.
- Deuxièmement, une restriction est posée sur le prédicat : Il doit être satisfait pour chaque **attribut** $A \in X$ plutôt que d'être satisfait par **l'ensemble d'attributs** X.

Notons que deux prédicats peuvent être syntaxiquement différents et être pourtant équivalents. Nous donnons dans la suite la définition de l'équivalence entre deux prédicats :

Définition 5:
Soit $s \in C_X$ une sémantique caractérisée par un ensemble c et deux prédicats $Pred_1$ et $Pred_2$. Les deux prédicats $Pred_1$ et $Pred_2$ sont dits ***équivalents****, notés par $Pred_1 \equiv Pred_2$, si et seulement si pour toute relation r sur U et pour tout ensemble d'attributs $X \subseteq U$, nous avons :*
$\{r' \in c(r) \mid Pred_1(X, r')\ est\ vrai\} = \{r' \in c(r) \mid Pred_2(X, r')\ est\ vrai\}$.

Comme nous le souhaitions, nous avons maintenant une équivalence entre cette nouvelle classe de sémantiques C_A et l'ensemble des sémantiques bien-formées :

Théorème 1:
Soit $s \in C_X$ une sémantique. La sémantique s est bien-formée si et seulement si $s \in C_A$.

Preuve
Soit $s \in C_X$ une sémantique. Nous devons d'abord montrer que si $s \in C_A$ alors s est bien-formée puis si s est bien-formée alors $s \in C_A$ ou de façon équivalente si $s \notin C_A$ alors s n'est pas bien-formée.

Lemme 4:
Soit $s \in C_X$ une sémantique. Si $s \in C_A$ alors s est bien-formée.

Preuve
(Si) Soit $s \in C_A$ une sémantique. Nous devons montrer que s est bien-formée i.e. que le système d'axiomes d'Armstrong est juste et complet pour s. Cette preuve est basée sur la preuve classique de justesse et de complétude du système d'Armstrong pour les dépendances fonctionnelles [94].

Lemme 5:
Le système d'axiomes d'Armstrong est juste pour s.

Preuve
Soit F un ensemble de règles sur U. Nous devons montrer que si $F \vdash X \to Y$ alors $F \models_s X \to Y$, i.e. soit r une relation sur U telle que $r \models_s F$, si $F \vdash X \to Y$ alors $r \models_s X \to Y$.

1. *(réflexivité) Nous devons montrer que si $X \subseteq Y \subseteq U$ alors $r \models_s Y \to X$.*
   Soient $X \subseteq Y \subseteq U$ et $r' \in c(r)$ tel que $\forall A \in Y$, $Pred(A, r')$ est vrai. Puisque $X \subseteq Y$, alors $\forall A \in X$, $Pred(A, r')$ est vrai et ainsi nous avons bien $r \models_s Y \to X$.
2. *(augmentation) Nous devons montrer que si $F \vdash X \to Y$ et $W \subseteq U$, alors $r \models_s XW \to YW$.*
   Soit $r' \in c(r)$ tel que $\forall A \in X \cup W$, $Pred(A, r')$ est vrai. Si $F \vdash X \to Y$, alors nous avons $\forall B \in Y$, $Pred(B, r')$ est vrai et donc $r \models_s XW \to YW$.
3. *(transitivité) Nous devons montrer que si $F \vdash X \to Y$ et $F \vdash Y \to Z$ alors $r \models_s X \to Z$.*
   Soit $r' \in c(r)$ tel que $\forall A \in X$, $Pred(A, r')$ est vrai. Si $F \vdash X \to Y$ et $F \vdash Y \to Z$, alors $\forall B \in Y$, $Pred(B, r')$ est vrai et $\forall C \in Z$, $Pred(C, r')$ est vrai respectivement. Nous avons donc bien $r \models_s X \to Z$.

Notons que cette preuve est possible grâce à la restriction posée sur le prédicat, qui doit être satisfait pour chaque attribut $A \in X$.

Lemme 6:

Le système d'axiomes d'Armstrong est complet pour s.

Preuve
Nous devons montrer que si $F \models_s X \rightarrow Y$ alors $F \vdash X \rightarrow Y$ ou de façon équivalente, si $F \nvdash X \rightarrow Y$ alors $F \not\models_s X \rightarrow Y$. Supposons alors que $F \nvdash X \rightarrow Y$, c'est suffisant de donner une relation r telle que $r \models_s F$ mais $r \not\models_s X \rightarrow Y$.

Notons que pour exhiber un contre-exemple r, il n'est pas nécessaire d'expliciter les données. En effet, soit r une relation telle que $\exists r' \in c(r)$ tel que $\forall A \in X^+$, $Pred(A, r)$ est vrai et $\forall B \in U \setminus X^+$, $Pred(B, r)$ est faux. Rappelons que $X^+ = \{A \in U \mid F \vdash X \rightarrow A\}$. Il est toujours possible de construire une telle relation r puisque la sémantique s appartient à la classe C_A, ce n'est donc pas une sémantique pathologique (cf section 6.2).
Si nous considérons juste la sous-relation r', nous avons, de par sa construction : $r' \not\models_s V \rightarrow W$ si et seulement si $V \in X^+$ et $\exists A \in W$ tel que $A \in U \setminus X^+$. Sinon, $r' \models_s V \rightarrow W$.

Premièrement, nous devons montrer que $r' \models_s F$. Supposons le contraire i.e. $r' \not\models_s F$ et ainsi, $\exists V \rightarrow W \in F$ telle que $r' \not\models_s V \rightarrow W$. Par la construction de r', nous avons $V \subseteq X^+$ et $\exists A \in W$ tel que $A \in U \setminus X^+$. Puisque $V \in X^+$, nous avons $F \vdash X \rightarrow V$ et puisque $F \vdash V \rightarrow W$, nous avons $F \vdash V \rightarrow A$. Ainsi, par transitivité, $F \vdash X \rightarrow A$ et ainsi $A \in X^+$. Ceci mène à une contradiction puisque $A \in W$, et donc $r' \models_s F$.

Deuxièmement, nous devons montrer que $r' \not\models_s X \rightarrow Y$. Supposons le contraire i.e. $r' \models_s X \rightarrow Y$. Par la construction de r', nous avons $Y \subseteq X^+$ et ainsi $F \vdash X \rightarrow Y$. Ceci mène à une contradiction puisque nous avons supposé que $F \nvdash X \rightarrow Y$, et donc $r' \not\models_s X \rightarrow Y$.

La seconde partie de la preuve ("seulement si") est plus surprenante puisqu'elle dit que toute sémantique bien-formée appartient à la classe C_A :

Lemme 7:
Soit $s \in C_X$ une sémantique. Si la sémantique s est bien-formée, alors $s \in C_A$.

Preuve
(Seulement si) Nous devons maintenant montrer que si s est bien-formée, alors $s \in C_A$ ou de façon équivalente, si $s \notin C_A$ alors s n'est pas bien-formée.

Supposons que $s \notin C_A$, deux cas sont alors possibles : Soit les deux prédicats ne sont pas équivalents, soit ils sont équivalents avec la restriction suivante : Il n'existe pas de prédicat équivalent pouvant être formulé comme une condition sur chaque attribut.

Considérons le premier cas, i.e. $Pred_1$ et $Pred_2$ ne sont pas équivalents : Dans ce cas, il existe une relation r et un ensemble d'attributs $Y \subseteq U$ tels que $\{r' \in c(r) \mid Pred_1(Y, r')$ est vrai$\} \neq \{r' \in c(r) \mid Pred_2(Y, r')$ est vrai$\}$.

Deux cas sont alors possibles :

– $\exists r' \in c(r)$ *tel que* $Pred_1(Y, r')$ *est vrai et* $Pred_2(Y, r')$ *est faux :*
Supposons que s soit bien-formée. Par réflexivité, nous avons $\forall X \subseteq Y$, $r \models_s Y \to X$ *et donc* $r \models_s Y \to Y$ *i.e.* $\forall r' \in c(r)$, *si* $Pred_1(Y, r')$ *est vrai alors* $Pred_2(Y, r')$ *est vrai ce qui mène à une contradiction.*
– $\exists r' \in c(r)$ *tel que* $Pred_1(Y, r')$ *est faux et* $Pred_2(Y, r')$ *est vrai :*
Sans perte de généralité, supposons qu'il existe $X \in U \setminus Y$ *et* $Z \in U \setminus Y$ *tel que* $Pred_1(X, r')$ *est vrai et* $Pred_2(Z, r')$ *est faux, comme décrit dans la table 6.2 (clairement, une telle relation r existe toujours lorsque s appartient à* C_X, *sinon s appartiendrait à* C_P*). Ainsi, nous avons* $r' \models_s X \to Y$ *et* $r' \models_s Y \to Z$.

r	X	Y	Z	...
...	...	...	...	...
$r'\{$	P_1 vrai	P_1 faux et P_2 vrai	P_2 faux	...
...	...	...	...	...

TABLE 6.2 – Relation exemple

Supposons maintenant que s est bien-formée. Par transitivité, nous devrions avoir $r' \models_s X \to Z$, *ce qui est faux est mène à une contradiction.*

Finalement, nous avons montré que si $Pred_1$ *et* $Pred_2$ *ne sont pas équivalents, alors s n'est pas bien-formée.*

Maintenant, considérons le second cas, i.e. $Pred_1 \equiv Pred_2$ *mais ils ne sont pas équivalents à un prédicat* $Pred'(Y) = [\forall A \in Y, Pred_1(A)]$ *: Dans ce cas, il existe une relation r et un ensemble d'attributs* $Y \subseteq U$ *tels que* $\{r' \in c(r) \mid Pred_1(Y, r')$ *est vrai*$\} \neq \{r' \in c(r) \mid \forall A \in Y, Pred_1(A, r')$ *est vrai*$\}$. *Nous pouvons alors montrer que les axiomes de réflexivité et de transitivité ne sont pas justes dans* r'. *La preuve est similaire aux précédentes :*

Deux cas sont possibles :
– $\exists r' \in c(r)$ *tel que* $Pred_1(Y, r')$ *est vrai et* $\exists A \in Y Pred_1(A, r')$ *est faux :*
Supposons que s soit bien-formée alors par l'axiome de réflexivité, nous avons $\forall A \in Y$, $r \models_s Y \to A$ *i.e.* $\forall r' \in c(r)$, *si* $Pred_1(Y, r')$ *est vrai alors* $Pred_1(A, r')$ *est vrai* $\forall A \in Y$, *ce qui mène à une contradiction.*
– $\exists r' \in c(r)$ *tel que* $\forall A \in Y Pred_1(A, r')$ *est vrai mais* $Pred_1(Y, r')$ *est faux :*
Pour toute sémantique s, il est possible de construire une relation r'' *sur* $\{XYZ\}$ *telle que* $r''[Y] = r'[Y]$, $Pred_1(X, r'')$ *est vrai et* $\forall A \in Z Pred_2(A, r'')$ *est faux. Par la construction de* r'', *nous avons* $r'' \models_s X \to Y$ *et* $r'' \models_s Y \to Z$. *Supposons que s soit bien-formée alors par l'axiome de transitivité, nous devrions avoir* $r'' \models_s X \to Z$, *ce qui est faux.*
Finalement, nous avons montré que si $Pred_1 = Pred_2$ *mais* $Pred_1 \neq Pred'$ *avec* $Pred'(Y) = \forall A \in Y, Pred_1(A)$, *alors s n'est pas bien-formée.*

Ce théorème permet donc de montrer rapidement qu'une sémantique est ou non bien-formée. En effet, il n'est plus nécessaire de montrer la justesse et la complétude du système

d'axiomes d'Armstrong pour chaque sémantique mais il suffit de montrer que la sémantique appartient à la classe C_A.

Exemple 9:
Les sémantiques s_1 et s_3 présentées dans l'exemple 5 peuvent être caractérisées par les ensembles et les prédicats suivants, ce qui suffit pour montrer que ces trois sémantiques sont bien-formées :
- *Sémantique s_1*
  - $c(r) = \{\{t\} \mid t \in r\}$.
  - $Pred(A, \{t\}) = [\ t[A] \geq 2.0\ ]$.

- *Sémantique s_3*
  - $c(r) = \{\{t_i, t_j\} \mid t_i, t_j \in r$ AND $t_i[a_8] = t_j[a_8] =$ 'c_1' AND $t_i \neq t_j\ \}$.
  - $Pred(A, \{t_i, t_j\}) = [\ |t_j[A] - t_i[A]| > 3.0\ ]$.

Par contre, la sémantique s_2 n'est pas bien-formée puisque les deux prédicats $Pred_1$ et $Pred_2$ ne sont pas équivalents.

De même, la sémantique s_4 n'est pas bien-formée puisque les prédicats $Pred_1 = Pred_2$ ne sont pas équivalents à un prédicat défini sur un seul attribut.

Nous avons défini un cadre théorique permettant d'inclure un grand nombre de sémantiques : l'ensemble des sémantiques bien-formées. Le théorème 1 nous permet de savoir rapidement si une sémantique peut ou non être incluse dans ce cadre global.

6.5 Indices de qualité des règles

Jusque là, nous avons introduit uniquement les règles **satisfaites** pour une sémantique et dans une relation données. Ainsi, les règles contredites par une seule sous-relation r' ne sont pas prises en compte, même si elles sont vraies pour un grand nombre de ces sous-relations. Pour prendre en considération ce type de règles, de nombreux indices de qualité ont été proposés, notamment pour les règles d'association.

Plus précisément, deux types d'indices permettent de caractériser les règles, les mesures d'erreurs et les mesures de qualité :
- Les **mesures d'erreur** comme par exemple la confiance définie pour les règles d'association [20] ou l'erreur g_3 définis pour les dépendances fonctionnelles [66], permettent d'ajouter aux règles exactes, les règles approximatives, i.e. celles qui sont **presque** satisfaites. Ces mesures sont très intéressantes puisqu'elles permettent notamment de prendre en compte le bruit dans les données.
- Les **mesures de qualité** comme le support, la dépendance ou le taux informationnel [79, 20, 32, 90, 31, 92], permettent au contraire de limiter le nombre de règles en

ne sélectionnant celles qui semblent les plus pertinentes. Ces mesures permettent de donner aux experts que les règles qui semblent les plus intéressantes, les plus surprenantes par rapport aux indices donnés.

Le cadre théorique que nous proposons nous permet d'étendre la définition de la plupart des indices existants à toute sémantique bien-formée.

En effet, la plupart des indices de qualité pour une règle $X \rightarrow Y$ sont calculés à partir de quatre valeurs : n, n_X, n_Y, n_{XY} dont la définition peut être adaptée à l'ensemble des sémantiques bien-formées :

- n : nombre de sous-ensembles $r' \in c(r)$.
- n_X : nombre de sous-ensembles $r' \in c(r)$ tels que $\forall A \in X$, $Pred(A, r')$ est vrai.
- n_Y : nombre de sous-ensembles $r' \in c(r)$ tels que $\forall A \in Y$, $Pred(A, r')$ est vrai.
- n_{XY} : nombre de sous-ensembles $r' \in c(r)$ tels que $\forall A \in X \cup Y$, $Pred(A, r')$ est vrai.

Exemple 10:
Voici par exemple les calculs réalisés pour la sémantique s_1 et la relation r (cf exemple 4) :
- *n : nombre de tuples $t \in r$.*
- *n_X : nombre de tuples $t \in r$ tels que $\forall A \in X$, $t[A] \geq 2.0$.*
- *n_Y : nombre de tuples $t \in r$ tels que $\forall A \in Y$, $t[A] \geq 2.0$.*
- *n_{XY} : nombre de tuples $t \in r$ tels que $\forall A \in X \cup Y$, $t[A] \geq 2.0$.*

Par exemple, nous avons $n = 8$, $n_{a_1} = 2$, $n_{a_2} = 4$ et $n_{a_1 a_2} = 2$.

A partir de ces paramètres, plusieurs indices de qualité peuvent être calculés pour chaque règle $X \rightarrow Y$ comme par exemple le support, la confiance [20], le lift [32], le leverage [79] ou la conviction [32] :

Support $(X \rightarrow Y)$ = Support $(Y \rightarrow X) = P(XY) = n_{XY}/n$.
: Le support correspond à la probabilité que X et Y soient simultanément satisfaits. Plus le support est grand, plus la règle est fréquente.

Confiance $(X \rightarrow Y) = P(XY|X) = n_{XY}/n_X$.
: La confiance correspond à la probabilité que Y soit satisfait sachant que X est satisfait. Lorsque la confiance est égale à 1 (ou 100%), la règle est dite **exacte**, sinon elle est dite **approximative**.

Lift $(X \rightarrow Y)$ = Lift $(Y \rightarrow X) = P(XY)/P(X)P(Y) = (n_{XY} * n)/(n_X * n_Y)$.
: Le lift mesure la dépendance entre X et Y. Il correspond au rapport entre la probabilité réelle d'avoir X et Y satisfaits et la probabilité attendue si X et Y étaient statistiquement indépendants.

Leverage $(X \rightarrow Y)$ = Leverage $(Y \rightarrow X)$ = P(XY) - P(X)*P(Y) = $(n_{XY}/n) - ((n_X/n) * (n_Y/n))$.
: Le leverage mesure aussi la dépendance entre X et Y, mais mesure cette fois la différence entre les deux probabilités.

Conviction $(X \rightarrow Y)$ = (P(X)*P(nonY)) / P(X et nonY) = $(n_X * (n - n_Y))/(n * (n_X - n_{XY}))$.

La conviction compare la probabilité attendue d'avoir X satisfait et Y non satisfait s'ils étaient indépendants avec la probabilité réelle d'avoir X satisfait et Y non satisfait. Un lift de 1, un leverage de 0 et une conviction de 1 signifient que les deux variables sont statistiquement indépendantes. Notons que pour les règles exactes, la conviction ne peut être calculée puisque P(X et nonY) est égale à 0.

Exemple 11:

Pour la sémantique s_1 *et la relation* r*, nous avons* $n = 8$*,* $n_{a_1} = 2$*,* $n_{a_2} = 4$ *et* $n_{a_1 a_2} = 2$*. Voici par exemple les indices pour les règles* $a_1 \rightarrow a_2$ *et* $a_2 \rightarrow a_1$ *:*

- *Support* $(a_1 \rightarrow a_2)$ = *Support* $(a_2 \rightarrow a_1)$ = $2/8 = 25\%$.
- *Confiance* $(a_1 \rightarrow a_2)$ = $2/2 = 100\%$.
- *Confiance* $(a_2 \rightarrow a_1)$ = $2/4 = 50\%$.
- *Lift* $(a_1 \rightarrow a_2)$ = *Lift* $(a_2 \rightarrow a_1)$ = $(2 * 8)/(2 * 4) = 2$.
- *Leverage* $(a_1 \rightarrow a_2)$ = *Leverage* $(a_2 \rightarrow a_1)$ = $(2/8) - ((2/8) * (4/8)) = 1/8 = 0.125$.
- *Conviction* $(a_2 \rightarrow a_1)$ = $(4 * (8 - 2))/(8 * (4 - 2)) = 1.5$.

Dans notre travail, nous avons choisi de ne pas intégrer les divers indices de qualité et d'erreur dans la définition générique des sémantiques proposée. Comme nous venons de le voir, ils peuvent en effet être calculés pour toute sémantique bien-formée et à notre sens, ils ne donnent pas la signification proprement dite des règles. De plus, nous ne souhaitons pas définir autant de sémantiques qu'il y a d'indices. Nous verrons dans la section suivante que nous générons une couverture des règles exactes mais aussi une couverture des règles approximatives. Les indices sont alors intégrés à posteriori pour filtrer et trier les règles obtenues.

Chapitre 7

Différentes sémantiques pour les données d'expression

Sommaire

Nous présentons dans ce chapitre, différentes sémantiques bien-formées spécifiquement adaptées aux données d'expression de gènes.

7.1 Etude des niveaux d'expression entre gènes

Pour illustrer cette première sémantique qui permet d'étudier les niveaux d'expression des gènes, considérons l'exemple suivant :

Exemple 12:
Soit la relation r_1 présentée dans la section 3.2 représentant les niveaux d'expression de 5 gènes pour 7 échantillons et redonnée dans la table 7.1.

Supposons qu'un gène est ***fortement exprimé*** *si son niveau d'expression est compris entre 1.0 et 2.0 (où 2.0 est la valeur max). Nous pouvons remarquer alors que chaque fois que le gène a_2 est fortement exprimé dans la relation r_1 (i.e. pour les échantillons t_1, t_2, t_5 et t_6), alors le gène a_3 est aussi fortement exprimé. Ceci peut se voir sur le graphique 7.1 représentant les niveaux d'expression des gènes a_2 et a_3.*

Ainsi, la règle $a_2 \rightarrow a_3$ est satisfaite dans r_1 avec la sémantique suivante : $a \rightarrow b$ si chaque fois que le gène a est fortement exprimé alors le gène b est aussi fortement exprimé.

r_1	a_1	a_2	a_3	a_4	a_5	*classe*
t_1	1.7	1.4	1.7	0.9	−1.9	c_1
t_2	1.8	1.6	1.9	0.8	2.0	c_1
t_3	−0.5	0.4	−1.4	1.0	0.4	c_1
t_4	−0.4	0.1	−0.4	0.8	−1.0	c_1
t_5	−1.4	1.8	1.3	1.9	0.1	c_2
t_6	−1.3	1.7	1.4	1.8	1.7	c_2
t_7	−1.4	−1.4	1.9	1.8	−2.0	c_2

TABLE 7.1 – Relation r_1

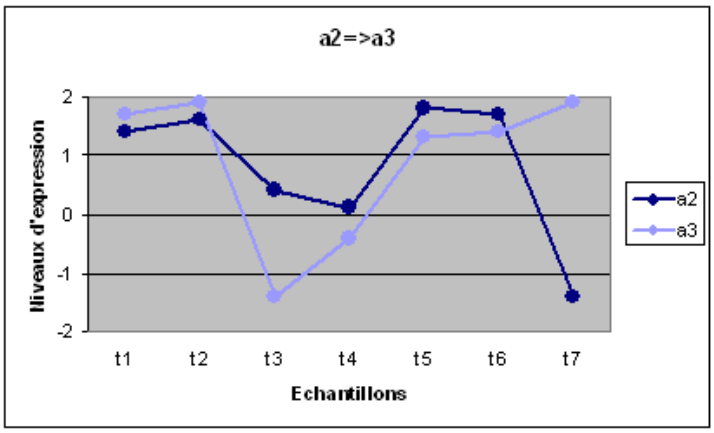


FIGURE 7.1 – Niveaux d'expression des gènes a_2 et a_3 dans la relation r_1

Nous pouvons remarquer par exemple qu'avec cette sémantique, la règle $a_3 \rightarrow a_2$ n'est pas satisfaite dans la relation r_1 puisque pour l'échantillon t_7, le gène a_3 est fortement exprimé tandis que le gène a_2 ne l'est pas.

Plus formellement, cette sémantique peut être définie de façon générale comme suit :

Définition 6:
Soit s_n la sémantique caractérisée par la contrainte et le prédicat suivants :
- $c(r) = \{\{t\} \mid t \in r\}$.
- $Pred(A, \{t\}) = [\epsilon_1^A \leq t[A] \leq \epsilon_2^A]$.

Nous avons alors le résultat suivant :

Propriété 1:
La sémantique s_n est bien-formée.

Preuve

Par le théorème 1, le résultat est évident puisque $s_n \in C_A$.

Les seuils ϵ_1^A et ϵ_2^A peuvent être identiques à chaque gène comme dans l'exemple 12 où $\epsilon_1 = 1.0$ et $\epsilon_2 = 2.0$ mais ils peuvent également être différents pour chaque gène (cf section 7.4).

Cette sémantique est proche des règles d'association, elle évite simplement une phase explicite de discrétisation.

Cette sémantique permet donc aux biologistes de découvrir des règles entre gènes fortement exprimés ou faiblement exprimés selon les seuils choisis. Nous pouvons également proposer des variantes de cette sémantique en considérant uniquement certains tuples de la relation au lieu de considérer tous les tuples de la relation. Par exemple, il est possible de se concentrer sur une seule classe e.g. c_1 et ne considérer ainsi que les tuples dont la classe est c_1 :

Définition 7:
Soit $s_n^{c_1}$ *la sémantique caractérisée par la contrainte et le prédicat suivants :*
– $c(r) = \{\{t\} \mid t \in r \text{ AND } class[t] = c_1\}$.
– $Pred(A, \{t\}) = [\epsilon_1^A \leq t[A] \leq \epsilon_2^A]$.

Propriété 2:
La sémantique $s_n^{c_1}$ *est bien-formée.*

Preuve
Par le théorème 1, le résultat est évident puisque $s_n^{c_1} \in C_A$.

Exemple 13:
Par exemple, la règle $a_3 \rightarrow a_2$ *qui n'était pas satisfaite dans la relation* r_1 *avec la sémantique* s_n*, est maintenant satisfaite avec la sémantique* $s_n^{c_1}$.

De la même façon, nous pouvons définir une sémantique pour chaque classe. Ceci nous permet par exemple de comparer les règles obtenues pour telle classe par rapport aux règles obtenues pour telle autre classe.

7.2 Etude de la variation des niveaux d'expression

Pour illustrer cette nouvelle sémantique qui permet d'étudier cette fois les **variations** des niveaux d'expression des gènes, considérons l'exemple suivant :

Exemple 14:
Soit la relation r_1 *donnée dans la table 7.1. Supposons que l'expression d'un gène* ***ne varie***

pas *entre deux échantillons si la différence de ses niveaux d'expression est comprise entre* 0.0 *et* 0.2 *(compte tenu du bruit présent dans les données). Nous pouvons remarquer alors que chaque fois que le gène* a_1 *ne varie pas entre deux échantillons de la relation* r_1 *(i.e. entre les échantillons* $t_1 - t_2$*,* $t_3 - t_4$*,* $t_5 - t_6$ *et* $t_6 - t_7$*), alors le gène* a_4 *ne varie pas non plus entre ces deux échantillons. Ceci peut se voir sur le graphique 7.2 représentant les niveaux d'expression des gènes* a_1 *et* a_4*.*

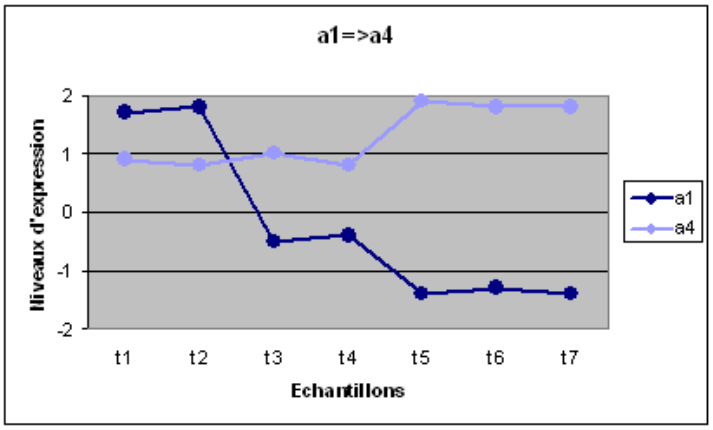


FIGURE 7.2 – Niveaux d'expression des gènes a_1 et a_4 dans la relation r_1

Ainsi, la règle $a_1 \rightarrow a_4$ *est satisfaite dans* r_1 *avec la sémantique suivante :* $a \rightarrow b$ *si chaque fois que le gène* a *ne varie pas entre deux échantillons, alors le gène* b *ne varie pas non plus entre ces deux échantillons.*

Nous pouvons remarquer par exemple qu'avec cette sémantique, la règle $a_4 \rightarrow a_1$ *n'est pas satisfaite dans la relation* r_1 *puisque entre les échantillons* t_2 *et* t_3*, le gène* a_4 *ne varie pas tandis que le gène* a_1 *varie.*

Plus formellement, cette sémantique peut être définie de façon générale comme suit :

Définition 8:
Soit s_v *la sémantique caractérisée par la contrainte et le prédicat suivants :*
- $c(r) = \{\{t_i, t_j\} \mid t_i, t_j \in r\}$.
- $Pred(A, \{t_i, t_j\}) = [\epsilon_1^A \leq |t_i[A] - t_j[A]| \leq \epsilon_2^A]$.

Nous avons alors le résultat suivant :

Propriété 3:
La sémantique s_v *est bien-formée.*

Preuve
Par le théorème 1, le résultat est évident puisque $s_v \in C_A$.

Comme précédemment, les seuils ϵ_1^A et ϵ_2^A peuvent être identiques à chaque gène comme dans l'exemple 14 où $\epsilon_1 = 0.0$ et $\epsilon_2 = 0.2$ ou bien différents pour chaque gène (cf section 7.4).

Cette sémantique prend en compte les lignes deux à deux, comme pour les dépendances fonctionnelles. Une DF $X \rightarrow Y$ peut être reformulée de la façon suivante : "A des valeurs égales de X correspondent des valeurs égales de Y". Mais pour les données d'expression, puisque la plupart des valeurs de la relation diffèrent l'une de l'autre, la satisfaction des DF a été relâchée. Avec cette sémantique, $X \rightarrow Y$ peut être reformulée de la façon suivante : "A des valeurs proches de X correspondent des valeurs proches de Y".
Notons que la satisfaction classique des DF est réalisée quand $\epsilon_1 = \epsilon_2 = 0$.

Cette sémantique permet donc aux biologistes de découvrir des règles entre gènes qui ne varient pas ou qui varient selon les seuils choisis. Comme précédemment, nous pouvons proposer des variantes de cette sémantique en considérant uniquement certains couples de tuples de la relation au lieu de considérer tous les couples de tuples.

Par exemple, nous pouvons nous concentrer sur une seule classe e.g. c_1 et ne considérer ainsi que les couples de tuples dont la classe est c_1 :

Définition 9:
Soit $s_v^{c_1}$ *la sémantique caractérisée par la contrainte et le prédicat suivants :*
– $c(r) = \{\{t_i, t_j\} \mid t_i, t_j \in r \text{ AND } class[t_i] = c_1 \text{ AND } class[t_j] = c_1\}$.
– $Pred(A, \{t_i, t_j\}) = [\epsilon_1^A \leq |t_i[A] - t_j[A]| \leq \epsilon_2^A]$.

Propriété 4:
La sémantique $s_v^{c_1}$ *est bien-formée.*

Preuve
Par le théorème 1, le résultat est évident puisque $s_v^{c_1} \in C_A$.

Exemple 15:
Par exemple, la règle $a_4 \rightarrow a_1$ *qui n'était pas satisfaite dans la relation* r_1 *avec la sémantique* s_v*, est maintenant satisfaite avec la sémantique* $s_v^{c_2}$.

De la même façon, nous pouvons définir une sémantique pour chaque classe. Ceci nous permet comme précédemment de comparer les règles obtenues pour telle classe par rapport aux règles obtenues pour telle autre classe.

Nous pouvons également ne considérer que les couples de tuples appartenant à une même classe. Ainsi, $a \rightarrow b$ si chaque fois que le gène a ne varie pas entre deux échantillons d'une même classe, alors le gène b ne varie pas non plus entre ces deux échantillons. Plus formellement :

Définition 10:
Soit $s_v^{c_i}$ la sémantique caractérisée par la contrainte et le prédicat suivants :
– $c(r) = \{\{t_i, t_j\} \mid t_i, t_j \in r \text{ AND } class[t_i] = class[t_j]\}$.
– $Pred(A, \{t_i, t_j\}) = [\epsilon_1^A \leq |t_i[A] - t_j[A]| \leq \epsilon_2^A]$.

Propriété 5:
La sémantique $s_v^{c_i}$ est bien-formée.

Preuve
Par le théorème 1, le résultat est évident puisque $s_v^{c_i} \in C_A$.

Cette sémantique permet d'étudier la variation des niveaux d'expression des gènes à l'intérieur des différentes classes d'échantillons.

Nous pouvons également ne considérer que les couples de tuples appartenant à différentes classes. Ainsi, $a \rightarrow b$ si chaque fois que le gène a ne varie pas entre deux échantillons appartenant à des classes différentes, alors le gène b ne varie pas non plus entre ces deux échantillons. Plus formellement :

Définition 11:
Soit $s_v^{c_d}$ la sémantique caractérisée par la contrainte et le prédicat suivants :
– $c(r) = \{\{t_i, t_j\} \mid t_i, t_j \in r \text{ AND } class[t_i] \neq class[t_j]\}$.
– $Pred(A, \{t_i, t_j\}) = [\epsilon_1^A \leq |t_i[A] - t_j[A]| \leq \epsilon_2^A]$.

Propriété 6:
La sémantique $s_v^{c_d}$ est bien-formée.

Preuve
Par le théorème 1, le résultat est évident puisque $s_v^{c_d} \in C_A$.

Cette sémantique permet d'étudier la variation des niveaux d'expression des gènes non plus à l'intérieur des différentes classes d'échantillons mais entre les différentes classes.

7.3 Etude de l'évolution des niveaux d'expression

Pour illustrer cette nouvelle sémantique qui permet d'étudier cette fois l'**évolution** des niveaux d'expression des gènes, considérons l'exemple suivant :

Exemple 16:
Soit la relation r_2 présentée dans la section 3.2 et redonnée dans la table 7.2, représentant les niveaux d'expression de 3 gènes pour une cellule à 6 instants consécutifs.

r_2	a_1	a_2	a_3	*temps*
t_1	0.4	−1.5	0.4	0
t_2	1.5	−0.3	−2.0	1
t_3	−0.7	0.8	1.6	2
t_4	0.4	1.8	−0.9	3
t_5	−1.4	0.5	−0.1	4
t_6	1.9	1.7	2.0	5

TABLE 7.2 – Relation r_2

Supposons que l'expression d'un gène ***croît*** *significativement entre deux instants consécutifs si la différence de ses niveaux d'expression est comprise entre* 1.0 *et* 4.0 *entre les échantillons correspondant (*4.0 *étant l'évolution maximum). Nous pouvons remarquer alors que chaque fois que l'expression du gène a_1 croît entre deux instants consécutifs (i.e. entre les échantillons $t_1 - t_2$, $t_3 - t_4$ et $t_5 - t_6$), alors l'expression du gène a_2 croît également entre ces deux instants. Ceci peut se voir sur le graphique 7.3 représentant les niveaux d'expression des gènes a_1 et a_2 dans la relation r_2.*

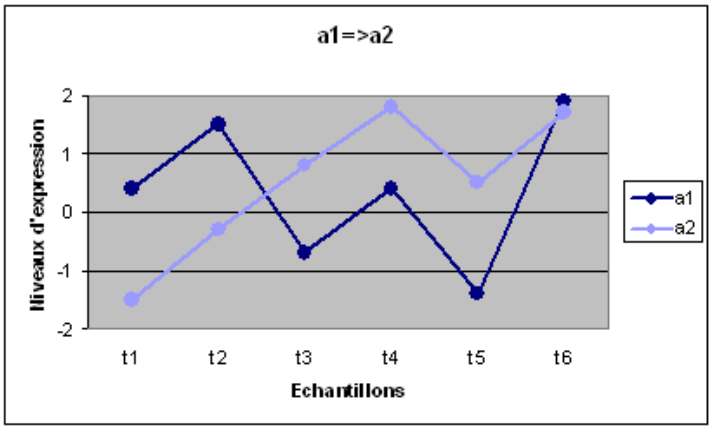


FIGURE 7.3 – Niveaux d'expression des gènes a_1 et a_2 dans la relation r_2

Ainsi, la règle $a_1 \rightarrow a_2$ est satisfaite dans r_2 avec la sémantique suivante : $a \rightarrow b$ si chaque fois que le gène a croît entre deux instants consécutifs, alors le gène b croît également entre ces deux instants.

Nous pouvons remarquer par exemple qu'avec cette sémantique, la règle $a_2 \rightarrow a_1$ n'est pas satisfaite dans la relation r_2 puisque entre les échantillons $t_2 - t_3$, le gène a_2 croît tandis que le gène a_1 ne croît pas.

Plus formellement, cette sémantique peut être définie de façon générale comme suit :

Définition 12:
Soit s_e la sémantique caractérisée par la contrainte et le prédicat suivants :
- $c(r) = \{\{t_i, t_j\} \mid t_i, t_j \in r$ AND $time[t_j] = time[t_i] + 1\}$.
- $Pred(A, \{t_i, t_j\}) = [\epsilon_1^A \leq t_j[A] - t_i[A] \leq \epsilon_2^A]$.

Nous avons alors le résultat suivant :

Propriété 7:
La sémantique s_e est bien-formée.

Preuve
Par le théorème 1, le résultat est évident puisque $s_e \in C_A$.

Cette sémantique prend également en compte les tuples deux par deux mais ne considère pas toutes les paires possibles de tuples. Ici, nous nous plaçons dans le cadre de données temporelles et nous nous intéressons uniquement aux couples de tuples correspondant à des instants consécutifs.

Cette sémantique permet donc aux biologistes de découvrir des règles entre gènes qui croissent ou qui décroissent selon les seuils choisis. Si plusieurs classes sont définies comme dans la relation r_3 présentée dans la section 3.2, nous pouvons également nous restreindre à certains couples de tuples.

Par exemple, nous pouvons nous concentrer sur une seule classe e.g. c_1 et ne considérer ainsi que les couples de tuples dont la classe est c_1 :

Définition 13:
Soit $s_e^{c_1}$ la sémantique caractérisée par la contrainte et le prédicat suivants :
- $c(r) = \{\{t_i, t_j\} \mid t_i, t_j \in r$ AND $time[t_j] = time[t_i]+1$ AND $class[t_i] = c_1$ AND $class[t_j] = c_1\}$.
- $Pred(A, \{t_i, t_j\}) = [\epsilon_1^A \leq t_j[A] - t_i[A] \leq \epsilon_2^A]$.

Propriété 8:
La sémantique $s_e^{c_1}$ est bien-formée.

Preuve
Par le théorème 1, le résultat est évident puisque $s_e^{c_1} \in C_A$.

Exemple 17:
Considérons par exemple la relation r_3 redonnée dans la table 7.3, représentant les niveaux d'expression de 5 gènes pour 3 cellules différentes, chacune à 3 instants consécutifs.

r_3	a_1	a_2	a_3	a_4	a_5	*cellule*	*temps*
t_1	−0.4	0.3	−2.0	−1.5	1.4	c_1	0
t_2	0.3	−1.9	−0.5	1.4	−0.4	c_2	0
t_3	1.1	−0.5	−1.9	1.2	0.7	c_3	0
t_4	0.9	1.5	−0.6	−0.3	1.7	c_1	1
t_5	1.8	−0.3	0.6	−0.5	−1.1	c_2	1
t_6	−0.5	1.8	1.2	1.1	−0.5	c_3	1
t_7	0.7	1.3	2.0	1.8	0.4	c_1	2
t_8	0.2	1.5	1.8	1.9	−1.1	c_2	2
t_9	−0.2	0.4	−0.5	−1.5	1.9	c_3	2

TABLE 7.3 – Relation r_3

Par exemple, la règle $a_1 \rightarrow a_4$ est satisfaite avec la sémantique $s_e^{c_1}$. En effet, chaque fois que a_1 croît entre deux échantillons de la classe c_1 (i.e. entre les échantillons $t_1 - t_4$), alors a_4 croît également.

De la même façon, nous pouvons définir une sémantique pour chaque classe. Ceci nous permet comme précédemment de comparer les règles obtenues pour telle classe par rapport aux règles obtenues pour telle autre classe.

Nous pouvons également ne considérer que les couples de tuples appartenant à une même classe. Ainsi, $a \rightarrow b$ si chaque fois que le gène a croît entre deux échantillons d'une même classe, alors le gène b croît également entre ces deux échantillons. Plus formellement :

Définition 14:
Soit $s_e^{c_i}$ la sémantique caractérisée par la contrainte et le prédicat suivants :
– $c(r) = \{\{t_i, t_j\} \mid t_i, t_j \in r$ AND $time[t_{i+1}] = time[t_i] + 1$ AND $class[t_i] = class[t_j]\}$.
– $Pred(A, \{t_i, t_j\}) = [\epsilon_1^A \leq t_j[A] - t_i[A] \leq \epsilon_2^A]$.

Propriété 9:
La sémantique $s_e^{c_i}$ est bien-formée.

Preuve
Par le théorème 1, le résultat est évident puisque $s_e^{c_i} \in C_A$.

Exemple 18:
Dans la relation r_3 donnée dans la table 7.3, la règle $a_2 \rightarrow a_3$ est satisfaite avec la sémantique $s_e^{c_i}$. En effet, chaque fois que a_2 croît entre deux échantillons de la même classe (i.e. entre les échantillons $t_1 - t_4$ pour la cellule c_1, $t_2 - t_5$ et $t_5 - t_8$ pour la cellule c_2 et $t_3 - t_6$ pour la cellule c_3), alors a_3 croît également. Nous pouvons remarquer que la règle $a_2 \rightarrow a_3$ est aussi satisfaite avec les sémantiques $s_e^{c_1}$, $s_e^{c_2}$ et $s_e^{c_3}$ puisque pour ces sémantiques le nombre de couples de tuples pris en compte est inférieur.

Nous pourrions également ne considérer que les couples de tuples appartenant à diffé-

rentes classes. Mais dans le cadre des données temporelles, cela ne semble pas pertinent.

En fonction des caractéristiques du jeu de données étudié (données temporelles, présence de classes ou non...) et en fonction des objectifs de l'étude, les sémantiques les plus adaptées pourront être proposées.

7.4 Choix des seuils

Pour généraliser la définition des sémantiques présentées dans la section précédente, nous avons introduit des seuils ϵ_1^A et ϵ_2^A.

Ces seuils peuvent être **identiques** à chaque gène i.e. $\forall A \in U$, $\epsilon_1^A = \epsilon_1$ et $\epsilon_2^A = \epsilon_2$. Dans ce cas, l'utilisateur spécifie les seuils qu'ils souhaitent étudier en fonction de ces objectifs i.e. s'ils s'intéressent plus particulièrement aux gènes fortement exprimés, aux gènes qui croissent...

Il est également possible de spécifier des seuils différents pour chaque gène. Ces seuils peuvent être calculés en fonction des valeurs prises par chacun des gènes.

Pour la sémantique s_n et ses variantes, les seuils ϵ_1^A et ϵ_2^A peuvent par exemple être choisis de telle sorte que l'intervalle $[\epsilon_1^A; \epsilon_2^A]$ contienne les niveaux d'expression les plus élevés. Par exemple, le seuil ϵ_2^A est égal au niveau d'expression maximum pris par le gène A et le seuil ϵ_1^A est calculé de façon à ce que pour 25% des échantillons, le niveau d'expression du gène A soit compris entre ϵ_1^A et ϵ_2^A. L'intervalle $[\epsilon_1^A; \epsilon_2^A]$ contiendra alors les 25% niveaux d'expression les plus élevés.

Pour la sémantique s_v et ses variantes, les seuils ϵ_1^A et ϵ_2^A peuvent par exemple être choisis de telle sorte que l'intervalle $[\epsilon_1^A; \epsilon_2^A]$ contienne les variations des niveaux d'expression les plus faibles.

Enfin, pour la sémantique s_e et ses variantes, les seuils ϵ_1^A et ϵ_2^A peuvent par exemple être choisis de telle sorte que l'intervalle $[\epsilon_1^A; \epsilon_2^A]$ contienne les évolutions des niveaux d'expression les plus élevées.

Chapitre 8

Génération des règles

Sommaire

Le problème de découverte des règles a été largement étudié dans le contexte des règles d'association [20] et des dépendances fonctionnelles [76, 39, 61]. Notre objectif ici n'est pas de développer de nouveaux algorithmes pour chaque sémantique mais au contraire de proposer une méthode pouvant s'appliquer à toute sémantique bien-formée.

8.1 Préliminaires

Deux principales techniques existent pour la découverte des règles :

– Les techniques utilisées pour les règles d'association [104] permettant d'obtenir par exemple une couverture minimum des règles [54].

– Les techniques utilisées plutôt pour les dépendances fonctionnelles [76, 39] permettant d'obtenir la couverture canonique des règles.

Les méthodes utilisées pour la découverte des règles d'association consistent à identifier tout d'abord les motifs fréquents, puis à trouver les règles dont le support et la confiance sont supérieurs à des seuils définis par l'expert. Or, la spécificité des données d'expression est que le nombre de gènes est généralement très grand par rapport au nombre d'échantillons. Dans le deuxième projet (cf section 5.2), nous avons par exemple 10 échantillons pour plus de 40000 gènes. Dans ce contexte, la notion même de support minimum n'a

que peu de sens, une valeur de seuil minimum étant difficile voire impossible à spécifier. De plus, le nombre de motifs peut être exponentiel dans le nombre de gènes, ce type d'énumération n'est donc pas ou peu envisageable.

Nous avons alors choisi naturellement les méthodes utilisées pour les dépendances fonctionnelles qui n'imposent pas le choix d'un seuil de support minimum.

L'idée ici est donc de généraliser le processus de découverte des dépendances fonctionnelles à toute sémantique bien-formée. Ce processus nous permet de générer une **couverture canonique** des règles exactes mais également une couverture pour les règles approximatives, il s'agit de la **couverture de Gottlob et Libkin** [51].

Nous détaillons dans les sections suivantes le processus de découverte des règles qui se compose de deux principales étapes :

- Soient r une relation définie sur U et s une sémantique bien-formée. Déterminer une base du système de fermeture associé à $F_s(r)$, l'ensemble des règles satisfaites dans r pour la sémantique s i.e. $F_s(r) = \{X \rightarrow Y \mid r \models_s X \rightarrow Y,\ avec\ X, Y \subseteq U\}$.

- Générer la couverture canonique et la couverture de Gottlob et Libkin.
  Pour une relation r sur U et une sémantique bien-formée s donnée, la couverture canonique des règles satisfaites dans r regroupe l'ensemble des règles $X \rightarrow A$ telles que $r \models_s X \rightarrow A$ et $\nexists Z \subset X$ tel que $r \models_s Z \rightarrow A$. Ceci correspond aux plus petites parties gauches des règles ayant un unique attribut en partie droite.
  Lors du calcul de la couverture canonique, nous obtenons sans calcul supplémentaire, les plus grandes règles approximatives correspondant à la couverture de Gottlob et Libkin. Celle-ci regroupe l'ensemble des règles $X \rightarrow A$ telles que $r \not\models_s X \rightarrow A$ et $\forall Z \supset X$, $r \models_s Z \rightarrow A$ i.e. les plus grandes parties gauches qui n'impliquent pas A.

La première étape est primordiale puisque c'est durant cette étape qu'a lieu l'accès aux données. Le cadre que nous proposons nous permet de calculer de façon identique une base du système de fermeture pour toute sémantique bien-formée.

La seconde étape qui part de la base est totalement identique quelle que soit la sémantique utilisée pour calculer cette base.

8.2 Calcul d'une base du système de fermeture

L'objectif ici est d'étendre à toute sémantique bien-formée le calcul d'une base du système de fermeture. Rappelons tout d'abord la définition d'un système de fermeture :

8.2 Calcul d'une base du système de fermeture

Définition 15:
Soient U un ensemble fini et $C \subseteq 2^U$. L'ensemble C est un ***système de fermeture*** *sur U si $U \in C$ et $\forall X, Y \in C$, $X \cap Y \in C$.*

Dans la suite, nous étendons à toute sémantique bien-formée un résultat donné dans [46] :

Définition 16:
Soient r une relation définie sur U, s une sémantique bien-formée et $F_s(r)$ l'ensemble des règles satisfaites dans r pour la sémantique s.
Soit $C(F_s(r))$ l'ensemble défini de la façon suivante :

$$C(F_s(r)) = \{Z \subseteq U \mid Z \text{ respecte } F_s(r)\}$$

Avec Z respecte $F_s(r)$ si $\forall X \rightarrow Y \in F_s(r)$, $X \not\subseteq Z$ ou $Y \subseteq Z$.

Autrement dit, $C(F_s(r))$ contient **tous** les ensembles d'attributs Z ne contredisant pas les règles $X \rightarrow Y$ telles que $r \models_s X \rightarrow Y$. Notons que $F_s(r)$ et $C(F_s(r))$ sont deux représentations différentes des règles satisfaites dans r.

Nous avons alors la propriété suivante :

Propriété 10:
L'ensemble $C(F_s(r))$ est un ***système de fermeture*** *par rapport à $F_s(r)$.*

Preuve
Pour montrer que $C(F_s(r))$ est un système de fermeture, nous devons montrer que $U \in C(F_s(r))$ et $\forall X, Y \in C(F_s(r))$, $X \cap Y \in C(F_s(r))$.
- *Montrons tout d'abord que $U \in C(F_s(r))$:*
  Pour cela, il faut montrer que U respecte $F_s(r)$. Soit $X \rightarrow Y \in F_s(r)$, la relation r est définie sur U, donc $X, Y \subseteq U$ donc U respecte $F_s(r)$.
- *Soient $X, Y \subseteq C(F_s(r))$. Montrons que $X \cap Y \subseteq C(F_s(r))$:*
  Pour cela, il faut montrer que $X \cap Y$ respecte $F_s(r)$ i.e. soit $V \rightarrow W \in F_s(r)$, il faut montrer que $V \not\subseteq X \cap Y$ ou $W \subseteq X \cap Y$. Or comme $X \subseteq C(F_s(r))$, $V \not\subseteq X$ ou $W \subseteq X$ et comme $Y \subseteq C(F_s(r))$, $V \not\subseteq Y$ ou $W \subset Y$. Donc si $V \not\subseteq X$ ou $V \not\subseteq Y$, alors $V \not\subseteq X \cap Y$ et si $W \subseteq X$ et $W \subseteq Y$, alors $W \subseteq X \cap Y$.

L'ensemble $C(F_s(r))$ est donc bien un système de fermeture.

Rappelons maintenant la définition des inf-irréductibles d'un système de fermeture puis d'une base d'un système de fermeture :

Définition 17:
Soient U un ensemble d'attributs et C un système de fermeture sur U.

L'ensemble $Inf(C) = \{Z \subseteq C \mid \forall X, Y \subseteq C$ *tels que*$X \cap Y = Z$ *alors* $Z = X$ *ou* $Z = Y\}$ *est l'ensemble des* ***inf-irréductibles*** *de* C.

L'ensemble des inf-irréductibles correspond au sous-ensemble unique et minimal de C tel que chaque élément de C peut être exprimé comme une intersection des éléments de Inf.

Définition 18:
Soient U *un ensemble d'attributs,* C *un système de fermeture sur* U *et* $B \subseteq U$. *L'ensemble* B *est une* ***base*** *du système de fermeture* C *si* $Inf(C) \subseteq B \subseteq C$ *où* $Inf(C)$ *est l'ensemble des inf-irréductibles de* C.

Nous étendons ici un résultat obtenu dans le cadre des dépendances fonctionnelles [27] à toute sémantique bien-formée pour le calcul d'une base du système de fermeture :

Définition 19:
Soient r *une relation définie sur* U *et* s *une sémantique bien-formée caractérisée par l'ensemble* c *et le prédicat* $Pred$.
Soit $B_s(r)$ *l'ensemble défini de la façon suivante :*

$$B_s(r) = \bigcup_{r' \in c(r)} \{A \in U \mid Pred(A, r') \text{ est vrai}\}.$$

Nous voyons clairement dans cette définition, la nécessité d'avoir $s \in C_A$ i.e. le prédicat défini au niveau de chaque attribut. Nous avons la propriété suivante :

Propriété 11:
$B_s(r)$ *est une* ***base*** *du système de fermeture* $C(F_s(r))$.

Preuve
Nous devons tout d'abord montrer que $B_s(r) \subseteq C(F_s(r))$ *puis que* $Inf_s(r) \subseteq B_s(r)$, *où* $Inf_s(r)$ *est l'ensemble des inf-irréductibles du système de fermeture* $C(F_s(r))$.

- *Soit* $Z \in B_s(r)$ *i.e.* $\exists r* \subseteq r$ *vérifiant* $c(r*)$ *tel que* $\forall A \in Z$, $Pred(A, r*)$ *est vrai et* $\forall A \in U \setminus Z$, $Pred(A, r*)$ *est faux. Nous devons montrer que* $Z \in C(F_s(r))$. *Supposons le contraire i.e.* $Z \notin C(F_s(r))$. *Dans ce cas,* $\exists X \subseteq Z$ *et* $Y \not\subseteq Z$ *tels que* $X \rightarrow Y \in F_s(r)$. *Nous devrions alors avoir* $\forall A \in Y$, $Pred(A, r*)$ *est vrai puisque* $X \subseteq Z$ *et* $\forall A \in Z$, $Pred(A, r*)$ *est vrai. Or ceci mène à une contradiction puisque* $Y \not\subseteq Z$.
- *Soit* $Z \in Inf_s(r)$, *nous devons montrer que* $Z \in B_s(r)$. *Comme* $Z \in C(F_s(r))$, $\forall X \in U \setminus Z$, $r \not\models_s Z \rightarrow X$ *i.e.* $\forall X \in U \setminus Z$, $\exists Z_X \in B_s(r)$ *tel que* $Z \subseteq Z_X$ *et* $X \notin Z_X$. *Or puisque* $Z = \bigcap_{X \in U \setminus Z}(Z_X)$ *et puisque* $Z \in Inf_s(r)$, $\exists X \in U \setminus Z$ *tel que* $Z = Z_X$. *Nous avons alors* $Z \in B_s(r)$.

Finalement, nous avons montré que $Inf_s(r) \subseteq B_s(r) \subseteq C(F_s(r))$, donc l'ensemble $B_s(r)$ est bien une base du système de fermeture $C(F_s(r))$.

Le calcul de la base $B_s(r)$ est une étape primordiale du processus de génération puisque c'est lors de cette étape qu'a lieu l'accès aux données.

Exemple 19:
Voici une base du système de fermeture pour l'ensemble des règles satisfaites dans la relation r_1 pour la sémantique s_v :

$$B_{s_v}(r_1) = \bigcup_{\{t_i,t_j\}\in r} \{\ A \in U \mid 0.0 \leq |t_i[A] - t_j[A]| \leq 0.2\ \}$$

$$= \{\{a_1a_2a_3a_4\}, \{a_4\}, \{\}, \{a_3a_5\}, \{a_2\}, \{a_3\}, \{a_1a_4\}\}.$$

8.3 Inférence des règles

A partir d'une base du système de fermeture, nous devons maintenant inférer des couvertures des règles. Il faut noter que pour le calcul de la couverture canonique et de la couverture de Gottlob et Libkin, toutes les étapes sont réalisées de façon identique quelle que soit la sémantique bien-formée utilisée pour le calcul de la base.

Pour générer la couverture canonique, nous utilisons un processus utilisé pour l'inférence des dépendances fonctionnelles [76, 39, 73] que nous étendons ici à toute sémantique bien-formée.

Le résumé de ce processus pour une relation r et une sémantique bien-formée s est donné dans la figure 8.1 et détaillé dans la suite. A partir d'une relation r et d'une sémantique bien-formée s, nous obtenons la couverture canonique (CC) et la couverture de Gottlob et Libkin (CGL). Ce processus est identique quelles que soient la sémantique et la relation étudiée.

Tout d'abord, il faut voir que chaque élément $X \in B_s(r)$ nous dit que toutes les règles de la forme $X \rightarrow A$ avec $A \in U \setminus X$ sont fausses dans la relation r pour la sémantique s, et également les règles de la forme $Y \rightarrow A$ avec $Y \subseteq X$.
De plus, soit $A \in U \setminus X$, si la règle $X \rightarrow A$ est fausse dans la relation r pour la sémantique s alors $\exists Z \in B_s(r)$ tel que $X \subseteq Z$ et $A \notin Z$.

Donc pour un attribut donné A, l'ensemble des plus grands éléments $X \in B_s(r)$ tels que $A \notin X$ représente les plus grandes parties gauches des règles telles que $r \not\models_s X \rightarrow A$. Nous commençons donc par calculer cet ensemble pour tous les attributs $A \in U$:

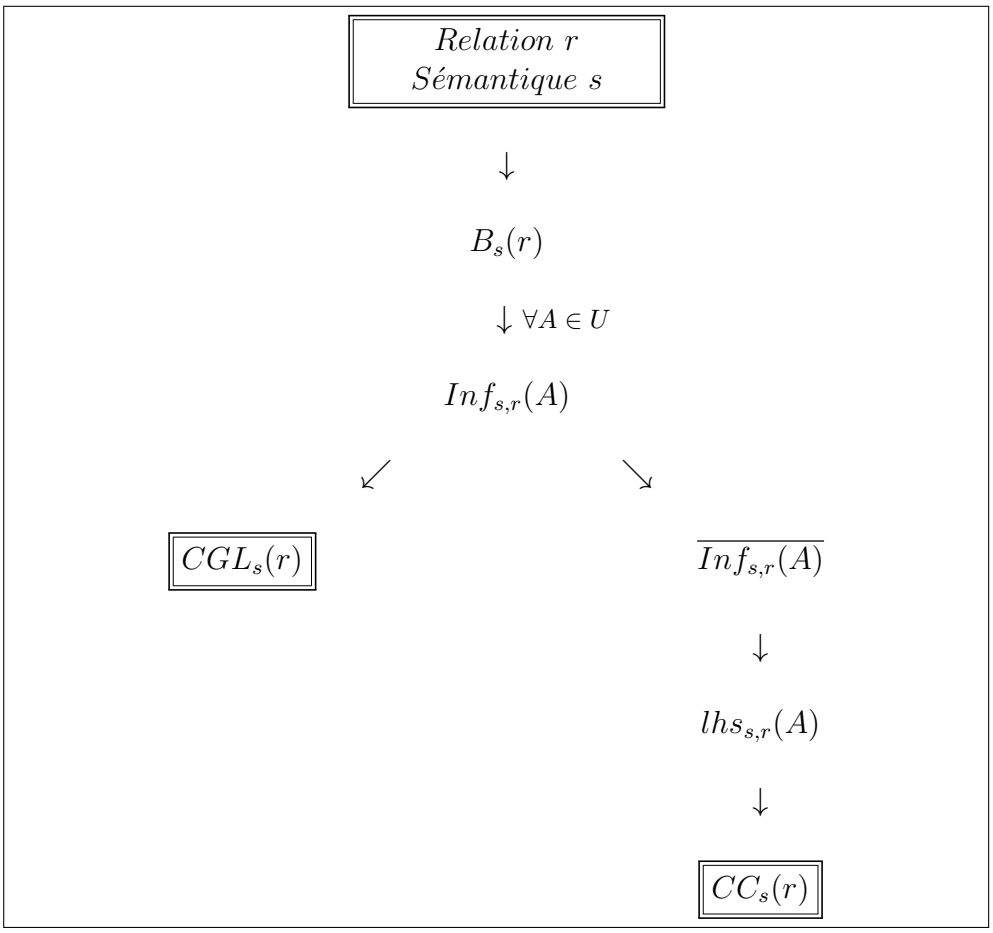


TABLE 8.1 – Génération des règles

Définition 20:
Soient s une sémantique bien-formée, r une relation sur U et $A \in U$ un attribut. Soit $Inf_{s,r}(A)$ l'ensemble défini comme suit :

$$Inf_{s,r}(A) = \max_{\subseteq}\{X \in B_s(r) \mid A \notin X\}.$$

Remarquons que cet ensemble est suffisant pour calculer la couverture de Gottlob et Libkin, qui regroupe l'ensemble des règles $X \to A$ telles que $r \not\models_s X \to A$ et $\nexists Z \supset X$ tel que $r \not\models_s Z \to A$:

Définition 21:
Soient s une sémantique bien-formée et r une relation sur U.
La couverture de Gottlob et Libkin de $F_s(r)$, notée $CGL_s(r)$ est donnée par :

$$CGL_s(r) = \bigcup_{A \in U} \{X \to A \mid X \in Inf_{s,r}(A)\}$$

Rappelons que pour la couverture canonique, nous cherchons les plus petites parties gauches des règles ayant un unique attribut en partie droite. Pour cela, nous utilisons ensuite une relation bien connue entre l'ensemble des maximaux n'impliquant pas A et l'ensemble des minimaux impliquant A [76, 39] :

Définition 22:
Soient s une sémantique bien-formée, r une relation sur U et $A \in U$ un attribut.
Soit $lhs_{s,r}(A)$ l'ensemble défini comme suit :

$$lhs_{s,r}(A) = \min_{\subseteq}\{X \subseteq U \mid \forall Y \in Inf_{s,r}(A), X \not\subseteq Y\}.$$

Nous avons alors le résultat suivant [76, 39] :

Théorème 2:
Soient s une sémantique bien-formée, r une relation sur U et $A \in U$ un attribut. Nous avons :

$$lhs_{s,r}(A) = transmin(\overline{Inf_{s,r}(A)}),$$

où

$$\overline{Inf_{s,r}(A)} = \{U \setminus X \mid X \in Inf_{s,r}(A)\}.$$

Rappelons que les transversaux minimaux d'un hypergraphe $H = (V, E)$ notés $transmin(H)$ où V est l'ensemble des sommets de H et E est l'ensemble des arêtes de H, sont les plus petits sous-ensembles de V qui interceptent tous les éléments de E.

Une fois ces calculs réalisés pour tous les attributs $A \in U$, nous obtenons la couverture canonique de l'ensemble des règles satisfaites dans r pour la sémantique s :

Définition 23:
Soient s une sémantique bien-formée et r une relation sur U.
La couverture canonique de $F_s(r)$, notée $CC_s(r)$ est donnée par :

$$CC_s(r) = \bigcup_{A \in U} \{X \rightarrow A \mid X \in lhs_{s,r}(A)\}$$

L'étape clé de ce processus est le calcul des transversaux minimaux, qui est exponentiel en nombre d'attributs. Pour l'implémentation de cette étape, nous utilisons un algorithme basé sur la proposition faite dans [43].

Quelques optimisations Parmi l'ensemble des règles générées, certaines n'apportent aucune connaissance aux experts. Il est alors possible d'éviter certains calculs inutiles :

- Les règles du type $A \rightarrow A$ appartiennent toutes à la couverture canonique mais sont toutes triviales. Lors du calcul des complémentaires pour un attribut A, nous pouvons alors enlever cet attribut i.e. :

$$\bigcup_{X \in Inf_{s,r}(A)} \{U \setminus \{X \cup A\}\}.$$

- Les règles $X \rightarrow A$ de la couverture canonique telles que $n_X = 0$ i.e. $\nexists Z \in B_s(r)$ tel que $X \subseteq Z$ (cf section 6.5), n'apparaissent jamais dans la relation. En fait, ces règles ne sont jamais contredites, elles sont donc exactes mais elles ne sont jamais vérifiées non plus, aucune conclusion intéressante ne peut donc être tirée de ce type de règles.
  Pour éviter de générer des règles $X \rightarrow A$ où $n_X = 0$ et $|X| = 1$, nous pouvons tout d'abord ajouter la restriction suivante lors du calcul des complémentaires :

$$\bigcup_{X \in Inf_{s,r}(A)} \{U \setminus \{X \cup A \cup B \mid n_B = 0\}\}.$$

  Et pour les cas où $|X| > 1$, nous ajoutons une restriction sur X lors du calcul de la couverture canonique :

$$\bigcup_{A \in U} \{X \rightarrow A \mid X \in lhs_{s,r}(A) \; et \; n_X > 0\}.$$

  Notons que pour les règles $X \rightarrow A$ de la couverture de Gottlob et Libkin, n_X est toujours strictement positif puisque $X \in B_s(r)$.
- De la même façon, les règles $X \rightarrow A$ telles que $n_A = 0$ ne sont pas pertinentes. Nous limiterons alors les calculs aux attributs suivants :

$$\forall A \in U \mid n_A > 0.$$

Exemple 20:
Voici la couverture canonique et la couverture de Gottlob et Libkin pour la sémantique s_v dans la relation r_1 :

1. $B_{s_v}(r_1) = \{\{a_1a_2a_3a_4\}, \{a_4\}, \{\}, \{a_3a_5\}, \{a_2\}, \{a_3\}, \{a_1a_4\}\}$
2. *Attribut a_1 :*
   (a) $Inf_{s_v,r_1}(a_1) = \{\{a_4\}, \{a_3a_5\}, \{a_2\}\}$
   (b) $\overline{Inf_{s_v,r_1}(a_1)} = \{\{a_2a_3a_5\}, \{a_2a_4\}, \{a_3a_4a_5\}\}$
   (c) $lhs_{s_v,r_1}(a_1) = \{\{a_2a_3\}, \{a_2a_4\}, \{a_3a_4\}\}$
3. *Attribut a_2 :*
   (a) $Inf_{s_v,r_1}(a_2) = \{\{a_3a_5\}, \{a_1a_4\}\}$
   (b) $\overline{Inf_{s_v,r_1}(a_2)} = \{\{a_1a_4\}, \{a_3a_5\}\}$
   (c) $lhs_{s_v,r_1}(a_2) = \{\{a_1a_3\}, \{a_3a_4\}\}$
4. *Attribut a_3 :*
   (a) $Inf_{s_v,r_1}(a_3) = \{\{a_2\}, \{a_1a_4\}\}$
   (b) $\overline{Inf_{s_v,r_1}(a_3)} = \{\{a_1a_4a_5\}, \{a_2a_5\}\}$
   (c) $lhs_{s_v,r_1}(a_3) = \{\{a_5\}, \{a_1a_2\}, \{a_2a_4\}\}$
5. *Attribut a_4 :*
   (a) $Inf_{s_v,r_1}(a_4) = \{\{a_3a_5\}, \{a_2\}\}$
   (b) $\overline{Inf_{s_v,r_1}(a_4)} = \{\{a_1a_2\}, \{a_1a_3a_5\}\}$
   (c) $lhs_{s_v,r_1}(a_4) = \{\{a_1\}, \{a_2a_3\}\}$
6. *Attribut a_5 :*
   (a) $Inf_{s_v,r_1}(a_5) = \{\{a_1a_2a_3a_4\}\}$
   (b) $\overline{Inf_{s_v,r_1}(a_5)} = \{\{\}\}$
   (c) $lhs_{s_v,r_1}(a_5) = \Phi$
7. $CGL_{s_v}(r_1) = \{a_4 \rightarrow a_1, a_3a_5 \rightarrow a_1, a_2 \rightarrow a_1, a_3a_5 \rightarrow a_2, a_1a_4 \rightarrow a_2, a_2 \rightarrow a_3, a_1a_4 \rightarrow a_3, a_3a_5 \rightarrow a_4, a_2 \rightarrow a_4, a_1a_2a_3a_4 \rightarrow a_5\}$
8. $CC_{s_v}(r_1) = \{a_2a_3 \rightarrow a_1, a_2a_4 \rightarrow a_1, a_3a_4 \rightarrow a_1, a_1a_3 \rightarrow a_2, a_3a_4 \rightarrow a_2, a_5 \rightarrow a_3, a_1a_2 \rightarrow a_3, a_2a_4 \rightarrow a_3, a_1 \rightarrow a_4, a_2a_3 \rightarrow a_4\}$

8.4 Prise en compte d'attributs centraux

Nous avons décrit dans la section précédente, une méthode pour calculer la couverture canonique de l'ensemble des règles satisfaites pour une relation et une sémantique données. Il faut noter cependant que le nombre de règles peut être très élevé pour certaines relations et certaines sémantiques. Ainsi, pour une relation donnée, générer toutes

les règles satisfaites avec toutes les sémantiques adaptées à cette relation, peut être très long.

Une première méthode pour remédier à ce problème est de filtrer les règles en fonction de divers indices de qualité (cf section 6.5). Une seconde méthode consiste à limiter les calculs à certains attributs d'intérêt, que nous appellerons **attributs centraux**. Pour certains domaines d'application, les experts peuvent en effet être intéressés plus particulièrement par quelques attributs. Ceci est notamment le cas pour l'étude des données d'expression, puisque les biologistes s'intéressent souvent à quelques gènes en particulier.

Dans ce contexte, nous proposons une approche qui permet de générer les règles pour lesquelles les attributs centraux jouent un rôle majeur, i.e. les règles dont la partie gauche ou la partie droite est un attribut central. Il s'agit d'un processus interactif dans lequel l'utilisateur a un rôle important puisqu'il définit les attributs centraux.

Dans la suite, nous notons I l'ensemble des attributs centraux et $CC_{s,I}(r)$ le sous-ensemble de $CC_s(r)$ contenant les règles dont la partie gauche ou la partie droite est un attribut central.

Définition 24:
Soient s une sémantique bien-formée, r une relation sur U et $I \subseteq U$ l'ensemble des attributs centraux.
L'ensemble des règles de la base canonique $CC_s(r)$ dont la partie gauche ou la partie droite est un attribut central, noté $CC_{s,I}(r)$, est donné par :

$$CC_{s,I}(r) = \bigcup_{B \in I} \{X \rightarrow B \in CC_s(r)\} \cup \bigcup_{B \in I} \{B \rightarrow A \in CC_s(r)\}.$$

Tout d'abord, pour calculer l'ensemble des règles de la base canonique dont la partie droite est un des attributs centraux, il suffit de n'appliquer le processus précédent qu'aux attributs centraux. C'est donc une restriction triviale. L'ensemble des règles de la couverture de Gottlob et Libkin dont la partie droite est un des attributs centraux sera noté $CGL_{s,I}(r)$.

Deuxièmement, pour identifier les règles de la base canonique dont la partie gauche est un attribut central, nous définissons le nouvel ensemble suivant :

Définition 25:
Soient s une sémantique bien-formée, r une relation sur U et $B \in U$ un attribut. Soit $rhs_{s,r}(B)$ l'ensemble défini comme suit :

$$rhs_{s,r}(B) = \{A \in U \mid B \rightarrow A \in CC_s(r)\}.$$

Or si la règle $B \rightarrow A \in CC_s(r)$, cela signifie que A appartient à tous les éléments de la base qui contiennent l'attribut B. Notons toutefois, que les parties gauches des règles de la base canonique sont minimales i.e. que si $B \rightarrow A \in CC_s(r)$, alors $\Phi \rightarrow A \notin CC_s(r)$.

Nous avons alors le résultat suivant :

Propriété 12:
Soient s une sémantique bien-formée, r une relation sur U et $B \in U$ un attribut. Nous avons :

$$rhs_{s,r}(B) = \bigcap\{X \in B_s(r) \mid B \in X\} \setminus \{T \in U \mid \forall X \in B_s(r), T \in X\}.$$

Preuve
Tout d'abord, si $B \rightarrow A \in CC_s(r)$ alors $r \models_s B \rightarrow A$ et $B \rightarrow A$ est minimale i.e. $r \not\models_s \phi \rightarrow A$. Nous avons donc $\forall X \in B_s(r) \mid B \in X$, $A \in X$ et $\exists X \in B_s(r) \mid A \notin X$. Ce qui montre bien que si $B \rightarrow A \in CC_s(r)$ alors $A \in \bigcap\{X \in B_s(r) \mid B \in X\} \setminus \{T \in U \mid \forall X \in B_s(r), T \in X\}$.

Nous devons montrer ensuite que si $A \in \bigcap\{X \in B_s(r) \mid B \in X\} \setminus \{T \in U \mid \forall X \in B_s(r), T \in X\}$ alors $B \rightarrow A \in CC_s(r)$. Comme $A \in \bigcap\{X \in B_s(r) \mid B \in X\}$, la règle $B \rightarrow A$ est exacte. De plus, comme $A \notin \{T \in U \mid \forall X \in B_s(r), T \in X\}$, nous avons $r \not\models_s \phi \rightarrow A$, donc $B \rightarrow A$ est minimale et appartient donc à $CC_s(r)$.

Ce nouvel ensemble nous permet d'éviter le calcul des transversaux pour les attributs qui ne sont pas centraux.

Finalement, le nouveau processus de découverte des règles avec prise en compte des attributs centraux I est illustré dans la figure 8.2.

Quelques optimisations Comme dans le cas général, nous évitons la génération des règles n'apportant aucune connaissance aux experts :

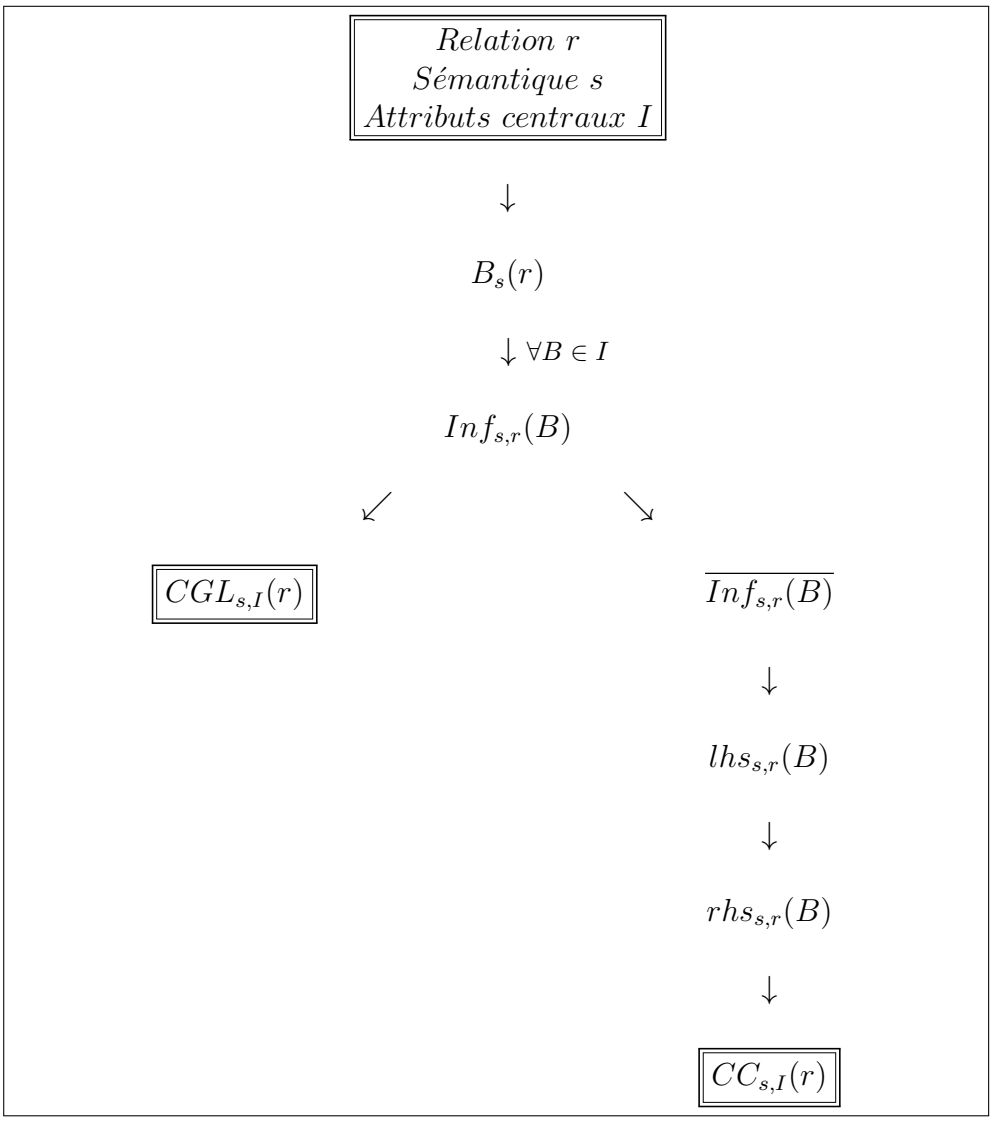


TABLE 8.2 – Génération des règles avec prise en compte d'attributs centraux

– Le calcul des règles du type $B \rightarrow B$ est évité lors du calcul des complémentaires, où nous enlevons l'attribut B traité i.e. :

$$\bigcup_{X \in Inf_{s,r}(B)} \{U \setminus \{X \cup B\}\}.$$

Et également pour les règles avec un élément $B \in C$ en partie gauche :

$$\bigcup_{B \in C} \{B \rightarrow A \mid A \in rhs_{s,r}(B) \; et \; A \neq B\}.$$

– Le calcul des règles $X \rightarrow B$ telles que $n_X = 0$ est évité en ajoutant les restrictions suivantes :

$$\bigcup_{X \in Inf_{s,r}(B)} \{U \setminus \{X \cup B \cup A \mid n_A = 0\}\}.$$

$$\bigcup_{B \in C} \{X \rightarrow B \mid X \in lhs_{s,r}(B) \; et \; n_X > 0\}.$$

– Le calcul des règles du type $X \rightarrow B$ ou $B \rightarrow A$ telles que $n_B = 0$ est évité en ne considérant que les attributs centraux suivants :

$$\forall B \in C \mid n_B > 0.$$

Notons que pour les règles canoniques de la forme $B \rightarrow A$, n_A est toujours strictement positif puisque $A \in rhs_{s,r}(B)$.

Notons de plus que si l'utilisateur désigne tous les attributs comme centraux i.e. $I = U$, alors nous avons bien $CC_{s,U}(r) = CC_s(r)$, mais le calcul des rhs est inutile puisque dans ce cas, $\bigcup_{B \in I}\{B \rightarrow A \in CC_s(r)\} \subseteq \bigcup_{B \in I}\{X \rightarrow B \in CC_s(r)\}$. De même $CGL_{s,U}(r) = CGL_s(r)$.

Enfin, les règles du type $I_1 \rightarrow I_2$ où I_1 et I_2 sont deux attributs centraux seront calculées deux fois ($I_2 \in rhs_{s,r}(I_1)$ et $I_1 \in lhs_{s,r}(I_2)$). Pour éviter des calculs inutiles, nous excluons dans le calcul des rhs tous les attributs centraux.

Exemple 21:
Voici les nouveaux calculs effectués pour la sémantique s_v et la relation r_1 et $I = \{a_1, a_3\}$ l'ensemble des attributs centraux :

1. $B_{s_v}(r_1) = \{\{a_1a_2a_3a_4\}, \{a_4\}, \{\}, \{a_3a_5\}, \{a_2\}, \{a_3\}, \{a_1a_4\}\}$
2. *Attribut a_1 :*
   (a) $Inf_{s_v,r_1}(a_1) = \{\{a_4\}, \{a_3a_5\}, \{a_2\}\}$
   (b) $\overline{Inf_{s_v,r_1}(a_1)} = \{\{a_2a_3a_5\}, \{a_2a_4\}, \{a_3a_4a_5\}\}$
   (c) $lhs_{s_v,r_1}(a_1) = \{\{a_2a_3\}, \{a_2a_4\}, \{a_3a_4\}\}$
   (d) $rhs_{s_v,r_1}(a_1) = \bigcap\{\{a_1a_2a_3a_4\}, \{a_1a_4\}\} = \{a_1a_4\}$
3. *Attribut a_3 :*
   (a) $Inf_{s_v,r_1}(a_3) = \{\{a_2\}, \{a_1a_4\}\}$
   (b) $\overline{Inf_{s_v,r_1}(a_3)} = \{\{a_1a_4a_5\}, \{a_2a_5\}\}$
   (c) $lhs_{s_v,r_1}(a_3) = \{\{a_5\}, \{a_1a_2\}, \{a_2a_4\}\}$
   (d) $rhs_{s_v,r_1}(a_3) = \bigcap\{\{a_2a_3a_4\}, \{a_3a_5\}, \{a_3\}\} = \{a_3\}$
4. $CGL_{s_v,I}(r_1) = \{a_4 \rightarrow a_1, a_3a_5 \rightarrow a_1, a_2 \rightarrow a_1, a_2 \rightarrow a_3, a_1a_4 \rightarrow a_3\}$
5. $CC_{s_v,I}(r_1) = \{a_2a_3 \rightarrow a_1, a_2a_4 \rightarrow a_1, a_3a_4 \rightarrow a_1, a_1 \rightarrow a_4, a_5 \rightarrow a_3, a_1a_2 \rightarrow a_3, a_2a_4 \rightarrow a_3\}$

Troisième partie

Vers la reconstruction de réseaux de gènes

Chapitre 9

Des règles aux réseaux de gènes

Sommaire

Nous avons présenté dans la partie II notre approche basée sur la découverte de différents types de règles entre gènes. Dans cette partie, nous faisons le lien entre l'approche proposée et la reconstruction de réseaux de gènes à partir de données d'expression.

9.1 Les réseaux de régulation

Les réseaux de régulation (cf figure 9.1) ont pour objectif de modéliser les processus biologiques qui contrôlent le développement d'un caractère phénotypique particulier d'un organisme. La compréhension du fonctionnement de ces réseaux permettra d'appréhender les perturbations liées à des pathologies et de sélectionner les cibles thérapeutiques les plus pertinentes pour traiter efficacement les patients.

Les réseaux de régulation se présentent sous forme de graphes et représentent des interactions entre diverses entités biologiques dans une cellule [25]. Les entités peuvent être de différents types : gènes, ARN, protéines, métabolites... Plusieurs types d'interactions sont également possibles en fonction de la nature des entités qu'elles associent : réaction chimique, catalyse d'une réaction par une enzyme, régulation de l'expression d'un gène...

Divers types de réseaux plus spécifiques peuvent être étudiés : les réseaux d'interaction

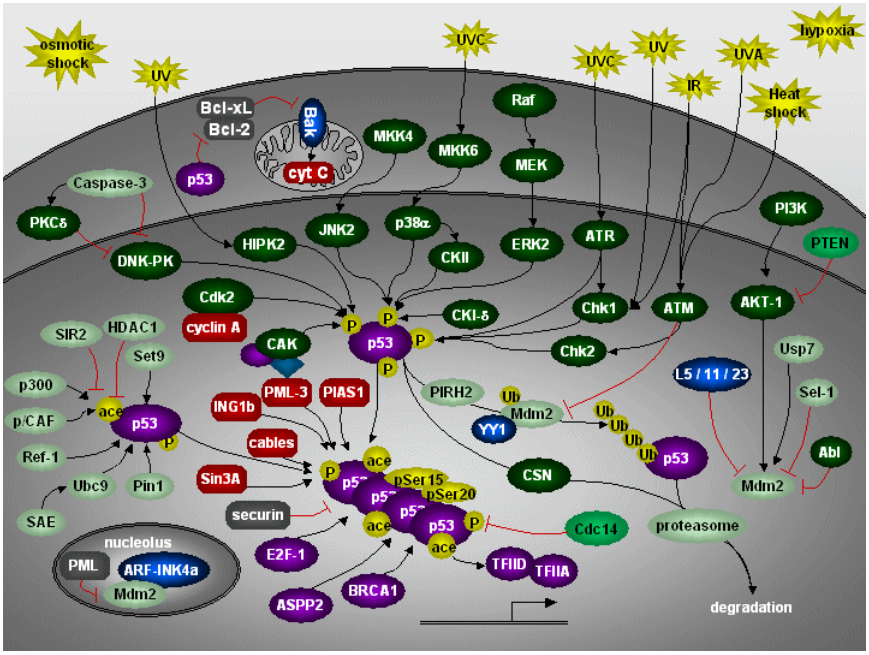


FIGURE 9.1 – Exemple de réseau biologique

protéine-protéine, les réseaux de gènes, les réseaux métaboliques, les réseaux de signalisation... L'objectif étant de comprendre les interactions dans et entre ces différents types de réseaux. En fonction des données disponibles, de la nature des interactions et de la taille des réseaux, différentes méthodes sont utilisées pour inférer ces réseaux.

9.1.1 Reconstruction de réseaux de gènes

Suite au succès rencontré par les techniques de puces à ADN pour mesurer l'expression des gènes à grande échelle, l'inférence des réseaux de gènes à partir de ces données d'expression (cf figure 9.2) a suscité depuis quelques années un intérêt croissant [101, 86, 58, 57, 38].

L'objectif de la reconstruction de réseaux de gènes est de proposer à partir de données expérimentales, des interactions probables entre les gènes, qui pourront être ensuite plus profondément validées avec par exemple des expérimentations plus poussées.

De manière générale, les relations entre les gènes sont rarement directes, i.e. qu'ils n'interagissent pas physiquement. Une relation entre deux gènes A et B dans un réseau, signifie simplement l'idée qu'un changement dans l'activité du gène A causera un changement dans l'activité du gène B, ce changement étant le résultat d'une succession de modifications d'activité au niveau des produits associés aux deux gènes.

Le principe est donc d'extraire des interactions entre les gènes à partir de données

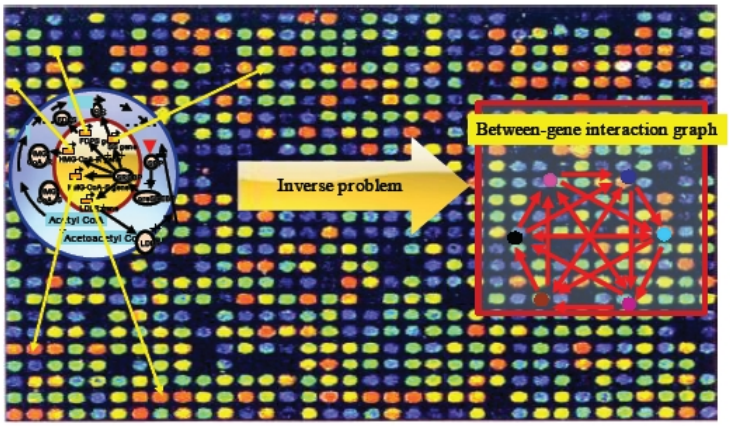


FIGURE 9.2 – Reconstruction de réseaux de gènes à partir de données d'expression

d'expression temporelles ou non, en utilisant des méthodes d'inférence efficaces. Plusieurs approches ont été proposées pour inférer les réseaux de gènes à partir des données d'expression, parmi lesquelles les réseaux booléens, les réseaux bayésiens, l'analyse des corrélations ou les équations différentielles.

Pour les réseaux booléens, les données d'expression temporelles sont binarisées, ainsi à un instant donné, un gène pourra être soit dans l'état ON pour exprimer qu'il est actif, soit dans l'état OFF pour exprimer qu'il est inactif. L'objectif est alors de trouver des règles logiques permettant de déterminer l'état d'un gène à l'instant $t+1$, à partir de l'état de ce gène et des autres gènes à l'instant t. Plusieurs méthodes d'inférence des règles logiques, permettant d'obtenir les réseaux, ont été proposées [72, 21, 64].

Les réseaux bayésiens se basent sur le calcul de probabilités [45, 62], en partant de la formule de Bayès sur les probabilités conditionnelles. Pour chaque couple de gènes $\{g_i, g_j\}$, la probabilité d'avoir g_i sachant g_j est calculée. Si cette probabilité est importante, alors on aura une relation de causalité entre ces deux gènes. L'objectif est de trouver pour un jeu de données, le modèle avec le meilleur score i.e. celui ayant la plus grande probabilité d'être correct avec les données étudiées.

Une autre méthode consiste à calculer la corrélation entre chaque paire de gènes, basé sur le coefficient de corrélation de Pearson. Si la valeur absolue de la corrélation est supérieure à un certain seuil alors il existera dans le réseau une relation positive ou négative (selon le signe) entre les deux gènes. Ces réseaux sont connus sous le nom de "Relevance Networks" [34].

Pour les équations différentielles [40], l'évolution du niveau d'expression de chaque gène dans le temps est supposée fonction linéaire des niveaux d'expression des autres gènes. Les coefficients linéaires obtenus représentent l'influence de chaque gène sur la régulation du gène étudié. Le signe du coefficient donne le sens de la relation (positive ou négative)

et sa valeur absolue donne la force de l'interaction.

D'autres méthodes d'inférence ainsi que des variantes des méthodes présentées ont émergé ces dernières années, comme les réseaux booléens temporels [85] ou probabilistes [84], les réseaux bayésiens dynamiques... L'inférence de règles d'association pour la reconstruction de réseaux de gènes a également été développée dans plusieurs articles [63, 37, 36].

Il faut bien noter que l'inférence de réseaux de gènes sur la base des seules données d'expression n'est qu'une étape dans tout le processus de caractérisation de réseaux de régulation impliqués dans des processus cellulaires spécifiques comme le cycle cellulaire, des voies de différenciation cellulaire, ou encore de leurs dérèglements pathologiques.

Toutefois la combinaison de ces différentes méthodes d'inférence avec des données fonctionnelles diverses (fonctions connues ou prédites des gènes, données massives sur les interactions moléculaires protéine-protéine, ou protéine/ADN, etc.) devrait s'avérer très utile dans les prochaines années.

9.1.2 Positionnement de l'approche proposée

Ce qu'il faut retenir est que la signification de la relation entre les gènes est très différente d'une méthode à l'autre et la comparaison entre toutes ces méthodes est difficile. Il est important lorsque nous étudions un réseau de gènes de garder en mémoire la méthode qui a été utilisée pour le construire. L'objectif commun de toutes ces méthodes est bien de décrire des régularités observées dans les données.

L'approche présentée dans la partie II a pour but de découvrir différents types de règles entre attributs. L'originalité de notre travail réside dans le fait que nous proposons aux utilisateurs une approche globale pouvant inclure plusieurs sémantiques pour les règles, adaptées à plusieurs types de données.

Dans le cadre des données d'expression, cette approche entre dans le cadre de la reconstruction de réseaux de gènes puisque l'objectif est bien de décrire des régularités observées dans les données et de les proposer aux biologistes.

Diverses sémantiques spécialement adaptées aux données d'expression de gènes ont été présentées dans le chapitre 7. Ces sémantiques sont particulièrement intéressantes pour les biologistes et doivent être choisies en fonction de leurs objectifs et des données étudiées.

Nous montrons dans la section suivante comment les règles peuvent être visualisées sous forme de graphes présentant les diverses relations découvertes dans les données. Nous présentons ensuite l'outil qui a été implémenté avec les différentes étapes du processus de découverte des réseaux. Enfin dans la dernière section, nous présentons les résultats obtenus à partir de ce travail sur les différents projets présentés dans le chapitre 1.

9.2 Visualisation des règles

Pour faciliter l'interprétation des règles par les experts, nous souhaitons proposer une visualisation conviviale des règles générées. Plusieurs approches de visualisation des règles ont été proposées, essentiellement pour les règles d'association. Nous rappelons tout d'abord les méthodes les plus couramment rencontrées puis nous présentons l'approche que nous avons choisie et qui permet de "mélanger" différentes sémantiques.

9.2.1 Etat de l'art

Pour les règles d'association, trois représentations sont couramment rencontrées : la représentation textuelle, la représentation par matrice 2D ou 3D ainsi que la représentation par graphe.

Représentation textuelle La représentation la plus classique et la plus répandue est la représentation textuelle des règles. Le support et la confiance sont donnés pour chaque règle ainsi que des indices de qualité supplémentaires pour certains logiciels (cf par exemple la figure 9.3 pour le logiciel Tanagra [82]). Il est généralement possible de trier les règles en fonction des indices donnés.

Number of rules : 51

N°	Antecedent	Consequent	Lift	Support	Confidence
1	"irradiat='no'" - "inv-nodes='0-2'"	"node-caps='no'" - "Class='no-recurrence-events'"	1,325	0,507	0,792
2	"node-caps='no'" - "Class='no-recurrence-events'"	"irradiat='no'" - "inv-nodes='0-2'"	1,325	0,507	0,848
3	"node-caps='no'" - "irradiat='no'"	"inv-nodes='0-2'" - "Class='no-recurrence-events'"	1,321	0,507	0,771
4	"inv-nodes='0-2'" - "Class='no-recurrence-events'"	"node-caps='no'" - "irradiat='no'"	1,321	0,507	0,868
5	"node-caps='no'" - "breast='left'"	"irradiat='no'" - "inv-nodes='0-2'"	1,320	0,343	0,845
6	"inv-nodes='0-2'" - "breast='left'"	"node-caps='no'" - "irradiat='no'"	1,296	0,343	0,852
7	"node-caps='no'" - "irradiat='no'" - "breast='left'"	"inv-nodes='0-2'"	1,290	0,343	0,961
8	"node-caps='no'" - "irradiat='no'" - "Class='no-recurrence-events'"	"inv-nodes='0-2'"	1,289	0,507	0,960
9	"node-caps='no'" - "breast='left'"	"inv-nodes='0-2'"	1,273	0,385	0,948
10	"irradiat='no'" - "inv-nodes='0-2'" - "Class='no-recurrence-events'"	"node-caps='no'"	1,271	0,507	0,986
11	"node-caps='no'" - "irradiat='no'"	"inv-nodes='0-2'"	1,264	0,619	0,941
12	"inv-nodes='0-2'"	"node-caps='no'" - "irradiat='no'"	1,264	0,619	0,831
13	"irradiat='no'" - "Class='no-recurrence-events'"	"node-caps='no'" - "inv-nodes='0-2'"	1,258	0,507	0,884
14	"node-caps='no'" - "Class='no-recurrence-events'"	"inv-nodes='0-2'"	1,256	0,559	0,936
15	"inv-nodes='0-2'"	"node-caps='no'" - "Class='no-recurrence-events'"	1,256	0,559	0,751

FIGURE 9.3 – Représentation textuelle des règles sous Tanagra

Cette première méthode permet de représenter facilement un grand nombre de règles mais ne donne pas une vision globale de toutes ces règles.

Représentation par matrice 2D ou 3D Deux types de représentation par matrice ont émergées : les matrices "itemset-itemset" et les matrices "itemset-règle" (un itemset est un ensemble d'attributs).

Pour les matrices itemset-itemset, les colonnes correspondent aux parties gauches des règles et les lignes correspondent aux parties droites. Les indices de qualité sont représentés grâce à des effets de couleur, de forme ou de taille (cf par exemple la figure 9.4 pour le logiciel DBMiner [56]).

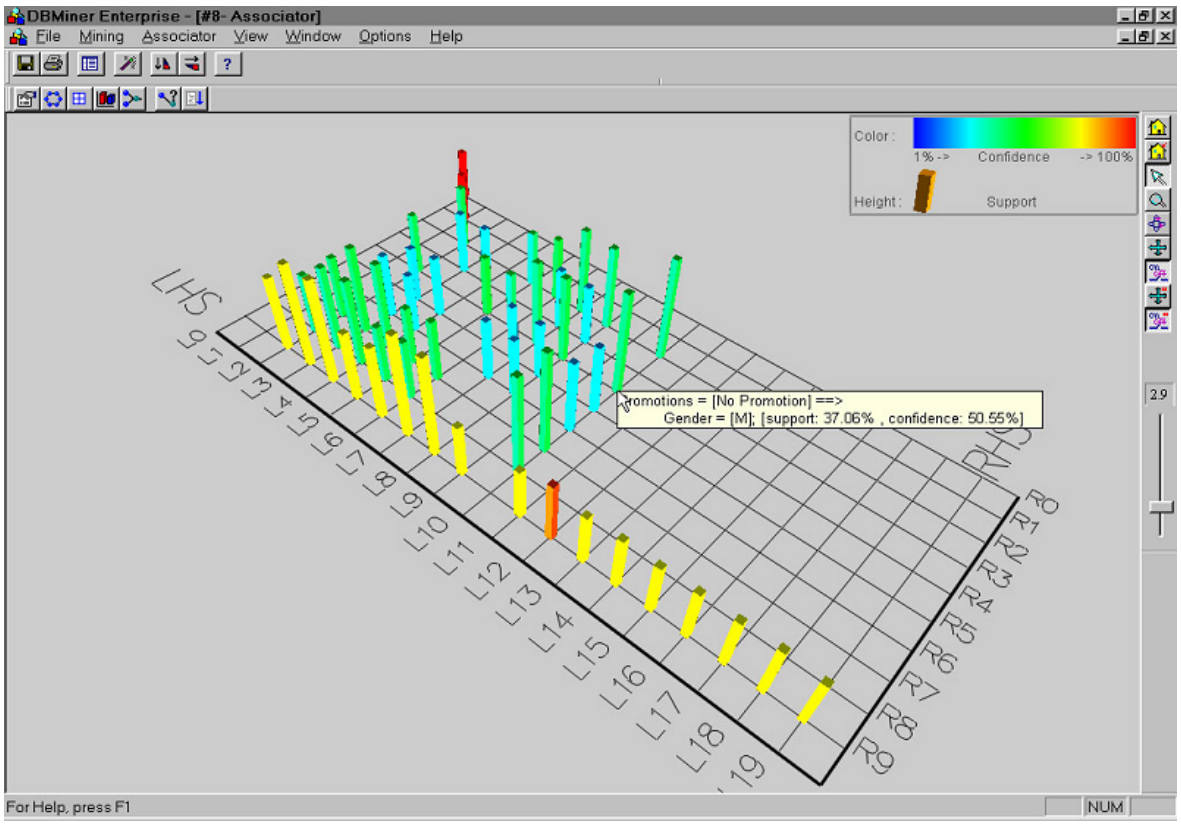


FIGURE 9.4 – Représentation par matrice itemset-itemset sous DBMiner

Pour les matrices itemset-règle [100], les colonnes correspondent aux attributs qui composent les parties droites et gauches des règles et à chaque ligne correspond une règle. Deux couleurs sont utilisées pour distinguer les parties droites des parties gauches. Deux lignes sont également ajoutées pour visualiser le support et la confiance.

Ces approches nécessitent un effort considérable de la part des experts puisque cette visualisation n'est à priori pas très intuitive.

Représentation par graphe Les règles sont ici représentées dans un graphe où les nœuds représentent les parties gauches et droites des règles et les arcs représentent les règles entre ces parties. Les indices de qualité peuvent être mis en avant en augmentant la taille des nœuds ou des arêtes, en utilisant plusieurs couleurs ou plusieurs formes (cf par exemple la figure 9.5 pour le logiciel DBMiner [56]).

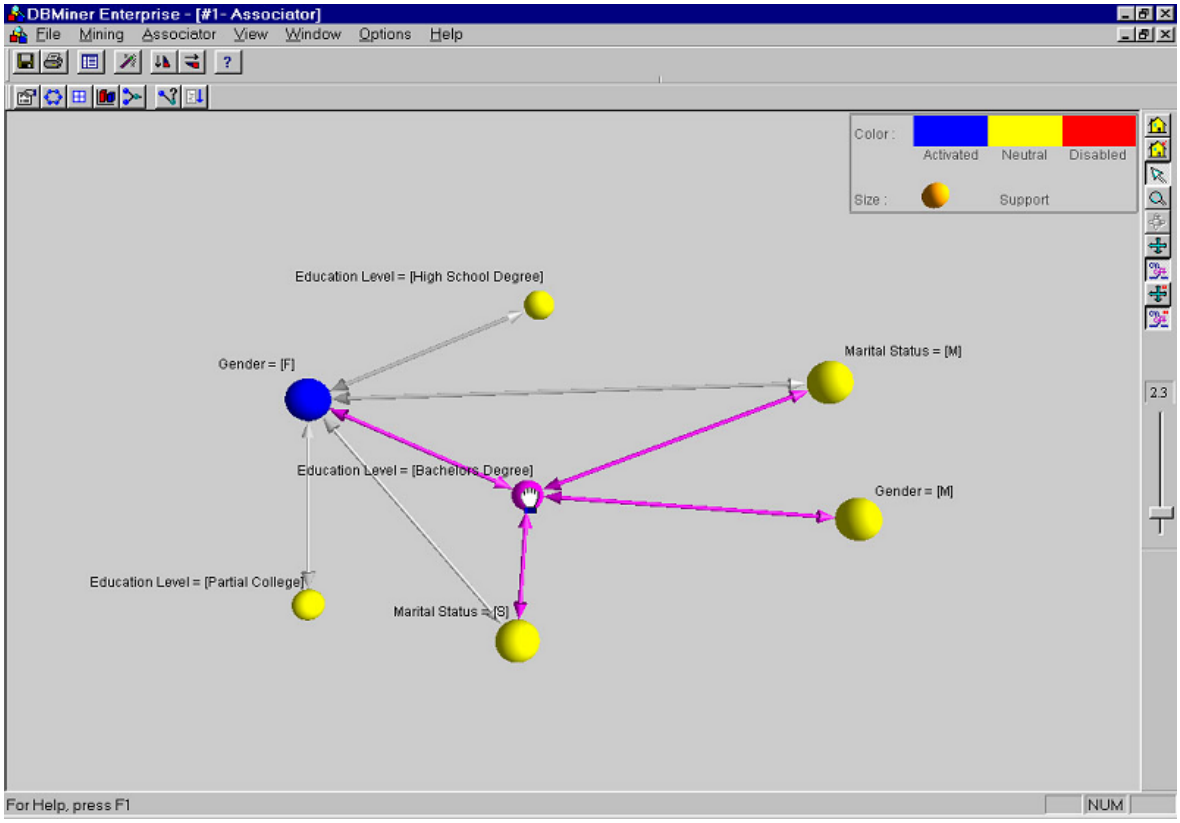


FIGURE 9.5 – Représentation par graphe sous DBMiner

Ces méthodes de visualisation sont très intéressantes dès lors que le nombre de règles n'est pas très important. Lorsque celui-ci augmente, les arcs se mélangent rapidement et la visualisation est vite illisible.

Toutefois, ces approches de visualisation sont très intuitives et sont particulièrement familières aux biologistes puisque ces approches se rencontrent fréquemment, notamment pour l'analyse de réseaux de régulation.

D'autres méthodes de visualisation ont été proposées comme par exemple la représentation 3D par réalité virtuelle dans le logiciel ARvis [30].

9.2.2 Vers des réseaux globaux

Pour choisir la meilleure représentation des règles en fonction du domaine d'application, plusieurs critères peuvent être pris en compte :

- Visualisation intuitive : La visualisation est-elle intuitive pour les experts du domaine ?
- Vue globale des règles : La représentation permet-elle de donner une vue globale des règles, regroupant les règles avec les mêmes parties gauches, les mêmes parties droites... ?
- Nombre important de règles : Lorsque le nombre de règles est important, la représentation est-elle toujours lisible ?
- Nombre important d'attributs : La représentation permet-elle aux règles d'avoir beaucoup d'attributs en partie gauche et en partie droite ?
- Présentation des indices de qualité : Les indices de qualité sont-ils faciles à visualiser ?
- Mise en avant des attributs centraux : Est-il possible de mettre facilement en avant certains attributs ?

Dans le tableau 9.1, nous avons regroupé les caractéristiques des différentes approches de visualisation décrites précédemment. Pour chaque approche et pour chaque critère, nous avons indiqué s'il s'agissait plutôt d'un point fort (+ à +++) ou plutôt d'un point faible (-).

	Représentation textuelle	Matrice 2D ou 3D	Représentation par graphe
Visualisation intuitive	++	-	+++
Vue globale des règles	-	+	+++
Nombre important de règles	++	+	-
Nombre important d'attributs	++	+	-
Présentation des indices de qualité	+++	++	+
Mise en avant des attributs centraux	+	+	+++

TABLE 9.1 – Différentes méthodes de visualisation des règles

Dans le cadre des données d'expression, nous souhaitons avant tout réaliser un outil convivial pour les utilisateurs. La meilleure méthode de visualisation est donc celle qui paraîtra la plus facile à interpréter pour les biologistes.

D'autre part, nous proposons de générer uniquement les règles de la base canonique et celles de la base de Gottlob et Libkin. De plus, une phase de post-traitement des règles permet de filtrer les règles en fonction de cinq indices de qualité. Ces deux aspects permettent de diminuer considérablement le nombre des règles pour ne garder que les règles les plus intéressantes pour les experts. De plus, cette phase de post-traitement nous amène à penser qu'il n'est pas utile de visualiser des indices directement mais il serait plutôt préférable de proposer une méthode qui permette en un clic de visualiser rapidement tous les indices pour la règle ou l'attribut sélectionnés.

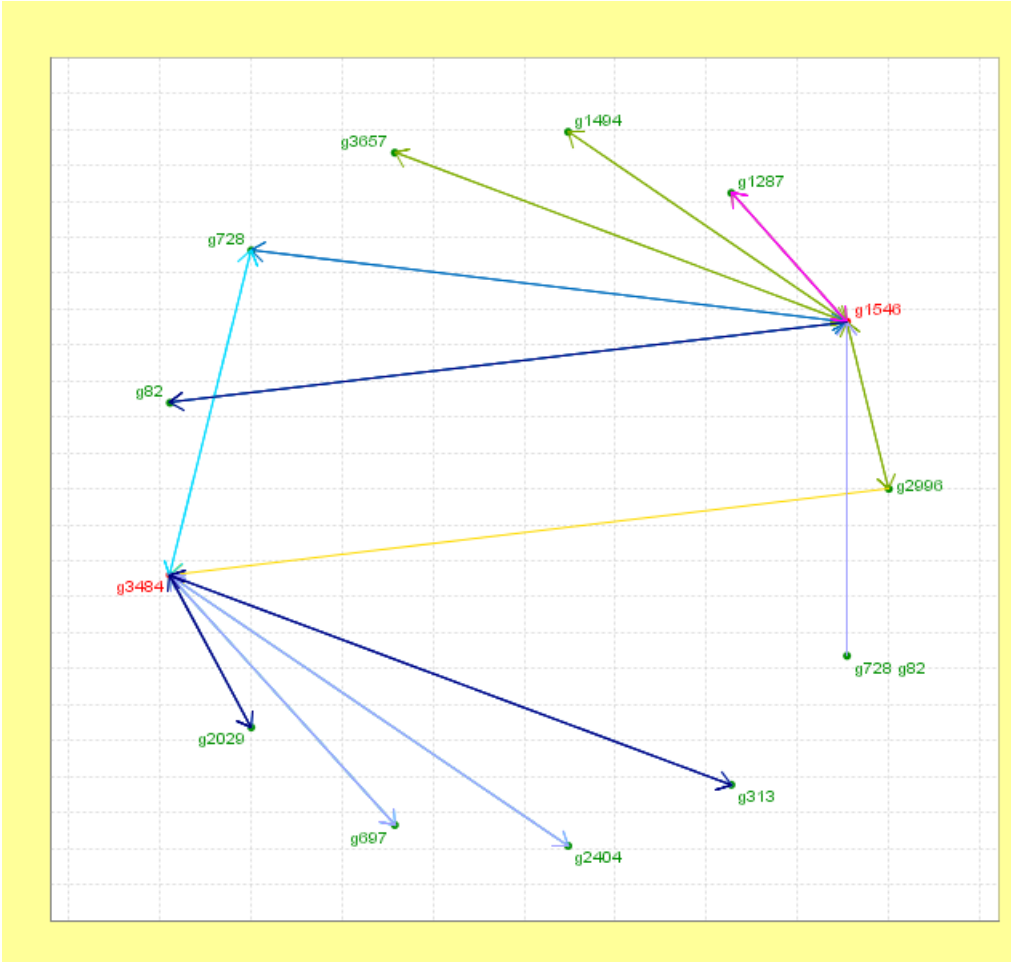


FIGURE 9.6 – Réseau global avec plusieurs sémantiques et deux gènes centraux

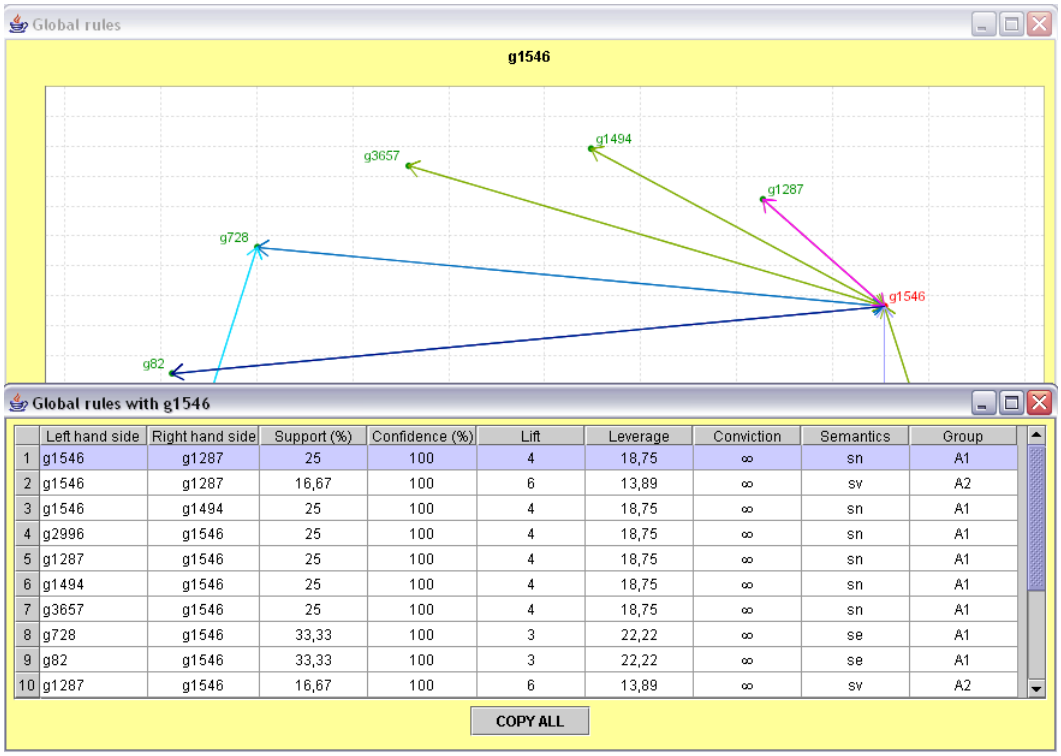
Dans le cadre spécifique de la découverte de règles entre gènes, l'interprétation des règles est une étape particulièrement délicate et très difficile pour les biologistes, puisqu'une règle entre deux gènes impliquent également divers produits associés (protéines, facteurs de transcription...). C'est pourquoi les biologistes sont rarement intéressés par les règles avec plus de 3 ou 4 attributs en partie gauche, l'interprétation devenant très vite impossible.

Enfin, nous souhaitons offrir la possibilité aux utilisateurs de choisir avant la génération des règles, quelques gènes d'intérêts qu'ils voudront voir apparaître en parties gauches ou droites des règles (cf section 8.4). Les biologistes pourront ainsi visualiser l'ensemble des règles associées à ces quelques gènes qui les intéressent plus particulièrement.

Nous avons donc opté tout naturellement pour une visualisation par graphe qui nous semble être la méthode la plus facile à interpréter et la plus intuitive pour les experts, notre objectif étant aussi de nous positionner dans le cadre de la reconstruction de réseaux de gènes (cf section 9.1).

L'originalité de notre approche réside dans le fait que nous proposons des graphes avec **plusieurs sémantiques**, que nous appellerons **réseaux globaux**. Chaque sémantique est caractérisée par une couleur particulière. De plus, nous avons choisi de représenter les différents attributs composant les parties droites et gauches des règles, sous forme d'un cercle pour une meilleure lisibilité. Enfin, les gènes centraux sont colorés en rouge, permettant ainsi de les repérer plus facilement (cf figure 9.6).

Nous proposons également en un simple clic de visualiser les indices associés aux règles impliquant le gène sélectionné (cf figure 9.7).



	Left hand side	Right hand side	Support (%)	Confidence (%)	Lift	Leverage	Conviction	Semantics	Group
1	g1546	g1287	25	100	4	18,75	∞	sn	A1
2	g1546	g1287	16,67	100	6	13,89	∞	sv	A2
3	g1546	g1494	25	100	4	18,75	∞	sn	A1
4	g2996	g1546	25	100	4	18,75	∞	sn	A1
5	g1287	g1546	25	100	4	18,75	∞	sn	A1
6	g1494	g1546	25	100	4	18,75	∞	sn	A1
7	g3657	g1546	25	100	4	18,75	∞	sn	A1
8	g728	g1546	33,33	100	3	22,22	∞	se	A1
9	g82	g1546	33,33	100	3	22,22	∞	se	A1
10	g1287	g1546	16,67	100	6	13,89	∞	sv	A2

FIGURE 9.7 – Règles associées au gène spécifié

Chapitre 10

Module RG

Sommaire

L'approche proposée a été développée spécifiquement pour les données d'expression de gènes et dans le but de faciliter son utilisation par les biologistes. Pour cela, un nouveau module appelé RG (Rule Generation) a été intégré à un logiciel gratuit et open-source consacré à l'analyse de données d'expression de gènes, le logiciel MeV (MultiExperiment-Viewer) [83] (cf figure 10.1).

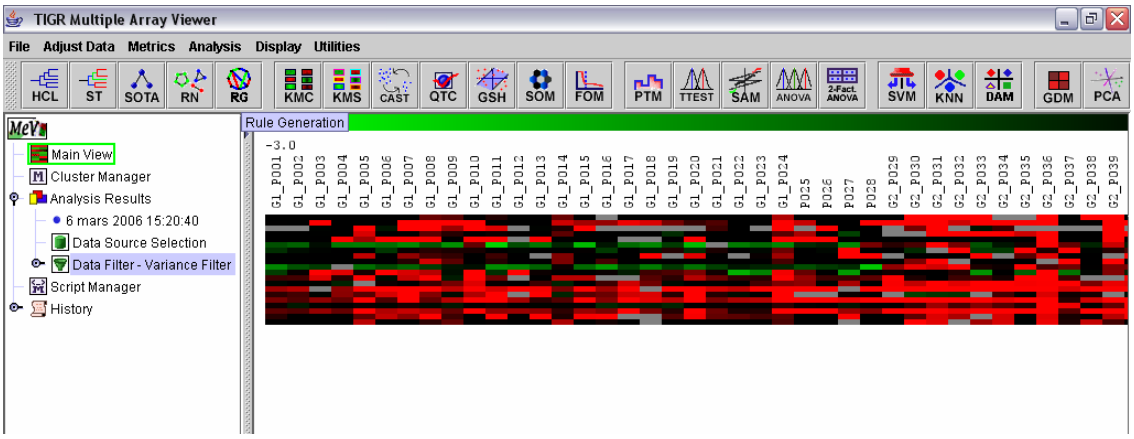


FIGURE 10.1 – Module RG intégré au logiciel MeV

Cet outil fait partie d'une suite de quatre logiciels, appelée TM4, développée par The

Institute for Genomic Research (TIGR). Cette suite est entièrement consacrée au traitement des données issues de biopuces : stockage, analyse des images, normalisation, analyse statistique et informatique des données... Le logiciel MeV est l'application consacrée à cette dernière étape, sa convivialité fait de cet outil, un des plus utilisés actuellement.

La version étendue du logiciel MeV avec le module RG est disponible sur le site http ://www.isima.fr/agier/GeneRules.

Les points-clés de ce module sont les suivants :
- Une interface conviviale permettant de choisir parmi plusieurs sémantiques.
- La possibilité de spécifier des gènes d'intérêts dits "gènes centraux".
- Le calcul de 5 indices de qualité pur chaque règle générée.
- La visualisation de réseaux globaux incluant diverses sémantiques.

La découverte des règles au sein du module RG s'opère en six étapes, décrites dans les sections suivantes :

1. Caractérisation des données (données temporelles, présence de groupes d'échantillons...)
2. Choix des gènes centraux
3. Choix des sémantiques adaptées au jeu de données étudié
4. Génération des règles
5. Post-traitement des règles
6. Visualisation des réseaux de gènes

10.1 Caractérisation des données

Tout d'abord, il est possible de préciser si les échantillons appartiennent à différents groupes et si les données sont temporelles (cf figure 10.2).

Si c'est le cas, une nouvelle fenêtre s'ouvre permettant d'attribuer à chaque échantillon une classe et/ou un temps. Par exemple, pour la relation r_3 présentée dans la section 3.2, il faut spécifier les caractéristiques des données comme décrit dans la figure 10.3.

En fonction de ces caractéristiques, ne seront proposées ensuite que les sémantiques adaptées aux données. Par exemple, si les données ne sont pas temporelles, la sémantique s_e et ses variantes ne seront pas accessibles, puisque celles-ci n'auraient aucun sens.

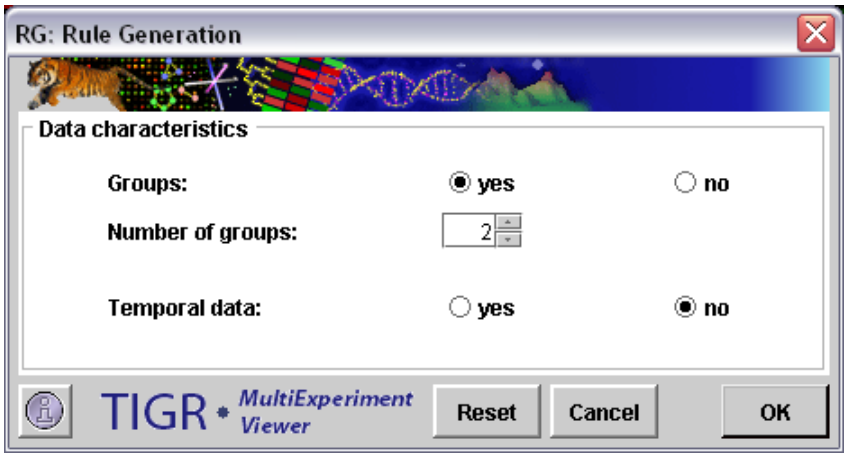


FIGURE 10.2 – Caractérisation des données

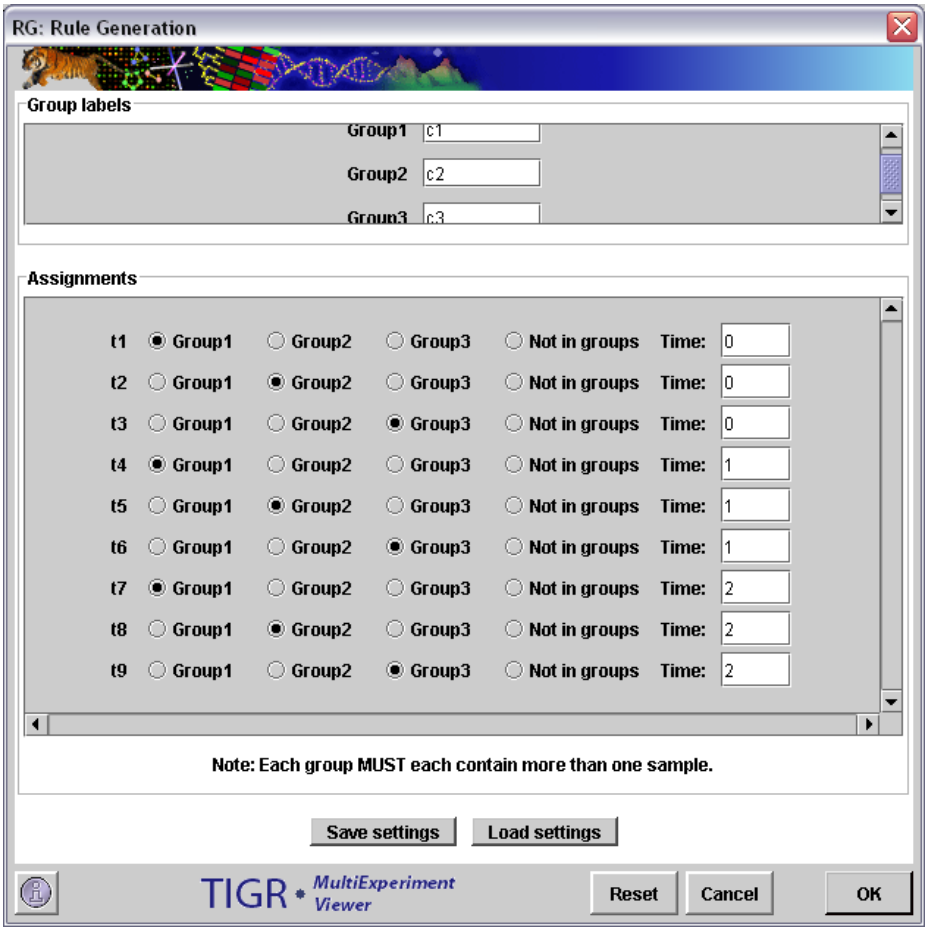


FIGURE 10.3 – Caractérisation des données pour la relation r_3

10.2 Choix des gènes centraux

Une fois les caractéristiques des données spécifiées, il faut choisir les gènes centraux ainsi que les sémantiques (cf figure 10.4).

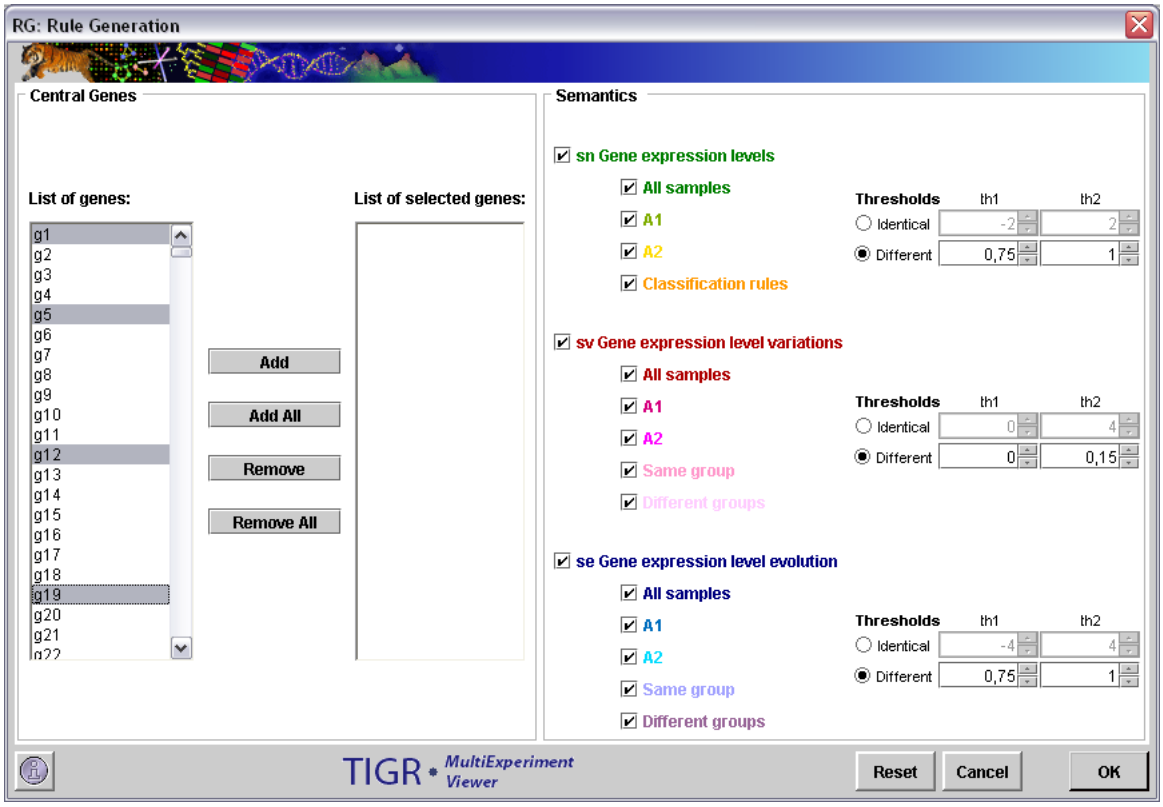


FIGURE 10.4 – Choix des gènes centraux et des sémantiques

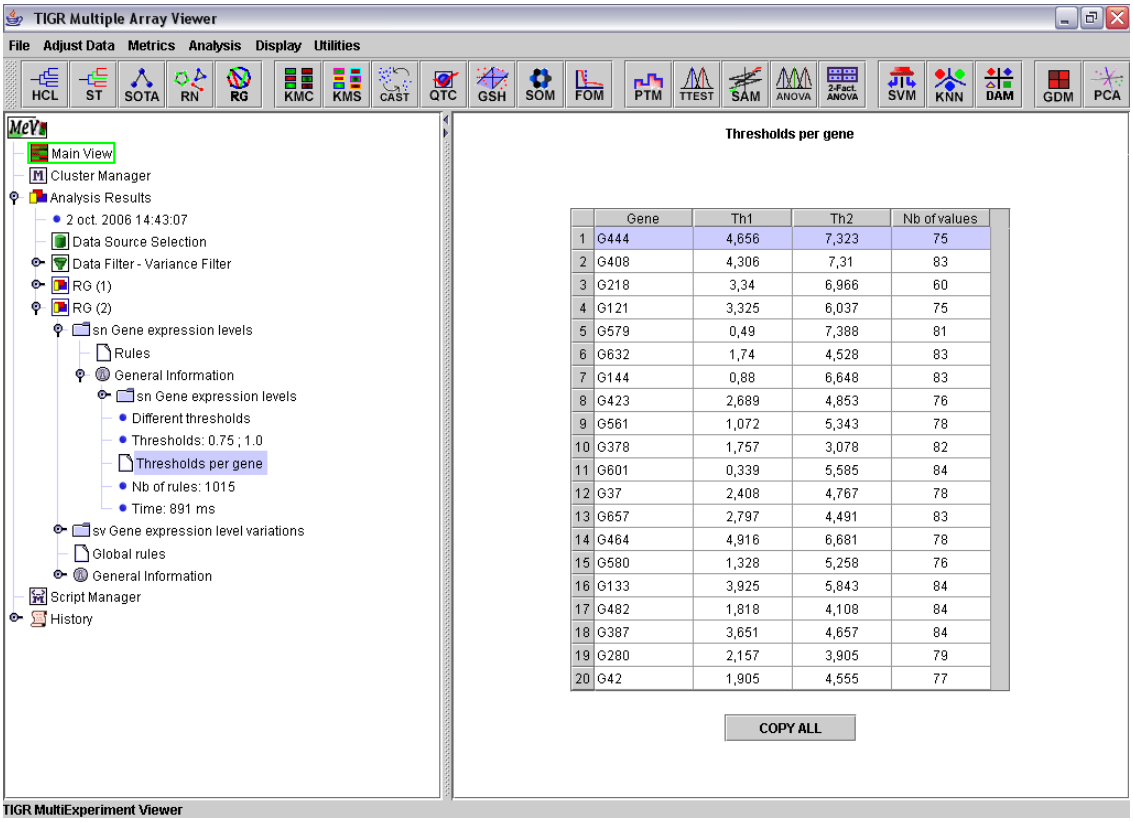
Les gènes centraux correspondent aux gènes d'intérêts, les règles générées seront celles pour lesquelles les gènes centraux jouent vraiment un rôle i.e. les règles dont la partie gauche ou la partie droite est un gène central.

10.3 Choix des sémantiques

Se fait ensuite le choix des sémantiques et des seuils associés en fonction des objectifs de l'étude. Les différentes sémantiques présentées dans le chapitre 7 sont proposées. Il est également possible de générer des règles de la forme $g_1 g_2 \rightarrow Group1$ ou $Group2 \rightarrow g_1$. Ces règles permettent de caractériser les différents groupes et facilitent ainsi l'interprétation des réseaux obtenus.

Par défaut, pour toutes les sémantiques, les seuils sont différents pour chaque gène

et calculés automatiquement. Les seuils calculés peuvent ensuite être visualisés (cf figure 10.5).



	Gene	Th1	Th2	Nb of values
1	G444	4,656	7,323	75
2	G408	4,306	7,31	83
3	G218	3,34	6,966	60
4	G121	3,325	6,037	75
5	G579	0,49	7,388	81
6	G632	1,74	4,528	83
7	G144	0,88	6,648	83
8	G423	2,689	4,853	76
9	G561	1,072	5,343	78
10	G378	1,757	3,078	82
11	G601	0,339	5,585	84
12	G37	2,408	4,767	78
13	G657	2,797	4,491	83
14	G464	4,916	6,681	78
15	G580	1,328	5,258	76
16	G133	3,925	5,843	84
17	G482	1,818	4,108	84
18	G387	3,651	4,657	84
19	G280	2,157	3,905	79
20	G42	1,905	4,555	77

FIGURE 10.5 – Seuils calculés automatiquement pour la sémantique s_n

10.4 Génération des règles

Une fois les gènes centraux et les sémantiques choisies en fonction des objectifs de l'étude, les couvertures des règles sont générées : la couverture canonique pour les règles exactes et la couverture de Gottlob et Libkin pour les règles approximatives (cf section 8.2). La génération des règles a été implémenté en C++ tandis que les interfaces ont été développées sous Java. Les expérimentations ont été faites sur les données présentées dans la section 5, l'accent n'a pas été mis sur les comparaisons expérimentales des algorithmes.

10.5 Post-traitement des règles

Une fois les règles générées, un filtre est possible à partir des cinq indices de qualité présentés dans la section 6.5 : support, confiance, lift, leverage et conviction (cf figure 10.6).

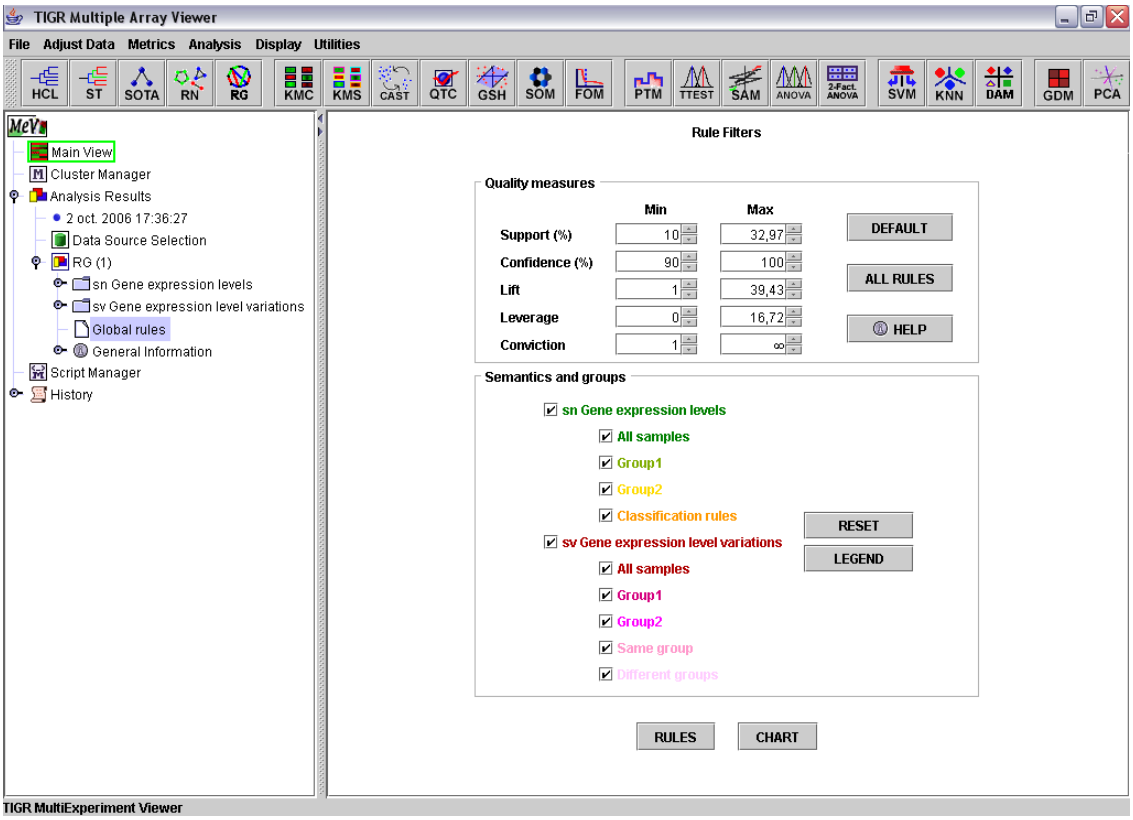


FIGURE 10.6 – Filtrage des règles

De manière générale, nous pouvons dire que plus les indices sont élevés plus la règle est intéressante, mais tous les indices n'évoluent pas de la même façon.

10.6 Visualisation des réseaux de gènes

Deux modes de visualisation sont possibles :

– **Représentation textuelle** (cf figure 10.7) :
Les règles obtenues peuvent être visualisées sous forme textuelle avec le détail des différents indices de qualité ainsi que les sémantiques correspondantes.

Global rules

	Left hand side	Right hand side	Support (%)	Confidence (%)	Lift	Leverage	Conviction	Semantics	Group
1	G42	G444	38,04	100	2,03	19,3	∞	sv	Group1
2	G444 G121	G561	32,97	100	2,03	16,72	∞	sv	Group1
3	G121 G561	G444	32,97	100	2,03	16,72	∞	sv	Group1
4	G121 G42	G561	28,26	100	2,03	14,34	∞	sv	Group1
5	G561	G579	28,25	100	1,56	10,09	∞	sv	Group2
6	G657	Group2	23,81	100	1,5	7,94	∞	sn	Class. rules
7	G378	Group2	23,81	100	1,5	7,94	∞	sn	Class. rules
8	G482	Group2	22,62	100	1,5	7,54	∞	sn	Class. rules
9	G632	Group2	22,62	100	1,5	7,54	∞	sn	Class. rules
10	G121	Group2	22,62	100	1,5	7,54	∞	sn	Class. rules
11	G444	Group2	21,43	100	1,5	7,14	∞	sn	Class. rules
12	G579 G280	G580	20,83	100	2,4	12,15	∞	sn	Group1
13	G580 G280	G579	20,83	100	2,4	12,15	∞	sn	Group1

COPY ALL

FIGURE 10.7 – Représentation textuelle des règles

– **Représentation graphique** (cf figure 10.8) :
Les règles peuvent également être visualisées sous forme de **réseau global** regroupant différentes sémantiques représentées par diverses couleurs.

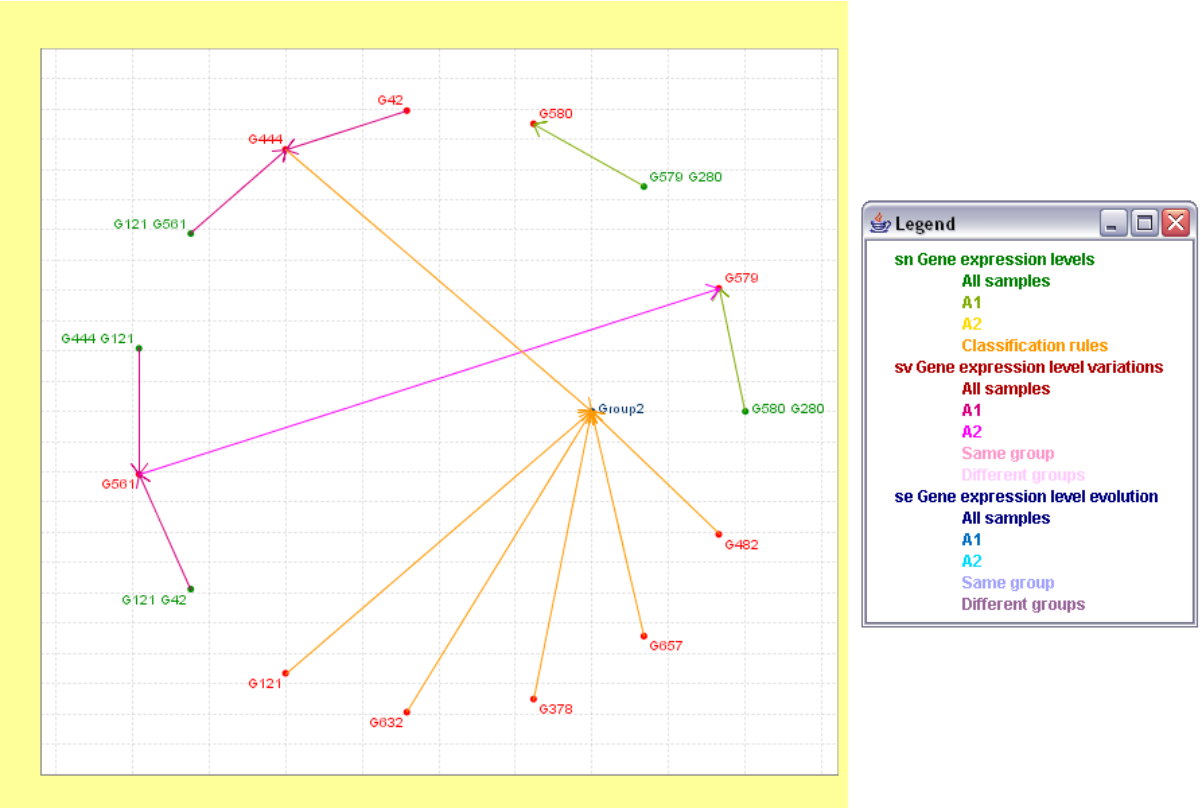


FIGURE 10.8 – Réseau global avec plusieurs sémantiques

Chapitre 11

Mise en œuvre

Sommaire

Dans ce chapitre, nous présentons les premiers réseaux obtenus pour les deux applications présentées dans l'introduction.

11.1 Etude de l'envahissement ganglionnaire

L'objectif de cette étude était d'identifier des gènes permettant de caractériser le statut ganglionnaire de patientes atteintes de cancer du sein. Dans la section 5.1, plusieurs listes de gènes ont été découvertes ayant chacune un fort pouvoir discriminant. L'objectif était donc pour chacune de ces listes, d'étudier les relations entre les différents gènes.

Pour des raisons de confidentialité, les noms des gènes ne peuvent être cités ici. Toutefois, nous donnons dans la figure 11.1, un réseau obtenu à partir de données publiques [95] de tumeurs mammaires similaires aux données étudiées dans le projet. Ces données correspondent aux profils transcriptomiques de 34 patientes atteintes de cancer du sein pour lesquelles nous disposons du statut des récepteurs hormonaux.

A partir de 24 gènes connus pour être corrélés avec le statut des récepteurs hormonaux, les règles à partir des gènes faiblement exprimés ont par exemple été découvertes. Les règles les plus intéressantes en fonction des différents indices calculés ont ensuite été sélectionnées. Ces règles sont données dans la figure 11.2.

Les résultats obtenus révèlent des interactions entre gènes bien connues des biologistes. Par exemple, la règle : $ESR1 \rightarrow TFF1$ est une relation bien connue pour ce type de tumeurs : le gène ESR1 code pour un récepteur nucléaire, une super-famille de facteurs

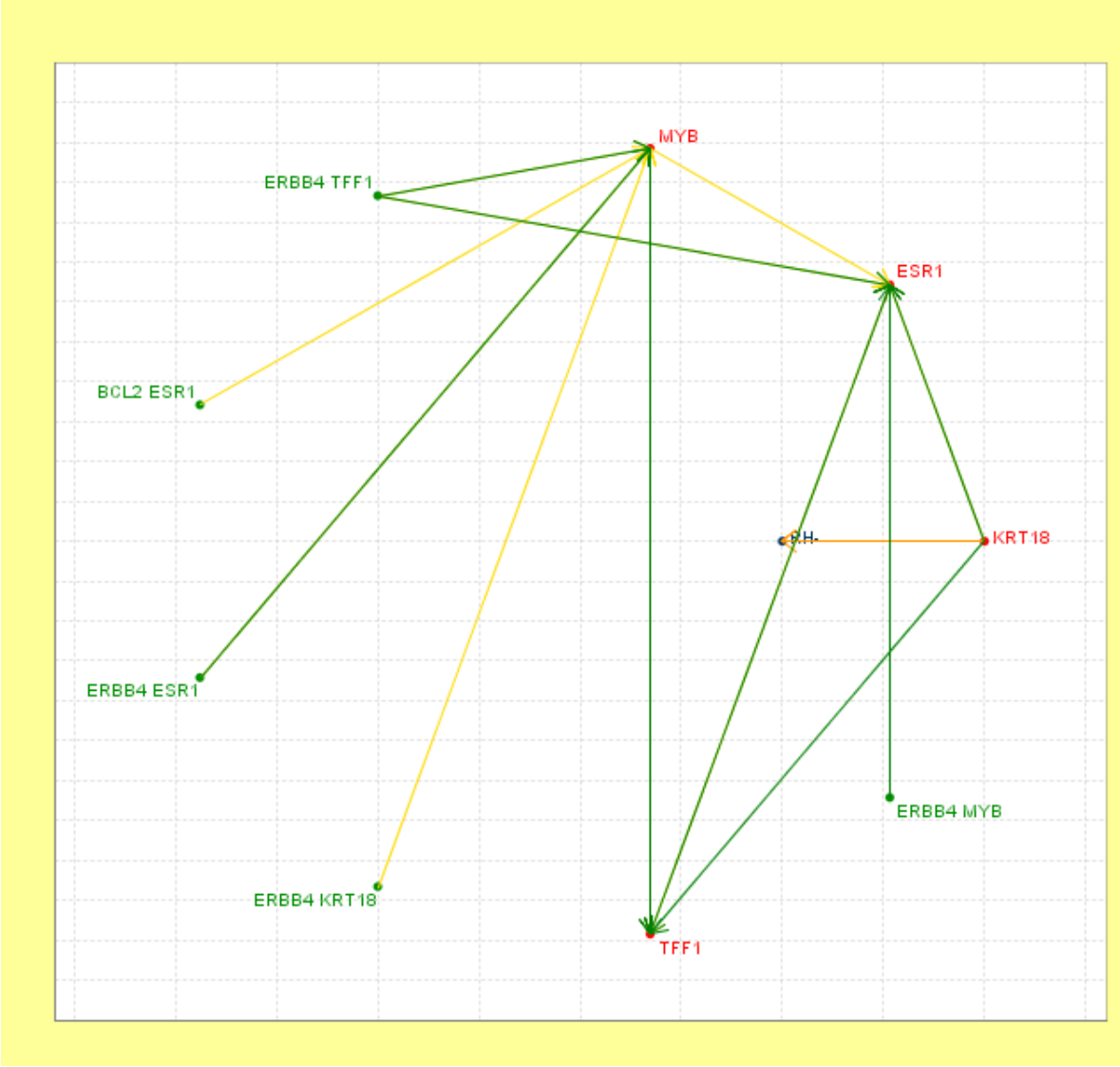
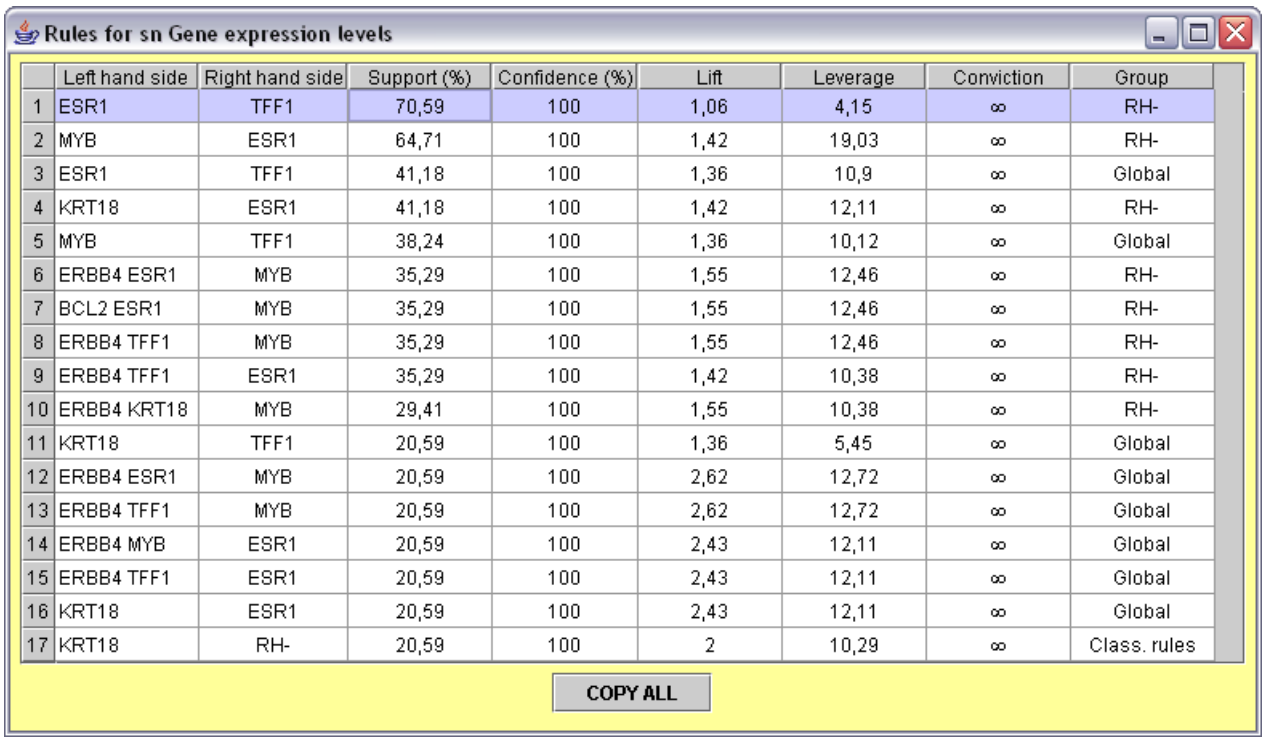


FIGURE 11.1 – Réseau obtenu pour la sémantique s_n



Rules for sn Gene expression levels

	Left hand side	Right hand side	Support (%)	Confidence (%)	Lift	Leverage	Conviction	Group
1	ESR1	TFF1	70,59	100	1,06	4,15	∞	RH-
2	MYB	ESR1	64,71	100	1,42	19,03	∞	RH-
3	ESR1	TFF1	41,18	100	1,36	10,9	∞	Global
4	KRT18	ESR1	41,18	100	1,42	12,11	∞	RH-
5	MYB	TFF1	38,24	100	1,36	10,12	∞	Global
6	ERBB4 ESR1	MYB	35,29	100	1,55	12,46	∞	RH-
7	BCL2 ESR1	MYB	35,29	100	1,55	12,46	∞	RH-
8	ERBB4 TFF1	MYB	35,29	100	1,55	12,46	∞	RH-
9	ERBB4 TFF1	ESR1	35,29	100	1,42	10,38	∞	RH-
10	ERBB4 KRT18	MYB	29,41	100	1,55	10,38	∞	RH-
11	KRT18	TFF1	20,59	100	1,36	5,45	∞	Global
12	ERBB4 ESR1	MYB	20,59	100	2,62	12,72	∞	Global
13	ERBB4 TFF1	MYB	20,59	100	2,62	12,72	∞	Global
14	ERBB4 MYB	ESR1	20,59	100	2,43	12,11	∞	Global
15	ERBB4 TFF1	ESR1	20,59	100	2,43	12,11	∞	Global
16	KRT18	ESR1	20,59	100	2,43	12,11	∞	Global
17	KRT18	RH-	20,59	100	2	10,29	∞	Class. rules

COPY ALL

FIGURE 11.2 – Différentes règles obtenues

transcriptionnels activés par un ligand, qui modulent l'expression spécifique de gènes. TFF1 (ou pS2) est un gène induit par les oestrogènes et impliqué dans divers processus biologiques. En l'absence d'expression de ESR1, les oestrogènes ne peuvent réguler le niveau d'ARN messager de ce gène. Nous pouvons de plus remarquer que cette règle est d'autant plus intéressante chez les tumeurs RH- i.e. les tumeurs dont la réceptivité hormonale est inactive ce qui est tout à fait logique. De la même façon, les autres règles ont également des interprétations biologiques [53, 52, 88].

11.2 Etude de la sensibilité au docétaxel

Dans la section 5.2, 101 gènes permettant de mettre en évidence un profil d'expression caractéristique de la réponse thérapeutique au docétaxel ont été identifiés. A partir de ces gènes, des réseaux ont été générés permettant de déterminer des interactions intéressantes entre les gènes.

Par exemple, le réseau donné dans la figure 11.3 permet de mettre en évidence à partir de plusieurs sémantiques, les relations entre d'une part les gènes sur-exprimés chez les patientes qui ont résisté au docétaxel et d'autre part les gènes sur-exprimés chez les patientes qui ont été sensibles au docétaxel.

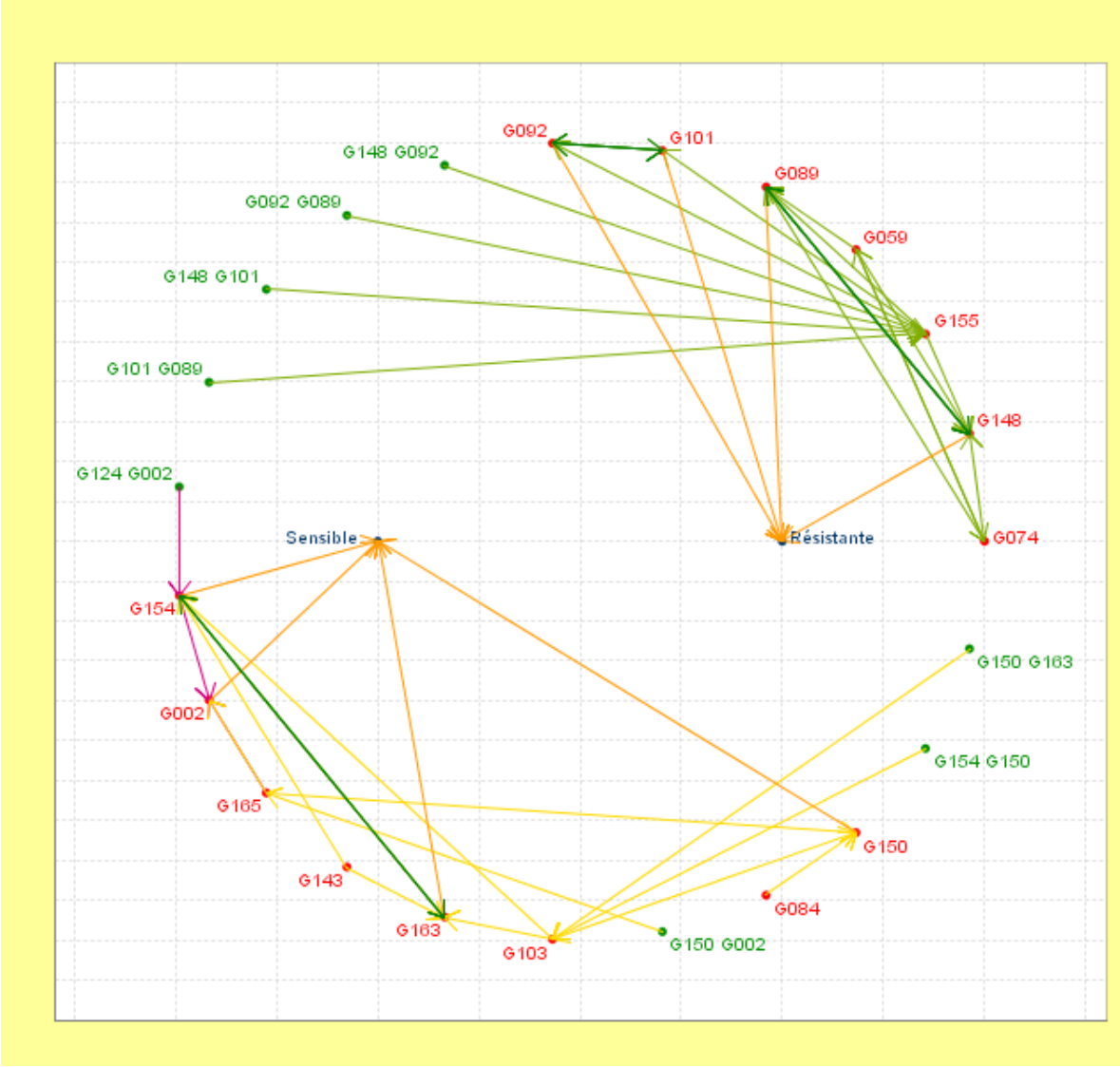


FIGURE 11.3 – Réseau global pour l'étude de la sensibilité au docétaxel

11.2 Etude de la sensibilité au docétaxel

Les règles obtenues présentant les plus grand supports sont données dans la figure11.4.

Global rules

	Left hand side	Right hand side	Support (%)	Confidence (%)	Lift	Leverage	Conviction	Semantics	Group
1	G154	G163	60	100	1,67	24	∞	sn	Sensible
2	G163	G154	60	100	1,67	24	∞	sn	Sensible
3	G148	G089	60	100	1,67	24	∞	sn	Résistante
4	G092	G101	60	100	1,67	24	∞	sn	Résistante
5	G101	G092	60	100	1,67	24	∞	sn	Résistante
6	G089	G148	60	100	1,67	24	∞	sn	Résistante
7	G150 G163	G103	40	100	2,5	24	∞	sn	Sensible
8	G154 G150	G103	40	100	2,5	24	∞	sn	Sensible
9	G103	G163	40	100	1,67	16	∞	sn	Sensible
10	G143	G163	40	100	1,67	16	∞	sn	Sensible
11	G165	G002	40	100	1,67	16	∞	sn	Sensible
12	G103	G150	40	100	1,67	16	∞	sn	Sensible
13	G165	G150	40	100	1,67	16	∞	sn	Sensible
14	G084	G150	40	100	1,67	16	∞	sn	Sensible
15	G150 G002	G165	40	100	2,5	24	∞	sn	Sensible
16	G103	G154	40	100	1,67	16	∞	sn	Sensible
17	G143	G154	40	100	1,67	16	∞	sn	Sensible
18	G165	G002	40	100	1,43	12	∞	sv	Résistante
19	G154	G002	40	100	1,43	12	∞	sv	Résistante
20	G059	G089	40	100	1,67	16	∞	sn	Résistante
21	G155	G089	40	100	1,67	16	∞	sn	Résistante
22	G074	G089	40	100	1,67	16	∞	sn	Résistante
23	G074	G059	40	100	2,5	24	∞	sn	Résistante
24	G101 G089	G155	40	100	2,5	24	∞	sn	Résistante
25	G148 G101	G155	40	100	2,5	24	∞	sn	Résistante

COPY ALL

FIGURE 11.4 – Différentes règles obtenues

La validation biologique de ces résultats n'en est qu'au début et consiste pour l'instant à retrouver des règles connues dans la littérature afin de valider l'approche proposée. L'étape suivante consistera à essayer d'expliquer d'un point de vue biologique de nouvelles règles jusqu'ici inconnues.

Quatrième partie

Conclusion

L'objectif de la post-génomique est d'étudier les fonctions des gènes et leurs interactions. Dans le cadre de cette thèse, nous nous sommes intéressés plus particulièrement à des données de tumeurs mammaires. Deux projets ont été principalement développés, le premier sur l'étude du statut ganglionnaire des tumeurs et le deuxième sur la sensibilité au docétaxel. L'objectif à terme de ces projets est d'améliorer le diagnostic et le pronostic du cancer du sein.

Cette thèse se place plus particulièrement dans le domaine de la fouille de données d'expression de gènes et s'articule autour de deux parties. L'objectif de la première partie était de proposer les méthodes et outils les plus adaptés pour pré-traiter et analyser l'ensemble des données d'expression de ces différents projets, en assurant une veille technologique. La seconde partie plus prospective consistait à mettre en place une nouvelle technique de reconstruction de réseaux de gènes basée sur la notion de règles entre gènes.

Un nouveau processus de pré-traitement a été proposé pour les biopuces TM/CS DIAGNOGENE utilisées pour l'étude du statut ganglionnaire dont l'originalité est la prise en compte de différents réglages de scanner afin de choisir le réglage le plus intéressant pour chaque spot ainsi qu'une méthode permettant de s'assurer de la reproductibilité des réplicats d'un même gène. Des outils existants sont utilisés pour certains points, d'autres plus novateurs ont nécessité l'implémentation d'un nouvel outil.

Dans le cadre des deux principaux projets, plusieurs analyses ont été réalisées et ont permis d'identifier des gènes permettant de mettre en évidence un profil d'expression caractéristique du statut ganglionnaire et de la réponse thérapeutique au docétaxel. Les résultats de cette première partie ont ainsi pu donner quelques pistes pour essayer de déterminer la fonction des différents gènes.

Le second objectif de la post-génomique est de découvrir les relations entre les gènes. La reconstruction de réseaux de gènes à partir de données d'expression est une piste explorée par de nombreuses équipes actuellement.

Dans ce travail de thèse, nous nous sommes intéressés plus particulièrement à la notion de règles entre gènes et avons proposé une approche permettant d'inclure plusieurs sémantiques pour les règles. Ce choix se justifie par notre domaine d'application, puisque compte tenu des différents objectifs biologiques possibles, il peut être utile de découvrir différents types de règles entre gènes et ne pas se restreindre à une seule sémantique.

L'originalité de ce travail est de proposer un cadre global pouvant inclure un grand nombre de sémantiques pour les règles et d'utiliser des méthodes identiques de génération, de post-traitement et de visualisation des règles pour toutes les sémantiques proposées. Il est possible dans ce cadre de considérer des données temporelles ou non, selon la sémantique choisie ou bien de prendre en compte des informations externes comme la présence de groupes d'échantillons.

Pour formaliser le cadre, nous avons introduit la notion de "sémantiques bien-formées", i.e. les sémantiques pour lesquelles les axiomes d'Armstrong sont justes et complets. Nous avons également donné un résultat important permettant de savoir simplement si une sémantique est ou non bien-formée grâce à des restrictions syntaxiques.

Nous proposons également une visualisation des règles sous forme de graphes. L'originalité de notre approche est d'une part de superposer des règles avec des sémantiques différentes au sein d'un même support visuel et d'autre part de ne générer que les règles qui impliquent des gènes dits "centraux", spécifiés en amont par les experts.

L'approche proposée a été développée spécifiquement pour les données d'expression dans un nouveau module appelé RG (Rule Generation), intégré à un logiciel gratuit et open-source consacré à l'analyse de données d'expression de gènes, le logiciel MeV (MultiExperimentViewer) [83]. Ce module est disponible sur internet.

L'objectif de ce travail étant d'offrir aux biologistes la possibilité de générer différents types de réseaux, il serait tout à fait intéressant d'inclure d'autres méthodes d'inférence de réseaux, comme celles présentées dans la section 9.1. L'idée est de proposer aux biologistes un outil complet, permettant de générer à partir d'un même jeu de données, différents réseaux et mesurer ainsi la "force" des relations obtenues. Il est également envisageable de permettre aux utilisateurs de définir de nouvelles sémantiques en fonction de leurs propres besoins, et ceci de façon interactive directement à partir du module RG.

Nous avons fait le choix du système d'axiomes d'Armstrong. Il serait intéressant d'étudier d'autres règles d'inférence [74, 69] permettant peut-être d'ajouter des sémantiques non permises dans le cadre proposé. Les sémantiques avec deux prédicats différents par exemple, pourraient dans certains cas être intéressantes.

Bibliographie

[1] Agilent Technologies. www.agilent.com.

[2] BioDiscovery. www.biodiscovery.com.

[3] Centre Jean Perrin. www.cjp.fr.

[4] Clinical Tools Inc. www.clinicaltools.com.

[5] DIAGNOGENE. www.diagnogene.fr.

[6] LIMOS. www.isima.fr/limos.

[7] Molecular Devices. www.moleculardevices.com.

[8] National Human Genome Research Institute. www.genome.gov/glossary.cfm.

[9] Plate-forme transcriptome de l'Ecole Normale Supérieure. www.transcriptome.ens.fr.

[10] SAS. www.sas.com.

[11] The Institute for Genomic Research. www.tigr.org.

[12] M. Agier. Différents types de règles pour la reconstruction de réseaux de gènes à partir de données d'expression. *Numéro spécial de la revue I3 Information - Interaction - Intelligence*, à paraître, 2006.

[13] M. Agier, V. Chabaud, J.-M. Petit, V. Sylvain, C. D'Incan, V. Vidal, and Y.-J. Bignon. Towards meaningful rules between genes from gene expression data. In *poster, MGED*, Aix en Provence, France, 2003.

[14] M. Agier and J.-M. Petit. Defining, mining and reasoning on rules in tabular data. In *Proceedings of the* 2^{nd} *Franco-Japanese Workshop on Information Search, Integration and Personalization*, Lyon, France, 2005.

[15] M. Agier and J.-M. Petit. A new and useful syntactic restriction on rule semantics for tabular data. In *Proceedings of the* 21^{th} *Bases de Données Avancées*, pages 135–150, Saint-Malo, France, 2005.

[16] M. Agier and J.-M. Petit. Notion de sémantiques bien-formées pour les règles. In *Proceedings of the Conférence d'Extraction et de Gestion des Connaissances*, volume 1, pages 19–30, Paris, France, 2005. Cépaduès Editions.

[17] M. Agier, J.-M. Petit, V. Chabaud, C. Pradeyrol, Y.-J. Bignon, and V. Vidal. Vers différents types de règles pour les données d'expression de gènes-Application à des données de tumeurs mammaires. In *Proceedings of the* 22^{th} *Congrès INFORSID*, pages 351–367, Biarritz, France, 2004.

[18] M. Agier, J.-M. Petit, and E. Suzuki. Towards ad-hoc rule semantics for gene expression data. In *Proceedings of the International Symposium on Methodologies for Intelligent Systems*, pages 494–503, Saratoga Springs, New-York, USA, 2005. Springer-Verlag.

[19] M. Agier, J.-M. Petit, and E. Suzuki. Unifying framework for rule semantics : Application to gene expression data. *Fundamenta Informaticae, Special issue on "Best Papers from ISMIS 2005"*, à paraître, 2006.

[20] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 207–216, Washington DC, USA, 1993. ACM Press.

[21] T. Akutsu, S. Miyano, and S. Kuhara. Algorithms for identifying boolean networks and related biological networks based on matrix multiplication and fingerprint function. *Journal of Computational Biology*, 7 :331–343, 2000.

[22] O. Alter, P. Brown, and D. Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences*, 97(18) :10101–10106, 2000.

[23] W. Armstrong. Dependency structures of data base relationships. In *Proceedings of the IFIP Congress*, pages 580–583, 1974.

[24] S. Attoor, E. Dougherty, Y. Chen, M. Bittner, and J. Trent. Which is better for cdna-microarray-based classification : ratios or direct intensities. *Bioinformatics*, 20(16) :2513–2520, 2004.

[25] Z. Bar-Joseph, G. K. Gerber, T. I. Lee, N. J. Rinaldi, J. Y. Yoo, F. Robert, D. B. Gordon, E. Fraenkel, T. S. Jaakkola, R. A. Young, and D. K. Gifford. Computational discovery of gene modules and regulatory networks. *Nature Biotechnology*, 21(11) :1337–1342, 2003.

[26] C. Beeri and P. Berstein. Computational problems related to the design of normal form relation schemes. *ACM Transaction on Database Systems*, 4(1) :30–59, 1979.

[27] C. Beeri, M. Dowd, R. Fagin, and R. Statman. On the structure of Armstrong relations for functional dependencies. *Journal of the ACM*, 31(1) :30–46, 1984.

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

[29] M. Black and R. Doerge. Calculation of the minimum number of replicate spots required for detection of significant gene expression fold change in microarray experiments. *Bioinformatics*, 18(12) :1609–1616, 2002.

[30] J. Blanchard, F. Guillet, and H. Briand. A user-driven and quality-oriented visualization for mining association rules. In *Proceedings of the IEEE International Conference on Data Mining*, pages 493–496, 2003.

[31] J. Blanchard, F. Guillet, R. Gras, and H. Briand. Mesurer la qualité des règles et de leurs contraposées avec le taux informationnel tic. In *Proceedings of the Conférence*

d'Extraction et de Gestion des Connaissances, volume 1, pages 287–98, Clermont-Ferrand, France, 2004. Cépaduès Editions.

[32] S. Brin, R. Motwani, J. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 255–264, 1997.

[33] M. Brown, W. Grundy, D. Lin, N. Chistianini, C. Sugnet, T. Furey, M. Ares, and D. Haussler. Knowledge-based analysis of microarray gene expression data using support vector machines. *Proceedings of the National Academy of Sciences*, 97(1) :262–267, 2000.

[34] A. Butte, P. Tamayo, D. Slonim, T. Golub, and I. Kohane. Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proceedings of the National Academy of Sciences*, 97 :12182–12186, 2000.

[35] W. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74 :829–836, 1979.

[36] G. Cong, A. Tung, X. Xu, F. Pan, and J. Yang. Farmer : Finding interesting rule groups in microarray datasets. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 143–154, 2004.

[37] C. Creighton and S. Hanash. Mining gene expression databases for association rules. *Bioinformatics*, 19 :79–86, 2003.

[38] H. de Jong and D. Ropers. Qualitative approaches towards the analysis of genetic regulatory networks. *System Modeling in Cellular Biology : From Concepts to Nuts and Bolts*, pages 125–148, 2006.

[39] J. Demetrovics and V. Thi. Some remarks on generating Armstrong and inferring functional dependencies relation. *Acta Cybernetica*, 12(2) :167–180, 1995.

[40] P. D'Haeseleer, S. Liang, and R. Somogyi. Gene expression data analysis and modelling. In *Proceedings of the Pacific Symposium on Biocomputing*, Hawai, 1999.

[41] S. Dudoit, Y. H. Yang, M. J. Callow, and T. P. Speed. Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments. Technical report, 578, August 2000.

[42] M. Eisen, P. Spellman, P. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25) :14863–14868, 1998.

[43] T. Eiter, G. Gottlob, and K. Makino. New results on monotone dualization and generating hypergraph transversals. In *Proceedings of the* 34^{th} *ACM Symposium on Theory Of Computing*, pages 14 – 22, Montreal, Quebec, Canada, 2002. ACM Press.

[44] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17(3) :37–54, 1996.

[45] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using bayesian networks to analyze expression data. *Journal of Computational Biology*, 7 :601–620, 2000.

[46] B. Ganter and R. Wille. *Formal Concept Analysis*. Springer-Verlag, 1999.

[47] Y. Ge, S. Dudoit, and T. Speed. Resampling-based multiple testing for microarray data analysis. *Technical Report 633, Department of Statistics, University of California, Berkeley*, 2003.

[48] R. Gentleman, V. Carey, W. Huber, R. Irizarry, and S. Dudoit, editors. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Statistics for Biology and Health. Springer, 2005.

[49] G. Golfier, M. T. Dang, L. Dauphinot, E. Graison, J. Rossier, and M.-C. Potier. Varan : a web server for variability analysis of dna microarray experiments. *Bioinformatics*, 20(10) :1641–1643, 2004.

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

[51] G. Gottlob and L. Libkin. Investigations on Armstrong relations, dependency inference, and excluded functional dependencies. *Acta Cybernetica*, 9(4) :385–402, 1990.

[52] J. Gudas, R. Klein, M. Oka, and K. Cowan. Posttranscriptional regulation of the c-myb proto-oncogene in estrogen receptor-positive breast cancer cells. *Clinical Cancer Research*, 1(2) :235–43, 1995.

[53] M. Guerin, Z.-M. Sheng, N. Andrieu, and G. Riou. Strong association between c-myb and oestrogen-receptor expression in human breast cancer. *Oncogene*, 5(1) :131–135, 1990.

[54] J.-L. Guigues and V. Duquenne. Familles minimales d'implications informatives résultant d'un tableau de données binaires. *Math. Sci. Humaines*, 24(95) :5–18, 1986.

[55] M. A. Hall. Correlation-based feature selection for discrete and numeric class machine learning. In *Proceedings of the* 17^{th} *International Conference on Machine Learning*, pages 359–366. Morgan Kaufmann, San Francisco, CA, 2000.

[56] J. Han, Y. Fu, W. Wang, J. Chiang, W. Gong, K. Koperski, D. Li, Y. Lu, A. Rajan, N. Stefanovic, B. Xia, and O. R. Zaïane. Dbminer : A system for mining knowledge in large relational databases. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 250–255, 1996.

[57] A. Hartemink. Reverse engineering gene regulatory networks. *Nature Biotechnology*, 23(5) :554–555, 2005.

[58] P. Haverty, M. Frith, and Z. Weng. Carrie web service : automated transcriptional regulatory network inference and interactive analysis. *Nucleic Acids Research*, 32(Web-Server-Issue) :213–216, 2004.

[59] J. Herrero, F. Al-Shahrour, R. Díaz-Uriarte, Á. Mateos, J. M. Vaquerizas, J. Santoyo, and J. Dopazo. Gepas : a web-based resource for microarray gene expression data analysis. *Nucleic Acids Research*, 31(13) :3461–3467, 2003.

[60] S. Holm. A simple sequentially rejective bonferroni test procedure. *Scandinavian Journal of Statistics*, 6 :65–70, 1979.

[61] Y. Huhtala, J. Kärkkäinen, P. Porkka, and H. Toivonen. Efficient discovery of functional and approximate dependencies using partitions. In *Proceedings of the 14^{th} IEEE International Conference on Data Engineering*, pages 392–401, 1998.

[62] D. Husmeier. Reverse engineering of genetic networks with bayesian networks. *Biochemical Society Transactions*, 31 :1516–1518, 2003.

[63] A. Icev, C. Ruiz, and E. Ryder. Distance-enhanced association rules for gene expression. In *Proceedings of the Workshop on Data Mining in Bioinformatics BIOKDD*, Washington DC, USA, 2003.

[64] T. Ideker, V. Thorsson, and R. Karp. Discovery of regulatory interactions through perturbation : Inference and experimental design. In *Proceedings of the Pacific Symposium on Biocomputing*, volume 5, pages 302–313, 2000.

[65] S.-H. Jung, H. Bang, and S. Young. Sample size calculation for multiple testing in microarray data analysis. *Biostatistics*, 6(1) :157–169, 2005.

[66] J. Kivinen and H. Mannila. Approximate inference of functional dependencies from relations. *Theoritical Computer Science*, 149(1) :129–149, 1995.

[67] T. Kohonen. *Self-Organizing Maps*. Springer-Verlag, 1997.

[68] D. Komura, H. Nakamura1, S. Tsutsumi, H. Aburatani, and S. Ihara. Multidimensional support vector machines for visualization of gene expression data. *Bioinformatics*, 21(4) :439–444, 2005.

[69] M. Kryszkiewicz. Concise representations of association rules. In *Proceedings of the ESF Exploratory Workshop on Pattern Detection and Discovery*, pages 92–109, London, UK, 2002. Springer-Verlag.

[70] K. E. Lee, N. Sha, E. R. Dougherty, M. Vannucci, and B. K. Mallick. Gene selection : a bayesian variable selection approach. *Bioinformatics*, 19(1) :90–97, 2003.

[71] M.-L. Lee, F. Kuo, G. Whitmorei, and J. Sklar. Importance of replication in microarray gene expression studies : Statistical methods and evidence from repetitive cdna hybridizations. *Proceedings of the National Academy of Sciences*, 97(18) :9834–9839, 2000.

[72] S. Liang, S. Fuhrman, and R. Somogyi. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. In *Proceedings of the Pacific Symposium on Biocomputing*, pages 18–29, 1998.

[73] S. Lopes, J.-M. Petit, and L. Lakhal. Functional and approximate dependencies mining : Databases and FCA point of view. *Journal of Experimental and Theoretical Artificial Intelligence*, 14(2/3) :93–114, 2002.

[74] V. Luong. The representative basis for association rules. In *Proceedings of the IEEE International Conference on Data Mining*, pages 639–640, 2001.

[75] D. Maier. Minimum covers in the relational database model. *Journal of the ACM*, 27(4) :664–674, 1980.

[76] H. Mannila and K.-J. Räihä. Algorithms for inferring functional dependencies from relations. *Data and Knowledge Engineering*, 12(1) :83–99, 1994.

[77] J. Mary, G. Mercier, J.-P. Comet, A. Cornuéjols, C. Froidevaux, and M. Dutreix. *Informatique pour l'analyse du transcriptome*, chapter Utilisation d'une méthode d'estimation d'attributs pour l'analyse du transcriptome de cellules de levures exposées à de faibles doses de radiation, pages 189–205. Hermès, 2004.

[78] N. L. Meur, G. Lamirault, A. Bihouee, M. Steenman, H. Bedrine-Ferran, R. Teusan, G. RamsteinAnd, and J. Leger. A dynamic, web-accessible resource to process raw microarray scan data into consolidated gene expression values : importance of replication. *Nucleic Acids Research*, 32(18) :5349–58, 2004.

[79] G. Piatetsky-Shapiro. Discovery, analysis, and presentation of strong rules. In *Knowledge Discovery in Databases*, pages 229–248. AAAI/MIT Press, 1991.

[80] G. Piatetsky-Shapiro and P. Tamayo. Microarray data mining : Facing the challenges. In *SIGKDD Explorations, Special Issue on Microarray Data Mining*, 2003.

[81] J. Quackenbush. Microarray data normalization and transformation. *Nature Genetics*, 32 :496–501, 2002.

[82] R. Rakotomalala. Tanagra : un logiciel gratuit pour l'enseignement et la recherche. In *Proceedings of the Conférence d'Extraction et de Gestion des Connaissances*, volume 2, pages 697–702, 2005.

[83] A. Saeed, V. Sharov, J. White, J. Li, W. Liang, N. Bhagabati, J. Braisted, M. Klapa, T. Currier, M. Thiagarajan, A. Sturn, M. Snuffin, A. Rezantsev, D. Popov, A. Ryltsov, E. Kostukovich, I. Borisovsky, Z. Liu, A. Vinsavich, V. Trush, and J. Quackenbush. TM4 : a free, open-source system for microarray data management and analysis. *Biotechniques*, 34(2) :374–78, 2003.

[84] I. Shmulevich, E. Dougherty, S. Kim, and W. Zhang. Probabilistic boolean networks : A rule-based uncertainty model for gene regulatory networks. *Bioinformatics*, 18(2) :261–274, 2002.

[85] A. Silvescu and V. Honavar. Temporal boolean network models of genetic networks and their inference from gene expression time series. *Complex Systems*, 13 :54–70, 2001.

[86] L. Soinov, M. Krestyaninova, and A. Brazma. Towards reconstruction of gene networks from expression data by supervised learning. *Genome Biology*, 4 :R6.1–R6.10, 2003.

[87] P. Soularue and X. Gidrol. Puces à adn. *Techniques de l'Ingénieur*, 6 :1–10, 2002.

[88] F. Spyratos, C. Andrieu, D. Vidaud, M. Briffod, M. Vidaud, R. Lidereau, and I. Bieche. Ccnd1 mrna overexpression is highly related to estrogen receptor positivity but not to proliferative markers in primary breast cancer. *International Journal of Biological Markers*, 15(3) :210–214, 2000.

[89] G. Stolovitzky. Gene selection in microarray data : the elephant, the blind men and our algorithms. *Current Opinion in Structural Biology*, 13(3) :370–376, June 2003.

[90] E. Suzuki and Y. Kodratoff. Discovery of surprising exception rules based on intensity of implication. In *Proceedings of the* 2^{nd} *European Conference on Principles of Knowledge Discovery in Databases*, pages 10–18, 1998.

[91] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhudagger, S. Kitareewandagger, E. Dmitrovskydagger, E. Lander, and T. Golub. Interpreting patterns of gene expression with self-organizing maps : Methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences*, 96(6) :2907–2912, 1999.

[92] P.-N. Tan, V. Kumar, and J. Srivastava. Selecting the right objective measure for association analysis. *Information Systems*, 29(4) :293–313, 2004.

[93] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99(10) :6567–6572, 2002.

[94] J. Ullman. *Principles of Database Systems*. Computer Science Press, second edition, 1982.

[95] L. van't Veer, H. Dai, M. Vijver, Y. He, A. Hart, M. Mao, H. Peterse, K. Kooy, M. Marton, A. Witteveen, G. Schreiber, R. Kerkhoven, C. Roberts, P. Linsley, R. Bernards, and S. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871) :530–536, 2002.

[96] V. Vinciotti, R. Khanin, D. D'Alimonte, X. Liu, N. Cattini, G. Hotchkiss, G. Bucca, O. de Jesus, J. Rasaiyaah, C. P. Smith, P. Kellam, and E. Wit. An experimental evaluation of a loop versus a reference design for two-channel microarrays. *Bioinformatics*, 21(4) :492–501, 2005.

[97] P. Westfall and S. Young. Resampling-based multiple testing : examples and methods for p-value adjustment. *New York : Wiley*, 1999.

[98] J. M. Wettenhall and G. K. Smyth. limmagui : A graphical user interface for linear modeling of microarray data. *Bioinformatics*, 20(18) :3705–3706, 2004.

[99] I. Witten and E. Frank. *Data Mining : Practical machine learning tools and techniques*. 2nd Edition, Morgan Kaufmann, 2005.

[100] P. C. Wong, P. Whitney, and J. Thomas. Visualizing association rules for text mining. In *Proceedings of the IEEE Symposium on Information Visualization*, pages 120–123, 1999.

[101] X. Wu, Y. Ye, and L. Zhang. Graphical modeling based gene interaction analysis for microarray data. *SIGKDD Explorations*, 5(2) :91–100, 2003.

[102] I. V. Yang, E. Chen, J. P. Hasseman, W. Liang, B. C. Frank, S. Wang, V. Sharov, A. I. Saeed, J. White, J. Li, N. H. Lee, T. J. Yeatman, and J. Quackenbush. Within the fold : assessing differential expression measures and reproducibility in microarray assays. *Genome Biology*, 3 :research00, 2002.

[103] Y. Yang, S. Dudoit, D. Lin, V. Peng, J. Ngai, and T. Speed. Normalization for cdna microarray data : A robust composite method adressing single and multiple slide systematic variation. *Nucleic Acids Research*, 30(4) :e15.1–e15.10, 2002.

[104] M. Zaki. Generating non-redundant association rules. In *Proceedings of the* 6^{th} *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 34–43. ACM Press, 2000.

[105] A. Zien, J. Fluck, R. Zimmer, and T. Lengauer. Microarrays : How many do you need ? *Journal of Computational Biology*, 10(3-4) :653–667, 2003.

TITRE DE LA THÈSE EN FRANÇAIS :

De l'analyse de données d'expression à la reconstruction de réseaux de gènes

RESUMÉ DE LA THÈSE EN FRANÇAIS

Le premier aspect de ce travail de thèse concerne le pré-traitement et l'analyse de données d'expression dans le cadre de deux principaux projets dont l'objectif global est d'améliorer le diagnostic et le pronostic du cancer du sein et de mieux cibler les traitements. Un processus de pré-traitement des données d'expression optimisé a été mis en place et plusieurs analyses ont été réalisées et ont permis d'identifier des gènes permettant de mettre en évidence un profil d'expression caractéristique du statut ganglionnaire des patientes et de la réponse thérapeutique à un traitement chimiothérapeutique particulier, le docétaxel.
Une nouvelle technique de reconstruction de réseaux de gènes basée sur la notion de règles entre gènes a ensuite été proposée, l'idée étant d'offrir aux biologistes la possibilité de choisir parmi plusieurs sémantiques, le sens des règles qu'ils souhaitent générer. L'originalité de ce travail est de proposer un cadre global pouvant inclure un grand nombre de sémantiques pour les règles et d'utiliser des méthodes identiques de génération, de post-traitement et de visualisation pour toutes les sémantiques proposées. La notion de sémantiques bien-formées i.e. pour lesquelles les axiomes d'Armstrong sont justes et complets, est introduite. Un résultat est également donné permettant de savoir simplement si une sémantique est ou non bien-formée. Une visualisation des règles sous forme de réseaux globaux i.e. incluant plusieurs sémantiques est proposée. Finalement, cette approche a été développée sous forme d'un module appelé RG intégré à un logiciel d'analyse de données d'expression de gènes, le logiciel MeV de TIGR.

RESUMÉ DE LA THÈSE EN ANGLAIS

The first aspect of this work concerns the pre-processing and analysis of gene expression data in the context of two main projects whose overall aim is to improve the diagnosis and the prognosis of breast cancer and better target treatments. A optimized pre-processing workflow of expression data was established and several analysis have been conducted and have identified genes allowing to point out a characteristic expression profile of nodal status of patients and therapeutic response to a particular chemotherapy, the docetaxel.
A new technique for reconstruction of gene networks based on the notion of rules between genes was then proposed, the idea is to offer biologists the possibility to choose among several semantic meanings, the rules they want to generate. The originality of this work is to propose a comprehensive framework that could include a large number of semantic for rules and use identical methods of generation, post-processing and visualization for all proposed semantics. The notion of well-formed semantics i.e. for which Armstrong's axioms are sound and complete, is introduced. A result is also given to know simply whether or not a semantic is well-formed. A visualization of the rules in the form of global networks i.e. including several semantic is proposed. Finally, this approach has been developed as a module called RG integrated into a software dedicated to gene expression data analysis, the TIGR MeV software.

TITRE DE LA THÈSE EN ANGLAIS :

From gene expression data analysis to gene network reconstruction

PROPOSITION DE MOTS-CLÉS :

1 DATA MINING
2 DECOUVERTE DE REGLES
3 AXIOMES D'ARMSTRONG
4 RESEAUX DE GENES
5 DONNES D'EXPRESSION
6 PRE-TRAITEMENT
7 CANCER DU SEIN
8