Les réseaux bayésiens : classification et recherche de réseaux locaux en cancérologie

Emmanuel Prestat

N^o d'ordre : 65-2010 Année 2010

Thèse

Présentée

devant l'Université Claude Bernard - Lyon 1

pour l'obtention

du Diplôme de Doctorat
(arrêté du 7 août 2006)

et soutenue publiquement le
25 mai 2010

par

Emmanuel Prestat

Les Réseaux Bayésiens : classification et recherche de réseaux locaux en cancérologie

---

Directeur de thèse : Christian Gautier

Jury :	Pr Alexandre Aussem, Université Lyon 1,	Examinateur
	Pr Pascale Cohen, Université Lyon 1,	Examinatrice
	Dr Philippe Dessen, Université Paris-XI,	Rapporteur
	Pr Christian Gautier, Université Lyon 1,	Directeur
	Pr Philippe Leray, Université de Nantes,	Rapporteur
	Dr Marie-France Sagot, Université Lyon 1,	Présidente

UNIVERSITÉ CLAUDE BERNARD-LYON 1

Président de l'Université	**M. le Professeur L. COLLET**
Vice-Président du Conseil Scientifique	M. le Professeur J-F. MORNEX
Vice-Président du Conseil d'Administration	M. le Professeur G. ANNAT
Vice-Président du Conseil des Etudes et de la Vie Universitaire	M. le Professeur D. SIMON
Secrétaire Général	M. G. GAY

COMPOSANTES SANTÉ

UFR de Médecine Lyon-Est – Claude Bernard	Directeur : M. le Professeur J. ETIENNE
UFR de Médecine Lyon-Sud – Charles Mérieux	Directeur : M. le Professeur F-N. GILLY
UFR d'Ontologie	Directeur : M. le professeur D. BOURGEOIS
Institut des Sciences Pharmaceutiques et Biologiques	Directeur : M. le Professeur F. LOCHER
Institut des Sciences et Techniques de Réadaptation	Directeur : M. le Professeur Y. MATILLON
Département de Formation et Centre de Recherche en Biologie Humaine	Directeur : M. le Professeur P. FARGE

COMPOSANTES SCIENCES ET TECHNOLOGIE

Faculté des Sciences et Technologies	Directeur : M. le Professeur F. GIERES
UFR Sciences et Techniques des Activités Physiques et Sportives	Directeur : M. C. Collignon
Observatoire de Lyon	Directeur : M. B. Guiderdoni
Institut des Sciences et des Techniques de l'Ingénieur de Lyon	Directeur : M. le Professeur J. LIETO
Institut Universitaire de Technologie A	Directeur : M. le Professeur C. COULET
Institut Universitaire de Technologie B	Directeur : M. le Professeur R. LAMARTINE
Institut de Science Financière et d'Assurances	Directeur : M. le Professeur J-C. AUGROS

Remerciements

Je tiens en premier lieu à remercier Christian, mon directeur de thèse, de m'avoir proposé de travailler sur ce sujet innovant en bioinformatique. Tu es une preuve que l'accession à de nombreuses et importantes responsabilités (direction du laboratoire, de la plateforme Rhône-Alpes de Bio-Informatique, de l'UFR puis du département de Biologie de Lyon, présidence de la section Ecologie du CNRS, membre du CA de l'Université Lyon 1, et j'en oublie sûrement) n'est pas forcément motivée par une ambition personnelle. Avec ces quelques années dans ton entourage, j'ai pu constater que ton moteur (pourtant très énergivore) était avant tout l'envie de promouvoir la biologie et la bioinformatique tant au niveau de la recherche que de l'enseignement, ce qui force mon admiration.

Je voudrais aussi remercier très chaudement Marie-France, qui m'a accueilli dans son équipe de recherche. Tu réussis l'alchimie de former un groupe performant, cosmopolite (nous ouvrant, nous, « petits français », à d'autres cultures et à la recherche collaborative au sens premier) et ce, avec une bienveillance sans limite. L'affirmation « Marie-France est la grande spécialiste des réseaux » contient paraît-il un double sens. Tu es aussi spécialisée en combinatoire, alors que tu ne sais pas compter ton temps. Tu carburent aux relations humaines, ce qui semble sacrément efficace ! Tu n'es pas seulement une chercheuse renommée internationalement, tu es également le pense-bête de Christian : je te remercie aussi pour ça.

Je remercie les membres (restants) de mon jury : Philippe Leray et Philippe Dessen, pour avoir accepté de rapporter et évaluer mon travail ; ainsi que mes examinateurs et collaborateurs : Pascale Cohen et Alexandre Aussem pour leurs expertises, avec qui ce fut de surcroit un plaisir de travailler, en espérant continuer dans de futurs projets.

J'ai eu l'occasion pendant cette thèse d'enseigner à l'Université Lyon 1 en Licence et en Master, ainsi qu'à l'INSA de Lyon. Je remercie les responsables de modules Dominique Mouchiroud, Hubert Charles et Catherine Cerutti de m'avoir fait confiance. J'ai particulièrement aimé cette activité qui à mon sens fait partie intégrante du métier de chercheur.

Je remercie mon responsable actuel à l'Ecole Centrale de Lyon, Tim Vogel, de me faire confiance pour des sujets de recherche passionnants, et accessoirement d'avoir accepté de m'embaucher en sachant que la rédaction de mon mémoire de thèse n'était pas terminée !

Je remercie les secrétaires et assistantes du LBBE : Nathalie, Gaëlle, Agnès, Isabelle, Misou, le labo a de la chance de vous avoir (ou de vous avoir eu pour Agnès et Misou). Je remercie les infomen : Bruno, Stéphane et Lionel. Vous maniez le bourne shell et le

gigaflop aussi bien que l'humour : ces deux aspects valent des remerciements.

Je remercie tous les baobabs que j'ai côtoyés durant ces années, tout d'abord mes compagnons de bureau : Ludo, Vicente, Claire, Vincent L., Leo et les Patricia. J'ai adoré travagoler en votre compagnie, vous êtes devenus des amis. Je remercie Elise, Janisse, Amélie qui ont également beaucoup compté pour moi ces années. Je vais aussi garder de bons souvenirs de l'humour et la finesse de Paulo, qui n'a rien à envier aux ENArques quant à sa connaissance de l'administration française ! Je remercie Marilia pour sa personnalité unique (des algorithmiciennes qui jouent à l'OL, ça en dit déjà long) et pour tes cours de Portugais « do Brasiou » animés ! Je remercie Pierre, le Baffie de la recherche, tu dégaines plus vite que ton ombre, tu vannes tellement que t'as fini par trouver un poste en Bretagne ! Je remercie Franck et Vincent, des fortes personnalités venues au renfort du groupe. Je remercie aussi les baobabs-connectés Emilie, Sophie, Perrine, Thibaut, Céline, Lauranne et Aurélie : j'ai même cru à un moment qu'ils étaient dans notre équipe...

Je remercie Delphine, qui m'a supporté un certain temps sur les bancs de la fac... Je remercie Vincent D., Gab et Raquel, pour les bons moments passés ensemble. Je remercie Simon et Navrate : leur humour n'ont d'égal que leur grain de folie, je suis très client !

Je remercie aussi mes amis hors labo : Claire, Marc, Guigui, Juju, Véro, Nico, Nive, Tony, Pez, Boubou, Chal, Fanny R., Florian P., Fanny B., Rodolphe, Florian C., Olive, Caro, Thomas, Mat, Cla, Lool et Alex. Je les conseille à tout le monde, ce sont de très bons amis !

Je remercie Stan de m'avoir cité dans ses remerciements.

Je remercie ma maman, qui a élevé seule deux garçons. Tu ne cesses de m'impressionner par tant d'énergie, dont bénéficient tes proches, tes patients, et tes innombrables projets.

Je remercie mon frère Sylvain, de mettre autant d'énergie dans la recherche à ne plus en dépenser : tu m'épateras toujours par tes concrétisations. Je remercie Delphine pour avoir réussi à te canaliser.

Enfin je remercie celle qui aura été successivement une copine, ma copine, ma coturne, ma colocataire, ma copropriétaire, ma co-PACS, ma co-parent et bientôt ma conjointe. Merci à toi, Elo, de m'avoir toujours soutenu. Tu es mon capital confiance. J'ai le bonheur d'avancer dans la vie avec toi et nos filles, Lise et Diane : ma première ambition est que l'on continue sur cette lancée !

Je ne peux terminer sans exprimer une pensée à mon papa, disparu tellement jeune...

Table des matières

Introduction

La biologie a connu d'importants évènements depuis le début du XX^{e} siècle. Un bouleversement particulièrement important de la discipline est né de la redécouverte des lois de Mendel par *de Vries*, *Correns* et *Tschermak* qu'ils énoncent en 1900, 35 ans après le moine autrichien (ses archives avaient été brûlées par son successeur peu après sa mort). Ce qu'avait compris Gregor Mendel en croisant des petits pois de caractères différents (peau lisse ou fripée par exemple), c'est que la peau des petits pois était dépendante de la combinaison d'objets biologiques pouvant exister en plusieurs versions. Chacun d'eux est transmis par un parent, par le biais des cellules sexuelles qui ont la particularité de n'avoir qu'une copie de l'ensemble de ces objets. Il remarqua que certaines versions de ces objets pouvaient cacher l'effet visible de son homologue chez un individu. Il en déduit alors un ensemble de règles associants des traits de caractère à des combinaisons de ces versions, en fonction des générations issues de croisements contrôlés. Ces « lois » constituèrent un grand pas dans la réponse à la question : « comment s'explique l'hérédité ? ».

Peu de temps après, *Thomas Morgan* avec pour modèle d'étude la drosophile, avait compris que ce « support de caractères » pouvait connaître des changements, des mutations, expliquant les différentes versions de ces éléments biologiques. Cette découverte permettait d'expliquer une partie des différences morphologiques infra-spécifiques voire populationelles. De plus, cela ajoutait la dimension « évolution » à la dimension « hérédité » que l'on avait déjà attribuée à ces objets. En 1902, l'utilisation du microscope pour étudier le noyau de la cellule révéla des éléments particuliers, appelés les chromosomes, dont le lien avec les objets de Mendel a été constaté. Ces derniers ont été nommés « gène » (*i.e.* engendre, donne naissance) en 1909 par le danois *Wilhem Johannsen*, soulignant ainsi leur rôle dans l'hérédité.

Ces découvertes successives qui sont les prémices de ce qu'on appellera plus tard la génétique constituent un évènement de l'histoire de la biologie. Enfin on commençait à comprendre certaines différences entre individus, la manière dont elles étaient transmises de générations en générations, même si les explications mécanistiques n'ad-

viendront que plus tard. Il faudra attendre en effet la découverte de la composition des chromosomes en 1944 donc celle de l'ADN (ou « acide desoxyribonucléique »), puis de la structure de cette molécule par [Watson et Crick, 1953] pour que les découvertes précédentes deviennent les piliers de la génétique en tant que discipline. On la nomme aussi : « biologie moléculaire » dont le champ d'action est moins générique qu'il n'y paraît (l'étude des protéines n'en fait pas partie). Il manquait encore un fondement à cette discipline : le lien entre l'ADN (le support de l'information génétique) et les protéiness dont on connaissait déjà leur rôle fonctionnel. La réponse arrive une dizaine d'années plus tard grâce aux travaux combinés de *Nirenberg*, *Holley* et *Khorana* sur le code génétique, résolvant ainsi la correspondance entre la composition génétique et la composition protéique, expression fonctionnelle du gène.

Enfin, le socle était complété. Non seulement on allait beaucoup plus loin dans la compréhension de la biologie, mais en plus on manipulait du matériel génétique : c'est l'ère du génie génétique. Les chercheurs étaient désormais capables de couper des fragments d'ADN, de les coller, de les transférer dans d'autres cellules, d'autres organismes, de les cloner. Le développement de méthodes de séquençage automatique de l'ADN à la fin des années 70 [Sanger *et al.*, 1977, Maxam et Gilbert, 1977] ont permis d'aller encore plus loin, jusqu'à aujourd'hui, où dominent le haut-débit, de séquences complètes de génomes, de recherches exhaustives de polymorphisme génétique, *etc.*

En effet les outils récents élaborés en biologie expérimentale suscitent un enthousiasme des biologistes, pharmacologues et médecins en premier lieu. La composante principale de ces outils est l'analyse d'objets biologiques (acides nucléiques et protéiques) en grandes quantités et de manière extrêmement véloce : dépassant de plusieurs ordres de grandeurs ce que l'on était capable de produire depuis le séquençage de Sanger ou Maxam. Ces innovations technologiques, qui vont souvent de paire avec le séquençage des génomes, permettent d'en apprendre de plus en plus sur le programme (le génome) et son exécution (les mécanismes cellulaires). En effet, sur le plan des génomes, le séquençage de l'ADN a considérablement évolué ces dernières années, drainant derrière lui autant de nouvelles problématiques que de résultats.

Parmi ces nouvelles problématiques, il existe celle de l'analyse des résultats d'expériences, de façon à ce qu'ils soient manipulables et compréhensibles. Les génomes sont des objets assez complexes de part leur taille et leur variabilité. Si on veut les analyser pour mieux comprendre leur origine (évolution), leur fonctionnement, le lien avec les traits d'histoire de vie, les symptômes d'une maladie, les sensibilités différentielles à des mutations cancérigènes, à des traitements, *etc.*, on doit adapter, revoir, innover

du point de vue de la méthodologie. Les méthodes en question prennent place dans ce que l'on appelle la bioinformatique. Plus précisément, la masse de données qui est générée depuis ce début de millénaire implique une meilleure gestion des bases de données (stockage, rapidité, accès), d'améliorer les algorithmes de recherche de similarité entre séquences, les méthodes statistiques pour analyser le caractère singulier ou non d'une séquence, d'un motif, dans des conditions particulières, de proposer de nouvelles formes de représentations des connaissances, de se servir de ces connaissances de façon prospective, d'imaginer des *scenarii*.

En se plaçant du point de vue de la recherche médicale, on se rend compte que les progrès technologiques associés aux connaissances liées aux génomes ont joué un grand rôle. En effet, les promesses d'avancées dans le diagnostic et les thérapies grâce à ce que l'on peut obtenir de ces nouvelles données est immense ; par exemple, en ce qui concerne les cancers qui depuis 2005 sont la première cause de mortalité due à une maladie en France [Aouba *et al.*, 2007]. Beaucoup de cancers sont caractérisés par des mutations génétiques induisant un dérèglement dans le développement et la croissance des cellules dont l'ADN est muté. On comprend alors les espoirs que suscite la biologie moléculaire dans l'essor des recherches contre le cancer. S'ajoutent à cela les phénomènes qui font varier la mutabilité des génomes qui peuvent aussi être d'origine génétique. En conséquences, beaucoup d'investissements dans l'avancement des connaissances sur le génome humain proviennent des problématiques de médecine et en majeur partie de l'oncologie (on pourra aussi citer d'autres maladies comme les diabètes et les maladies cardio-vasculaires).

Les cancers sont un bon exemple pour illustrer à la fois tout ce que l'on peut espérer de la connaissance des génomes et de leur dynamique, mais aussi de leur complexité et de l'immensité des efforts à consentir pour aller plus loin : en effet, il est possible d'identifier des gènes comme étant reliés à certains types de cancer, mais de là à fournir des explications mécanistiques associées, cela peut s'avérer très complexe. Prenons le cas des gènes *BRCA1* [Miki *et al.*, 1994] et *BRCA2* [Wooster *et al.*, 1995] dont les mutations sont connues pour être liées aux cancers du sein et ovarien. Les observations et les statistiques sont là : d'après une méta-analyse [Antoniou *et al.*, 2003] la détection d'une mutation de *BRCA1* ou *BRCA2* chez une personne indique que cette personne aurait environ 65% ou 45% de risque respectivement d'être atteinte d'un cancer du sein ou des ovaires. Sachant que le cancer du sein est le premier cancer en terme de prévalence chez les femmes du Monde occidental, c'est un résultat particulièrement intéressant. Le fait est que depuis la découverte de ces deux gènes il y a quinze ans,

on s'est aussi rendu compte de la difficulté à localiser les mutations cancérigènes dans ces gènes (beaucoup de positions mutantes observées, mais peu sont récurrentes). De plus, les fonctions des protéines correspondantes ne sont toujours pas clairement déterminées : on est conscient qu'elles jouent un rôle dans la réparation de l'ADN, mais on ne sait pas exactement lequel. D'autres informations concernant les domaines protéiques de *BRCA1* et *BRCA2*, ainsi que le réseau d'interaction protéique pour former un complexe fonctionnel avec celles-ci sont encore floues aujourd'hui. Cela complique beaucoup la problématique de départ, où on aurait pu simplement ambitionner de localiser la mutation, proposer une méthode de détection dédiée et rapide de celle-ci, voire développer un médicament ciblé directement à partir de la connaissance de cette séquence, ou de sa participation dans les voies métaboliques. Au contraire, il faut prendre en compte les facteurs responsables de ces mutations ainsi que leur nature (*e.g.* les sites mono-nucléotidiques polymorphes ou « SNPs »), les aspects mécanistiques (protéines associées) impliquant des études à larges échelles pour trouver des profiles d'expression, d'interaction entre protéines, pour finalement élargir l'échelle d'approche de cette problématique. Dans beaucoup d'autres maladies, même lorsque des gènes qui leur sont associés ne sont pas si clairement identifiés, on a bien compris l'existence de composantes génétiques (ne serait-ce que par leur caractère héréditaire) mais aussi de la complexité de celle-ci. C'est pourquoi les études à grande échelle sont devenues systématiques que ce soit au niveau du génome, du transcriptome ou du protéome, par les techniques (entre autres) de séquençage, de biopuces ou de spectrométrie de masse, et qu'on essaye de relier ces informations entre-elles ainsi qu'aux connaissances existantes sur les voies métaboliques. Des développements bioinformatiques ont suivis l'évolution de ces données.

Depuis que [Schena *et al.*, 1995] ont innové avec les biopuces permettant la mesure d'un transcriptome en une expérience, nombre de bioinformaticiens, statisticiens et algorithmiciens se sont attelés à développer des outils permettant d'analyser ces données dont la quantité d'informations (en particulier la dimension « nombre de variables ») a augmenté de façon exponentielle : tandis qu'à cette période on était habitué à analyser simultanément au maximum quelques dizaines de variables, on est rapidement passé à quelques milliers puis maintenant quelques millions. Pourtant, le nombre de répétitions biologiques et expérimentales n'ont elles pas ou peu augmenté. L'accent a été mis sur le traitement statistique pour minimiser des biais d'ordre instrumental (grâce à la mesure ou l'inférence du signal n'ayant pas une origine biologique), ainsi que dans le but de comparer différentes biopuces, quantifier les erreurs liées aux tests multiples, *etc.* Il

existe de nombreuses problématiques justifiant l'utilisation de biopuces pour l'analyse du transcriptome. L'une d'entre-elles est la mise en évidence des profils d'expression (c'est-à-dire des combinaisons d'expression de gènes) associés à des conditions biologiques particulières (par exemple un type tumoral). Dans ce cas, on se place dans des problématiques de classification.

La classification est un problème qui se traite de façon très différente, selon que l'on utilise une base d'exemples contenant des mesures faites en association avec des conditions connues (les classes) ou que l'on a pas accès à ces informations. Dans le premier cas on parlera de classification supervisée (les analyses discriminantes en font partie) et dans le deuxième cas de classification non-supervisée (classification hiérarchique, K-means, *etc.*), l'anglicisme « clusterisation » est alors souvent utilisé. Les deux familles sont très souvent mises en œuvre en fonction du contexte. Si on revient à l'utilisation des données de transcriptome dans un contexte médical, l'une des priorités est le diagnostic. Dans ce cas, il est nécessaire d'évaluer la technologie des « biopuces » en la comparant à d'autres techniques à partir de données cliniques. Plusieurs études basées sur l'analyse utilisant des méthodes de classification supervisées ont dans ce cas montré l'intérêt des biopuces dans ces problématiques. Citons l'expérience menée par [Golub *et al.*, 1999] qui a démontré l'intérêt des puces dans la typologie de leucémies aiguës, ou [van't Veer *et al.*, 2002] qui a mis en avant cette technologie dans le pronostic clinique de patientes atteintes par un cancer du sein.

Ces analyses ont fait date dans le domaine, mais il est apparu que ce n'est pas toujours simple d'obtenir des classifications aussi précises. On note aussi l'aspect « boîte noire » et le manque d'interprétabilité mécanistique associé aux méthodes de classifications statistiques utilisées. Dans ce contexte, ce travail propose d'investir un champ relativement nouveau des statistiques et probabilités basés sur les graphes. On propose dans ce travail d'investiguer ce type de méthodes pour : rapprocher graphiquement des gènes ayant un « comportement » associé (en terme d'expression), utiliser la construction du graphe pour proposer une signature génique, voire de mettre en oeuvre la classification d'une condition biologique observée ou de chercher des interprétations biologiques à partir du modèle visuel reconstruit.

Ce manuscrit est divisé en trois grandes parties. La première introduit le champ de la bioformatique, des concepts de biologie, avec un accent sur l'obtention des données à haut-débit et les réseaux cellulaires. Dans la deuxième partie, on présente quelques modèles mathématiques et statistiques qui sont aussi des modèles graphiques, dans le but de préciser la stratégie de modélisation employée dans ce travail de thèse. Le modèle

choisi, *i.e.* les Réseaux Bayésiens sont ensuite définis et leur utilisation potentielle est détaillée. La troisième partie est celle des résultats : les Réseaux Bayésiens ont été mis en œuvre dans diverses contextes, tous « connectés » à des problématiques de cancer. On propose dans un premier temps une stratégie d'accélération de la recherche d'un graphe global à l'aide de ce modèle, puis on les emploie pour résoudre des problèmes de classifications de tumeurs et patients (à plusieurs niveaux : la classification à proprement parler ou la sélection de gènes), et enfin on montre leur utilité dans la recherche d'un réseau local au voisinage d'un gène d'intérêt.

Première partie

Généralités et contexte biologiques

Chapitre 1

La modélisation en biologie

Sommaire

1.1 La modélisation : généralités

Pour comprendre ce qu'est la modélisation (en biologie), il faut revenir (pour ce qui concerne le Monde occidental) à Georges-Louis Leclerc, dit Buffon. Ce savant du XVIII^e siècle a étudié de nombreux animaux (pour ce qui est de la biologie) et minéraux. Il est le premier à avoir systématisé à grande échelle l'association « observation » et « expérience », en témoigne son œuvre principale : l'« histoire naturelle, générale, et particulière » [Buffon *et al.*, 1749]. En plus du caractère conséquent et la popularité de ces trente-six volumes quasiment comparables à l'encyclopédie de Diderot, il a fait quelques remarques intéressantes. Outre ses recommandations expliquant l'intérêt de l'expérience, de la dissection pour comparer les êtres vivants, il fût un des premiers à faire le lien de la filiation entre l'homme et le singe qu'il qualifiait comme une dégénération de l'homme (certains disent qu'il voulait s'éviter les foudres du clergé). A la lumière des connaissances actuelles, d'abord issues des théories de Lamarck puis de Darwin, cela parait un peu rétrograde comme façon de voir ce lien entre l'homme et le singe, néanmoins, un siècle avant les théories du transformisme et de l'évolution, dans un monde où le créationnisme fait tellement l'unanimité que le mot n'existe pas, c'est un fait notable et pas très connu, même dans le monde scientifique. Bien-sûr, Buffon n'a pas inventé l'expérience, mais il l'a complètement systématisée et intégrée à sa démarche d'étude d'objets naturels.

Cette démarche expérimentale, si on la résume, est constituée d'une phase d'observation, puis d'un raisonnement menant à des hypothèses pouvant expliquer l'observé, et enfin de la phase expérimentale permettant de conforter ou de rejeter ces hypothèses. Le raisonnement passe par une représentation intellectuelle de phénomènes réels. Il est extrêmement rare de pouvoir intégrer toute la complexité du monde réel dans un raisonnement, c'est pourquoi, consciemment ou pas, on simplifie la réalité pour la rendre intelligible. Cela a l'avantage de simplifier aussi l'expérience testant l'hypothèse. En biologie « moderne », l'étape de raisonnement est intégrée dans un processus dit de « modélisation », et notre représentation du réel, que l'on pourra appeler dans notre domaine le naturel, est le modèle. Le modèle issu d'une observation biologique est donc une formalisation et une simplification de phénomènes naturels.

Pour éviter tout *quiproquo*, on ne parle ici que de modèles abstraits, que l'on pourra opposer, dans notre domaine, au modèle biologique qui correspond à un choix d'espèces ou de populations qui satisfont des critères de recherche.

Le modèle permet de confronter ses idées à la réalité, en le testant sur de nouvelles observations biologiques par exemple, ou en simulant des phénomènes naturels à l'aide

de ce modèle. Dans les deux cas, cela constitue une forme de validation du raisonnement.

Le modèle, selon sa forme, a l'avantage de permettre bien plus que de structurer et communiquer un raisonnement. Il est, par essence, aussi modelable. Cette plasticité est utilisée pour tenter d'améliorer les connaissances grâce aux simulations. En effet, il est souvent très facile de changer la valeur des paramètres dans un modèle et d'observer le résultat en simulant des données. On peut donc à la fois juger de celui-ci (et comme il résume la connaissance du système, c'est la connaissance que l'on affine) et l'améliorer car on retiendra les paramètres générant des résultats de simulations les plus réalistes.

Enfin, un modèle peut être utilisé pour établir des prédictions. Si on considère un échantillon comme étant issu d'un système comparable à celui pour lequel le modèle est valide, on pourra estimer des paramètres auxquels on n'a pas toujours accès par des mesures.

1.2 La bioinformatique

La bioinformatique est l'utilisation et le développement d'approches mathématiques ou informatiques pour répondre à une question biologique. Selon les sensibilités des bioinformaticiens, on peut dénombrer un certain nombre de disciplines : génomique comparative, biomathématiques, biostatistiques, protéomique, biologie structurale, biologie combinatoire, biologie systémique ou biologie théorique.

En pratique, cette compartimentation est source de confusions, et cause parfois des rivalités qui ne sont constructives en terme de visibilité ni dans le monde scientifique, ni dans la société civile. Il est donc plus simple de tout regrouper sous le terme « bioinformatique ».

On a tendance à réduire la bioinformatique à son caractère utilitaire : elle l'est, elle développe d'ailleurs des outils devenus indispensables en biologie ; et les bioinformaticiens mettent généralement en avant cette casquette d'« aide à la recherche et aux développements en biologie ». Cependant c'est également un champ de recherche à part entière, et les thématiques telles que la phylogénie, l'analyse de séquences, la recherche de fonctions moléculaires, *etc.* en sont de bons exemples.

1.3 Conclusion

Quelque-soit le champ de la bioinformatique, on utilise et développe des modèles. Cela peut aller du modèle statistique, au modèle mathématique, à des modèles de bases de données. La bioinformatique est une discipline devenue désormais indispensable dans le monde de la biologie : elle enregistre toutes les connaissances biologiques actuelles dans des banques de données mondiales, permet de résoudre des structures 3D de protéines, de classer des malades, de prendre des décisions, d'optimiser, de simuler des expériences, d'étudier la dynamique de systèmes dans le temps et dans l'espace *etc.* Elle permet de répondre, grâce au séquençage, à beaucoup de question sur l'évolution : les connaissances, les mécanismes impliqués, sont de plus en plus nombreux. Les données générées par des expériences biologiques sont tellement massives que la bioinformatique est aujourd'hui obligatoire pour les intégrer, et les rendre intelligibles, bref, pour que les informations deviennent des connaissances.

Chapitre 2

Masses de données biologiques et enjeux thérapeutiques

Sommaire

2.1 Introduction

Depuis la fin des années 90, les méthodes d'analyses en biologie moléculaire ont considérablement évolué. Les projets de séquençage complet de génomes, l'utilisation de puces à ADN, les études quantitatives, épidémiologiques en masse, les procédés de séparation des protéines, des détections d'interactions entre elles ou avec les acides nucléiques, ont changé la donne. Le facteur commun de ces nouveautés est l'exhaustivité de ces méthodes. Il est désormais possible d'« observer » le génome non-seulement dans sa globalité, mais aussi à une résolution de l'ordre du nucléotide. Cela cristallise le passage de la génétique à la génomique. On fait souvent référence à la génétique lorsque l'on étudie un nombre restreint de gènes, dont les protéines correspondantes ont une fonction que l'on peut mettre en évidence de façon isolées ou presque. Au contraire, on parle de génomique quand on prend en compte la globalité du génome, c'est-à-dire l'ensemble des molécules d'ADN (ou d'ARN pour certains virus) appartenant à un organisme vivant. Cela inclut les gènes, mais aussi les régions non transcrites. Lorsqu'on s'intéresse aux gènes, on peut chercher à extraire de ce type d'analyses un réseau génique d'interactions (c'est l'objet de ce travail de thèse). Dans ce mémoire, on s'intéresse à une partie du génome : les gènes protéiques.

Naturellement, s'atteler à la génomique qui suggère une prise en compte d'un nombre de gènes beaucoup plus important que dans la génétique, a pour conséquence l'étude d'une plus grande complexité dès lors que l'on s'intéresse à l'effet de leur comportement combiné, par exemple, sur un phénotype (voir figure 2.1).

Aussi, l'avancement des connaissances dans ce domaine a drainé un ensemble de disciplines vers des points de vue plus larges, correspondants à des angles d'attaque par la globalité que l'on rassemble désormais sous la dénomination « omique ». Le lien avec la génomique est établi soit parce que les gènes y participent (protéomique, voire métabolisme indirectement), soit parce que le génome englobe d'autres points de vue plus spécifiques (polymorphisme transcriptionel, focalisation sur les phases de lecture). Le suffixe « ome » est de plus en plus utilisé, et la tendance est de l'employer dès lors qu'on recherche l'exhaustivité dans une étude de biologie cellulaire ou moléculaire. A titre d'exemples, on a :

- la **génomique** : étude des génomes ;
- la **transcriptomique** : étude de l'ensemble des transcrits dans une condition biologique donnée (cf. paragraphe 2.3) ;
- la **protéomique**, étude de l'ensemble des protéines présentes dans une condition biologique donnée ;

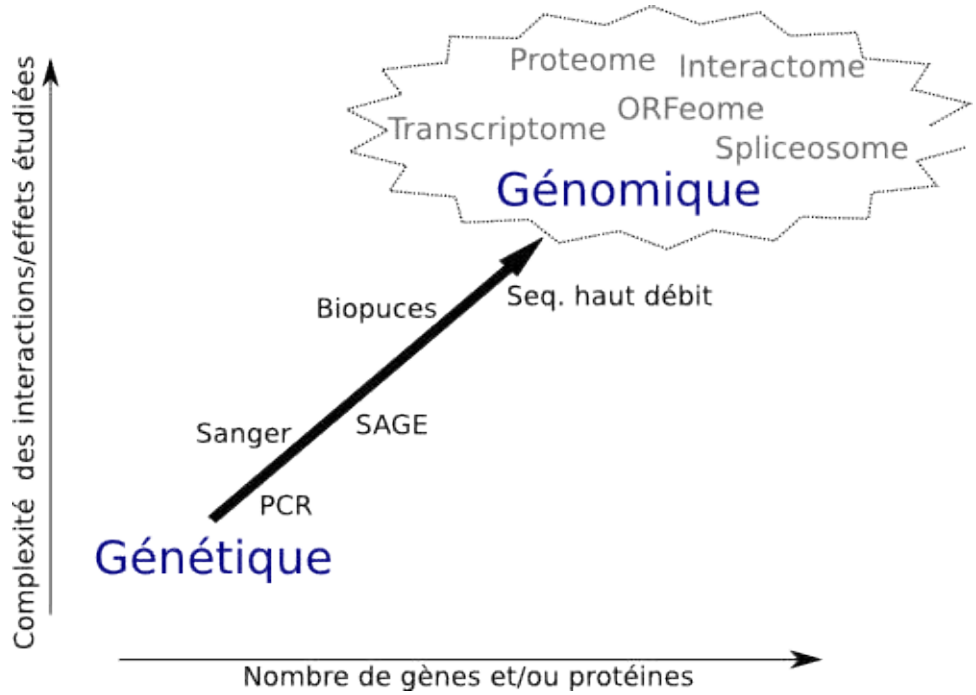


FIGURE 2.1 – **Génétique et génomique** : la flèche représente l'évolution des technologies permettant d'accéder au niveau d'analyse « génomique » et autres disciplines liées.

- l'**interactome**, ensemble des interactions entre protéines ;
- la **métabolomique**, étude de l'ensemble des métabolites dans un organisme vivant ;
- le « **spliceosome** » répertoriant tous les isoformes de protéines issues de transcrits alternatifs, les procaryotes ne disposant pas d'intron, il n'y a donc pas d'épissage et de transcrits alternatifs chez ces organismes ;
- l'**ORFeome** correspondant à l'ensemble des phases ouvertes de lectures ;
- la **métagénomique** : étude du génome rapporté à un échantillon incluant généralement beaucoup d'organismes (échantillon de sol, d'eau naturelle, flore digestive, *etc.*) ;
- le **résistome**, ensemble des gènes ayant une fonction de résistance à une molécule particulière.

L'objectif de ces approches à larges échelles est de capturer des combinaisons biologiques auxquelles on ne peut accéder à l'aide d'approches locales. L'enjeu est le développement de méthodes capable d'intégrer toutes les données à ces fins.

2.2 Les réseaux d'interactions cellulaires

Une manière compréhensible de représenter les connaissances sur les interactions moléculaires dans une cellule est de faire un schéma reliant des entités (une molécule ou un gène) par des traits ou des flèches qui représentent des relations. Cette représentation des connaissances est effectuée par un effort d'abstraction de phénomènes réels destinée à être support de réflexion et d'interprétation. Il s'agit donc de modélisation. Lorsqu'un formalisme est défini pour ces schémas, et que par conséquent on peut les formuler mathématiquement, on appellera ces objets des graphes. On a alors l'avantage de pouvoir recourir à tout l'attirail mathématique pour analyser ou inférer ces graphes, les utiliser afin d'effectuer des simulations, bref : modéliser.

Il est en pratique impossible d'organiser toutes les connaissances cellulaires actuelles dans un graphe. Elles sont nombreuses, pas encore assez étayées et ne sont pas selon leur nature toujours présentables de la même manière. C'est pourquoi on s'intéresse généralement à des « sous-réseaux cellulaires » qui sont liés en réalité, mais que l'on représente (en particulier en modélisation mathématique) de façon séparée.

Historiquement, les premiers réseaux cellulaires que l'on a dessiné sont des voies métaboliques (celle de la glycolyse pour la première, consistant en la dégradation du glucose, impliquant la formation du pyruvate et le relâchement d'énergie sous forme de molécules d'ATP). Puis on s'est intéressé à d'autres formes d'interactions, comme celles qui permettent la transduction du signal. On s'est aussi rendu compte qu'il existait un système de régulation génique que l'on représente par des réseaux de régulation génique. Il existe une terminologie adaptée à l'échelle des interactions étudiées.

Ces « sous-réseaux » sont donc définis par le type de processus étudié : la biosynthèse et dégradation de molécules (métabolisme) la régulation des gènes par eux-mêmes (régulation génique), la transmission moléculaire de l'information (transduction du signal).

En pratique, on adapte le vocabulaire à l'échelle de l'étude considérée. Lorsqu'on s'intéresse à un processus local on parlera plutôt de voie (voie métabolique, voie de signalisation, voie de régulation). Au contraire, si on veut décrire un processus global, on appelle souvent ce graphe d'interactions un réseau.

2.2.1 Les réseaux métaboliques

Un réseau métabolique est une formalisation d'un métabolisme. Il représente l'ensemble des réactions métaboliques dans un organisme. Celles-ci se caractérisent par la transformation de substrats (molécules d'entrée) en produits (molécules de sortie). En général, ces réactions sont catalysées par des molécules tierces, qui aident la réaction à se produire, sans pour autant que ces catalyseurs soient transformés. La plupart de ces catalyseurs sont des protéines particulières, les enzymes, produites majoritairement par l'organisme en question.

Par la suite, les produits de ces réactions peuvent être utilisés par d'autres réactions comme substrat (ou cofacteur, ou catalyseur si le produit est une enzyme, *etc.*) et ainsi de suite. C'est l'ensemble de ces informations dans un organisme ou une cellule que l'on peut tenter de représenter sous la forme d'un réseau métabolique. Généralement, on utilise soit des graphes bipartis (exemple en figure 2.2) soit des hypergraphes pour dessiner de tels objets.

Les graphes bipartis ont deux types de sommets symbolisant soit une réaction (avec le ou les enzymes associé(s)) soit un métabolite (substrats et produits de la réaction). Les hypergraphes ont eux la particularité d'être composés d'hyperarêtes : arêtes pouvant avoir plusieurs entrées et/ou plusieurs sorties, elles correspondent chacune à une réaction. A l'aide de ces graphes, certaines ambiguïtés pouvant être introduites par

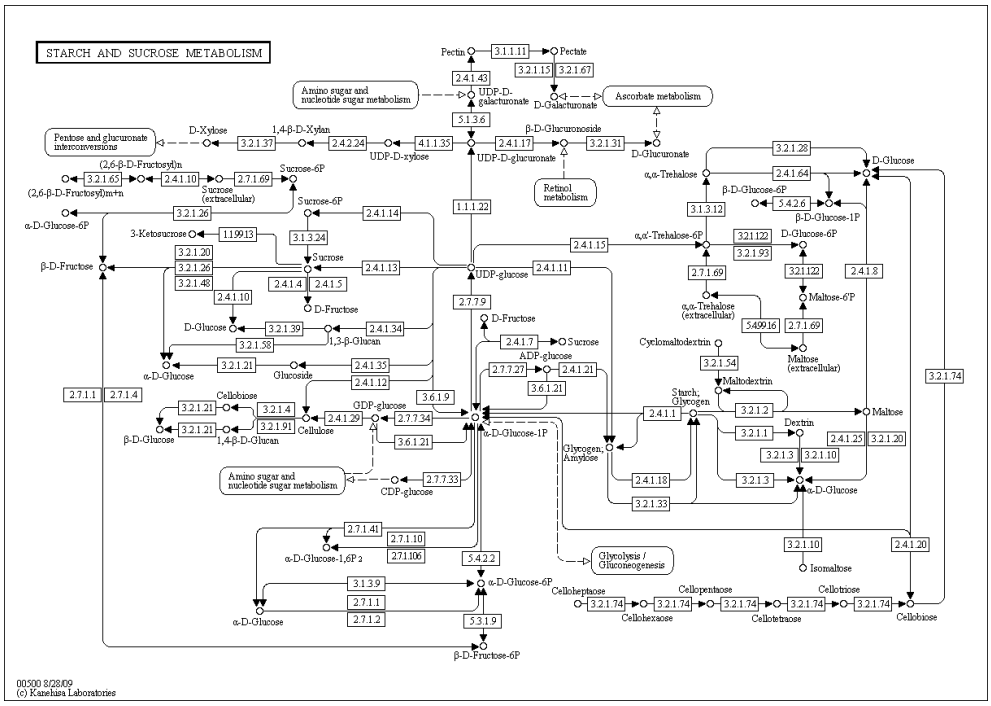


FIGURE 2.2 – **Carte métabolique de la biosynthèse du sucre et de l'amidon** issue de *KEGG* [Ogata *et al.*, 1999]. Cette carte est sous la forme d'un graphe biparti. Les sommets ronds représentent les métabolites (substrats et produits), les sommets rectangulaires représentent les enzymes. Les identifiants de ces enzymes sont ici des numeros EC, nomenclature liée à la fonction de l'enzyme en question.

l'utilisation de graphes dirigés simples peuvent être levées.

2.2.2 Les réseaux de régulation génique

Ils font l'objet de ce travail de thèse. Les interactions entre gènes régulateurs et gènes régulés forment un réseau : le réseau de régulation génique. Un exemple en figure 2.3 de [Hanahan et Weinberg, 2000] illustre l'état des connaissances en 2000 de la régulation des gènes impliqués dans des processus liés au cancer.

Les réseaux étudiés ici sont des réseaux de régulation de la transcription. Le paragraphe 2.3 abordant la transcriptomique détaille ce type de régulation. Lorsqu'on s'intéresse plus généralement à la régulation de l'activation d'un gène, des mécanismes (non approfondis dans ce projet) différents de ceux qui interviennent dans la transcription entrent également en jeu. On compte parmi eux des mécanismes agissant au niveau de la chromatine (on parlera de mécanismes épigénétiques), au niveau de l'ARN messager (par exemple l'effet répresseur des microARN), de la traduction ou encore

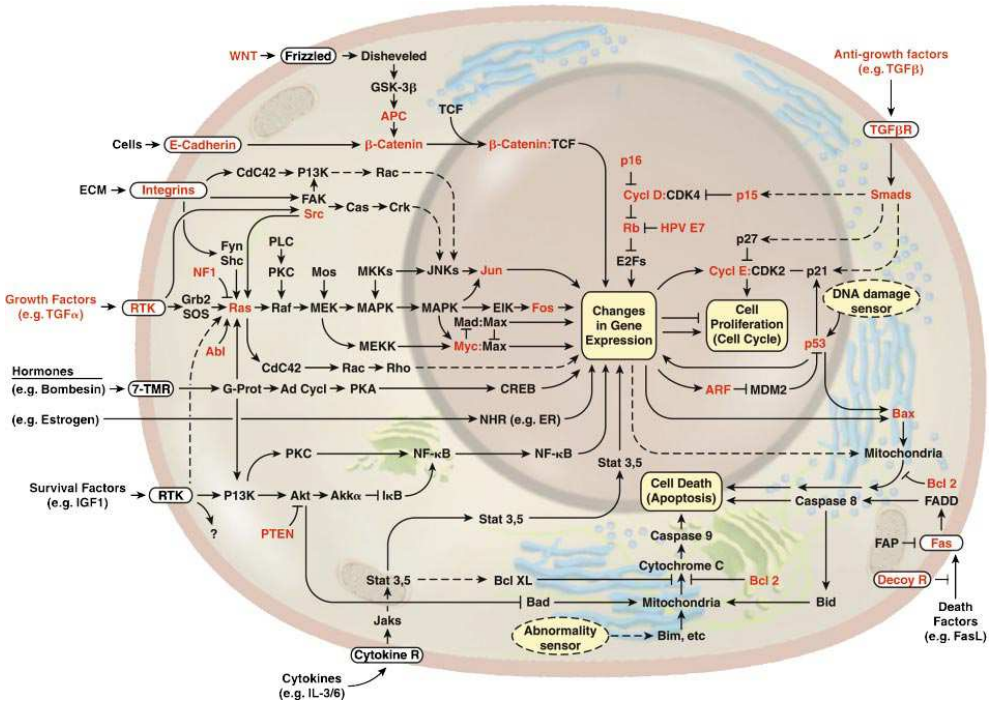


FIGURE 2.3 – **Réseau de régulation de gènes impliqués dans le cancer.** Le formalisme est emprunté aux électriciens. Les transistors sont ici les protéines, et les électrons des phosphates ou lipides. Les connecteurs de type « → » signifient « activation », et les connecteurs de type « ⊣ » signifient « inhibition ».

des processus agissant au niveau de la protéine.

2.2.3 Les réseaux d'interactions protéine-protéine

Les interactions entre protéines peuvent avoir plusieurs formes différentes. On pourra distinguer des interactions de transport (une protéine peut participer au transport d'une protéine tierce d'un compartiment cellulaire à l'autre), de formation de complexe (par exemple l'hémoglobine est un complexe formé par plusieurs protéines), d'interactions courtes menant à la modification de l'une d'entre elles (par exemple la phosphorilation). Une illustration de ce type de réseau est donnée en figure 2.4.

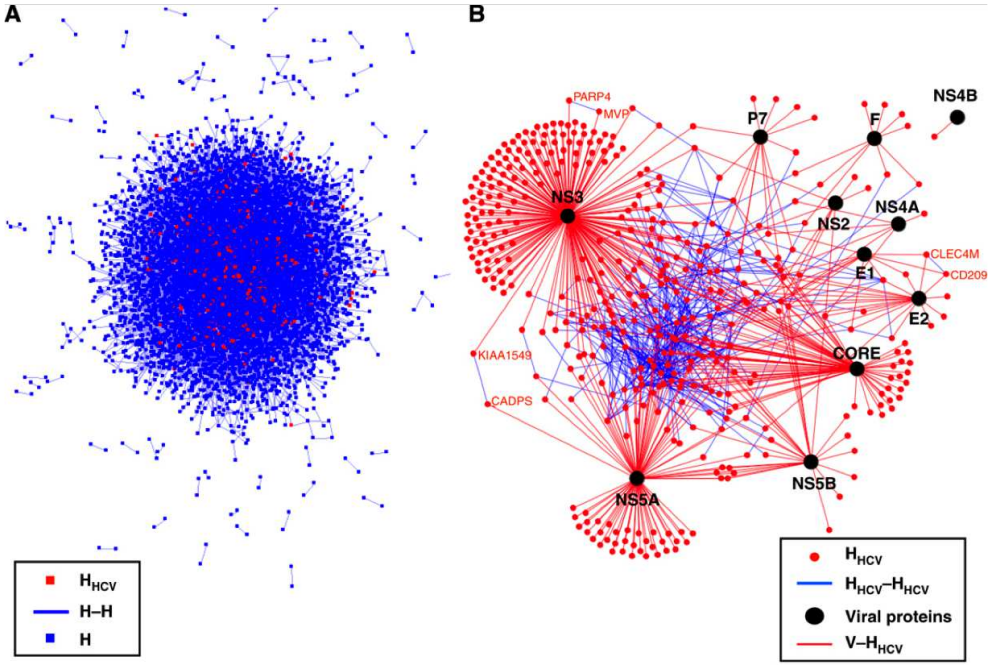


FIGURE 2.4 – **Réseau d'interaction entre protéines du virus de l'hépatite C et hôte humain** extrait par [de Chassey *et al.*, 2008]. A : réseau d'interaction global. B : mise en relief des protéines virales interagissant avec beaucoup de protéines humaines.

Afin d'obtenir des informations sur ce type d'interactions, plusieurs techniques sont mises en œuvre, comme la technique de co-immunoprécipitation, ou de levure double-hybride.

2.2.4 Les réseaux de signalisation

Un réseau de signalisation regroupe un ensemble de voies de signalisation (on parle aussi de « transduction du signal »). Celles-ci correspondent à un ensemble de proces-

sus faisant suivre un signal extracellulaire qui peut être d'ordre hormonal, nerveux, immunitaire, la façon dont ce signal est intégré, transformé, restitué. Un exemple de voie de signalisation (initiée par les récepteurs de lymphocytes T) est illustré en figure 2.5.

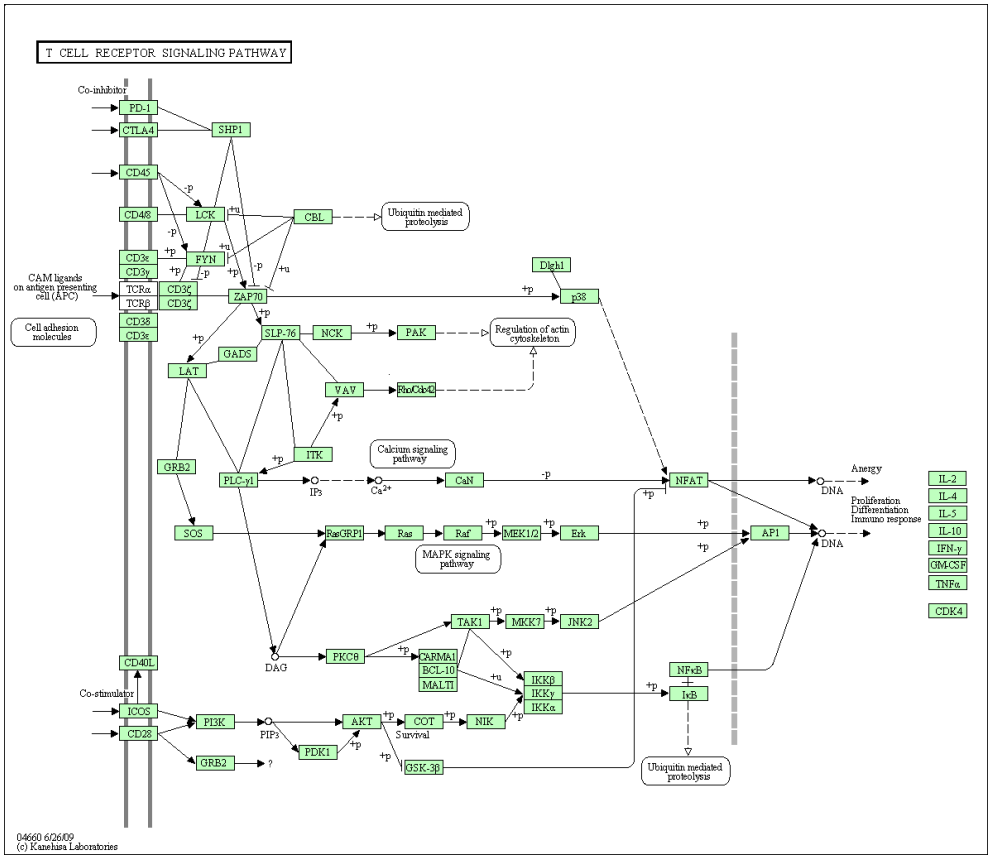


FIGURE 2.5 – **Voie de signalisation du récepteur de lymphocyte T** issue de *KEGG* [Ogata *et al.*, 1999]. Le signal extracellulaire est la reconnaissance d'une molécule d'adhésion cellulaire par une cellule présentant un antigène correspondant. Les rectangles représentent des gènes, les ronds d'autres molécules, les flèches les interactions entre elles. P et u signifient respectivement phosphorilation et ubiquitination. Cette voie aboutit à la réponse immunitaire.

Récemment, [Sackmann *et al.*, 2006] ont appliqué le formalisme des réseaux de Petri (voir le paragraphe 3.3) dans le but de modéliser ce type de réseau.

2.2.5 Conclusion

La figure 2.6 replace ces sous-réseaux plus globalement dans la cellule.

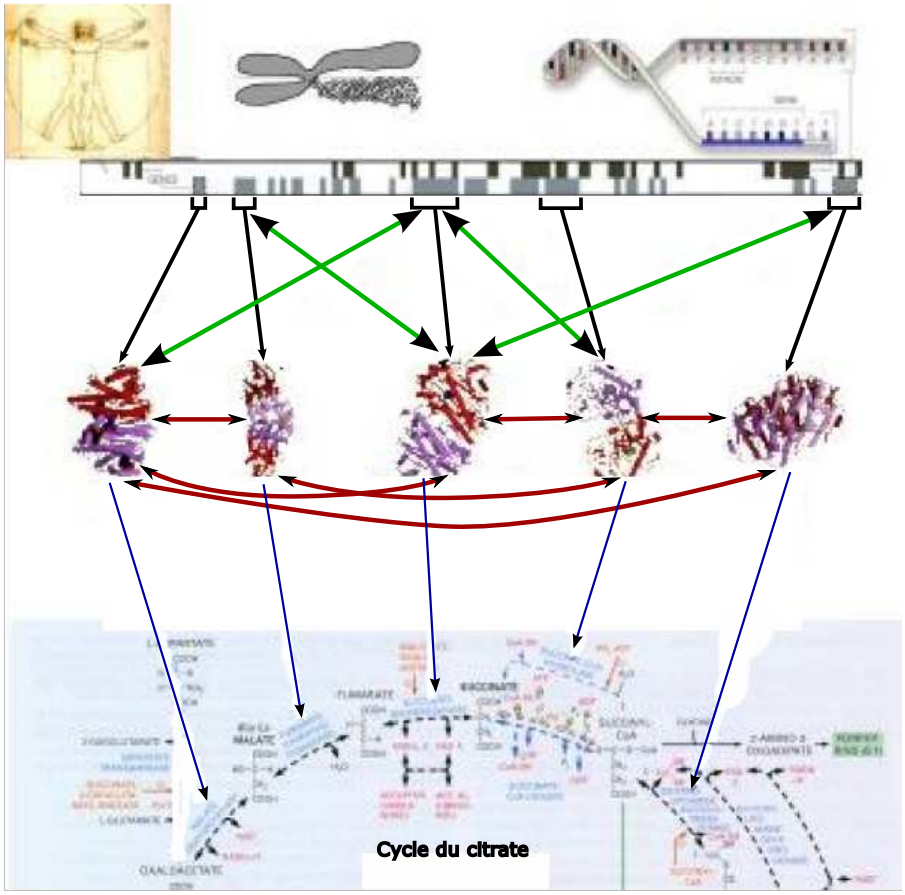


FIGURE 2.6 – **Réseau cellulaire**. Respectivement, les flèches noires représentent des évènements de transcription/traduction, vertes des régulations de gènes par des facteurs de transcriptions protéiques (forment un réseau de régulation génique), rouges des interactions entre protéines (réseau d'interaction protéique), bleu l'action catalytique de complexes protéiques dans le réseau métabolique.

La frontière entre tous ces réseaux cellulaires n'est pas toujours aussi nette. Par exemple, et cela se voie sur la figure 2.5, les réseaux de signalisation font intervenir à la fois des processus de régulation génique, et des interactions entre protéines.

La recherche de ces réseaux soulève beaucoup d'espoir dans la compréhension de mécanismes biologiques, et dans le développement de traitements en médecine.

2.3 La transcriptomique

2.3.1 Définition et généralités

On appelle « transcriptome » un instantané de tous les transcrits présents dans un tissu cellulaire, informations quantitatives comprises. Le « transcrit » est une molécule résultante du processus de transcription, consistant à polymériser des ribonucléotides, de façon à recopier la partie de la molécule d'acide desoxyribonucléique (ou « ADN ») bicaténaire bornée par des signaux promoteurs et terminateurs (figure 2.7). La séquence du transcrit, qui est une molécule monocaténaire, est complémentaire au brin d'ADN qui sert de matrice lors de la transcription. Après maturation, un transcrit est communément appelé « ARN ».

La synthèse d'ARN est catalysée grâce à une ARN polymérase. L'ARN polymérase est recrutée quand plusieurs conditions sont remplies (voir la figure 2.8), comme la présence d'un promoteur sur l'ADN et de différents facteurs de transcription.

2.3.2 Transcription : comparaison entre procaryotes et eucaryotes

On peut être étonné de la similarité de la machinerie cellulaire faisant qu'un gène s'exprime en ARN puis parfois en protéine entre les procaryotes et les eucaryotes considérant le fossé taxonomique séparant ces groupes. Il n'en demeure pas moins quelques différences concernant la transcription dont les plus notoires sont :

- une seule ARN polymérase existe chez les procaryotes, trois chez les eucaryotes ;
- une séquence consensus est reconnue dans les promoteurs : boîte de Pribnow [TATAAT] chez les procaryotes, et boîte TATA [TATA] chez les eucaryotes ;
- les transcrits subissent une étape de mâturation chez les eucaryotes : ajout d'une

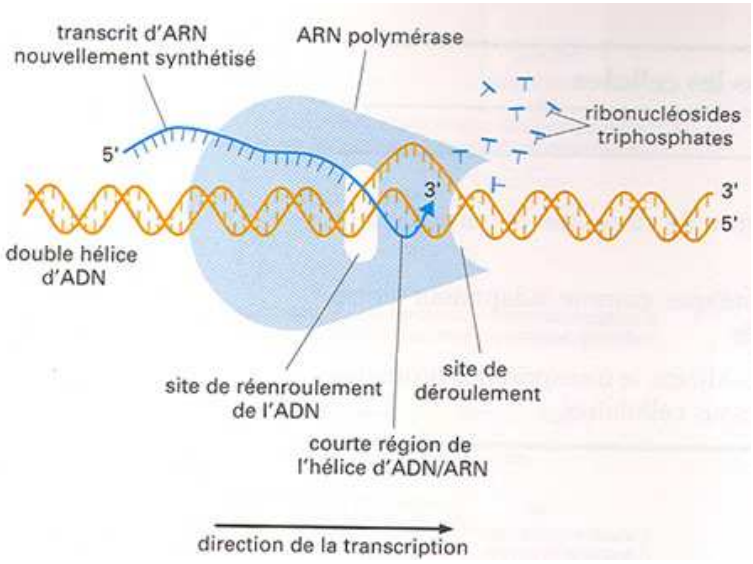


FIGURE 2.7 – **La transcription**. L'ARN polymérase lit un des deux brins de l'ADN (le brin transcrit) et polymérise les ribonucléosides de façon à produire une molécule d'ARN complémentaire à ce brin, par conséquent similaire (à la nature des sucres et des bases azotées près) au brin matrice.

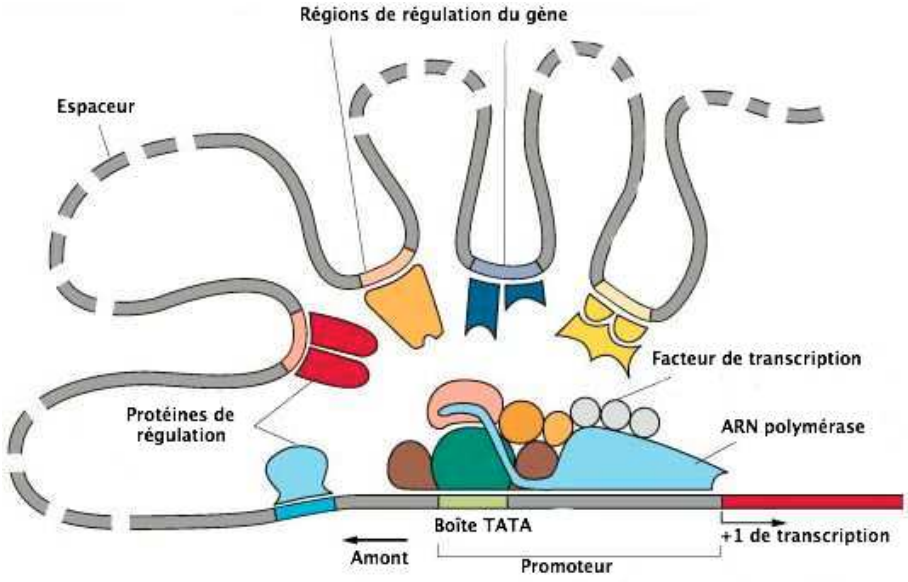


FIGURE 2.8 – **Initiation de la transcription.** L'ARN polymérase est mise en place au niveau du promoteur. Cela se produit lorsque certaines protéines, les facteurs de transcriptions, sont liées à des zones régulatrices de l'ADN en amont du promoteur.

queue poly-A, ajout d'une coiffe méthylguanosine et épissage (ablation de certaines parties du transcrit, ces régions transcrites non-codantes du gène sont appelés des introns, les parties codantes sont les exons).

a. Les facteurs de transcriptions : généralités et mise en évidence expérimentale

Un facteur de transcription est une protéine dotée d'au moins un site lui permettant de se lier à l'ADN, et agissant un contrôle sur l'expression du gène en aval du facteur par rapport au sens de la transcription. Ce contrôle peut être positif (il aura alors un rôle activateur), ou négatif (action inhibitrice). Cette protéine se lie à des régions spécifiques d'amplification (plus connues sous le nom anglais « enhancer ») ou directement sur le promotteur en amont du gène. Les facteurs de transcription peuvent réguler plusieurs gènes, et certains ne peuvent être actifs qu'à condition d'être liés à une protéine tierce, le tout formant un complexe protéine-protéine.

Afin d'identifier les régions où se forment des complexes ADN-protéine d'une protéine ciblée, [Ren *et al.*, 2000] ont mis au point un procédé nommé « chIP-on-chip » (signifiant « chromatine immuno-précipitation sur puce ») basé sur la technologie des biopuces décrite dans le paragraphe 2.4.2. Ce protocole, illustré par la figure 2.9, a pour principe l'isolation des fragments d'ADN en interaction avec la protéine testée, elle-même ciblée à l'aide d'anticorps spécifiques. L'anticorps se fixe à la protéine ciblée, l'ADN est alors découpé en petits fragments, puis filtré en faisant précipiter les anticorps (étape d'immunoprécipitation). Puis, l'étape de lavage permet de détacher le complexe [protéine cible - anticorps] du fragment d'ADN qui est récupéré et identifié à l'aide d'une puce, dont les sondes sont constituées d'un maximum de régions intergéniques du génome considéré, présumant que les facteurs de transcription sont forcément en amont des gènes.

2.3.3 Conclusion

Il existe différents types d'ARN, dont les plus connus sont les ARN ribosomiaux (ou « ARNr ») qui forment avec des protéines les complexes ribosomiques, et les ARN messagers (ou « ARNm ») qui peuvent par la suite être « traduits » (grâce aux complexes ribosomiques) pour former un polypeptide. Chez les eucaryotes, en moyenne, 75% des ARN présents dans le cytoplasme sont des ARNr, et seulement 3% sont des ARNm. Plus récemment, ont été mis en évidence chez les eucaryotes des petites séquences d'ARN (de l'ordre de 20 bases) jouant un rôle dans la maturation de l'ARNm

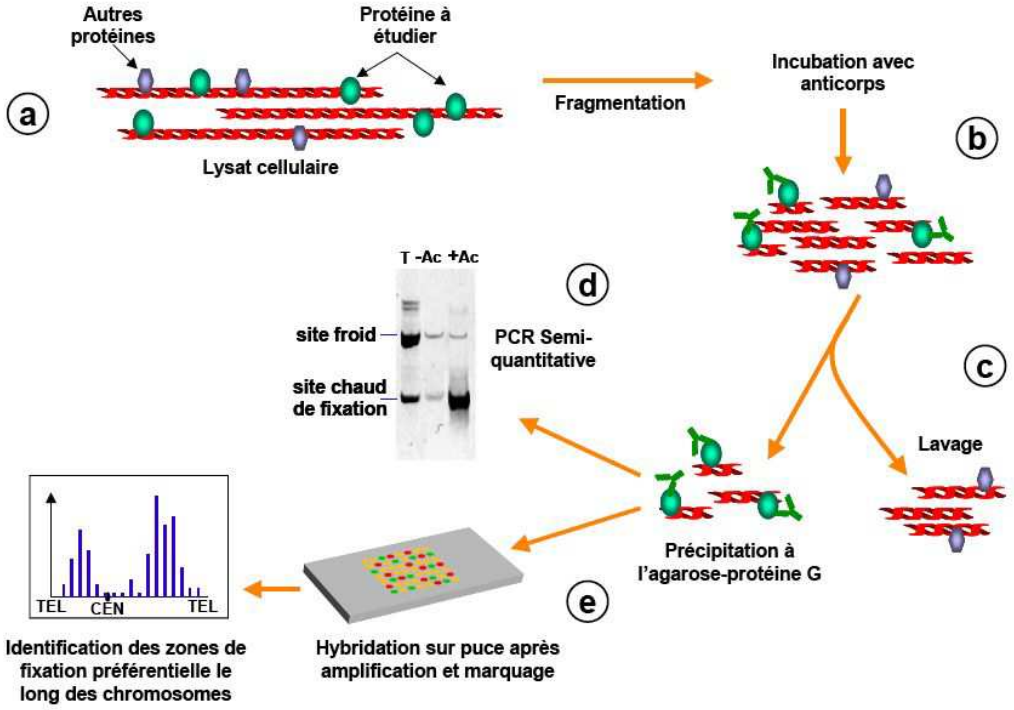


FIGURE 2.9 – **Principe des chIP-on-chip.** L'ADN (a) est fragmentée, puis les anticorps se fixant spécifiquement sur la protéine recherchée sont placés en incubation (b). Les complexes ADN-protéines d'intérêt sont ensuite isolés (c). Enfin, après dénaturation du complexe, les fragments d'ADN résultants sont caractérisés par PCR (d) et identifiés à l'aide d'une puce à ADN (e).

(les snoARN), ou leur dégradation/répression de la traduction (miARN et siARN).

Contrairement au génome qui est statique (sauf événements rares comme les réarrangements ou mutations, mais là on ne se place pas à l'échelle de temps d'un individu) le transcriptome a une dynamique :

- spatiale : tissulaire, voire cellulaire, si les cellules sont asynchrones ou si on parle d'organismes unicellulaires ;
- temporelle : dans une même cellule le transcriptome est variable.

Les mécanismes à l'origine de cette dynamique sont en grande partie basés sur la régulation de la transcription, via le recrutement de l'ARN polymérase II. Les facteurs de transcription sont très nombreux, et on ne les connaît pas tous. Lorsqu'on cherche à étudier la régulation de l'expression des gènes en transcrits, on recherche en général un « réseau de régulation génique ». La dynamique à la fois spatiale et temporelle et conjointement la diversité des facteurs stimulant ou inhibant la transcription font que le transcriptome est un système complexe : on en attend beaucoup de son exploration, qui de fait est difficile.

2.4 Méthodes de détection des transcrits

Différentes méthodes ont été développées pour détecter des transcrits à partir des années 1990. Certaines requièrent une connaissance des séquences de gènes dont on cherche à connaître l'expression (puces à ADN et SAGE). D'autres, basées sur le séquençage n'en ont par conséquence pas besoin.

2.4.1 Sans connaissance du génome complet

a. Librairies d'EST

Les ESTs [Adams *et al.*, 1991], ou « Expressed Sequence Tag », sont des petites séquences nucléotidiques (entre 200 et 800 paires de bases) issues du séquençage d'une librairie d'ADN complémentaire (ADNc). Un ADNc est une molécule d'ADN générée par un procédé de retrotranscription d'ARNm, mettant en jeu une enzyme (produite naturellement par les retrovirus) nommée la *transcriptase inverse*. A partir d'une molécule d'ARNm mâture, la transcriptase inverse génère un ADNc simple brin. Généralement,

ce procédé est couplé à une étape de PCR (« Polymerase Chain Reaction ») dans le but de multiplier les fragments d'ADNc. On parle alors de *Reverse Transcriptase PCR*, on obtient dans ce cas de multiples copies d'une séquence d'ADNc double-brin, correspondante à l'ARNm d'origine, et donc à la séquence du gène associé (mais sans les introns si on se place chez les eucaryotes, car l'ARNm utilisé est épissé). Une librairie d'ADNc est un ensemble d'ADNc générés à partir des ARNm d'un échantillon. Pour constituer une librairie d'EST, on séquence aléatoirement les ADNc dans les deux sens. Les séquences ne sont pas de bonne qualité (car elles ne sont lues qu'une seule fois) et redondantes. Cependant la méthode est rapide et peu onéreuse, et a le mérite d'être la première approche permettant d'avoir une vision du transcriptome d'un échantillon.

b. SAGE : « Serial Analysis of Gene Expression »

[Velculescu *et al.*, 1995] ont élaboré une méthode plus rapide et impliquant moins d'erreurs, car seule une très courte séquence (le « tag ») associée à un ARNm est lue, mais plusieurs fois. En effet, le principe repose sur le fait que statistiquement, une séquence (par exemple de 14 pb) suffise à l'identification de l'ARNm correspondant.

A partir d'une librairie d'ADNc, plusieurs étapes sont conduites pour obtenir un unique tag par ADNc, puis pour les accoler deux à deux, formant des ditags. Les ditags sont ensuite amplifiés par PCR puis concaténés. Enfin, les concaténats sont clonés dans un vecteur bactérien avant d'être séquencés. Le nombre de tags associé à un ARNm est en principe proportionnel à la quantité présente dans l'échantillon.

c. Differential Display

Cette technique imaginée par [Liang et Pardee, 1992], basée sur la comparaison de plusieurs populations d'ARNm, utilise dans un premier temps la PCR inverse avec une amorce poly-T complémentaire à la queue poly-A des ARNm (pour éviter d'amplifier les autres types d'ARN) et d'autres amorces aléatoires. La PCR inverse permet donc d'obtenir des fragments d'ADN complémentaire, en proportion au nombre de chaque ARNm au départ. Ensuite, par électrophorèse, les produits de la PCR migrent en fonction de la charge des ADNc. Chaque colonne du gel correspond à une des populations d'ARNm à comparer : il suffit de comparer les bandes pour observer des différences de quantités. On peut éventuellement séquencer l'ADNc contenu dans des bandes suscitant par leur différence un intérêt de l'expérimentateur.

d. Séquençage à haut-débit, RNA-SEQ

La très récente technique du RNA-seq introduite par [Nagalakshmi *et al.*, 2008] peut être considérée comme similaire à celle des ESTs, seulement la technologie de séquençage est différente. Elle utilise les techniques de séquençage de nouvelle génération beaucoup moins onéreuses que la méthode de Sanger, et n'a pas besoin de passer par l'étape clonage qui introduit un biais de représentativité des séquences.

2.4.2 A partir de la connaissance des séquences de génomes complets : Puces à ADN

Les puces à ADN [Schena *et al.*, 1995] constituent une technologie d'étude du transcriptome reposant sur une connaissance des séquences des gènes ciblés. Elles permettent une quantification simultanée de l'expression de tous les gènes connus. Cette technologie est particulièrement détaillée car les données utilisées dans ce travail sont issues de celle-ci.

a. La fabrication

Les puces à ADN sont sujettes à des développements et utilisations multiples, c'est pourquoi ce paragraphe ne peut être exhaustif : est énoncé ici simplement le principe de la technique telle qu'elle est le plus couramment utilisée.

Le support

Le support d'une biopuce est une lame de verre type « lame de microscope » traitée de façon à permettre la fixation des courtes séquences que sont les sondes.

Les sondes

Les sondes représentent ce que l'on connaît au départ. Ce sont des fragments d'ADN (généralement des oligonucléotides) dont on connaît la séquence allant de 15 à 60 bases azotées chacune. Ces séquences peuvent provenir de génomes bactériens, de plasmides, de produits d'amplification d'ADNc ou de segments génomiques, ou peuvent être d'origine synthétique. Ces différents types de sondes présentent des avantages et des inconvénients spécifiques. Elles sont amplifiées par PCR puis purifiées avant d'être fixées sur le support de la puce. Les sondes en solution sont déposées sur le support grâce à

un robot en un point que l'on appelle « spot ». Les robots sont capables de déposer plusieurs dizaines de milliers de spots sur un même support.

Elles peuvent aussi être fabriquées directement *in situ* (cas des sondes synthétiques).

Les cibles

Les cibles représentent ce que l'on cherche à identifier. Elles proviennent d'échantillons dont l'ADN ou l'ARN ont été extraits. Dans le cas des puces pour l'analyse du transcriptome ce sont souvent des molécules d'ADNc issues des ARNm à quantifier.

Incorporation des fluorochromes

Les fluorochromes sont des molécules qui servent à marquer les acides nucléiques cibles. Elles ont un spectre de longueurs d'ondes d'excitation et d'émission bien documenté. Cela permet l'excitation de fluorochromes spécifiques, même lorsque différents types de fluorochromes sont présents sur une même biopuce (cas des biopuces bifluorescentes, où généralement on cherchera à comparer deux conditions marquées chacune par une fluorescence). Les fluorochromes sont le plus souvent incorporés aux cibles pendant l'amplification par PCR, ou pendant une transcription artificielle (*in vitro*).

Hybridation sondes/cibles

Les cibles sont mises en contact avec les sondes déposées sur la lame, dans des conditions contrôlées de température et de salinité. Les conditions d'hybridation sont le produit du compromis entre spécificité et sensibilité de l'expérience.

Lavage

Cette étape permet d'éliminer toute cible marquée et non-hybridées.

La figure 2.10 résume les étapes de la fabrication d'une biopuce unifluorescente (*i.e.* un seul marquage fluorescent est effectué).

b. L'acquisition d'images

Un lecteur de fluorescence pour lames est employé pour acquérir les images. Les principaux éléments sont :

– un laser par type de fluorochrome à exciter (excitation des fluorochromes incorporés dans les acides nucléiques cibles) ;

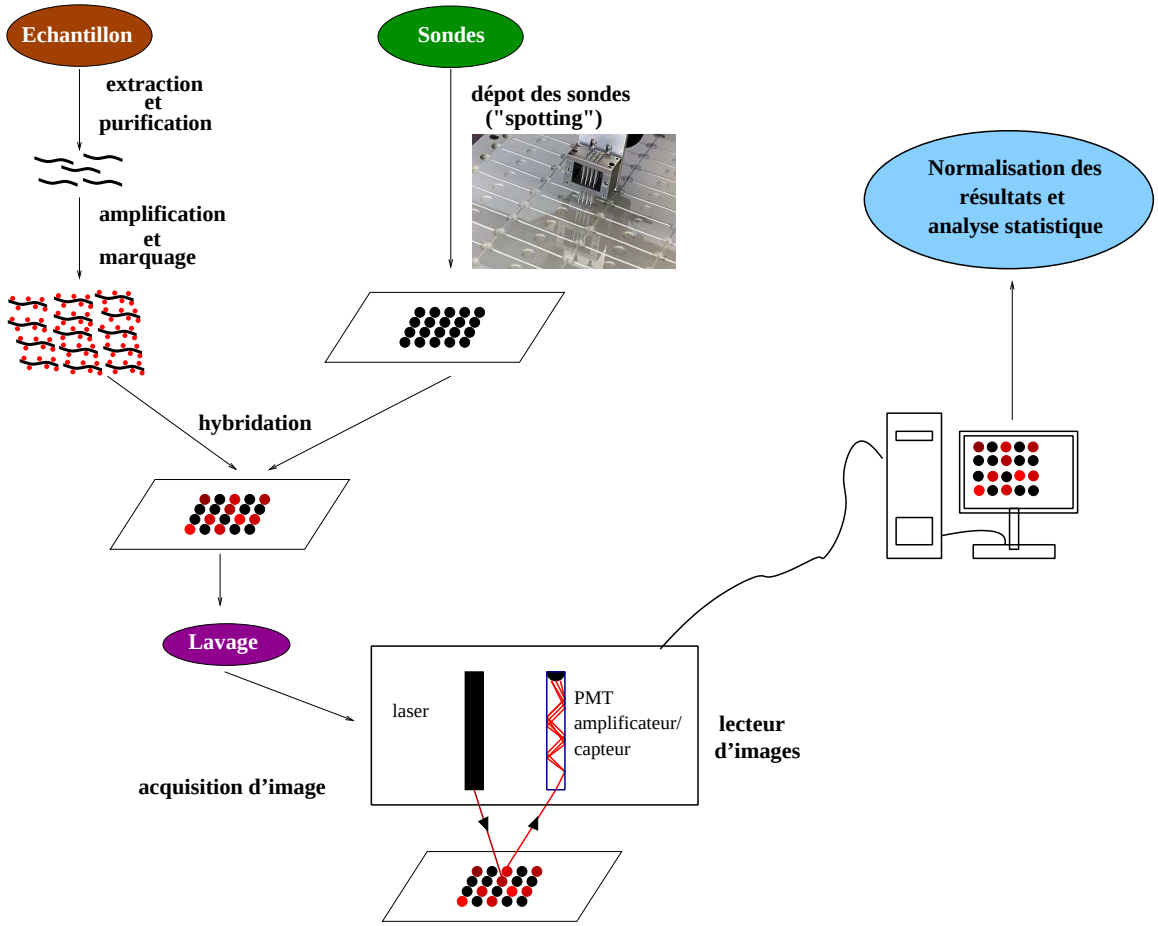


FIGURE 2.10 – **Puces à ADN : mise en œuvre**. L'ensemble des ARNm sont extraits de l'échantillon, puis marqués à l'aide de molécules fluorescentes. Parallèlement, des lames sur lesquelles ont été déposées les sondes sont préparées. Les ARNm marqués sont mis en présence avec les lames, ce qui constitue l'étape d'hybridation. Par complémentarité des séquences, les fragments cibles d'ARNm s'apparient avec les sondes correspondantes : plus les fragments sont nombreux, plus le nombre d'hybridations par spot de sondes est important. Après l'élimination des fragments non-appariés par lavage des lames, elles sont lues par un scanner utilisant un laser spécifique au spectre d'excitation de la molécule fluorescente. Les données brutes sont les images issues de ces lectures, par la suite traitées et analysées.

– un tube photomultiplicateur (capte et amplifie des signaux émis par l'état d'excitation des fluorochromes).

La puce est lue pour chacun des fluorochromes utilisés. On obtient donc une image par fluorochrome. Les images sont en niveaux de gris (65536 niveaux de gris = 16 bits = 2^{16}) et généralement la résolution est de 10 μm (taille du côté d'un pixel carré). Lorsqu'on utilise des puces bifluorescentes on acquiert deux images (une par couple fluorochrome/laser). C'est la comparaison des deux images qui révèle les différences d'expression.

c. Extraction de données des images

Un logiciel se charge ensuite d'extraire l'information pour chaque image. Il effectue dans un premier temps l'étape d'adressage (ou de localisation) qui consiste à repérer la position du centre de chaque spot, puis, dans un deuxième temps, il réalise ce que l'on appelle la segmentation : il repère le contour du spot (de manière plus ou moins fine selon la méthode). Enfin, le programme récupère les données d'intensités : moyenne, médiane, nombre de pixels, *etc.*, pour chaque spot, et localement autour du spot pour que l'on puisse ensuite déduire l'intensité du bruit de fond global, local, *etc.*

2.4.3 Traitement et analyse des données de puces à ADN dédiées à l'analyse du transcriptome

a. La normalisation

Une expérience de puces à ADN fait intervenir plusieurs étapes, chacune potentiellement source de biais. Certains d'entre eux peuvent être mesurés en particulier lorsque l'on introduit des sondes lors du « spotting » qui vont servir de témoin. La normalisation est le traitement qui vise à ajuster les données en tenant compte des effets des variations dues à la technologie plutôt qu'à des différences biologiques, ces dernières étant celles qui nous intéressent [Smyth *et al.*, 2003].

Généralement on distingue deux étapes de normalisation : la normalisation du bruit de fond (propre à chaque lame) et la normalisation inter-lames permettant une comparaison plus pertinente.

Procédure visant à retirer le bruit de fond d'une puce

Dans un premier temps, on essaye de mesurer le bruit de fond. Un bruit de fond est estimé soit grâce à des mesures directes (zones au voisinage des spots, ou dépôts

de solution tampon dans différentes localisations de la lame) soit en estimant la distribution par déconvolution de la distribution totale (signal + bruit de fond) comme cela est proposé par [Irizarry *et al.*, 2003] et [Wu *et al.*, 2004] (ces derniers tenant compte du biais induit par le contenu en GC), l'estimation est la seule approche pour certaines puces qui n'ont pas d'espace entres les objets fluorescents comme c'est le cas avec la technologie Affymetrix.

Normalisation à l'échelle d'une puce

Une fois le bruit de fond estimé et soustrait aux valeurs du signal, on peut s'intéresser à certains biais : en particulier, deux d'entre-eux ont été identifiés sur les puces avec des sondes déposées :

1. biais microplaque/aiguille : les microplaques sont constituées de dizaines de puits contenant les sondes avant leur dépôt par les aiguilles sur la lame. Les différences de concentrations de sondes dans chaque puits, ainsi que les différences de quantité prélevées puis déposées entre chacune des aiguilles sont source de biais. Il est possible d'en tenir compte dans le modèle statistique de normalisation, en utilisant par exemple celui de [Smyth *et al.*, 2003].
2. biais lié au marqueur fluorescent : tous les fluorochromes n'ont pas la même gamme dynamique (intervalle dans lequel le signal est proportionnel aux concentrations de molécules marquées), ou les mêmes seuils de détection, ou le même temps de demi-vie. Bien-sûr on essaye de régler le lecteur de lames fluorescentes de façon à s'affranchir de ce biais, mais c'est insuffisant. Généralement, un plan d'expérience en « flip-flop » où plusieurs lames sont utilisées dans le but d'échanger les marqueurs fluorescents attribués aux deux conditions, et une correction sous l'hypothèse d'un niveau global d'expression semblable entre les deux conditions biologiques est produite pour diminuer l'effet indésirable de ce biais.

Normalisation inter-puces

A partir de réplicats de biopuces, il a été montré qu'il était très difficile de mesurer des intensités comparables. On devrait, en faisant abstraction des biais expérimentaux, au minimum obtenir une moyenne et une dispersion d'intensité égales entre les différentes lames, seulement ce n'est pas le cas. Faisant l'hypothèse là-aussi que même avec des lames différentes (sauf cas spécifiques où on réprimerait la transcription dans un tissu ou une colonie de cellules) l'intensité globale (moyenne) et la dispersion devraient

être égales, des procédures de normalisation corrigeant ces éventuelles différences ont été proposées. Une première famille visant à aligner les médianes de toutes les puces a été proposée par [Yang *et al.*, 2002]. Par la suite, une deuxième famille alignant non-seulement la médiane, mais d'autre bornes de quantiles dans le but d'égaliser les dispersions ont été proposées par [Yang et NP, 2003] pour les biopuces bifluorescentes, et [Bolstad *et al.*, 2003] pour les biopuces unifluorescentes du type affymetrix. Plusieurs variantes de normalisation dite « par quantiles » existent, par exemple certaines normalisent les intensités globales mais pas les différences d'expressions entre deux conditions.

b. L'analyse statistique des données

Cette dernière étape est étroitement liée au plan d'expérience qui doit théoriquement être optimisé en fonction des hypothèses de départ. La base du plan d'expérience est d'organiser au moins quelques répétitions (cf. figure 2.11 de [Yang *et al.*, 2002]) malgré le surcoût engendré.

Tests d'hypothèse

Généralement l'hypothèse nulle H_0 testée pour chaque gène est soit :

- comparative : « il n'y a aucune différence d'expression entre les différentes conditions » , c'est systématiquement le cas avec les biopuces bifluorescentes ;
- quantitative : « le gène ne s'exprime pas ».

La première étape est de choisir entre test paramétrique et non-paramétrique. Les tests paramétriques gardent la dimension quantitative des mesures d'intensités, tandis que les tests non-paramétriques transforment ces mesures en rangs, donc en valeurs qualitatives ordinales. Cela a l'avantage de diminuer l'effet des « outliers », de ne pas se préoccuper de la normalité des distributions, mais a l'inconvénient d'accorder une importance égale quelque soit la différence entre deux intensités lorsque la différence de rang respectif ne varie pas. On considère généralement les tests non-paramétriques intéressants lorsque le nombre de répétitions croît.

Une fois la décision prise, le choix du test dépend du nombre d'effets testés (exemples : un traitement, une cinétique, un tissu) pouvant expliquer une variation du transcriptome. Lorsqu'il y a un seul facteur, on peut simplement procéder à des tests de Student, (ou Wilcoxon si on travaille sur les rangs). Si on teste plusieurs facteurs, on peut utiliser une approche par analyse de la variance multifactorielle (dont la variante non paramétrique est le test de Friedman).

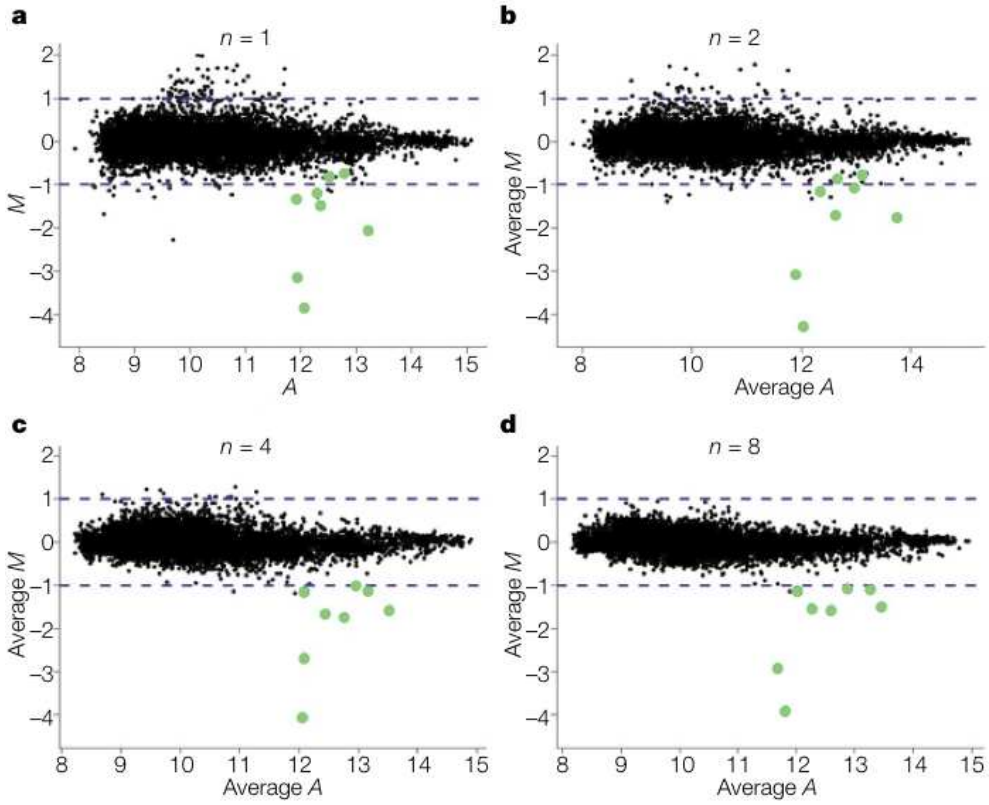


FIGURE 2.11 – **Importance des réplicats.** Les figures a, b, c et d sont des MA-plot (axe X : A=$\frac{1}{2}(log_2(I_1) + log_2(I_2))$ et axe Y : $M = log_2(I_2) - log_2(I_1)$), très utilisés pour comparer deux conditions. Elles illustrent respectivement des valeurs d'intensités I_c unique par condition, puis des valeurs moyennes sur deux, quatre et huit répétitions par condition. On observe une diminution de la dispersion avec le nombre de répétitions, du nombre de faux positifs (points noirs en dehors des valeurs de M comprises entre −1 et +1), et enfin on constate que certains gènes attendus comme différentiellement exprimés (en vert) parce qu'ils ont été KO pour une condition ne le sont qu'à partir de quelques répétitions.

Tests multiples

Pour chaque analyse, le nombre de tests est répété autant de fois qu'il y a de gènes ciblés par la puce. Cela pose le problème dit des « tests multiples ». En effet, quand on procède à des tests d'hypothèse, on calcule une p-valeur que l'on compare à un seuil correspondant à un risque dit de première espèce (risque α) afin de prendre la décision de rejeter ou d'accepter l'hypothèse nulle. Puisqu'on appelle un gène « positif » celui pour lequel H_0 est rejetée, que le risque α peut être interprété comme la probabilité de rejeter à tort l'hypothèse nulle, ce risque correspond alors à une probabilité d'obtenir un faux positif. Ce que l'on obtient à la fin de tous les tests est une liste de gènes « positifs » et on peut estimer le nombre de faux positifs en multipliant le nombre de gènes testés par le risque α. Par exemple, si on considère 30000 gènes et un risque $\alpha = 5\%$ on estime le nombre total de faux positifs à $0,05 \times 30000 = 1500$. Le problème est qu'en réalité, on s'intéresse à un nombre de gènes beaucoup plus restreint, et ce que l'on recherche est une estimation de ce taux parmi ces gènes choisis ayant les plus faibles p-valeurs : ça peut être de l'ordre de la cinquantaine, ou de la dizaine. Il existe pour cela des procédures de correction des p-valeurs « FDR (False Discovery Rate) » de [Benjamini et Hochberg, 1995] permettant de contrôler non-pas le nombre de faux positifs sur le total des gènes testés mais sur les gènes choisis. On peut préférer contrôler la probabilité d'obtenir au moins 1 faux positifs : c'est le « FWER (Family Wise Error Rate) » en procédant à une correction de Bonferroni [Simes, 1986]. La deuxième méthode étant très conservative, on ne détecte pas un grand nombre de gènes potentiellement intéressants, il est donc plus courant dans des problématiques de biopuces de contrôler le FDR.

Classification

Un objectif partagé par beaucoup de biologiste est de regrouper ensemble des gènes ayant un comportement qui se ressemble. C'est ce que l'on appelle de la classification. On distingue deux grandes familles de classification selon que l'on dispose d'informations sur l'appartenance de certains objets à des classes au préalable (classification supervisée) ou pas (classification non-supervisée). Généralement, les méthodes de classification s'appuient sur la minimisation de distance (type euclidienne, manhattan), ou à l'inverse la maximisation des similarités (corrélations, informations mutuelles). Une fois ce critère établi, on peut procéder à une classification de type hiérarchique (par exemple UPGMA) ou non (par exemple K-means).

Autres analyses

On a parlé jusqu'ici d'analyses univariées. Il peut être intéressant d'explorer les données de biopuces à l'aide d'analyses multivariées selon les problématiques. Par exemple, à l'aide d'une analyse en composantes principales, on peut chercher à comprendre comment se regroupent plusieurs gènes selon plusieurs conditions (à l'aide du cercle des corrélations), ou inversement comment la combinaison de plusieurs gènes peut expliquer certaines de ces conditions (à l'aide des plans factoriels). Une analyse en composantes principales inter-classe peut en plus chercher à expliquer par les niveaux d'expression la divergence entre classes connues. Il est aussi possible d'utiliser des analyses de correspondance, analyses canoniques, analyses en coinertie détaillées par [Dray *et al.*, 2003], *etc.* dans le but d'explorer un jeu de données de puces.

2.5 Conclusion

Dans ce chapitre on a décrit toute la diversité des approches lorsqu'il s'agit d'étudier un certain nombre d'interactions moléculaires nécessaires au fonctionnement cellulaire. On a vu que ces approches pouvaient diverger à la fois quantitativement (approches exhaustives ou non) et qualitativement selon le type de molécules ou interactions. D'après l'état des connaissances actuelles, il est encore trop tôt pour intégrer toutes ces informations ensemble telle que dans la figure 2.6 qui de plus est certainement incomplète.

Grâce à l'évolution de plusieurs technologies (ESTs, SAGE, biopuces), on a vu qu'une discipline a particulièrement murie cette dernière décennie : l'analyse du transcriptome. Etre capable de « facilement » caractériser des transcriptomes est très important et prometteur dans la compréhension de mécanismes cellulaires. Parmi ceux-ci, on peut s'intéresser à la mise en évidence, via l'identification de gènes fonctionnant ensemble, d'indices génomiques pouvant expliquer différents facteurs biologiques. L'analyse du transcriptome peut permettre aussi de chercher à alimenter nos connaissances sur le réseau de régulation génique. En effet il est *a priori* pertinent de penser que des gènes régulés par un même facteur de transcription sont co-exprimés dans une cellule. L'inverse, *i.e.* des gènes sont coexprimés s'ils sont régulés par un même facteur de transcription est beaucoup plus douteux puisqu'ils peuvent être exprimés ensemble

par le hasard. D'où l'intérêt de répéter les expériences, permettant la mise en exergue les « combinaisons gagnantes ».

Mais ce n'est pas suffisant car il faut aussi caractériser tous les facteurs de transcription. A ce propos, on décrit une méthode permettant de détecter des gènes régulés par un facteur de transcription à tester (méthode chIP-on-chip), néanmoins cette approche ne permet pas de détecter au préalable les facteurs de transcription potentiels. La recherche des facteurs de transcription est un problème ouvert en biologie.

Une autre approche dans la recherche de réseaux d'interactions est de considérer que celui-ci possède des propriétés modélisables dans un graphe, et de rechercher le meilleur graphe possible connaissant les données d'expressions. Une approche possible est celle des Réseaux Bayésiens et fait l'objet de cette thèse.

Deuxième partie

Des réseaux biologiques et des modèles en réseaux

Chapitre 3

Réseaux biologiques, mathématiques et statistiques

Sommaire

3.1 Introduction

Dans le chapitre précédent, on note que les réseaux biologiques auxquels on s'intéresse dans ce travail sont des réseaux cellulaires d'interactions moléculaires. Le type de réseau sur lequel on se focalise est le réseau de régulation génique, pouvant en principe être inféré au moins en partie avec des données de biopuces. L'idée de modéliser à l'aide d'un graphe d'interactions (ayant des propriétés mathématiques) un réseau biologique semble être naturelle. De surcroit, un graphe a l'avantage d'être visuel, et selon les propriétés modélisées (type d'indépendance par exemple) peut être plus ou moins puissant pour détecter des relations entre nos objets biologiques, ici l'expression des gènes. Ce chapitre a pour vocation de faire le point sur quelques modèles mathématiques graphiques, et d'argumenter notre choix pour répondre à nos attentes. La liste ici n'est pas exhaustive, elle représente simplement les modèles en réseau qui nous semble être les plus connus dans notre domaine : la bioinformatique. Après avoir introduit cinq types de modèles et présenté quelques-unes de leurs caractéristiques, une brève comparaison de ces dernières est effectuée.

3.2 Réseaux booléens et réseaux logiques généralisés

Dans un système, quand un objet peut prendre deux états (par exemple un gène peut être considéré comme activé ou non), il peut être modélisé à l'aide d'une variable booléenne. Si on se place à un temps t, on peut définir l'état dans lequel se trouve la variable i selon une fonction booléenne de la combinaison d'un ensemble de variables à $t-1$. Ce système forme un Réseau Booléen [Kauffman, 1969], qui est donc un modèle dynamique et à temps discrets. La figure 3.1 (reprise dans [de Jong, 2002]) montre un exemple de graphe booléen avec trois variables.

La recherche d'attracteurs, d'états et de cycles stables permet de dégager des propriétés locales d'un tel réseau. Une limite du modèle est que, comme leur nom l'indique, seuls deux états (activation ou non) peuvent être pris en compte pour une variable (un gène). De plus, les transitions se font toutes en simultanée. Une méthode généralisant les Réseaux Booléens (appelée « réseaux logiques généralisés ») a été développée par [Thomas, 1973] parallèlement aux Réseaux Booléens. Avec ce formalisme, les variables ne sont plus booléennes mais discrètes, et la mise à jour des sorties peut être asynchrone.

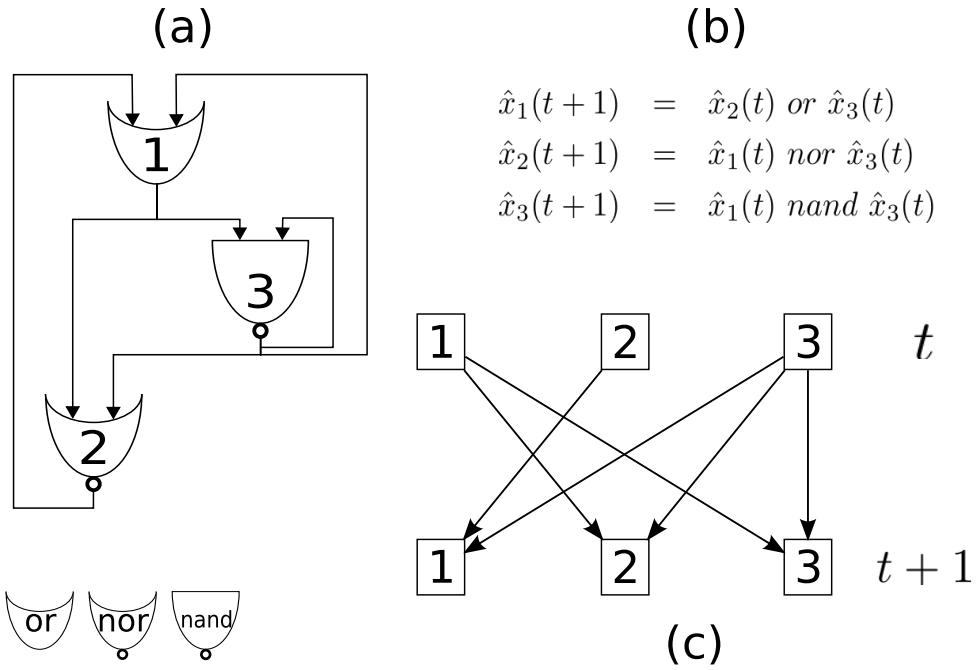


FIGURE 3.1 – Exemple de **Réseau Booléen**. (a) graphe booléen : la forme des sommets indique la nature de l'opérateur logique renvoyant la sortie (en bas) en fonction des entrées (en haut). (b) Equations du modèles (c). Schéma de connexion du réseau. Dans cet exemple, trois opérateurs booléens sont utilisés : (1) OR : renvoie 1 si l'une des entrées est égale à 1, (2) NOR : renvoie 0 si l'une des entrées est égale à 1, (3) NAND : renvoie 0 si toutes les entrées sont égales à 1.

3.3 Réseaux de Petri

Un Réseau de Petri est composé d'un graphe dirigé ayant deux types de nœuds (c'est donc un graphe bipartie). Certains sont nommés les « places », les autres sont les « transitions ». Les arêtes dirigées (donc les arcs) relient tous une place à une transition, ou l'inverse. Les arcs qui vont d'une place vers une transition sont les arcs d'entrée, ceux qui vont d'une transition vers une place sont les sorties. Plus formellement, un Réseau de Petri est formé de cinq éléments :

- $P = \{p_1, p_2, \ldots, p_m\}$ est un ensemble fini de places ;
- $T = \{t_1, t_2, \ldots, t_n\}$ est un ensemble fini de transitions ;
- $F \subseteq (P \times T) \cup (T \times P)$ est un ensemble d'arcs (relation de flux) ;
- $W : F \to \{1, 2, 3, \ldots\}$ est une fonction de poids ;
- $M_0 : P \to \{0, 1, 2, 3, \ldots\}$ est le marquage initial ;

avec $P \cap T = \varnothing$ et $P \cup T \neq \varnothing$.

L'exemple de la figure 3.2 montre une « exécution » d'un réseau qui représente la formation de la molécule d'eau. Cet exemple montre que ce formalisme est particulièrement adapté dans la modélisation de systèmes stœchiométriques, ce qui est le cas des réactions chimiques (donc biochimiques).

Le deuxième exemple (figure 3.3) montre le cas d'une place avec deux sorties. Si celle-ci ne possède qu'un jeton, celui-ci sera distribué au hasard soit vers une transition soit vers l'autre. Différentes itérations du système à partir de conditions initiales identiques peuvent alors mener à des systèmes différents à l'équilibre.

Les Réseaux de Petri semblent être intéressants, surtout pour modéliser des réseaux métaboliques comme le font par exemple [Sackmann *et al.*, 2006] déjà cités dans l'introduction des réseaux métaboliques dans le paragraphe 2.2.1. En effet les similarités (graphe biparti, stœchiométrie) avec ces réseaux biologiques sont intuitives, de plus ces modèles sont robustes lorsque les taux de réactions ne sont pas toujours connus. Cette dernière caractéristique n'est pas le cas par exemple des modèles en équations différentielles (dont on ne parle pas ici car ils ne sont pas intrinsèquement graphiques) souvent utilisés pour les réseaux métaboliques.

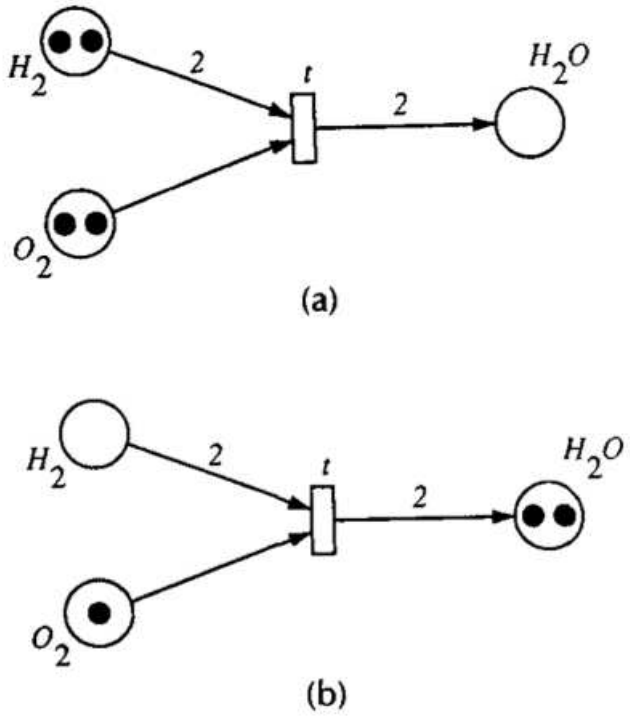


FIGURE 3.2 – **Réseau de Petri à (a) l'état initial, et (b) état final.** La seule transition représentée ici est convergente. Ce réseau représente une réaction celle de la formation d'eau à partir de molécules de dioxygène et de dihydrogène. Les chiffres étiquetant les arêtes renseignent la stœchiométrie de la réaction.

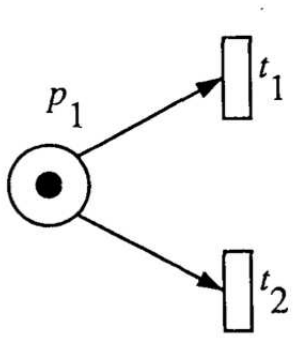


FIGURE 3.3 – **Réseau de Petri avec une connexion de sortie divergente.** L'unique jeton en p_1 passera soit par la transition t_1, soit t_2 lors de l'exécution du système.

3.4 Réseaux d'associations (ou de pertinence)

Ces réseaux n'ont pas de support mathématique, mais sont construits sur des bases statistiques. C'est pourquoi on les dénomme « réseaux statistiques ». Le principe de ces réseaux proposés par [Butte et Kohane, 2000] est de construire le modèle en évaluant un score pour chaque paire de variables, par exemple : l'expression des gènes. Le score peut être une corrélation :

$$cor(X,Y) = \frac{cov(X,Y)}{\sqrt{Var(X)Var(Y)}} \tag{3.1}$$

dans le cas où les variables sont discrètes, l'expression de la variance est :

$$Var(X) = \sum_i p(x_i)(x_i - \bar{x})^2 \tag{3.2}$$

et de la covariance :

$$Cov(X,Y) = \sum_i \sum_j p(x_i,y_j)(x_i - \bar{x})(y_j - \bar{y}) \tag{3.3}$$

ou, comme cela a été initialement proposé, un critère d'information mutuelle, différence entre l'entropie sommée des deux variables et l'entropie conjointe du patron d'expression des deux variables :

$$IM(X,Y) = H(X) + H(Y) - H(X,Y) \tag{3.4}$$

le critère d'entropie correspondant à :

$$H(X) = -\sum_i p(x_i) log_2(p(x_i)) \tag{3.5}$$

et d'entropie conjointe :

$$H(X,Y) = -\sum_i \sum_j p(x_i,y_j) log_2(p(x_i,y_j)) \tag{3.6}$$

Le critère d'information mutuel, peut être préféré au coefficient de corrélation lorsque la nature des relations n'est pas linéaire.

Une fois que le score pour tous les couples de gènes est obtenu, il est possible de représenter un graphe complet avec le score affecté à chaque nœud. Il suffit ensuite de choisir un seuil pour le score à partir duquel on supposera la réalité biologique de l'association. Dans la représentation, on pourra alors éliminer toutes les arêtes dont le score est inférieur à ce seuil.

Cette méthode de reconstruction de réseau a l'avantage de la simplicité, et semble convenir à notre problématique mettant en jeu des relations entre gènes (donc des entités de même nature). De plus, les scores sont calculables en un temps très court.

3.5 Modèles graphiques gaussiens

Tout comme précédemment, ce sont des « réseaux statistiques ». Les Modèles Graphiques Gaussiens [Whittaker, 1990] sont basés sur le calcul de coefficients de corrélation partielle, qui prennent en compte les autres variables du système : deux variables qui ne sont pas partiellement corrélées sont conditionnellement indépendantes, ajoutant ainsi une valeur informative importante, différentiant ce modèle du modèle décrit dans le paragraphe 3.4.

Les coefficients de corrélations partielles ρ_{ij} peuvent être, d'après [Edwards, 2000], calculés via l'inversion de la matrice des corrélations C :

$$\rho_{ij} = -\frac{C^{-1}_{ij}}{\sqrt{C^{-1}_{ii} C^{-1}_{jj}}} \tag{3.7}$$

La difficulté principale étant d'estimer la matrice des corrélations non-biaisée, dans le cas où le nombre de paramètres est bien plus important que le nombre de mesures (dit « $p >> n$ »). Ce problème, aussi appelé de « surparamétrisation » est très souvent rencontré lorsqu'on travaille avec des types de données dits « omique », en particulier le transcriptome tant son exploration à l'aide de puces à ADN est mise en œuvre. En effet, il faudrait 30000 réplicats pour arriver à un ratio $p = n$, et encore, on ne serait toujours pas dans un cas statistiquement parlant idéal ($n >> p$). Ce problème qui mobilise beau-

coup d'énergie peut être résumé par le fait que dans notre contexte expérimental, des corrélations entre intensités d'expression de gènes seront obligatoirement considérées comme non-nulles simplement par le fait du hasard (la surparamétrisation n'est pas un problème propre à la recherche de corrélations). Et d'un point de vue quantitatif, les valeurs absolues de ces corrélations auront tendance à être surestimées. C'est pourquoi, [Schafer et Strimmer, 2005] proposent une estimation sans biais de cette matrice par « rétrécissement », ce qui veut dire que les valeurs de corrélations seront diminuées en fonction de l'estimation d'un paramètre de rétrécissement $\hat{\lambda}^\star$:

$$C^\star_{ij} = \begin{cases} 1 & \text{si } i = j \\ C_{ij}.min(1, max(0, 1 - \hat{\lambda}^\star)) & \text{si } i \neq j \end{cases} \tag{3.8}$$

qui se calcule ainsi :

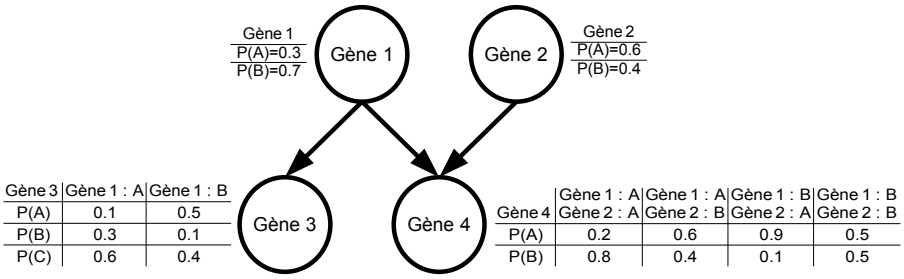
$$\hat{\lambda}^\star = \frac{\sum_{i \neq j} \widehat{Var}(C_{ij})}{\sum_{i \neq j} C^2_{ij}} \tag{3.9}$$

Cette correction de la matrice des corrélations a de plus le mérite de rendre la matrice définie positive, se trouvant être une condition d'inversibilité d'une matrice symétrique (comme la matrice des corrélations), étape nécessaire au calcul de la matrice des corrélations partielles.

Les Modèles Graphiques Gaussiens ont donc un avantage indéniable sur les Réseaux de Pertinence, car les variables conditionnellement indépendantes ne sont pas directement reliées. Ils sont néanmoins un peu plus difficiles à mettre en œuvre.

3.6 Réseaux Bayésiens

Les Réseaux Bayésiens sont un formalisme graphique définissant et simplifiant une loi conjointe de probabilités d'un modèle. C'est donc un modèle à la fois graphique et probabiliste. Les variables ont une distribution de probabilité conditionnées par l'état d'autres variables du modèle, si ces dernières relient les premières par des arcs (représentés par des flèches). A ces variables représentées par des nœuds dans le



Gène 3	Gène 1 : A	Gène 1 : B
P(A)	0.1	0.5
P(B)	0.3	0.1
P(C)	0.6	0.4

Gène 4	Gène 1 : A Gène 2 : A	Gène 1 : A Gène 2 : B	Gène 1 : B Gène 2 : A	Gène 1 : B Gène 2 : B
P(A)	0.2	0.6	0.9	0.5
P(B)	0.8	0.4	0.1	0.5

FIGURE 3.4 – **Représentation d'un Réseau Bayésien constitué de 4 variables.** Pour chacune d'entre elles, une table de probabilités conditionnelles est associée. Cette table précise les probabilités de chacun des états possibles de la variable conditionnées par les états des variables directement ascendantes.

graphe sont donc associées des distributions de probabilités conditionnelles. Les arcs indiquent simplement une dépendance statistique entre deux variables. Les Réseaux Bayésiens constituent un modèle graphique très en vogue depuis qu'ils ont été définis par [Pearl, 1988]. La section recherche de Microsoft a par ailleurs beaucoup contribué au développement de méthodes basées sur les Réseaux Bayésiens dans le milieu des années 1990 [Heckerman *et al.*, 1995, Chickering, 1995, Heckerman, 1997, Heckerman, 1998]. [Friedman *et al.*, 2000] a introduit cette modélisation dans des problématiques de biologie dans le but d'inférer des réseaux de régulation à partir de puces à ADN. Les variables dans un Réseau Bayésien peuvent être discrètes ou continues. Les Réseaux Bayésiens sont par essence statiques, mais ils sont facilement extensibles à des problèmes de modélisations dynamiques [Dean et Kanazawa, 1989, Dean et Wellman, 1991].

Faisant l'objet d'une attention particulière dans ce travail de thèse, les Réseaux Bayésiens seront abordés en détail dans le chapitre suivant.

3.7 Comparaison entre ces modèles graphiques

3.7.1 Généralités

On a décrit les principales caractéristiques de modèles graphiques parmi les plus utilisés en bioinformatique. Les principaux traits de ces modèles sont résumés par le tableau 3.1.

A propos de la dénomination, un peu trop générale des Modèles Graphiques Gaus-

TABLE 3.1 – **Comparaison de quelques modèles en réseaux.** Les caractéristiques comparées sont la propriété statique ou dynamique et l'encodage des indépendances conditionnelles.

Caractéristique	Rés. booléen	Rés. de Petri	RN	GGM	Réseau Bayésien
Statique/Dyn	Dyn	Dyn	Stat	Stat	Stat [a]
Ind. cond.	-	-	non	oui	oui

a. il existe une extension des Réseaux Bayésiens, qui est dynamique

siens. Un Réseau Bayésien est un modèle graphique, on peut aussi considérer qu'il est gaussien dès lors qu'il encode des variables suivant une distribution Normale. Cependant les « Modèles Graphiques Gaussiens » au sens strict décrits ici (dénommés « Graphical Gaussian Model » ou « GGM » dans la littérature internationale) sont des modèles graphiques non orientés. Cela est la principale différence du point de vue graphique avec les RB. Une conséquence directe de cela est l'interprétation du graphe en terme de dépendances. En effet, si on s'intéresse à la dépendance conditionnelle, un GGM étant représenté par un graphe non-orienté, il n'est pas possible de faire la distinction entre indépendance et dépendance conditionnelle tel que cela est illustré plus loin par la figure 4.1 avec les Réseaux Bayésiens. Avec les GGM, on peut dire que si deux variables sont reliées par une arête dans le graphe, elles sont dépendantes conditionnellement à toutes les autres, ce qui est déjà un pas dans la prise en compte de la complexité du problème par rapport aux Réseaux de Pertinence. La notion de dépendance conditionnelle introduite par les GGM est globale, alors qu'elle se trouve à une résolution locale dans les Réseaux Bayésiens.

3.7.2 Quantification de leur utilisation dans le monde scientifique

L'objet de ce travail de thèse n'est pas de faire une étude bibliométrique de l'exploration de différentes méthodes graphiques dans le monde, mais il peut être intéressant d'avoir une vision globale de leur utilisation. Pour cela le nombre de résultats générés par deux moteurs de recherche est compilé, lorsqu'on y saisit en requête [1] les cinq modèles respectivement.

Les deux moteurs utilisés sont *Google scholar* qui répertorie les publications scientifiques en général et *PubMed* qui regroupe les journaux ayant trait à la biologie et à la médecine. La première constatation que l'on peut faire, c'est la très faible présence

1. Les requêtes ont été faites en anglais, sous forme de texte exact, au singulier et au pluriel

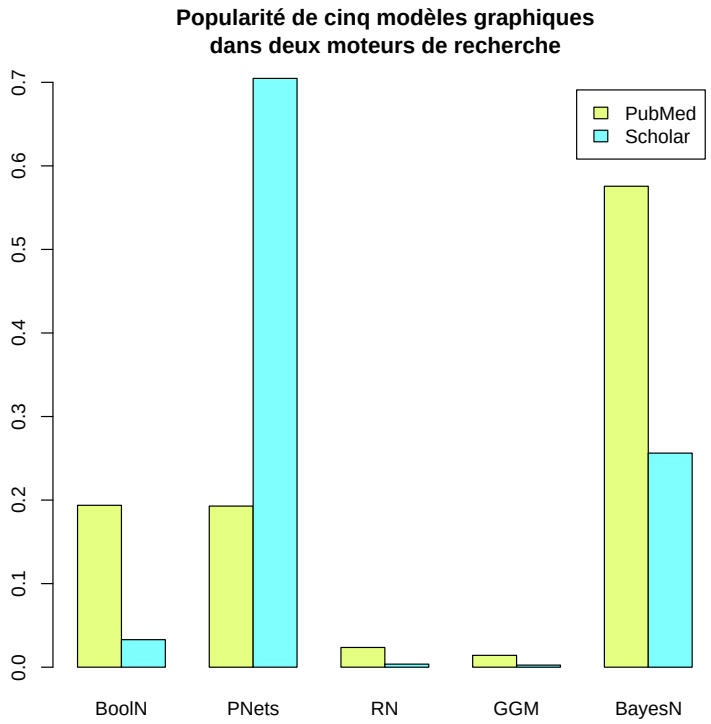


FIGURE 3.5 – **Diagramme en bâtons des fréquences de résultats dans deux moteurs de recherche pour les cinq méthodes** représentées en abscisses. BoolN : Réseau Booléen. PNets : Réseau de Petri. RN : Réseau de Pertinence. GGM : modèle graphique Gaussien. BayesN : Réseau Bayésien. Les nombres de résultats générés par les deux moteurs ont été normalisés de façon à ce que la somme pour chacun d'eux soit égale à 1

des Modèles Graphiques Gaussiens et Réseaux de Pertinence dans les résultats de ces requêtes. On note également que les Réseaux de Petri sont les plus présents dans la littérature scientifique accessible par le service de *Google* en s'octroyant près de 70% de la masse de résultats, largement devant les Réseaux Bayésiens avec environ 25%. Néanmoins, l'observation s'inverse lorsque l'on comptabilise les résultats générés par le moteur spécialisé en biologie *PubMed*. Les Réseaux Bayésiens sont présents à hauteur de 57%, tandis que les Réseaux de Petri ne représentent plus qu'un cinquième des résultats, tout comme les Réseaux Booléens qui y sont par conséquent beaucoup plus présents que dans les sciences en général (relativement au nombre total généré par chacun des moteurs).

On suppose à partir de ces observations que les Réseaux de Petri sont parmi ces méthodes les plus étudiées ou utilisées par les scientifiques en général, et que ce sont les Réseaux Bayésiens qui tiennent cette place (à moindre échelle) en biologie et en médecine. Tenant compte des 26 ans d'écart entre ces deux formalismes (Réseaux de Petri : 1962, réseaux Bayésiens : 1988), on peut même largement relativiser cette différence bibliométrique.

Les trois autres méthodes ont l'air d'être préférentiellement utilisées dans le domaine biomédical, et que même dans ces disciplines les Modèles Graphiques Gaussiens et les Réseaux de Pertinences n'y sont chacun que très peu (environ 2%) représentés.

Pour indication, il est intéressant de comparer les deux méthodes les plus étudiées en fonction de l'engouement (mesuré par le nombre de recherches dans *Google*) au cours du temps grâce à l'outil *Google Trends* (figure 3.6).

La première remarque est que les deux méthodes semblent susciter un intérêt à la baisse depuis 2004. Il faut bien-sûr prendre ces chiffres avec précautions, car cela peut être biaisé par l'utilisation du moteur de recherche qui pourrait être « cannibalisée » par celle de *Google Scholar* dont l'existence est plus récente. Mais ce qui est le plus remarquable dans ce graphe, est la superposition des deux courbes. Comme si l'intérêt pour une méthode était étroitement lié à l'intérêt pour l'autre. C'est d'autant plus surprenant qu'elles sont très différentes dans leurs fondements, ainsi que dans leur application (l'une est dynamique, l'autre est statique).

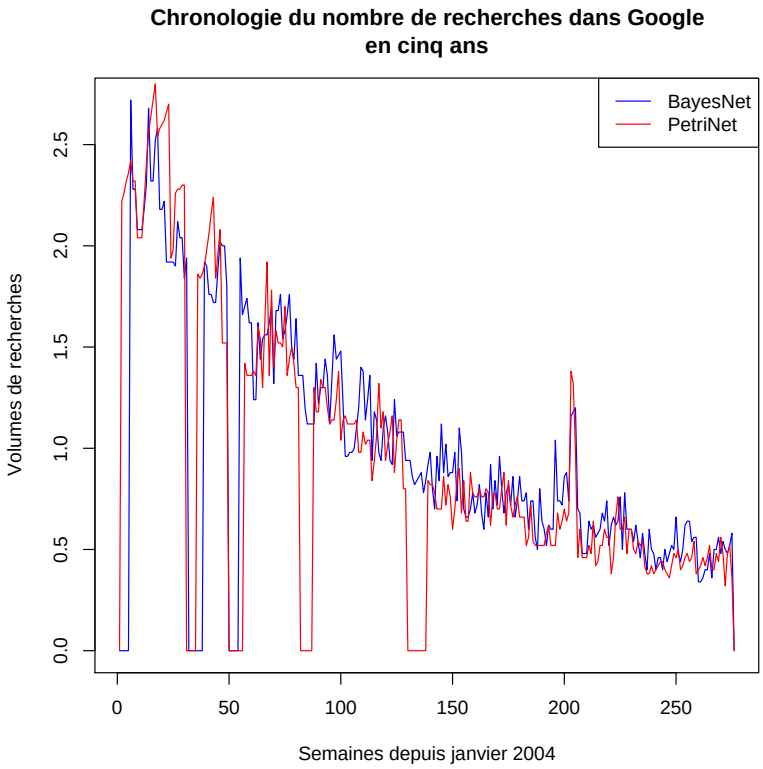


FIGURE 3.6 – **Comparaison chronologique des recherches en 5 ans.** Les données sont issues de Google Trends. La mise à l'échelle est effectuée de façon à ce que la plus grande aire sous l'une des courbes (ici les Réseaux Bayésiens) soit égale à 1 (pour les Réseaux de Petri, la valeur est quasiment identique : 0.94).

3.8 Conclusion

Nous avons vu que les Réseaux Booléens, et les Réseaux de Petri étaient des modèles dynamiques, donc adaptés par exemple dans la modélisation d'expérience de cinétique. Les Réseaux Booléens peuvent être employés à des fins de reconstruction de réseaux biologiques comme des réseaux génétiques si la dimension temps fait partie des données. Les Réseaux de Petri, parce que leur représentation est faite d'un graphe biparti et qu'ils sont capable de tenir compte de la stœchiométrie des réactions, sont particulièrement adaptés (et d'ailleurs utilisés) pour la modélisation de réseaux métaboliques. Les trois autres réseaux sont basés sur des modèles statiques. Cela est intéressant dans le présent contexte biologique car on ne s'intéresse pas ici à de données mesurées à plusieurs temps ou à plusieurs points de l'espace.

Pour choisir l'une des trois approches statiques, le compromis complexité (et temps) de mise en œuvre et finesse des résultats par les caractères informatifs des résultats obtenus a été le critère de décision. Supposant que l'originalité des résultats dépend en premier lieu de la finesse des relations obtenus dans le graphe, l'une des deux dernières options (Modèles Graphiques Gaussiens et Réseaux Bayésiens) semblait plus appropriée. Les Réseaux Bayésiens offrant des descriptions plus fines d'indépendances conditionnelles, et apportant en plus la possibilité de faire des prédictions de classe offerte par l'inférence, consistant en la mise à jour de certaines probabilités à partir de nouvelles observations, c'est cette approche qui a finalement été choisie pour modéliser graphiquement les réseaux de régulation géniques.

Le gap de complexité de mise en œuvre est cependant important. Bien que certaines conditions facilitent l'analyse à l'aide de ces modèles (type de données, base de données complètes), il n'en demeure pas moins différentes difficultés inhérentes à la dimension du jeu de données en général (beaucoup de gènes mesurés, donc beaucoup de variables) et la différence importante entre nombre n d'échantillons et nombre p de variables ($p >> n$) en particulier. Ce travail a pour objet l'exploration des Réseaux Bayésiens pour quelques applications liées à la connaissance du réseau de régulation, et propose quelques pistes dans le but de traverser la difficulté (relative) de mise en œuvre de ces derniers. Néanmoins tous les réseaux n'ont pas forcément été mis à l'écart dans ce projet : l'un des développements effectués dans ce travail, décrit dans le paragraphe 5, propose d'utiliser les Réseaux de Pertinences comme aide à la recherche de modèles en Réseaux Bayésiens.

Chapitre 4

Modélisation à partir de réseaux bayésiens

Sommaire

4.1 Réseau Bayésien : définitions et propriétés

4.1.1 Définitions

Soit un graphe dirigé sans circuit (ou « DAG » pour « *directed acyclic graph*) $G(X, E)$ où $X = \{X_1, X_2, ..., X_n\}$ est un ensemble de variables (les nœuds du graphe) et E un ensemble d'arcs. On note $\Theta = \{\theta_1, \theta_2, ..., \theta_n\}$ l'ensemble des distributions de probabilités tel que :

$$\theta_i = P(X_i|Pa(X_i)) \tag{4.1}$$

où $Pa(X_i)$ est l'ensemble des nœuds reliés à X_i par des arcs d'extrémité X_i (les nœuds parents de X_i). Alors on dit que $B(G, \Theta)$ est un Réseau Bayésien si et seulement si :

$$P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} \theta_i \tag{4.2}$$

Cette décomposition de la loi conjointe de probabilités en un produit de termes locaux est à l'origine de l'attraction suscitée par les Réseaux Bayésiens. C'est de la « compaction » de cette loi conjointe de probabilités qu'est né un nombre d'algorithmes permettant le calcul dans un système complexe probabilisé. Ces algorithmes permettent une utilisation typique des Réseaux Bayésiens : l'inférence.

Les distributions de probabilités associées à chacune des variables du modèle peuvent être soit continues, soit discrètes. En outre, un Réseau Bayésien peut à la fois contenir des variables continues et discrètes. Parce que les algorithmes sont plus nombreux, plus souvent implantés dans des librairies accessibles, et pour rester focaliser sur les objectifs du projet de recherche, n'ont été considérés et modélisés dans ce travail que des Réseaux Bayésiens à variables discrètes. Les paramètres des variables discrètes peuvent être résumés et représentés par des tableaux de probabilités conditionnées à toutes les combinaisons possibles des états des variables « parent ».

4.1.2 Propriétés

a. Indépendance conditionnelle et d-séparation

L'indépendance conditionnelle

Cette propriété est essentielle dans la compréhension et l'utilisation d'un Réseau Bayésien. Elle est le fondement de toutes les stratégies de recherche de structure de réseaux à partir d'un jeu de données. Deux variables X et Y représentées dans un graphe G sont conditionnellement indépendantes, si l'observation d'un ensemble de variables $\mathbf{Z}$ rend la variable X indépendante de Y. Cela est noté : $X \perp Y|\mathbf{Z}$. Ce qui revient à dire en terme de probabilités que : $P(X|Y, \mathbf{Z}) = P(X|\mathbf{Z})$.

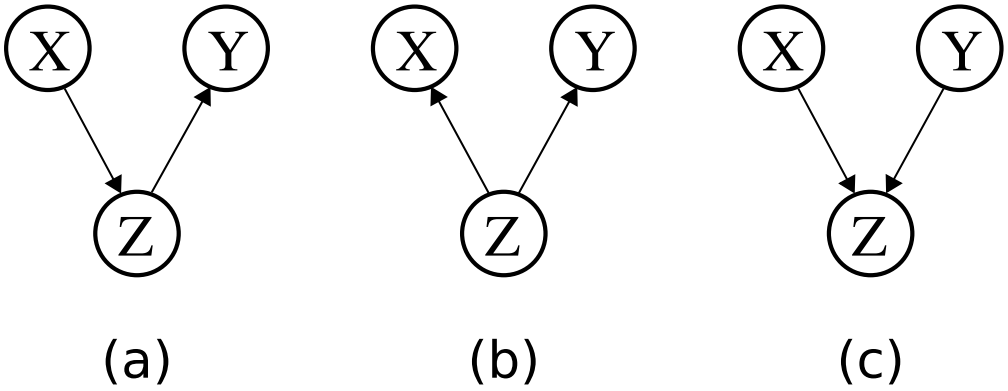


FIGURE 4.1 – Trois graphes exposant **l'indépendance conditionnelle**. (a) représente une conformation en chaîne : ce graphe illustre la même loi jointe que l'exemple divergent (b). La version convergente (c) encode une loi jointe différente. Cette dernière structure est communément appelée « V-structure ».

Dans l'exemple illustré en figure 4.1, l'ensemble des variables « intermédiaires » $\mathbf{Z}$ a été réduit à une seule variable Z par souci de simplicité. Sur le graphe bayésien représenté en (a), on observe que si X est la seule variable instanciée, cela va mettre à jour notre connaissance de Z, ce qui sera répercuté sur Y. X et Y ne sont donc pas indépendantes. Or, lorsque Z est observée, toute modification de X ne changera rien à ce que l'on attend sur Y. On comprend alors la notion d'indépendance conditionnelle. Toutes ces remarques sont applicables dans le cas (b) , qui forme donc un deuxième exemple d'indépendance conditionnelle. Cependant, le cas (c) est différent, voire inverse. En effet, si X est seule à être observée, alors la mise à jour de Z ne va pas modifier notre « certitude » de Y. Mais si Z est observée, alors l'observation de X va modifier notre connaissance de Y. Imaginons que Z représente l'humidité de l'herbe du parc de la

Tête d'Or, que X représente la mise en marche du système d'arrosage, et Z le fait qu'il ait plût à Lyon. Si Z est dans l'état « très humide » et que l'on apprend que les arrosoirs n'ont pas fonctionné depuis deux jours, alors on a de fortes présomptions sur le fait qu'il ait plût à Lyon. On parlera de *dépendances conditionnelles*. Il est à noter que ces notions sont généralisables à plus de deux enfants dans le cas (b) ou plus de deux parents dans le cas (c).

Supposons que les trois graphes de la figure 4.1 soient la représentation de trois Réseaux Bayésiens. On écrit la loi conjointe du cas (a) :

$$P(X,Y,Z) = P(X)P(Z|X)P(Y|Z) \tag{4.3}$$

D'après le théorème des probabilités conditionnelles :

$$P(X)P(Z|X) = P(X \cap Z) = P(Z)P(X|Z) \tag{4.4}$$

la loi conjointe de (a) peut être écrite ainsi :

$$P(X,Y,Z) = P(Z)P(X|Z)P(Y|Z) \tag{4.5}$$

On obtient donc la même loi conjointe que pour le cas (b). Au contraire, le graphe (c) :

$$P(X,Y,Z) = P(X)P(Y)P(Z|X,Y) \tag{4.6}$$

a une loi conjointe correspondante qui n'est pas identique aux cas (a) et (b).

En pratique, on observe des réalisations de la loi conjointe. On ne pourra donc jamais discriminer les cas (a) et (b).

La d-séparation :

Dans un Réseau Bayésien deux nœuds X et Y sont d-séparés par un ensemble de variables $\mathbf{Z}$ si tous les chemins (sans tenir compte de l'orientation) entre X et Y sont bloqués (en tenant compte de l'orientation). Concrètement, un chemin est bloqué s'il empêche X et Y d'être dépendants, donc si, en se référant à la figure 4.1 :

– dans un chemin de type (a) ou (b), on a une observation de Z ;
– dans un chemin de type (c), Z n'est pas observée.

On dit que $< G, \theta >$ satisfait la **condition de fidélité** si les d-séparations en G identifient *toutes et exclusivement* les indépendances conditionnelles dans θ, *i.e.*, $X \perp_\theta Y|\mathbf{Z}$ si et seulement si $X \perp_G Y|\mathbf{Z}$.

b. Equivalence de Markov

Les Réseaux Bayésiens ayant les mêmes lois conjointes de probabilités sont dits équivalents au sens de Markov. Cela signifie qu'ils représentent les mêmes indépendances conditionnelles. Tous les Réseaux Bayésiens ayant les mêmes nœuds et qui ont la même loi conjointe forment une classe d'équivalence. [Verma et Pearl, 1990] ont démontré que tous les graphes d'une classe d'équivalence de Réseaux Bayésiens ont en commun leur « squelette » (c'est à dire toutes les arêtes non-dirigées) et leurs V-structures.

c. Couverture et bordure de Markov

Dans un Réseau Bayésien, une couverture de Markov MB d'une variable cible T est un ensemble de variables tel que conditionnellement à la connaissance de MB, T est indépendante de toutes les autres variables du réseau.

L'ensemble des couvertures de Markov de T a un ensemble minimal appelé « bordure de Markov » de T. [Pearl, 1988] a démontré que cet ensemble minimal est représenté dans le graphe par les parents, les enfants, ainsi que les autres parents des enfants (qu'on appelera les « époux ») de la variable T. Un exemple illustré est donné en figure 4.2.

La couverture de Markov est particulièrement intéressante lorsqu'on focalise l'analyse sur une variable en particulier puisqu'elle renferme les variables les plus proches de la cible en terme de dépendance. On utilise cette couverture à plusieurs fins : de classification, de sélection de variables ou de recherche de structure locale.

4.1.3 La causalité

La causalité et sa représentation dans les Réseaux Bayésiens font débats parmi les spécialistes. Un consensus est à mon sentiment difficile, de part le caractère non formel de cette notion. Et justement, comme c'est une notion, il n'y a pas de définition mathématique de la causalité. Néanmoins la causalité est une idée généralement évocatrice. D'ailleurs, [Pearl, 1988] lui-même exprime le graphe d'un Réseau Bayésien comme un graphe causal. Et les Réseaux Bayésiens ont été nommés par beaucoup « Réseaux de

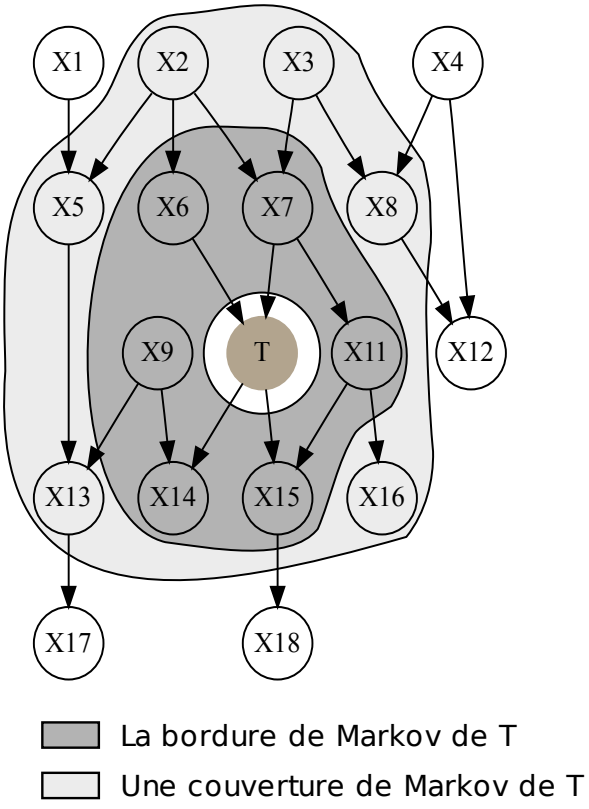


FIGURE 4.2 – Couverture et bordure de Markov

Causalité ». Quoiqu'il en soit, comme [Jensen, 1996], on peut au moins dire que la causalité a quelque chose à voir avec une action entrainant des changements d'états dans le monde réel. Une manière de mieux cerner ce qui pourrait correspondre à la notion de causalité est de prendre un exemple avec les variables corrélées A et B. Si le fait de fixer A change notre croyance en B alors on peut dire que A est une cause de B, si on ne peut décider de la variable « cause », alors on cherche une variable C pouvant être la cause de A et B, ceci se traduisant par une indépendance conditionnelle.

Prenons en exemple la régulation des gènes : si la protéine codée par le gène A est un facteur activant la transcription du gène B, et que l'on mesure la quantité de transcrits correspondants à ces deux gènes à plusieurs instants donnés, les mesures sont en principe toutes les deux corrélées. Si on inhibe l'expression du gène B, cela ne devrait pas modifier l'expression du gène A. Au contraire, si l'expression du gène A est inhibée, alors celle du gène B devrait par conséquence être très faible. Cela illustre la notion de « contrôle » inhérente à celle de la causalité. Le problème est qu'en pratique, un contrôle n'est pas toujours réalisable.

A propos du mot « croyance », les Réseaux Bayésiens peuvent aussi être nommés « réseaux de croyance » (il ne faut pas y voir pour autant quoique-ce soit de métaphysique).

4.2 Inférence dans un Réseau Bayésien

Un Réseau Bayésien est donc un cadre probabiliste formalisant les indépendances conditionnelles d'un système. La loi des probabilités conditionnelles permet de calculer des probabilités *a priori* pour toute variable du réseau. Mais ce que l'on appelle « inférence » dans un Réseau Bayésien, c'est l'estimation (ou la mise à jour) des distributions de certaines variables lorsque l'on fixe (avec une observation par exemple) des états pour d'autres variables. On peut inférer certaines variables en utilisant la loi des probabilités conditionnelles si la variable instanciée est en amont de la variable à inférer, ou le théorème de Bayes dans le cas inverse. Dans le cas où plusieurs variables sont instanciées, [Pearl, 1988] a développé des algorithmes dits de « propagation de messages » en appliquant les théorèmes des probabilités conditionnelles et de Bayes au voisinage des variables instanciées, et en propageant à leur tour les résultats de leur

mise à jour à leurs voisins et ainsi de suite. J'ai mentionné ici cette technique pour inférer dans un Réseau Bayésien à titre historique. Les cas qui posent problèmes (Réseaux Bayésiens de grande dimension) sont traités à l'aide d'heuristiques comme par exemple l'échantillonnage de *Gibbs*. Il y aurait beaucoup à dire sur l'inférence dans un Réseau Bayésien, mais ne faisant pas l'objet de ce travail de thèse, j'invite le lecteur à parcourir l'ouvrage de [Jensen et Nielsen, 2007].

4.3 Apprentissage de paramètres dans un Réseau Bayésien

Le problème de l'apprentissage de paramètres dans un Réseau Bayésien consiste à estimer les distributions de probabilités conditionnelles (donc à remplir tous les tableaux de probabilités conditionnelles) à partir d'un échantillonnage. Ce problème est divisible selon la nature des données, en particulier si toutes les épreuves ont été mesurées pour chacune des variables ou pas. En général, des stratégies de maximisation de vraisemblance selon les données sont adoptées, approches bayésiennes comprises.

4.3.1 Apprentissage de paramètres à partir d'un jeu de données « complet »

Idéalement, toutes les variables ont été observées pour chacune des expériences. Dans ce cas, l'apprentissage des paramètres se trouve être assez simple. En effet les paramètres les plus vraisemblables correspondent aux fréquences observées dans le jeu de données : c'est l'approche « fréquentiste ». On pourra aussi, si on a des raisons d'avoir une idée de ces probabilités, pondérer cette vraisemblance par un *a priori* : c'est l'approche bayésienne.

a. Approche fréquentiste

L'apprentissage statistique consiste à estimer les probabilités conditionnelles en fonction de la fréquence d'apparition des événements dans le jeu de données. C'est

l'approche par maximum de vraisemblance.

$$\hat{P}(X_i = x_k | pa(X_i) = x_j) = \frac{N_{ijk}}{\sum_k N_{ijk}} \tag{4.7}$$

avec $N_{i,j,k}$ le nombre de fois où la valeur k est observée pour X_i lorsque ses parents sont dans la configuration x_j.

b. Approche bayésienne

L'approche bayésienne est différente puisque l'on peut considérer une distribution *a priori* des paramètres. D'après le théorème de Bayes, la probabilité *a posteriori* est proportionnelle au produit de la probabilité *a priori* et de la vraisemblance :

$$P(\theta|D) \propto P(\theta)L(D|\theta) \tag{4.8}$$

Il est possible d'estimer le maximum de vraisemblance à l'aide de la méthode du « maximum *a posteriori* » (MAP) :

$$\hat{\theta}^{MAP} = argmax[P(\theta|D)] \tag{4.9}$$

Partant de l'hypothèse que les paramètres à estimer suivent une loi multinomiale, il est alors commode du point de vue écriture et calcul d'utiliser des distributions *a priori* conjuguées, en l'occurrence, la loi de Dirichlet de paramètres α (cette loi est la généralisation multivariée de la loi β, qui est donc elle-même conjuguée à loi binomiale) :

$$P(\theta) \propto \prod_{i=1}^{n} \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} \theta_{ijk}^{\alpha_{ijk}-1} \tag{4.10}$$

$$P(\theta|D) \propto \prod_{i=1}^{n} \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} \theta_{ijk}^{N_{ijk}+\alpha_{ijk}-1} \tag{4.11}$$

Ce qui conduit en passant en log-vraisemblance puis en développant à cet estimateur :

$$\hat{P}_{MAP}(X_i = x_k | pa(X_i) = x_j) = \frac{N_{ijk} + \alpha_{ijk} - 1}{\sum_{k=1}^{r_i}(N_{i,j,k} + \alpha_{ijk} - 1)} \tag{4.12}$$

La méthode du MAP n'est pas la seule méthode bayésienne. Il existe aussi une méthode approchante, celle de l'espérance *a posteriori* (EAP), où comme son nom l'indique l'espérance est calculée, et non pas le maximum de vraisemblance comme précédemment :

$$\hat{\theta}^{EAP} = E[P(\theta|D)] \tag{4.13}$$

Ce qui en développant, d'un point de vue calcul nous amène à :

$$\hat{P}_{EAP}(X_i = x_k | pa(X_i) = x_j) = \frac{N_{ijk} + \alpha_{ijk} - 1}{\sum_{k=1}^{r_i}(N_{i,j,k} + \alpha_{ijk})} \tag{4.14}$$

4.3.2 Apprentissage de paramètres à partir d'un jeu de données « incomplet »

Les données qui sont traitées dans ce travail (et beaucoup de données générées par les biotechnologies à haut-débit en général) ne sont pas sujettes à ce genre de désagrément, c'est pourquoi les méthodes ne seront pas détaillées. En effet, les données de puces peuvent être bruitées, de mauvaise qualité, des biais expérimentaux sont introduits, mais elles ne sont que très rarement incomplètes.

On notera que dans ce genre de situation, la démarche d'apprentissage des paramètres dépend de la nature plus ou moins aléatoire des données manquantes. Elles peuvent être totalement aléatoires (ne dépendent pas de la base de données), pseudo-aléatoires (dépendent des données observées) ou non-aléatoires. Dans les deux premiers cas, on a l'avantage de pouvoir estimer une distribution des données manquantes ; dans le dernier cas, il faut disposer d'autres informations.

Plusieurs approches ont été proposées pour cet apprentissage, nous ne citerons ici que la plus connue, basée sur l'algorithme EM permettant l'estimation de la log-vraisemblance des valeurs manquantes. L'algorithme EM est basé sur la répétition de deux étapes jusqu'à convergence du maximum de vraisemblance, après avoir initialisé les valeurs manquantes :

1. étape E, « espérance » : estimation des valeurs manquantes en calculant leur espérance selon les paramètres du modèle (ces valeurs sont initialisées aléatoirement lors de la première itération) ;
2. étape M, « maximisation » : estimation des paramètres par maximum de vraisemblance (de la même façon qu'avec les données complètes)

Une version de l'algorithme EM peut être adaptée à l'approche fréquentiste ou à l'approche bayésienne.

4.4 Apprentissage de la structure d'un Réseau Bayésien

Jusqu'à présent on a considéré que la structure du Réseau Bayésien (*i.e.* le graphe) était connue. Tandis que cette situation en biologie est rare en pratique, dans les autres cas on peut s'intéresser à la recherche de cette structure en faisant l'hypothèse que le système peut être modélisé par un Réseau Bayésien. On suppose alors que les variables choisies pour le systèmes suffisent à représenter toutes les indépendances du système biologique (c'est la **condition de suffisance causale**), et qu'il a tous les éléments dans les données pour les retrouver (**condition de fidélité**). Dès que le nombre de variables est conséquent, il est très rare qu'un graphe soit connu. La recherche d'un graphe bayésien peut être à dessein utilisée pour faire de l'inférence dans le graphe, mais la « simple » obtention de ce graphe peut aussi être un but en soi. Il est effectivement souvent intéressant d'avoir une image des relations entre variables qui prend en compte les propriétés d'indépendance conditionnelle inhérentes exclusivement à la modélisation par Réseaux Bayésiens. Retrouver une telle structure promet donc des résultats originaux.

Il faut néanmoins être conscient de deux limitations, la première étant liée directement au nombre dit « super-exponentiel » de graphes dirigés sans circuit en fonction du nombre de variables. Ceci peut être calculé grâce à l'expression de [Robinson, 1973, Robinson, 1977] :

$$G(n) = \sum_{k=1}^{n} (-1)^{k+1} \binom{n}{k} 2^{k(n-k)} G(n-k) \tag{4.15}$$

Par exemple, le nombre de graphes possibles avec trois nœuds est 25, sept nœuds environ 1 milliard, dix nœuds environ 4×10^{18}. A partir d'un certain nombre restreint de variables, il est impossible de parcourir ou de tester de façon exhaustive tout l'espace des graphes. En conséquence, cela constitue une limitation combinatoire, dont le contournement (difficile) passera par la limitation de l'espace de recherche. La deuxième limitation est liée à l'information au mieux contenue dans un jeu de données statiques. On ne peut pas déduire d'un tel jeu de données l'orientation des arcs. On peut cependant détecter certains d'entre-eux, en s'intéressant aux propriétés d'indépendance conditionnelle. En effet, rechercher une structure de Réseau Bayésien passe par l'estimation de sa loi conjointe de probabilité, et nous avons montré avec la figure 4.1 et les équations 4.1, 4.2 et 4.3 que la topologie (c) supportait une loi conjointe différente des structures (a) et (b). En exploitant ces différences, on est capable d'obtenir un graphe partiellement dirigé, même à partir de données statiques.

Il existe deux familles de méthodes d'apprentissage de la structure d'un réseau. La première est basée sur des tests d'indépendance (approche dite « sous contrainte »), la deuxième est basée sur le calcul d'un score.

4.4.1 Approche « sous contrainte »

Une première méthode pour rechercher la structure d'un réseau est de tester l'indépendance entre toutes les paires de variables du jeu de données. A partir d'un graphe complet, on décide alors de supprimer une arête non-dirigée entre les deux variables testées si le test les déclare significativement indépendantes ; puis on modifie le graphe avec des tests d'indépendance entre deux variables, conditionnés à un ensemble de variables tierces : l'indépendance conditionnelle entre deux variables est testée en appliquant le principe de la *d-séparation*.

Les tests généralement utilisés sont basés sur la statistique du χ^2 :

a. Test du χ^2

Soient deux variables discrètes X_i et X_j, ayant respectivement la possibilité d'être dans r_i et r_j états. Par définition, X_i et X_j sont indépendantes si et seulement si :

$$\forall x_i, x_j \quad P(X_i = x_i, X_j = x_j) = P(X_i = x_i)P(X_j = x_j) \tag{4.16}$$

Donc dans le jeu de données, si on estime $P(X_i)$ et $P(X_j)$ en calculant les fré-

quences, X_i et X_j sont en théorie indépendantes ($i.e\ X_i \perp X_j$) si pour tout état $X_i = x_i$ et $X_j = x_j$ on a :

$$E(O_{ij}) = T_{ij} = \frac{N_{x_i,.} N_{.,x_j}}{N} \tag{4.17}$$

avec $N_{x_i,.}$ le nombre de fois où X_i est observée dans la configuration x_i quelle que soit x_j et $N_{.,x_j}$ le nombre de fois où X_j est observée dans la configuration x_j quelle que soit x_i.

Le test du χ^2 d'indépendance compare le nombre d'éléments $T_{i,j}$ attendus sous l'hypothèse d'indépendance entre les deux variables X_i et X_j et le nombre d'éléments observés dans le jeu de données :

$$O_{ij} = N_{x_i,x_j} \tag{4.18}$$

La statistique en fonction des valeurs observées et des valeurs attendues sous l'hypothèse d'indépendance est calculée de la façon suivante :

$$S = \sum_{i=1}^{r_i} \sum_{j=1}^{r_j} \frac{(O_{ij} - T_{ij})^2}{T_{ij}} \tag{4.19}$$

L'approximation de la loi S par une loi de χ^2 permet la construction d'un test de l'indépendance entre X_i et X_j. Si la valeur observée est inférieure au seuil du test choisi en fonction d'un risque de première espèce α, on accepte l'hypothèse d'indépendance entre les deux variables. Une arête est alors ajoutée ou retirée du graphe en cours de construction selon la conclusion du test.

Un autre test très couramment utilisé dans ce contexte est basé sur la *G-statistique*, la statistique observée explicitée ci-après est supposée suivre comme précédemment une loi de χ^2. Il faut remarquer que le risque de première espèce α est le risque de mettre, à tort, une arête entre X_i et X_j : un risque faible correspond donc à une estimation avec peu d'arêtes.

$$G^2 = 2 \sum_{i=1}^{r_i} \sum_{j=1}^{r_j} O_{i,j}.ln \frac{O_{ij}.N}{T_{ij}} \tag{4.20}$$

Celle-ci présente l'avantage par rapport à la statistique S d'une meilleure robustesse.

Jusqu'à cette étape, le principe de la recherche de structure ne diffère pas d'une démarche de recherche de réseau de pertinence (introduit dans la section 3.4). La plus-value de l'approche par Réseau Bayésien en terme d'information contenue dans le graphe est apportée par l'étape suivante, grâce à l'application du principe de la *d-séparation*. Il s'agit de simplement adapter le calcul du χ^2 ou du G^2 en comptant N_{x_i} et N_{x_j} pour chaque configuration d'un ensemble de variables tierces $\mathbf{X}_k$ donnée. Cela revient à tester le critère d'indépendance suivant :

$$\begin{aligned} &\forall x_i, x_j, \mathbf{x}_k, \\ &\quad P(X_i = x_i, \quad X_j = x_j | \mathbf{X}_k = \mathbf{x}_k) = \\ &\qquad P(X_i = x_i | \mathbf{X}_k = \mathbf{x}_k) P(X_j = x_j | \mathbf{X}_k = \mathbf{x}_k) \end{aligned} \tag{4.21}$$

Dans le jeu de données, dans l'hypothèse où $X_i \perp X_j | \mathbf{X}_k$, on s'attend en théorie à ce que :

$$T_{ijk} = \frac{N_{x_i,\cdot,\mathbf{x}_k} N_{\cdot,x_j,\mathbf{x}_k}}{N_{\cdot,\cdot,\mathbf{x}_k}} \tag{4.22}$$

Les valeurs observées en fonction de $\mathbf{X}_k = \mathbf{x}_k$ sont :

$$O_{ijk} = N_{x_i,x_j,\mathbf{x}_k} \tag{4.23}$$

Le calcul de S est alors :

$$S = \sum_{i=1}^{r_i} \sum_{j=1}^{r_j} \sum_{k=1}^{r_k} \frac{(O_{ijk} - T_{ijk})^2}{T_{ijk}} \tag{4.24}$$

Parallèlement, si on choisit d'utiliser la *G-statistique*, on la calcule ainsi :

$$G^2 = 2 \sum_{i=1}^{r_i} \sum_{j=1}^{r_j} \sum_{k=1}^{r_k} O_{ijk} . ln \frac{O_{ijk} . N_{\cdot,\cdot,\mathbf{x}_k}}{T_{ijk}} \tag{4.25}$$

Selon le graphe obtenu à l'étape précédente et les résultats des tests « conditionnés » on pourra modifier le graphe de plusieurs façons.

– si X_i et X_j n'étaient pas reliées par une arête, et que l'indépendance conditionnée à l'ensemble de variables $\mathbf{X}_k$ n'est pas vérifiée, on déclare une « dépendance conditionnelle », ce qui se traduit dans le graphe par une convergence de X_i et X_j vers $\mathbf{X}_k$;

– si X_i et X_j étaient reliées par une arête, et que l'indépendance conditionnée à l'ensemble de variables $\mathbf{X}_k$ est vérifiée, on les déclare conditionnellement indépendantes, ce qui se traduit dans le graphe par une divergence ou une chaîne de X_i vers X_j en passant par $\mathbf{X}_k$. Comme on ne peut pas décider entre les deux, on se limitera à dessiner les arêtes sans direction, et à supprimer l'arête entre X_i et X_j.

On obtient au final un graphe partiellement dirigé représentant la classe d'équivalence de Markov regroupant tous les graphes dirigés supports de la même loi conjointe de probabilité. On peut éventuellement chercher à le diriger aléatoirement en évitant les cycles et un changement de classe markovienne.

Cette stratégie (nommée *SGS*) a été documentée par [Glymour *et al.*, 1991]. Elle présente une limite directement liée au nombre de variables maximum que l'on va vouloir tester dans l'ensemble de conditionnement ($\mathbf{X}_k$). En effet, d'une part le nombre de tests est exponentiel en fonction du nombre de variables, ce qui pose un problème combinatoire, et d'autre part le degré de liberté devient très faible rapidement lorsque la taille de $\mathbf{X}_k$ augmente, ce qui rend les tests inapplicables. Des heuristiques « utilisables » ont alors été proposées. L'algorithme *PC* de [Spirtes *et al.*, 2000] suit sensiblement le même principe, mais on choisit l'ordre du test (*i.e.* la taille de $\mathbf{X}_k$). Une stratégie nommée *IC* pour « Inductive Causation » similaire mais inverse (on part du graphe vide, et les arêtes sont ajoutées au fur et à mesure des tests) aux stratégies *SGS* et *PC* a été développée par [Verma et Pearl, 1991].

Quelques améliorations ont été apportées par la suite à ces approches. Par exemple, l'approche *BN-PC* pour « Bayesian Network Power Constructor » de [Cheng *et al.*, 1997] initialise la recherche à partir d'un graphe obtenu par la méthode d'arbre des poids maximum au lieu d'utiliser un graphe complet ou vide. Ce graphe est l'arbre dont la somme des poids affectés à chaque arête est maximisée. Ces poids peuvent être par exemple l'information mutuelle ou la corrélation.

Enfin, des méthodes plus spécifiques puisqu'elles se focalisent sur une variable cible ont été par la suite développées, pour rechercher directement la couverture de Markov de cette cible sans rechercher tout le réseau. Cette approche est décrite dans le paragraphe 6.3, dans un cadre de sélection de variables.

4.4.2 Approches basées sur le calcul d'un score

Précédemment, on procédait à des tests d'indépendances pour ajouter ou supprimer des arêtes ou des arcs. Les approches basées sur un score sont radicalement différentes

dans leur principe. On évalue un réseau en lui associant un score généralement basé sur la vraisemblance du graphe face aux observations. La stratégie de parcours de l'espace des graphes est la composante de l'approche qui choisit et propose le réseau à l'évaluation. Pour appliquer cette stratégie, on a donc besoin d'une fonction de calcul de score (on parle parfois « d'oracle », car c'est elle qui guide le reste), et d'une stratégie de parcours de l'espace des graphes.

Tout comme pour l'estimation des distributions, le calcul du score peut être soit basé sur le calcul d'une vraisemblance, soit c'est une probabilité *a posteriori* dans le cadre bayésien. On considère ici uniquement de Réseau Bayésien à variables multinomiales : comme l'apprentissage de structure passe par un apprentissage des paramètres à la volée, on considère qu'on veut veut obtenir *a posteriori* des distributions multinomiales. A cet effet, on introduit la notion de lois conjuguées. Les scores cités ici sont soit de la famille bayésienne, soit basés sur la recherche de parcimonie. Deux qualités essentielles pour un score sont d'une part leur décomposabilité, et d'autre part le respect des classes d'équivalence.

a. Décomposabilité et équivalence d'un score

Score décomposable

Un score S est dit décomposable s'il est égal à une somme ou un produit de scores locaux s, *i.e.* calculés à partir de chacune des variables et de leur(s) parent(s).

$$S(B) = \sum_{i=1}^{n} s(X_i, Pa(X_i)) \text{ ou } S(B) = \prod_{i=1}^{n} s(X_i, Pa(X_i)) \tag{4.26}$$

Score équivalent

Une fonction de score respecte les classes d'équivalence si elle attribue le même score à tous les graphes appartenant à une même classe d'équivalence de Markov.

b. Scores parcimonieux

Les scores parcimonieux formalisent le compromis entre vraisemblance et complexité d'un modèle. En effet, si le but est uniquement de maximiser une vraisemblance, celle-ci ayant une corrélation positive avec le nombre de probabilités à estimer, cela conduirait à privilégier de façon abusive les modèles ayant le plus de paramètres au détriment de ce que l'on recherche : un bon modèle simple. C'est pourquoi les fonctions de score

parcimonieuses ajoutent un terme de pénalité à ce calcul de vraisemblance, qui croît avec la complexité du modèle. Un tel critère d'information a d'abord été introduit par [Akaike, 1974] avec le score AIC, dont l'expression dans le cadre des Réseaux Bayésiens est introduite ci-après. Avant cela, nous doit proposer un calcul de la complexité d'un modèle bayésien.

Détermination de la complexité d'un Réseau Bayésien, introduction de la dimension du réseau

La complexité d'un Réseau Bayésien est une notion pouvant faire intervenir différents critères (le degré des nœuds, le nombre de modalités d'une variable, leur fonction de distribution). Une manière simple de formaliser cette complexité est de calculer la dimension du réseau, qui peut être écrite comme une somme de dimensions de chacune des variables :

$$Dim(B) = \sum_{i=1}^{n} Dim(X_i, B) \tag{4.27}$$

La dimension de chacune des variables $Dim(X_i, B)$ est le nombre de paramètres indépendants nécessaires pour la définir. Celui-ci dépend du nombre de configurations possibles r_i de X_i, ainsi que celui de l'ensemble de ses parents q_i :

$$Dim(X_i, B) = (r_i - 1)q_i \tag{4.28}$$

avec :

$$q_i = \begin{cases} 1 & \text{si } Pa(X_i) = \varnothing \\ \prod_{X_j \in Pa(X_i)} r_j & \text{si } Pa(X_i) \neq \varnothing \end{cases} \tag{4.29}$$

Exemple : calcul de la dimension du réseau $\mathbf{B}$ de la figure 4.3.

- $Dim(A, \mathbf{B}) = (2-1) \times 1 = 1$
- $Dim(B, \mathbf{B}) = (3-1) \times 1 = 2$
- $Dim(C, \mathbf{B}) = (2-1) \times (2 \times 3) = 6$
- $Dim(D, \mathbf{B}) = (3-1) \times 3 = 6$
- $Dim(\mathbf{B}) = 1 + 2 + 6 + 6 = 15$

Cette expression de la complexité pourra être utilisée comme terme de pénalité dans

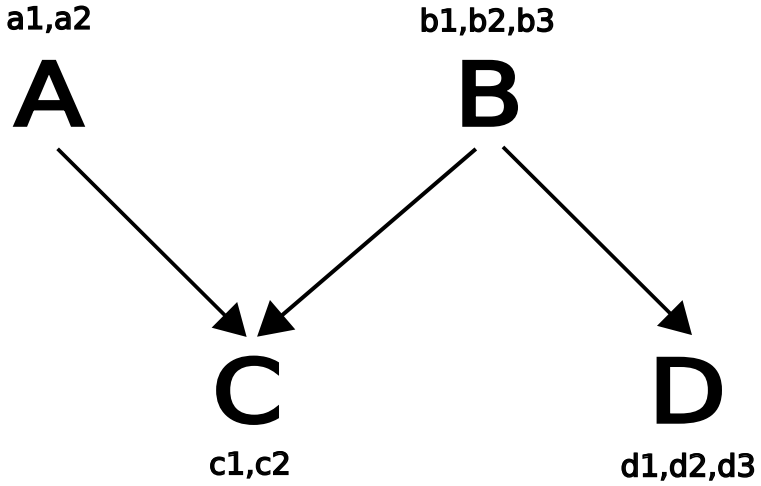


FIGURE 4.3 – **Représentation d'un Réseau Bayésien.** A et C ont deux états potentiels, B et D ont trois états potentiels.

le calcul d'un score utilisant la vraisemblance, considérant que la maximisation de celle-ci privilégie exagérément les modèles complexes (avec beaucoup de paramètres, et donc avec une grande dimension). Reste à bien équilibrer la valeur du terme de vraisemblance et celle du terme de pénalité.

Quelques scores basés sur la recherche de parcimonie

La vraisemblance des paramètres du réseau utilisée ici s'appuie sur toutes les données (l'équation 4.7 détaille l'estimation pour chacun des paramètres). Cette vraisemblance, pour un Réseau Bayésien B donné, s'écrit ainsi :

$$L(D|\Theta, B) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} \theta_{ijk}^{N_{ijk}} \tag{4.30}$$

ce qui en *log* s'écrit :

$$LL(D|\Theta, B) = \sum_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} N_{ijk}\ log\theta_{ijk} \tag{4.31}$$

et lorsque les paramètres sont estimés par maximum de vraisemblance :

$$LL(D|\Theta^{MV}, B) = \sum_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} N_{ijk}\ log\frac{N_{ijk}}{N_{ij}} \tag{4.32}$$

Score *AIC* Une première fonction de score suivant ce principe a été directement adaptée du critère d'information d' [Akaike, 1973] et a été en toute logique nommée le score *AIC*.

$$AIC(B|D) = LL(D|\Theta^{MV}, B) - Dim(B) \tag{4.33}$$

Score *BIC* Le critère *BIC* (« Bayesian Information Criterion ») proposé par [Schwarz, 1978] est une version modifiée du critère *AIC* de façon à ce qu'il converge vers un score bayésien lorsque le nombre de répétitions dans la base D tend vers l'infini. Mais contrairement aux « vrais » scores bayésiens, il ne tient pas compte de la probabilité de la structure *a priori*.

$$BIC(B|D) = LL(D|\Theta^{MV}, B) - \frac{1}{2}Dim(B)log\ N \tag{4.34}$$

Scores bayésiens

Pour définir un score bayésien, [Cooper et Herskovits, 1992] utilisent la remarque suivante :

$$\frac{P(G_1|D)}{P(G_2|D)} = \frac{\frac{P(G_1,D)}{P(D)}}{\frac{P(G_2,D)}{P(D)}} = \frac{P(G_1,D)}{P(G_2,D)} \tag{4.35}$$

Cela signifie que la variation de $P(G|D)$ en fonction de différents graphes et à partir d'une même base d'apprentissage est la même que la variation de $P(G,D)$. On peut donc s'intéresser à cette dernière valeur pour le calcul d'un score :

$$P(G,D) = P(G)P(D|G) \tag{4.36}$$

Les scores bayésiens dépendent d'un *a priori*. Ils sont aussi parcimonieux, seulement le terme pénalisant la complexité du graphe est moins explicite que pour les deux scores décrits précédemment. En effet, ce terme est pris en compte dans le calcul de la vraisemblance marginale du graphe $L(G|D) = P(D|G)$:

Vraisemblance marginale d'une structure

Les mêmes auteurs ont écrit l'expression de la vraisemblance d'une structure :

$$P(D|G) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(N_{ij} + r_{i-1})!} \prod_{k=1}^{r_i} N_{ijk}! \tag{4.37}$$

En prenant les *a priori* de Dirichlet sur les paramètres, le calcul de la vraisemblance marginale est simplifié (ceci évite l'utilisation des factorielles), cela mène au score BD :

Score BD (Bayesian Dirichlet)

$$BD = P(B) \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(N_{ij} + \alpha_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(N_{ijk} + \alpha_{ijk})}{\Gamma(\alpha_{ijk})} \tag{4.38}$$

Ce score est intéressant, car il est plus rapide à calculer que les précédents, il peut tenir compte de probabilités *a priori* sur les structures, il est décomposable, seulement il n'est pas équivalent. [Heckerman *et al.*, 1995] proposent une contrainte sur les *a priori* de Dirichlet, remédiant au problème de non-équivalence markovienne. Ils introduisent pour cela des probabilités conditionnées à leur valeur dans un graphe complètement connecté B_c, et le paramètre N' correspondant au nombre d'exemples équivalents défini directement par l'utilisateur :

$$\alpha_{ijk} = N' \times P(X_i = x_k, pa(Xi) = x_j | B_c) \tag{4.39}$$

Il existe d'autres scores, soit parcimonieux, soit bayésiens. Pour en citer quelques-uns, il existe le score MDL (Minimum Description Length), le score BDeu (simplifiant encore les (a priori) du score BDe) et le score $BD\gamma$ aussi appelé le score BDe généralisé.

c. Parcours de l'espace des graphes

Pour apprendre la structure d'un réseau à partir d'une base de donnée, le score ne suffit pas. Il faut « proposer » à ces fonctions des structures potentielles. Il suffirait donc de calculer un score pour tous les graphes possibles, puis de sélectionner la structure qui obtient le meilleur score. On a vu avec l'expression 4.15 de [Robinson, 1973] que le cas idéal, où les scores pour tous les graphes potentiels sont calculés est impossible, dès que le nombre de variables dépassait quelques dizaines. L'idée est de parcourir l'espace

des structures avec une stratégie de recherche la plus performante possible. Il existe plusieurs approches, que l'on peut combiner à des connaissances sur les variables du jeu de données.

L'utilisation d'un score décomposable est très fortement conseillée pour ne pas avoir à recalculer le score pour la totalité du graphe à chaque modification de celui-ci.

Parcours de l'espace de recherche sans information contextuelle

Heuristique de type recherche gloutonne La stratégie de la recherche gloutonne est à ma connaissance, actuellement la plus utilisée. Le principe est de partir d'un graphe quelconque (vide ou autre), de calculer le score de celui-ci, puis de le modifier. A partir d'un nœud pris au hasard, la modification proposée peut être l'ajout, la suppression, ou l'inversion d'un arc (voir l'illustration en figure 4.4), en évitant l'introduction d'un circuit, puis de recalculer le score. Si la valeur obtenue est supérieure à la valeur précédente, on garde le nouveau graphe puis on relance une nouvelle modification. On répète ces étapes jusqu'à obtenir un maximum, c'est-à-dire qu'aucune nouvelle modification ne conduit à un accroissement du score.

Une difficulté est que selon les choix opérés par l'algorithme au cours de son exploration de l'espace de recherche, la procédure conduit dans l'écrasante majorité des cas à un maximum local du score (*cf.* figure 4.5). Le contournement de ce problème est de recommencer la recherche en partant du graphe initial et de maximiser une nouvelle fois le score en proposant successivement de nouvelles structures, et ainsi de suite, sans jamais être garanti d'obtenir un maximum global.

[Chickering, 2002] propose une accélération de la procédure d'apprentissage de structure par la réduction de l'espace de recherche de l'algorithme en parcourant l'espace des graphes équivalents de Markov au lieu de tous les *DAG* : la méthode *GES* (« Greedy Equivalence Search »).

Pour éviter de « s'engouffrer » trop rapidement dans un maximum local, il est possible d'utiliser une approche qui une fois arrivée à un maximum est capable de revenir un peu en arrière, laissant ainsi la possibilité de tester d'autres chemins sans repartir de zéro. C'est la philosophie des algorithmes de recuit simulé.

Algorithmes de recuit simulé Cette fois l'analogie a été empruntée au monde de la métallurgie : après la cuisson d'un métal, celui-ci peut être refroidi lentement pour être ensuite recuit afin de modifier ses propriétés.

L'algorithme utilisé généralement suivant cette « philosophie » est celui de « Metropolis-

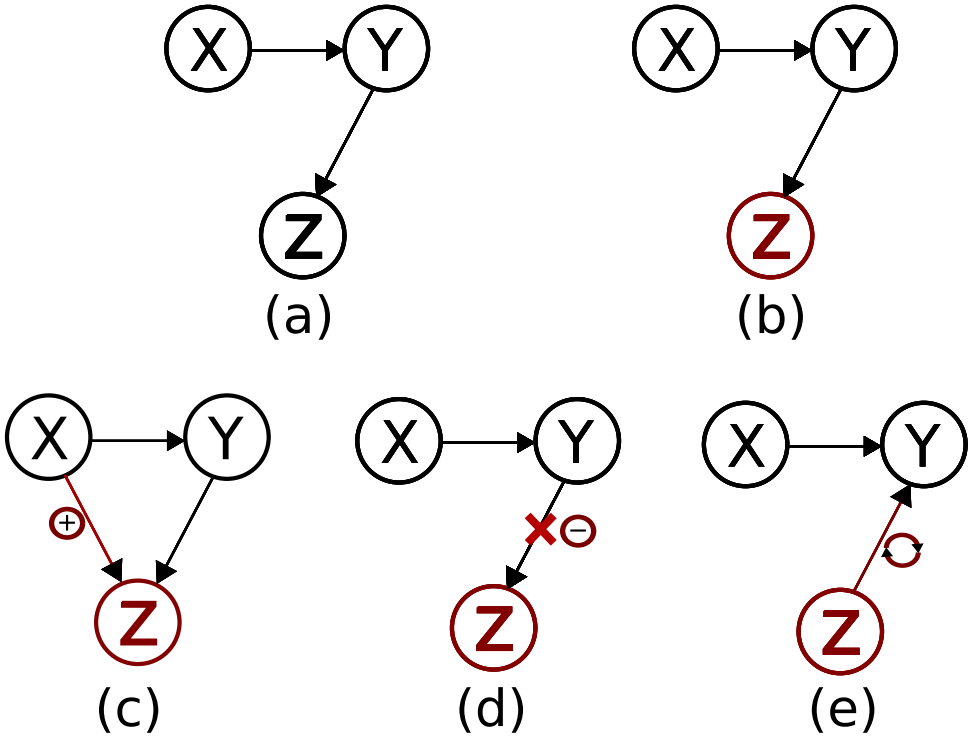


FIGURE 4.4 – **Représentation d'une itération de la recherche gloutonne** au parcours de l'espace de recherche d'un graphe bayésien. (a) représente le graphe à l'itération i. En (b), l'algorithme a choisi aléatoirement la variable Z dont la modification des connexions va être testée. (c) Addition, (d) délétion et (e) inversion d'un arc, sont trois propositions de graphe à l'itération $i + 1$ si on se focalise sur Z. Un score est attribué à chacune d'entre-elles, puis la procédure retiendra celle qui a le score le plus important pour l'itération suivante.

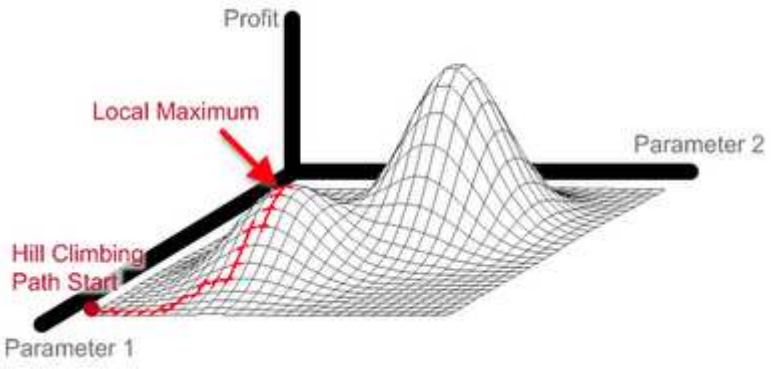


FIGURE 4.5 – **Parcours glouton de l'espace de recherche.** Selon le graphe initial, et le chemin emprunté par l'algorithme génétique, on ne peut éviter aucun maximum local.

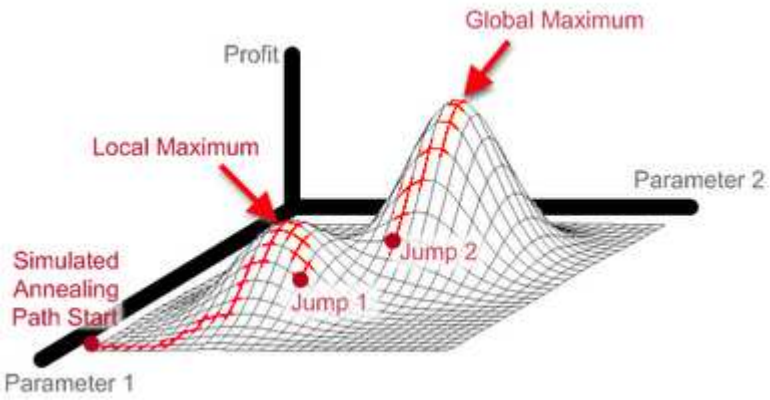


FIGURE 4.6 – **Parcours par recuits simulés de l'espace de recherche.** Le maximum local est évité grâce au « refroidissement », puis à la « rechauffe » permettant de ne pas perdre le bénéfice de tout le parcours de l'algorithme.

Hastings » [Metropolis *et al.*, 1953, Hastings, 1970]. Le comportement de cet algorithme est contrôlé par deux paramètres supplémentaires, que sont la température, et le refroidissement maximum.

Parcours de l'espace de recherche avec informations biologiques

Il peut être intéressant d'incorporer dans la stratégie de recherche des contraintes contextuelles, en l'occurrence biologiques, quand la possibilité de le faire est offerte. Voici deux exemples de mise en œuvre des Réseaux Bayésiens :

Couplage avec une recherche de motifs nucléotidiques (réseaux de régulations) Ce premier exemple [Tamada *et al.*, 2003] montre que l'on peut coupler inférence de facteur de transcription et inférence du graphe à l'aide de recherche de motifs et de la possibilité de modifier des probabilités *a priori* pendant l'apprentissage de cette structure. Les auteurs font l'hypothèse qu'après quelques itérations, si deux gènes sont proches dans le graphe, et qu'en plus ils partagent certains motifs de séquences, il est probable qu'ils soient régulés par un même facteur de transcription. La figure 4.7 résume bien la démarche.

Cette démarche est plutôt judicieuse et bien adaptée que ce soit du point de vue modélisation (mise à profit de la possibilité de renseigner des *a priori*) et d'un point de vue biologique (la recherche de facteurs de transcription en même temps que la recherche d'un réseau de régulation est plutôt pertinente).

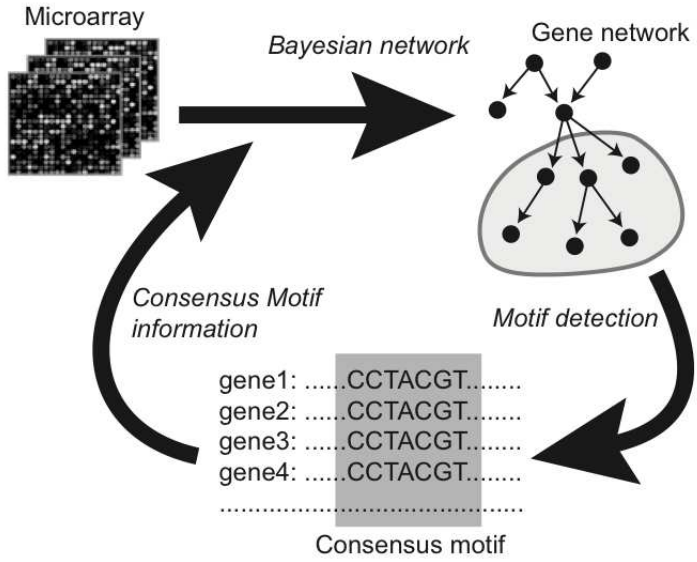


FIGURE 4.7 – **Recherche de facteurs de transcription, couplée à celle de la structure du Réseau Bayésien.** Au fur et à mesure de la reconstruction du graphe, des motifs communs sont recherchés entre séquences en amont de gènes voisins. Si tel est la situation, ces gènes sont considérés comme régulés par un même facteur de transcription et placés au même niveau hiérarchique.

Utilisation d'annotations fonctionnelles [Polanski *et al.*, 2007] calculent des log-vraisemblances de Réseaux Bayésiens en tenant compte de classes fonctionnelles présentes dans Gene Ontology, tout comme l'orientation des arcs.

4.5 Conclusion

Les Réseaux Bayésiens sont un cadre méthodologique semblant être adapté au problème de la modélisation de réseaux cellulaires à partir de données statiques. La description fine des indépendances et dépendances conditionnelles est séduisante quant à la recherche de résultats originaux, car elle outre-passe les simples relations de coexpressions que l'on peut déterminer avec des recherches de corrélations. Selon les problématiques, il peuvent être employés pour rechercher les effets de nouvelles observations dans le modèle (inférence), estimer des probabilités conditionnelles (apprentissage des paramètres), ou même rechercher une structure dans des données modélisables par de

tels réseaux. De plus, ces réseaux qui peuvent être bayésiens (!), permettent l'incorporation de probabilités *a priori* sur les paramètres ou sur la structure. Grâce à cela il est possible de préciser l'apprentissage en couplant par exemple d'autres méthodes comme la recherche de facteurs de transcription à l'aide de motifs, ou en incluant des connaissances extérieures provenant de banque d'annotations.

Troisième partie

Mise en œuvre des réseaux bayésiens : développements et résultats

Chapitre 5

Combiner deux méthodes de reconstruction

Sommaire

5.1 Introduction

Certaines stratégies d'apprentissage ont l'avantage de pouvoir être initialisées à partir d'un graphe. On pense tout de suite à l'utilisation de graphes issus d'une expertise, dans le but d'apprécier ou non la valeur ajoutée par l'apprentissage de structure, ou alors de compléter une connaissance par de l'apprentissage lorsque d'autres mesures peuvent être intégrées dans le système. On peut aussi proposer un graphe obtenu par une autre méthode, par exemple un réseau de pertinence, ou un modèle graphique gaussien. Dans ce cas, l'intérêt est d'accélérer la convergence vers le meilleur score possible en un temps raisonnable.

5.2 Données de transcriptome

5.2.1 Contexte médical

Le cancer du sein est le deuxième cancer le plus meurtrier après celui du poumon. Actuellement, un huitième des femmes est atteint d'un cancer du sein au cours de sa vie. Cette maladie présente une hétérogénéité à la fois génétique et histopathologique, et les mécanismes du développement de la maladie sont peu connus. La plupart des cancers du sein sont des carcinomes, *i.e.* des tumeurs malignes des épithelia. Moins de 1% sont des sarcomes, *i.e.* tissus conjonctifs, os, muscle ou graisse qui deviennent cancéreux.

L'amélioration de différents types de puces permet l'obtention d'un nombre tellement important d'informations sur l'expression, le polymorphisme, les interactions entre protéines que cela conduit à utiliser des stratégies d'analyses particulières, capables d'intégrer ces données en masse. On peut distinguer deux grandes familles de questions inhérentes à ce type d'expériences : (1) comment interagissent les molécules biologiques qui nous intéressent (protéines, gènes, ARNs) entre elles ? (2) peut-on prévoir le statut de certaines variables (par exemple la réponse physiologique d'un patient) à partir de la mesure d'autres variables qui sont liées ? Les Réseaux Bayésiens sont des modèles de graphes probabilistes qui peuvent servir à répondre à ce type de questions.

L'objet de ce travail est d'évaluer la capacité de l'apprentissage de structure, dans le but de capturer des interactions à partir des coexpressions mesurées sur chaque puce.

Le but est aussi de tester l'idée d'imposer une structure initiale à la procédure d'apprentissage, obtenue autrement et rapidement (voire instantanément), à partir d'une matrice de corrélations, que l'on transforme en matrice d'adjacences en fonction d'un seuil (à la manière des réseaux de pertinence décrits dans le paragraphe *cf. 3.4*).

5.2.2 Description des données explorées de cancer du sein

Ces données proviennent de l'étude publiée par [van't Veer *et al.*, 2002]. Les échantillons sont issus de tumeurs du sein primaires, chacune prélevée chez une patiente âgée de moins de 55 ans. Au total, 78 échantillons ont été prélevés :

- 34 de patientes ayant développé des métastases distantes en moins de 5 ans (1) ;
- 44 de patientes sans métastase durant les 5 ans de suivi (2).

Les puces utilisées dans cette expériences sont des puces bifluorescentes. Les 78 échantillons d'ARN ont été chacun hybridés sur une lame contre un pool de référence, fabriqué à partir d'un mélange homogène de ces 78 extractions. Deux puces ont été hybridées pour chacun des échantillons, pour appliquer la technique de coloration inversée (ou « flip-flop ») de façon à limiter le biais lié aux différentes natures de fluorochromes.

5.3 Recherche de réseaux à l'aide d'un graphe précalculé

5.3.1 Partitionnement des données utilisées dans la recherche de réseaux

Afin d'éviter les variables non-informatives, un premier filtre est appliqué pour directement éliminer des gènes dont l'intensité varie vraiment très peu entre les échantillons, et qui n'ont donc *a priori* pas d'intérêt dans cette analyse. Parmi les 25000 ARNm ciblés, 4355 ont une expression considérée significative : p-valeur $< 0,01$ pour au moins 5 tumeurs (ce sont les critères utilisés dans l'étude publiant ces données). L'apprentissage de Réseau Bayésien ci-après est donc effectué à partir de ces 4355 gènes.

5.3.2 Recherche d'un réseau de pertinence en utilisant le calcul de corrélations linaires

La matrice des corrélations est construite à partir des 4355 ARNm dont les quantités présentes ont été mesurées dans 78 échantillons. Par la suite, on applique une correction par rétrécissement de la matrice [1], comme cela est décrit dans le paragraphe 3.5.

On ne peut déduire de cette matrice directement un graphe car dans un Réseau Bayésien le graphe n'est pas valué : les arêtes ne portent pas de poids ou autre type de quantification. L'étape suivante est de déduire de la matrice de corrélation un graphe. Une façon de réaliser ceci est d'utiliser les valeurs de corrélation pour décider de la présence de chaque arête potentielle entre deux variables, après le choix d'un seuil de corrélation à partir duquel on considère l'existence ou pas d'une arête (cf. fin du paragraphe 3.4).

On cherche donc à obtenir une matrice d'adjacence dont les valeurs au croisement de chaque couple de variables, 0 ou 1, signalent l'absence ou la présence d'une arête entre les deux nœuds considérés. Naturellement, on pourrait utiliser une approche basée sur des tests d'hypothèse, évaluant la significativité de la différence entre chaque valeur de corrélation et 0. Or, lorsque cette démarche a été suivie et après avoir orienté le graphe, beaucoup de nœuds dépassaient la limite du nombre de parents maximum supporté par le programme de recherche de structure, c'est-à-dire 9.

Comme cette limitation d'origine combinatoire semblait incontournable, le critère de choix du seuil devient par conséquence une valeur de corrélation limitant le nombre de parents maximum à 9 pour chacune des variables. La notion de parent, contrairement à celle de la significativité d'une corrélation, implique la direction. Le seuil est donc calculé après l'orientation du graphe.

5.3.3 Conversion du graphe non-dirigé résultant de la matrice en un graphe dirigé sans circuit

A partir de la matrice des corrélations rétrécie, l'objectif est d'obtenir une matrice d'adjacence, assurant de représenter un graphe dirigé sans circuit. Pour cela l'idée du tri topologique de [Jarnagin, 1961] a été exploitée. Le principe de ce tri est d'ordonner les nœuds en fonction de leur position hiérarchique dans le graphe dirigé sans circuit. Il peut exister plusieurs ordres topologiques issus d'un même graphe. Une matrice dont

1. nous avons eu recours pour cela à la bibliothèque « corpcor » du logiciel *R* : http://strimmerlab.org/software/corpcor/

les lignes et les colonnes respectent l'ordre topologique est toujours triangulaire : seules les valeurs présentent dans le triangle supérieur peuvent être non-nulles. D'autre-part, toute matrice d'adjacence triangulaire correspond à un graphe dirigé sans circuit. En effet pour toute exploration en profondeur, on ne peut revenir en arrière, ce qui explique l'absence de cycle.

Comme aucun *a priori* sur la hiérarchie des gènes dans le réseau de régulation n'est disponible pour cette analyse, la matrice est triangulée tout simplement en annulant le triangle inférieur (après avoir permuté les lignes et les colonnes aléatoirement).

Le seuil de corrélation ségrégant la présence et l'absence d'un arc est choisi après triangulation de la matrice. Pour ce faire, la stratégie suivante a été implantée : en partant d'un seuil de valeur 0, tant qu'au moins une variable dispose de plus de 9 parents, la valeur $0,01$ est ajoutée à ce seuil.

5.3.4 Apprentissage de la structure du Réseau Bayésien

L'apprentissage de la structure du réseau est effectuée avec le programme *banjo*[2] développé par [Smith *et al.*, 2006]. Ce programme, écrit en langage *JAVA* implante deux stratégies de recherches : une heuristique de recherche gloutonne et la recherche par recuit simulé. Un score BDe est attribué à chaque graphe testé. En plus d'être libre et multiplateforme, ce programme a l'avantage de proposer d'imposer certaines contraintes topologiques : on peut forcer la présence ou l'absence de certains arcs et, ce qui est l'objectif ici, il est possible d'initialiser la procédure avec une structure choisie en entrée, et non le graphe vide (par défaut).

5.3.5 Apport de l'initialisation par un DAG issu d'un réseau de pertinence

Pour mesurer l'apport de cette proposition, l'apprentissage a été itéré plusieurs fois sur les 78 puces en faisant varier divers paramètres de l'heuristique :

- la présence ou non d'une structure en entrée ;
- la structure en entrée proposée n'est jamais la même, car elles sont chacune issues d'une triangulation de la matrice d'adjacence après permutation aléatoire des lignes et colonnes ;
- le temps maximum de calcul ;
- la stratégie de parcours de l'espace des graphes.

2. http://www.cs.duke.edu/~amink/software/banjo/

La métrique utilisée évaluant la cohérence du graphe bayésien selon les données pour cette comparaison est le score BDe.

La figure 5.1 montre en premier lieu l'impact de l'initialisation sur le score. Quel que soit le temps de calcul, l'avance conférée à la recherche par l'initialisation à l'aide d'un graphe de pertinence ne semble pas pouvoir être rattrapée en un temps de l'ordre de la journée par exemple (sur processeur ultraSPARC+ cadencé à $1,5GHz$ et en allouant $1,250Go$ de mémoire à la machine virtuelle de *JAVA*). On observe aussi qu'en comparaison du choix de l'initialisation, l'effet du temps (au moins à cette échelle) semble moindre. Une autre remarque est le peu de différence entre les deux stratégies de parcours de graphe dans ce cas d'utilisation. Pour un temps donné, les différents points correspondent à des exécutions avec des graphes issus de différentes permutations. On observe une petite variabilité du score selon cette initialisation, beaucoup moins importante que l'effet initialisation lui-même.

Indépendamment du critère d'initialisation de l'apprentissage de structure, la qualité globale de la reconstruction ne semble pas très bonne, ou pour le moins peu exploitable. En effet, les différents graphes issus de ces apprentissages sont très peu connectés, c'est-à-dire qu'ils sont constitués de nombreuses composantes connexes, la plupart composées de nœuds isolés, et pouvant aller jusqu'à une trentaine de nœuds. Il aurait été intéressant de confronter ces composantes connexes à des bases d'annotations, mais sans outil d'automatisation existant, et sans intérêt particulier pour une voie de régulation, il est très difficile d'investiguer parmi les centaines de composantes générées.

L'orientation aléatoire des réseaux de pertinence peuvent sembler de prime abord être un inconvénient de la méthode. Cela peut être vu autrement. On peut se servir des permutations comme un outil déstabilisant l'instanciation de façon à être plus spécifique. Autrement dit, il est intéressant de répéter un certain nombre de fois l'expérience avec des initialisations différentes, et considérer les arêtes les plus fréquentes comme plus fiables dans leur existence « réelle ». Cette approche est utilisée plus loin mais d'une façon différente, où la source de variabilité est soit la discrétisation des données, soit le partitionnement aléatoire des échantillons dans le jeu d'apprentissage.

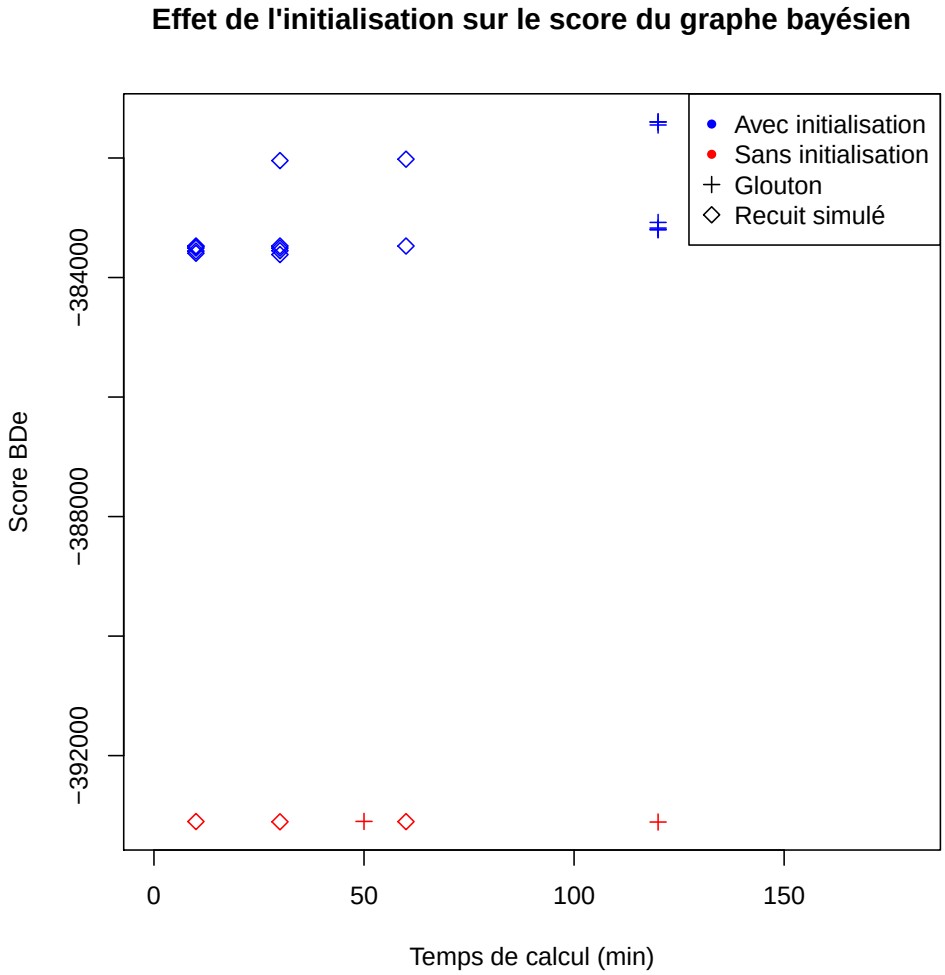


FIGURE 5.1 – **Initialisation et score bayésien.** Le score BDe est calculé sur des graphes obtenus de façons différentes : (1) la recherche est initialisée (points en bleu) ou pas (rouge) à l'aide d'un réseau de pertinence orienté aléatoirement, (2) la stratégie de recherche gloutonne (croix) ou de recuit simulé (carré), (3) le temps en abscisses.

5.4 Conclusion

Ce chapitre montre qu'il est possible de combiner plusieurs approches bénéficiant de la souplesse des Réseaux Bayésiens dans l'apprentissage de la structure. La démarche d'aide à l'apprentissage par une méthode moins fine mais instantanée semble donc être concluante. Le jeu de données utilisé pour tester cette approche a été choisi dans l'optique de conserver une cohérence dans le contexte biologique de ce projet. Néanmoins, la taille semble poser un problème d'interprétabilité. Il semble peu raisonnable de chercher à l'heure actuelle des réseaux globaux de régulation, car se pose le problème de la validation. C'est pourquoi dans les prochains chapitres, on ne considère plus la recherche de réseaux dans leur globalité. On se focalise soit sur des problèmes de classification, de sélections de variables semblant jouer un rôle prépondérant dans un contexte particulier, soit sur la recherche du voisinnage d'un gène d'intérêt.

Chapitre 6

Classer des malades selon leurs gènes

Sommaire

6.1 Introduction

En cancérologie, comme dans d'autres affections, la précision du diagnostic est un challenge perpétuel. Cela est d'autant plus vrai que pour un type de cancer (dont on aura déterminé le tissus atteint), il n'est pas toujours simple de distinguer les sous-types ou variantes.

Les données de génomique au sens large (données de transcriptome, de SNP, de nombres de copies, *etc.*) devenant de plus en plus accessibles, sont autant de matières premières permettant dans la mise en évidence *d*e structures et de mécanismes de la variante du cancer qui nous intéresse pour chaque individu atteint. On peut résumer ce problème de façon plus formelle ainsi : on dispose de nouveaux objets (mesures relatives aux gènes individuels), que l'on souhaite utiliser pour typer le plus précisément possible chaque patient. Cela peut donc se rapporter à un problème de classification. L'objet de ce chapitre, est d'examiner diverses possibilités qu'offre la modélisation par Réseaux Bayésiens pour résoudre ce type de problème.

Une manière d'aborder la question, peut-être la plus naturelle quand on connaît les potentialités de cette modélisation, est l'approche par inférence d'une variable symbolisant le statut du patient. On ajoute pour cela cette variable au modèle avant de procéder à l'apprentissage du réseau. C'est le cas d'utilisation qui est détaillé dans le paragraphe 6.2.

Les Réseaux Bayésiens peuvent donc directement servir à classer des patients. Il est également possible de mettre en œuvre les Réseaux Bayésiens en amont du problème de classification dans l'étape de sélection de gènes. Cette étape est souvent indispensable aux vues du grand nombre de gènes (variables) en rapport avec le nombre d'individus. A ces fins on utilise une méthode originale dans la partie 6.3, basée sur la recherche de la couverture de Markov (*cf.* paragraphe c.) d'une variable cible, qui est dans cette configuration la classe de référence. Une fois cette étape effectuée, rien n'empêche d'utiliser une procédure de classification autre que l'inférence par Réseaux Bayésiens, analyses discriminantes linéaires, ou non (ex : SVM).

6.2 Classification de patients atteints de leucémie

Avec le paragraphe 4.2, on a vu qu'il était possible d'utiliser le formalisme des Réseaux Bayésiens pour inférer les états de certaines variables que l'on ne peut pas forcément mesurer, lorsque d'autres le sont. Cette approche dans l'optique de classer des malades est mise en œuvre sur un jeu de données [Golub *et al.*, 1999] classique dans l'utilisation de données de puces d'expression dans la classification supervisée de patients.

6.2.1 Données de puces de patients atteints de leucémie aiguë

Les leucémies sont des cancers de cellules sanguines (on parle parfois de cancers du sang). Il existe plusieurs type de leucémies aiguës (forte vitesse de prolifération de cellules cancéreuses peu mâtures) selon l'origine des cellules malignes. Les leucémies aiguës lymphoblastiques (LLA) résultent de la modification de leucocytes de lignée lymphoïde (lignée évoluant vers les globules blanc) tandis que les leucémies myéloïdes aiguës (LMA) sont issues des autres leucocytes (lignées évoluant vers les macrophages, globules rouges ou plaquettes). La discrimination entre ces deux classes de leucémies passe par des analyses d'histochimie, d'immunophénotypage, d'analyses cytogénétiques et par l'expertise d'anatomopathologistes très expérimentés, le tout dans des laboratoires souvent distincts et très spécialisés, ce qui rend cette tâche lourde et coûteuse. De plus, même en suivant les bonnes pratiques, ces analyses sont toujours sujettes à des erreurs matérielles et humaines. Cette étude se propose d'explorer des données du transcriptome afin de classer les tumeurs. Les données de transcriptome ont été obtenues via des puces à ADN affymetrix de 72 patients. Le nombre de gènes ciblées par les sondes synthétisées sur ces biopuces est de 6817.

Dans le même esprit que la publication de [Golub *et al.*, 1999], l'objectif est d'utiliser ces biopuces pour caractériser le type tumoral (LLA ou LMA). Afin d'évaluer uniquement la méthode, la stratégie de sélection de gènes avant la recherche de structure du réseau de ces auteurs reste inchangée. Pour des raisons identiques, le partitionnement des données en jeu de test et jeu d'entraînement est le même :

- 38 échantillons de moëlle osseuse (27 LLA, 11 LMA) sont inclus dans le jeu de données d'entraînement permettant de construire le modèle prédicteur ;
- 34 échantillons (dont 10 issus de sang périphérique, plus facile à prélever mais rendant théoriquement le diagnostic plus difficile) permettent de tester la qualité de la classification supervisée.

6.2.2 Sélection de gènes

A partir des échantillons placés dans le jeu d'entraînement, les 50 gènes dont l'expression en ARNm est la plus corrélée avec le statut (LLA ou LMA) des tumeurs ont été conservés dans la suite de l'analyse.

6.2.3 Discrétisation des mesures d'intensité

Le modèle de Réseau Bayésien que l'on construit utilise des variables discrètes, alors que les données de biopuces sont continues. Le nombre d'états potentiels pour chaque variable est un paramètre sensible pour les temps de calcul de structures de Réseaux Bayésiens : la dimension du réseau augmente exponentiellement avec celui-ci. En conséquence, les variables sont discrétisées artificiellement en trois valeurs. Cette étape nécessite de faire des choix quant à sa réalisation. Il est en effet possible de procéder en une discrétisation :

- à intervalles constants : les intervalles de valeurs initiales se voyant chacun attribuer une valeur sont de même longueur ;
- par quantiles : la taille des intervalles est calculée de manière à ce qu'ils regroupent chacun autant d'échantillons que dans les autres ;
- globale : tous les gènes ont les mêmes intervalles ;
- locale : chaque gène a ses propres intervalles.

La stratégie de discrétisation adoptée ici est d'une part de prendre des intervalles constants car d'un point de vue biologique, rien ne justifie que la distribution d'expression d'un gène soit uniforme, et d'autre part locale, considérant que des gènes dont l'expression maximale est faible ont aussi leur importance.

D'autres stratégies de discrétisation plus fines existent. Elles sont exécutées en même temps que la procédure de classification de façon à optimiser certains indices de qualité de la classification. On ne s'est pas penché sur ce type de discrétisation « avancée » dans ce travail.

6.2.4 Apprentissage du réseau

Successivement, une procédure d'apprentissage de la structure et d'estimation des paramètres sont appliquées.

a. Apprentissage de la structure

L'apprentissage de la structure est exécuté à partir de 51 variables en entrée : 50 gènes et une variable pouvant prendre l'état LLA ou LMA. Le score BDe est combiné avec une stratégie de recherche gloutonne et le critère d'arrêt est de 24h (sur les différentes exécutions, on a atteint une asymptote au bout de 2 heures environ).

Le résultat de cette étape est visible en annexe (figure 7.8 qui représente le graphe ayant obtenu le meilleur score).

b. Apprentissage des paramètres

L'apprentissage des paramètres est effectué avec la méthode du maximum de vraisemblance, tel que cela a été décrit dans le paragraphe 4.3.1.

6.2.5 Inférence de la classe de patient : classification

Une fois le graphe et les paramètres obtenus, le Réseau Bayésien appris est complet. La classification est alors mise en œuvre grâce à la possibilité d'inférer directement l'état de la classe dans le graphe. La méthode employée pour cela est celle de l'arbre de jonction proposé par [Jensen, 1996] qui est le plus classique des algorithmes exacts dans ce type d'application.

L'estimation des paramètres et l'inférence ont été effectués à l'aide de la librairie *BNT*[1] développée par [Murphy, 2001] pour l'environnement *Matlab*[2].

On applique l'inférence sur la variable représentant la classe de la tumeur. Ainsi, la mise à jour de la probabilité correspond au critère de classification : en fonction du seuil de 50%, on classe la tumeur dont le type est inconnu en LLA ou LMA. Les résultats de cette inférence sont détaillés dans le tableau 6.1.

1. *BNT* : http://people.cs.ubc.ca/~murphyk/Software/BNT/bnt.html
2. *Matlab* : http://www.mathworks.com/

TABLE 6.1 – **Résultats de l'inférence dans le Réseau Bayésien** constitué des 50 variables les plus corrélées avec le type de leucémie aiguë. Pour chacun des 34 échantillons, on a ligne par ligne : (1) le type réel de la tumeur, la probabilité après inférence que la tumeur soit ALL (2), ou AML (3), et (4) la validité de la décision qui aurait été prise si on classait les tumeurs selon que la probabilité correspondante soit supérieure à 50% ou non.

ALL	**ALL**	ALL	ALL	ALL	ALL	ALL
0.999	0.436	0.807	0.602	1.000	0.982	0.993
0.001	0.564	0.193	0.398	0.000	0.018	0.007
OK	**ERR**	OK	OK	OK	OK	OK
ALL	ALL	ALL	ALL	ALL	**ALL**	ALL
0.956	0.996	0.711	0.797	0.816	0.213	0.981
0.044	0.004	0.289	0.203	0.184	0.787	0.019
OK	OK	OK	OK	OK	**ERR**	OK
ALL	ALL	ALL	**ALL**	ALL	ALL	AML
1.000	1.000	0.550	0.213	0.930	0.986	0.000
0.000	0.000	0.450	0.787	0.070	0.014	1.000
OK	OK	OK	**ERR**	OK	OK	OK
AML	AML	AML	AML	AML	AML	AML
0.002	0.018	0.000	0.010	0.006	0.213	0.010
0.998	0.982	1.000	0.990	0.994	0.787	0.990
OK	OK	OK	OK	OK	OK	OK
AML	AML	AML	AML	AML	AML	
0.024	0.018	0.050	0.000	0.012	0.000	
0.976	0.982	0.950	1.000	0.988	1.000	
OK	OK	OK	OK	OK	OK	

Selon le critère de seuillage à 50%, la méthode employée ne se trompe que 3 fois sur 34 échantillons testés. C'est à peu près comparable aux résultats de l'article publiant ces données (qui ne fait pas d'erreur, mais ne se prononce pas sur 5 échantillons). On a décidé ici de trancher à partir d'un seuil, mais on constate que l'inférence met à jour des probabilités qui peuvent très bien s'interpréter en tant que telles par un cancérologue. Par exemple, la deuxième colonne du tableau montre une probabilité de 56%. Avec un si faible niveau de certitude, il est conseillé de ne pas se prononcer quant au typage expérimental avant de proposer un traitement.

6.2.6 Conclusion

La possibilité d'utiliser les Réseaux Bayésiens dans une optique de classification est une stratégie attrayante. La mise en œuvre ici est payante puisque 91% des échantillons ont été bien classés. De surcroit, l'idée que cette dernière se base sur un graphe est séduisante d'un point de vue mécanistique, car elle prend en compte une structure de relations. De fait, on peut supposer que le classifieur a une dimension plus générique que des classifieurs basés sur des combinaisons linéaires entre variables. C'est pourquoi, on subodore qu'à l'aide d'un tel modèle, la classification n'est pas le résultat d'un apprentissage de la base par chœur si le modèle graphique correspond bien à une réalité biologique. Cependant, il est difficile de tester cet aspect compte-tenu des connaissances actuelles sur les réseaux cellulaires.

Remarque : quand on suit la démarche employée ici, on fait l'hypothèse que le réseau d'expression dans les tumeurs est semblable aux réseaux d'expression dans d'autres tissus. On peut néanmoins s'attendre à ce que la nature des relations entre les gènes diffère selon le type et l'environnement des cellules, que la perturbation due à la maladie fait que les connexions entre gènes ne sont pas les mêmes. Suivant cette lecture, on suppose qu'il existe autant de réseaux que de classes biologiques.

Si on voulait mettre en œuvre les techniques de classifications à l'aide de Réseaux Bayésiens tout en tenant compte de cette remarque, on pourrait abandonner la variable de classe que l'on inclue dans l'apprentissage et apprendre une structure par classe. Dans ce cas la classification pourrait être effectuée en appliquant le théorème de Bayes, c'est-à-dire en estimant la probabilité *a posteriori* de chacun des graphes après observation de nouveaux échantillons (dans le jeu de test) et classer l'échantillon test en fonction de la probabilité *a posteriori* la plus forte. Cette approche est appelée « multi-nets », et bien qu'intéressante dans son principe, n'a pas fait l'objet d'investigation dans ce projet.

6.3 Recherche d'une signature transcriptomique

Les données analysées dans cette application sont, comme dans le paragraphe 5.2 issues de patientes ayant eu un cancer du sein.

6.3.1 Contexte médical

a. Cancer du sein et traitement au tamoxifène

Comme environ deux tiers des cancers du sein sont hormono-dépendants (ont des récepteurs sensibles $ER+$ à l'œstrogène ou $PR+$ à la progestérone), et que l'hormone œstrogène est connue pour être un facteur de prolifération de la tumeur, une bonne stratégie de lutte contre ces cancers chez les patientes dont les tumeurs sont ER+ et/ou PR+ est de traiter avec un antagoniste de l'oestrogène : le tamoxifène. Ainsi, depuis trente ans, le traitement au tamoxifène chez ces patientes a permis une réduction de 51% de la récurrence (nombre de rechutes), et 28% de la mortalité. Seulement les interactions entre réseaux de signalisation souvent complexes et difficiles à appréhender expliquent que le tamoxifène peut dans certains tissus agir en tant qu'agoniste, et donc n'avoir aucun effet curatif, et peut même induire un cancer de l'utérus.

b. Données de transcriptome, de patientes traitées au tamoxifène

Les données publiées par [Chanrion *et al.*, 2008] ont été produites à partir de puces à ADN *Aminolink* issues de biopsies de tumeurs, appartenant à 132 patientes $ER+$ et/ou $PR+$. Les sondes sont constituées de 70 nucléotides, et ont été choisies de façon à cibler 21329 gènes humains.

6.3.2 Normalisation et discrétisation des intensités mesurées

Afin d'amoindrir les variabilités d'intensités qui n'auraient pas une origine biologique entre échantillons le bruit de fond mesuré autour des spots a été retiré au signal moyen de chacun des spots, puis elles ont été normalisées de façon à ce que les puces aient la même médiane. La normalisation a été effectuée à l'aide de la bibliothèque *marray* issue du projet *bioconductor*.

6.3.3 Réseaux Bayésiens et sélection de variables

Statistiquement, la réduction de la dimension des variables explicatives évite le problème de sur-apprentissage. Sans réduire la dimension au préalable, les méthodes statistiques standard en classification supervisée ne sont pas très performantes. Outre les Réseaux Bayésiens, les techniques dites de régularisation, telles que les machines à support vectoriel (SVM) et les modèles linéaires généralisés régularisés, semblent résister au problème de sur-apprentissage sans avoir besoin de réduire la dimension. Un grand nombre

d'algorithmes de sélection de variables est disponible dans la littérature mais rares sont les méthodes capables d'outrepasser des ordres de grandeur de plus en plus importants dans les données, donc du « passage à l'échelle ». Parmi celles-ci, les SVM sont des classifieurs très compétitifs et l'application des SVM aux variables sélectionnées a fait l'objet de plusieurs travaux, e.g., SVM Naive Weight Rank [Guyon *et al.*, 2006], Recursive Feature Elimination [Guyon *et al.*, 2002, Rakotomamonjy, 2003], Linear Programming-SVM [Fung et Mangasarian, 2004] et Approximation of the zeRO-norm Minimization [Weston *et al.*, 2003]. Les forêts aléatoires introduites par Breiman offrent également une méthode originale pour calculer une hiérarchie de ces variables. [Ishak, 2007] et [Tuleau, 2005] proposent des approches agrégeant des méthodes de classification (CART, SVM) pour sélectionner des gènes discriminants plusieurs conditions biologiques.

Dans la littérature bioinformatique, l'importance des gènes sélectionnés est traditionnellement validée à l'aide d'un classifieur supervisés (en général : SVM, Réseau Bayésien naïf) par validation croisée. Il est commun d'utiliser des tests statistiques pour contrôler la spécificité et maximiser la précision de façon à garantir un ensemble de gènes de qualité. Or [Nilsson *et al.*, 2006] montrent qu'aucune approche fondée sur les SVM ne garantit ce contrôle. D'autres travaux récents suggèrent que les forêts aléatoires sont plus efficaces que les SVM quand aucune sélection de variables n'est réalisée en amont du classifieur [Statnikov *et al.*, 2008]. Hormis le problème du passage à l'échelle, les comparaisons des méthodes de sélection de variables obtenue avec des grands jeux de données n'est pas nécessairement valide avec peu d'individus (< 150).

De plus, la performance n'est pas le seul critère à prendre en compte, la stabilité des résultats est primordiale car les gènes sont analysés ultérieurement dans les phases de validations longues et coûteuses (validation en laboratoire, recherche de leurs fonctions dans des bases d'annotation telles que Gene Ontology, analyse fonctionnelle...). [Tang *et al.*, 2007] montre expérimentalement que SVM-RFE est très sensible au « filter-out factor » d'où son instabilité avec peu d'exemples. [Saeys *et al.*, 2008] montre que les Forêts Aléatoires sont plus robustes et propose une technique ensembliste (consensus de sélection de variables). Plusieurs approches tentent de remédier à ce problème : [Havukkala et Vanderlooy, 2007] privilégie les approches par bootstrapping ; [Hanczar et Dougherty, 2008] identifie les patients pour lesquels le diagnostic est ambiguë.

Plus généralement, [Ma et Huang, 2008] et [Hua *et al.*, 2009] montrent, sur des comparaisons expérimentales sur de grandes bases génomiques, qu'aucune méthode ne surclasse les autres dans tous les scenarii en terme de performance. De plus, [Saeys *et al.*, 2008]

montrent que stabilité des résultats de classification et précision ne vont pas nécessairement de paire. Ajoutons à cela que les méthodes évoquées (e.g., SVM) ci-dessus sont des boîtes noires pour le biologiste. Les connaissances extraites sont masquées dans les équations non-linéaires sous-jacentes au modèle. Cet obstacle limite le champ des interactions entre le « modélisateur » (bioinformaticien) et les « utilisateurs finaux » (biologistes et médecins).

6.3.4 Sélection des variables

Concluant que la stratégie d'utilisation des Réseaux Bayésiens pour résoudre un tel problème était prometteuse et novatrice, c'est celle qui a été choisie afin de rechercher des variables liées de façon non-triviale à une condition biologique particulière. Celle-ci est utilisée par un algorithme de recherche de couverture de Markov [de Morais et Aussem, 2008, de Morais et Aussem, 2010] développé au sein d'une collaboration avec le laboratoire *LIESP* de *l'Université Lyon 1*. C'est un algorithme sous contrainte (*cf.* paragraphe 4.4.1). La statistique utilisée dans le but de tester l'indépendance et l'indépendance conditionnelle entre deux variables est la *G-statistique*. Cet algorithme est une extension du travail pionnier de [Nilsson *et al.*, 2007], accélérant la procédure dans le but de gérer des données de très grandes tailles.

Cette extension, appelée *MBOR* (Markov Boundary search using the OR condition), recherche la couverture de Markov de T en appliquant dans une première phase une réduction de l'ensemble des variables à celles qui sont potentiellement parents, enfants ou époux de T. Les recherches de sous-ensembles parents-enfants et époux sont faites séparément via respectivement des tests d'indépendance et d'indépendance conditionnelles (où l'ensemble de variables formant la condition est très réduit). On obtient un paquet de variables formant une sorte de super couverture de Markov. A partir de celle-ci, on applique l'algorithme *inter-IAPC* recherchant les parents et enfants de T parmi le paquet de variables en itérant des étapes d'ajout de vrai positifs et de retrait de faux positifs en procédant à des tests d'indépendance conditionnés à l'ensemble de la couverture de Markov en cours de construction. Cela est répété tant que la couverture de Markov de T varie. Enfin, *inter-IAPC* est appliquée aux enfants de T.

Le détail de cet algorithme ainsi que d'*inter-IAPC* sont présentés en annexe 7.6. L'implantation sous *Matlab* d'*MBOR* avec une limite de 30 gènes à la couverture de la variable de classe (exprimant la rechute ou non des patientes pendant cinq ans de suivi) est exécutée 250 fois sur les intensités des 132 biopuces. Chaque exécution est effectuée sur un sous-ensemble aléatoire incluant 90% des gènes dont l'expression est

mesurée. Cela permet de générer une liste de gènes plus stable. Le résultat compilant les 250 listes obtenues est visible dans la figure 6.1.

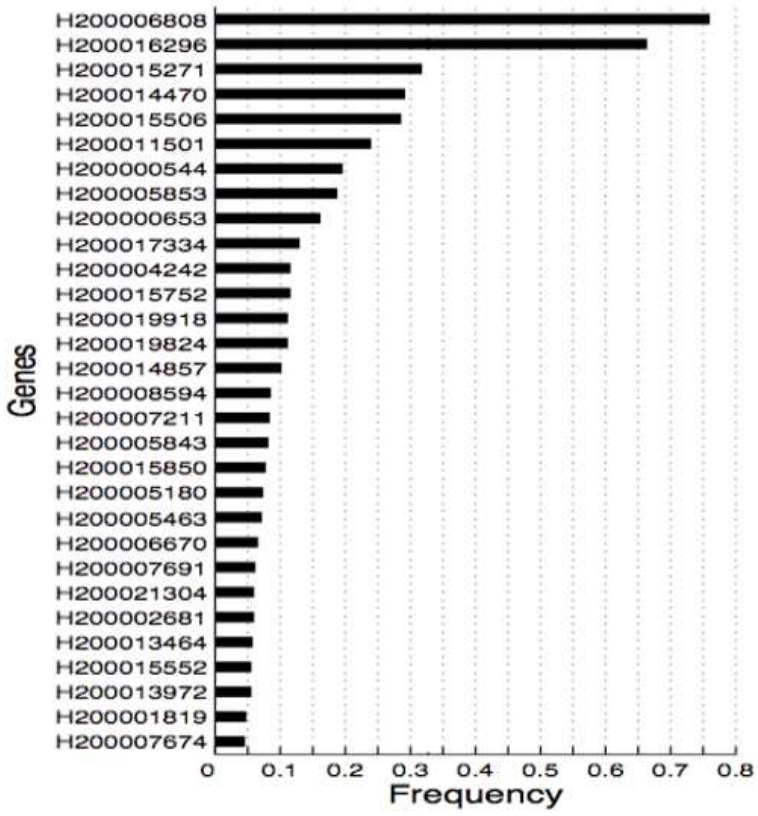


FIGURE 6.1 – Diagramme en bâtons des **30 gènes les plus souvent présents dans la couverture de Markov** de la classe de rechute ou non après 5 ans de suivi de patientes traitées au tamoxifène en adjuvant.

Le tableau 6.2 présente quelques annotations de ces 30 gènes pris dans le même ordre.

Les résultats obtenus ici sont biologiquement pertinents car certains de ces gènes ont déjà été identifiés comme ayant une expression discriminante. Parmi ceux-ci, une dizaine semblent particulièrement intéressants : **X65550**, **NM_000792**, NM_031423, **AK057339**, NM_005733, NM_001958, NM_001067, NM_002752, **NM_005654**, **NM_001238** sont connus pour être impliqués dans au moins un type de cancer, dont six (en gras) sont impliqués directement dans le cancer du sein ou ovarien (types de cancer très voisins).

Certains autres gènes sont hypothétiquement impliqués dans des mécanismes de transcription (NM_014630, NM_004891, NM_024650, BC007899) ; l'un d'entre-eux est un récepteur membranaire de la progestérone (jouant un rôle dans le cycle ovarien et la lactation), et les derniers n'ont pas de fonction connue, ou ne semblent pas liés au cancer ou à des particularités physiologiques chez la femme.

De plus, seul le gène NM_020038 a déjà été identifié par [Chanrion *et al.*, 2008] à l'aide d'une analyse discriminante améliorée : *PAM* [Tibshirani *et al.*, 2002]. Cela signifie que sur les 36 gènes de la signature issue de leur procédure de sélection, aucun des 14 gènes cités ci-dessus comme ayant une bonne cohérence selon les banques d'annotations avec le contexte médical n'est présent. La cohérence biologique et la nouveauté des résultats montre la potentielle plus-value de la méthode employée ici.

Les autres sont impliqués dans la prolifération cellulaire ou dans la mort programmée des cellules (apoptose) qui sont des mécanismes pouvant être liés à la transformation tumorale.

Cela constitue un bon argument sur l'efficacité de la méthode à la fois d'un point de vue biologique et algorithmique. L'utilisation de ces gènes pour procéder à la classification de patientes peut le confirmer.

6.3.5 Classification à l'aide de machines à support vectoriel et bayésien naïf

Afin de comparer les performances de cette signature de gènes avec les performances obtenues dans les résultats de [Chanrion *et al.*, 2008] en tant que classifieur, la liste de ces gènes a été utilisée dans deux procédures de classification : classification par SVM et Bayésien Naïf. L'utilisation du réseau en tant que classifieur n'a pas été mise en œuvre (comme cela avait été fait sur les données de leucémies). Ainsi, les classifieurs de types SVM ou Bayésien Naïf étant reconnus comme étant très efficaces en apprentissage supervisés, c'est uniquement le critère « sélection de variable » qui est au banc d'essai. Les données d'expressions pour chacun des gènes ont été binarisées par l'algorithme

CAIM [Kurgan et Cios, 2004] qui optimise dans chacune des partitions les seuils de discrétisation de façon à maximiser la fiabilité des tests d'indépendance. 23 biopuces supplémentaires d'autres patientes traitées au tamoxifène ont été classées, en prenant successivement le 1^{er}, puis le 2^{me}, 3^{me}, *etc.* jusqu'au dernier gène de la signature. Cela permet d'évaluer l'impact de chacun d'entre-eux dans la qualité de la classification. Ces classifications ont été effectuées à l'aide de l'outil *weka* [Hall *et al.*, 2009]. Les résultats de ce travail sont présentés dans la figure 6.2.

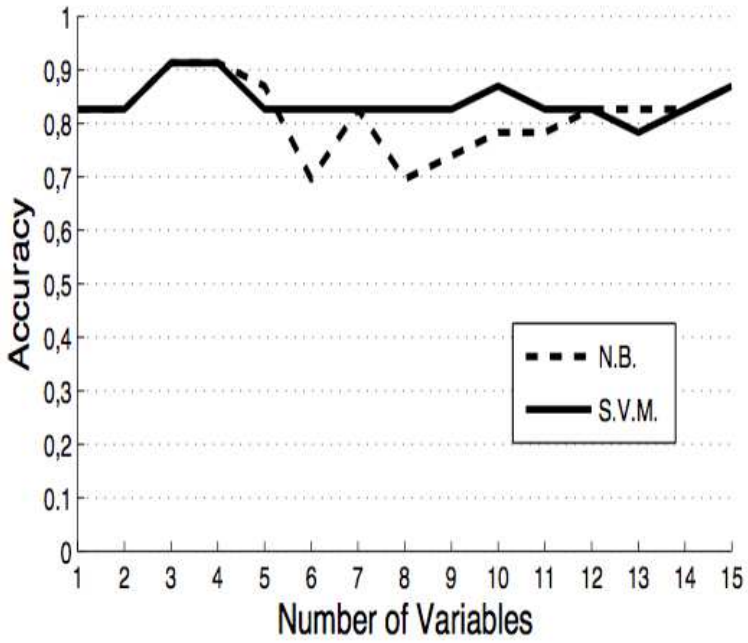


FIGURE 6.2 – **Précision des classifications en fonction des gènes (au cumul) en entrée du classifieur.** Le trait plein correspond aux résultats des SVM, les tiretés au Bayésien Naïf.

On observe qu'avec ces deux fonctions de classification, les résultats sont assez différents (globalement meilleurs à l'aide des SVM). On obtient avec ceux-ci des précisions supérieures à 82% dans la quasi-totalité des exécutions. Ils sont à comparer aux 78% de précision obtenus par les auteurs de l'article publiant ces données avec une signature comportant 36 gènes (nombre de gènes optimisant les résultats de la classification). Si on devait choisir de la même manière une signature optimisant ces résultats, on prendrait les 3 ou 4 premiers gènes de la liste, générant des classifications de 91% de précision, ce qui est largement plus satisfaisant.

6.4 Conclusion

Ce chapitre présente une mise en œuvre des Réseaux Bayésiens dans un contexte de classification. Ils sont applicables dans deux processus nécessaires pour classer des gènes : la classification en elle-même (premier exemple avec les données de leucémies aiguës), ou la sélection de gènes préalable (deuxième exemple avec les données de cancer du sein). Dans les deux cas, on a montré des aspects intéressants (en plus des résultats qui sont soient comparables, soient meilleurs que des résultats déjà obtenus avec ces mêmes biopuces). Quand ils sont mis en œuvre dans l'optique de procéder à de l'inférence, ce qui est intéressant est qu'on obtient des résultats sous forme de probabilités. Libre à l'expert de prendre une décision prenant en compte des incertitudes basées sur ces probabilités. Le graphe, peut aussi apporter des éléments de réponse à la question biologique sous-jacente à ces classifications. Un autre élément intéressant, mis en relief dans le deuxième exemple, où les Réseaux Bayésiens sont utilisés à des fins de sélection de gène : à aucun moment on a eu besoin de procéder à une étape de filtrage, contrairement à ce qui est fait dans la totalité des études sur des biopuces, où on cherche à éliminer de l'analyse des gènes dont l'expression varie peu. L'élimination de cette étape procure un gain de temps, et surtout en lisibilité et stabilité des résultats.

TABLE 6.2 – **Signature transcriptomique de 30 gènes** liée à la rechute ou non après 5 ans de suivi de patientes traitées au tamoxifène en adjuvant.

	Gene	Genbank	Description
1	MKI67	X65550	Antigen identified by monoclonal antibody Ki-67
2		AK055552	Homo sapiens cDNA FLJ30990 fis, clone HLUNG1000037
3	DIO1	NM_000792	Deiodinase, iodothyronine, type I
4	LOC81569	AK057339	Actin like protein
5	CCNE1	NM_001238	Cyclin E1
6	EEF1A2	NM_001958	Eukaryotic translation elongation factor 1 alpha 2
7	FLJ20778	NM_017957	Epsin 3
8		AK055770	Homo sapiens cDNA FLJ31208 fis, clone KIDNE2003373, moderately similar
9	EPPK1	NM_031308	Epiplakin 1
10	COL9A3	NM_001853	Collagen, type IX, alpha 3
11	HRLP5	AK055392	H-rev107-like protein 5
12	DRCTNNB1A	NM_032581	Down-regulated by Ctnnb1, a
13	HM13	BC008959	Histocompatibility (minor) 13
14		AK025020	Homo sapiens cDNA : FLJ21367 fis, clone COL03051
15		AK057821	Homo sapiens cDNA FLJ25092 fis, clone CBR00111
16	NUF2R	NM_031423	Hypothetical protein NUF2R
17	ALDH3B2	NM_000695	Aldehyde dehydrogenase 3 family, member B2
18		AK058158	Homo sapiens cDNA FLJ25429 fis, clone TST05630
19	RAB6KIFL	NM_005733	RAB6 interacting, kinesin-like (rabkinesin6)
20	FLJ11756	NM_024606	Hypothetical protein FLJ11756
21		BC011883	Homo sapiens, clone MGC :20120 IMAGE :3677070, mRNA, complete cds
22	NPL4	AK024398	Hypothetical protein FLJ20657
23	KRTHB6	NM_002284	Keratin, hair, basic, 6 (monilethrix)
24		BC004409	Homo sapiens, clone IMAGE :3638994, mRNA, partial cds
25	KIAA0186	NM_021067	KIAA0186 gene product
26		AK056211	Homo sapiens cDNA FLJ31649 fis, clone NT2RI2004078
27		AK025234	Homo sapiens cDNA : FLJ21581 fis, clone COL06796
28	NUDT2	NM_001161	Nudix (nucleoside diphosphate linked moiety X)-type motif 2
29	ABCC3	NM_020038	ATP-binding cassette, sub-family C (CFTR/MRP), member 3
30	CSAD	NM_015989	Cysteine sulfinic acid decarboxylase-related protein 2

Chapitre 7

Réseau local au voisinage d'un gène d'intérêt

Sommaire

7.1 Introduction

Une singularité des Réseaux Bayésiens est la possibilité de traiter des données statiques pour proposer un graphe reliant ces variables ayant des propriétés d'indépendance et de dépendance particulières. De plus, à l'aide de l'approche par la détection de la « couverture de Markov », on peut produire ce graphe au voisinage de gènes pour lesquels on a un intérêt particulier. Par rapport à des méthodes basées sur des similarités, l'approche proposée ici est complémentaire et permet de mettre en évidence des relations moins triviales, ce qui se traduit par la mise en évidence de nouveaux gènes pas toujours détectables autrement.

Avant cela, on a évalué comparativement cette méthode dans sa capacité à reconstruire un graphe d'interactions moléculaires réaliste, à partir de données simulées.

7.2 Performances comparées de trois stratégies d'apprentissage de structures

7.2.1 Réalisation d'un modèle de Réseaux Bayésiens à partir d'un graphe biologique et avec simulation des paramètres

Pour tester l'efficacité de l'utilisation de Réseaux Bayésiens dans l'inférence de structure de réseaux d'interactions entre gènes, un simulateur de réseaux a été développé. L'objectif est de générer à partir d'un réseau connu des données d'expressions puis d'essayer de retrouver le réseau en analysant ces données par la construction d'un Réseau Bayésien. La démarche générale du simulateur a été de choisir un réseau de gènes de topologie connue, décrite par exemple dans la base de connaissance *KEGG* [Ogata *et al.*, 1999], puis de générer des données d'expression des gènes à la fois bruitées (comme on s'y attend avec des données de puces) et cohérentes avec la topologie du réseau.

Le réseau de gènes sélectionné dans cet exemple est celui de l'apoptose : processus de mort cellulaire programmée. La plupart des cellules saines ont selon leur tissus d'appartenance une durée de vie limitée. Une altération dans le processus d'apoptose est connue pour être source de cancers.

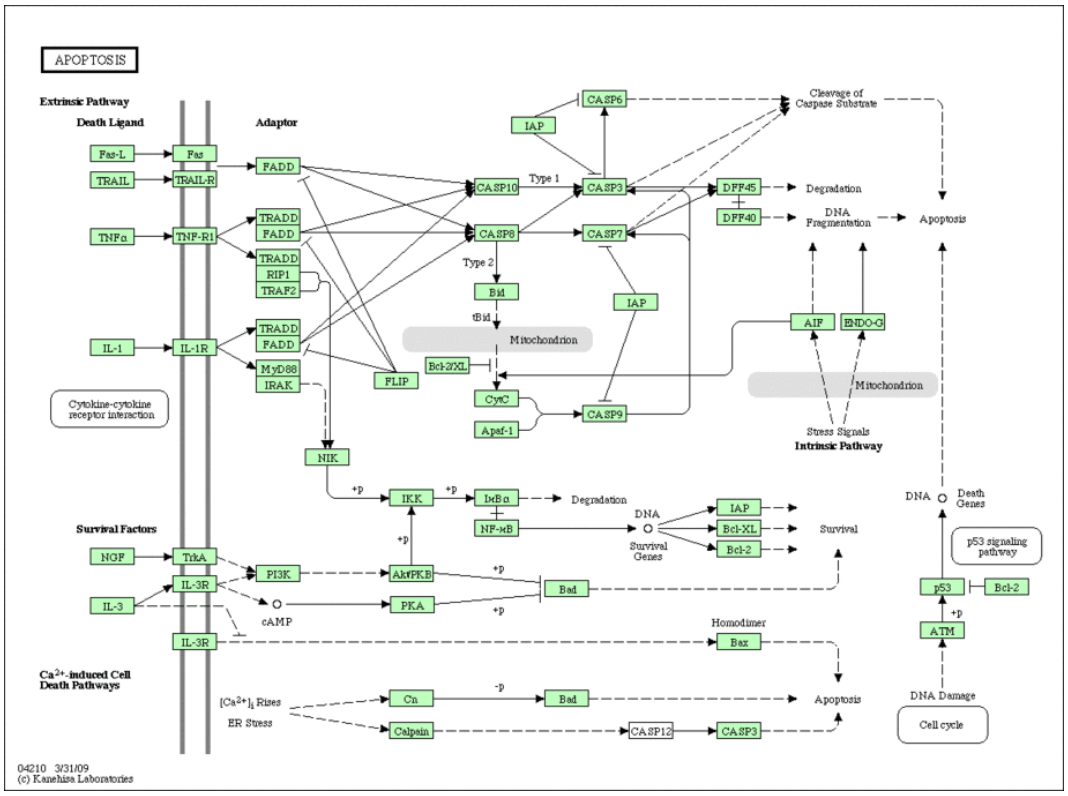


FIGURE 7.1 – **Voie de l'apoptose selon *KEGG*.** Voie humaine référencée sous l'identifiant **hsa04210**.

Dans la base *KEGG*, l'apoptose chez l'humain est référencée sous l'identifiant : **hsa04210**. Selon la localisation de l'activité des molécules de ce réseau, un même gène peut être représentée plusieurs fois (la voie de signalisation est illustrée en figure 7.1), or dans un Réseau Bayésien une variable ne peut être représentée que par un unique nœud.

Le graphe référencé dans la base de connaissance a donc été adapté, à la manière de [Calzolari *et al.*, 2007], en fusionnant les nœuds présents de façon multiple. Après cette transformation, on obtient le graphe présent en figure 7.2.

Pour que le Réseau Bayésien soit complet, les variables doivent être paramétrées (attribution des distributions conditionnelles). Le système comporte 47 nœuds. On leur donne chacun la possibilité d'être dans trois modalités, représentant les trois états de gènes :« non ou peu exprimé », « moyennement exprimé »ou « très exprimé » en ARN messagers. On fournit ensuite les distributions de probabilité de chaque variable « gène » dans chacun de ces trois états, étant données toutes les combinaisons possibles de configuration des nœuds parents.

Les variables sont discrètes, il est donc possible de représenter et renseigner ces

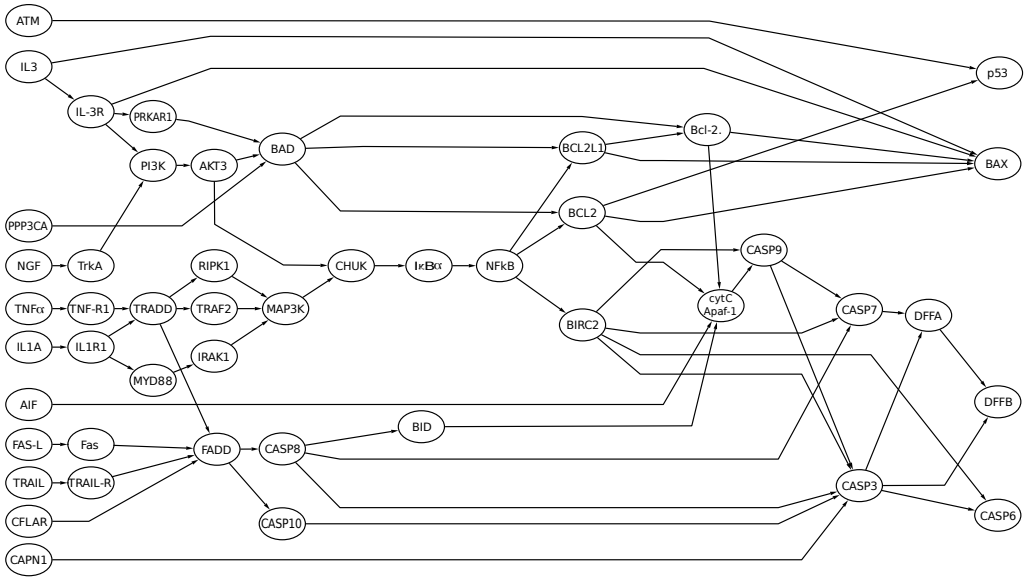


FIGURE 7.2 – **Apoptose selon KEGG**, après adaptation dans le but d'obtenir un DAG conforme aux Réseaux Bayésiens.

distributions à l'aide de tableaux de probabilités conditionnelles. Dans le but d'être fidèles avec la signification du graphe de l'apoptose, on se propose de ne pas remplir les tableaux de façon totalement aléatoire : il ne serait ni judicieux ni pertinent que les valeurs soient très proches de l'équiprobabilité traduisant une indépendance de la variable considérée avec ses parents. Dans le réseau de l'apoptose selon *KEGG* (figure 7.1), on constate que les flèches peuvent être soit en trait plein signifiant un effet activateur, soit en tiretés, traduisant une inhibition de l'activité de la protéine considérée par la protéine figurée via le nœud parent. Les probabilités générées dans les tables de probabilités conditionnelles par le simulateur tiennent compte de la nature de ces flèches, même si cette information n'est pas utilisée dans ce travail.

Il faut considérer une certaine limite à cette paramétrisation du réseau : pour chaque variable X_i, le nombre de cases t_i du tableau i à compléter est de :

$$t_i = n_i \prod_{X_j \in Pa(X_i)} n_j \tag{7.1}$$

avec n_i le nombre de modalités du nœud i. Comme pour toutes les variables le nombre de configurations possibles est constant et égal à 3, ce nombre devient :

$$t_i = 3 \times 3^{n_p} \tag{7.2}$$

Dans l'« adaptation » du réseau de l'apoptose tel qu'il est représenté en figure 7.2, on voit que les nœuds peuvent posséder de 0 (*ATM, IL3, PPP3CA, etc.*) à 5 parents (*BAX, CASP3*), cela pose le problème de l'intégration des effets des différents types de flèches sur les distributions conditionnelles. Le nombre de probabilités à renseigner varie entre 3 et 729, la somme pour tout le réseau étant de 2679. Il est alors difficile de paramétrer ce réseau manuellement. Pour le produire automatiquement, on se base sur le cas simple, où les flèches ont un effet additif : si deux flèches « activatrices » sont dirigées vers un même sommet, on considérera leur effet activateur comme plus important quand les gènes parents sont « très exprimés », que si l'un des deux seulement est « très exprimé » et ainsi de suite. Une façon de prendre en compte ces flèches de façon additive, est de représenter les modalités « non ou peu exprimé », « moyennement exprimé » ou « très exprimé » respectivement par 0, 1 et 2, et de moyenner tous les cas possibles entre les parents. On calcule ensuite une distance entre cette moyenne par combinaison de parents et de l'état de la variable considérée. Par exemple, si on se focalise sur le gène *DFFA* et ses deux parents *CASP7* et *CASP3*, on obtient pour chacun des i paramètres :

$$dist_i = \left| \frac{CASP7_i + CASP3_i}{2} - DFFA_i \right| \tag{7.3}$$

Toutes les 27 combinaisons possibles et les détails du calcul des paramètres correspondants sont dans le tableau 7.1.

Une fois la distance obtenue, on en déduit un score de similarité :

$$sim_i = (max(dist) - dist_i)^3 \tag{7.4}$$

L'expression de la similarité a été élevée au cube pour « renforcer » l'effet de l'activation ou de l'inhibition. En effet on estime qu'un gène même très peu exprimé pouvait malgré tout jouer son rôle d'activateur ou d'inhibiteur.

Afin de générer du bruit et ainsi se rapprocher des conditions expérimentales, nous ajoutons à cette valeur de similarité un terme d'erreur aléatoire, tiré d'une loi Normale de moyenne 0 et d'écart-type égal au dixième de la valeur de similarité maximale.

$$raw = sim_i + \epsilon \tag{7.5}$$

$$\text{avec} \quad \epsilon \rightsquigarrow \mathcal{N}\left(\mu = 0, \sigma = \frac{max(sim)}{10}\right) \tag{7.6}$$

Enfin, les valeurs (qui seront considérées et utilisées comme probabilités) sont ramenées entre 0 et 1 et la somme des probabilités de chaque état de l'enfant pour chaque configuration de l'ensemble des parents est ramenée à 1. Par exemple, dans le cas où *CASP7* est moyennement exprimé (état 2) et *CASP3* est très peu exprimé (état 1) la valeur de la probabilité que *DFFA* soit très exprimée (état 3) sera mise à l'échelle de la façon suivante :

$$norm_{2,1,3} = \frac{raw_{2,1,3}}{\sum_{i=1}^{3}(raw_{2,1,i})} \tag{7.7}$$

Les fonctions permettant de mettre en œuvre cette procédure de simulation de tables de probabilités conditionnelles ont été implantées en langage R.

7.2.2 Simulation de 200 échantillons respectant le modèle de Réseau Bayésien « apoptose »

La simulation de ces échantillons est produite en suivant la hiérarchie, c'est-à-dire en partant des racines, et en allant jusqu'aux feuilles du graphe dirigé. Voici un début de cas d'utilisation à partir de la figure 7.2 :

1. l'état 2 de *IL3* est tiré selon $P(IL3 = 2)$;
2. l'état 1 de *IL-3R* est tiré selon $P(\text{IL-3R} = 1 | IL3 = 2)$;
3. l'état 1 de *NGF* est tiré selon $P(NGF = 1)$;
4. l'état 2 de *TrkA* est tiré selon $P(TrkA = 2 | NGF = 1)$;
5. l'état 2 de *PI3K* est tiré selon $P(PI3K = 2 | \text{IL-3R} = 1, TrkA = 2)$;
6. et ainsi de suite.

Ces simulations ont été effectuées à 200 reprises à l'aide de la fonction *sample_ bnet* de la librairie *BNT*. On obtient au final un tableau comportant l'état 1, 2 ou 3 de chacune des variables pour les 200 échantillons.

7.2.3 Comparaison de trois procédures de recherche de topologie du réseau de signalisation de l'apoptose

Le jeu de données fabriqué a l'inconvénient de (par définition) ne pas être réel, cependant on a l'avantage de connaître le modèle en Réseaux Bayésiens sous-jacent. Le but est ici de comparer des procédures de reconstruction de graphe qui paraissent les plus adaptées à la recherche d'un réseau de régulation de l'ordre de 50 gènes. Les trois approches mises à l'épreuve ici sont les méthodes de reconstruction à partir :

- de score bayésien *BDe* associé au parcours glouton des DAG (GS) ou des classes d'équivalence de Markov (GES) ;
- de la recherche sous contrainte de la couverture de Markov (incluant dans ce contexte toutes les variables du graphes) *MBOR*. MBOR a été éxécuté avec un seuil α de la G-statistique variant dans $[0.01; 0,1]$ à chaque instanciation sur des centaines de tirages aléatoires de 90% des données.

Dans l'optique de comparer ces approches dans plusieurs conditions se rapprochant de la réalité, elles ont été appliquées sur les 50, 100 premiers échantillons (considérant que beaucoup de jeux de données de biopuces sont approximativement de cette taille), puis sur la totalité des 200 échantillons générés. Le tableau 7.2 expose les résultats de ces apprentissages de structure en terme de vrais positifs, faux positifs et faux négatifs.

	CASP7	CASP3	DFFA	dist	sim	raw	norm
1	0	0	0	0.00	8.00	8.95	0.81
2	1	0	0	0.50	3.38	4.50	0.52
3	2	0	0	1.00	1.00	0.75	0.13
4	0	1	0	0.50	3.38	2.48	0.37
5	1	1	0	1.00	1.00	1.53	0.18
6	2	1	0	1.50	0.12	-0.80	0.00
7	0	2	0	1.00	1.00	1.45	0.16
8	1	2	0	1.50	0.12	-0.01	0.08
9	2	2	0	2.00	0.00	0.04	0.07
10	0	0	1	1.00	1.00	0.68	0.12
11	1	0	1	0.50	3.38	3.10	0.38
12	2	0	1	0.00	8.00	7.44	0.67
13	0	1	1	0.50	3.38	3.61	0.50
14	1	1	1	0.00	8.00	7.84	0.66
15	2	1	1	0.50	3.38	5.37	0.60
16	0	2	1	0.00	8.00	8.98	0.71
17	1	2	1	0.50	3.38	3.22	0.39
18	2	2	1	1.00	1.00	1.81	0.23
19	0	0	2	2.00	0.00	-0.03	0.06
20	1	0	2	1.50	0.12	0.16	0.09
21	2	0	2	1.00	1.00	1.78	0.21
22	0	1	2	1.50	0.12	0.36	0.13
23	1	1	2	1.00	1.00	1.29	0.16
24	2	1	2	0.50	3.38	3.31	0.40
25	0	2	2	1.00	1.00	0.88	0.12
26	1	2	2	0.50	3.38	4.64	0.53
27	2	2	2	0.00	8.00	7.25	0.70

TABLE 7.1 – Calcul des paramètres pour un gène qui a deux parents, chacun pouvant avoir trois niveaux d'expression. dist : calcul de la distance, sim : de la similarité, raw : sim + erreur aléatoire, norm : valeur de raw ramenée entre 0 et 1.

TABLE 7.2 – **Résultats de la comparaison entre les trois méthodes d'apprentissage de structure** *GES*, *MBOR* et *GS* avec 50, 100 et 200 échantillons simulés. Le tableau de contingence à gauche expose pour chacune des méthodes le nombre vrais positifs (TP), de faux positifs (FP) et le nombre de faux négatifs (FN). Le tableau de droite montre les valeurs de sensibilité et de spécificité calculées à partir du tableau de contingence.

50	GES	MBOR	GS
TP	27	24	10
FP	22	8	36
FN	40	43	57
100	GES	MBOR	GS
TP	29	30	11
FP	4	1	2
FN	38	37	56
200	GES	MBOR	GS
TP	39	46	14
FP	0	3	3
FN	28	21	53

50	GES	MBOR	GS
Sensi	0.40	0.35	0.14
Speci	0.55	0.75	0.21
100	GES	MBOR	GS
Sensi	0.43	0.45	0.16
Speci	0.88	0.97	0.85
200	GES	MBOR	GS
Sensi	0.58	0.69	0.21
Speci	1	0.94	0.82

Les méthodes de reconstruction ne pouvant faire mieux que de proposer des graphes partiellement dirigés, on considère que la capacité de la méthode à déterminer la présence (ou l'absence) d'une arête d'une arête sans tenir compte de l'orientation est un bon résultat. L'attribution des *TP*, *FP* et *FN* est donc effectuée à partir du graphe reconstruit désorienté.

Une fois ces valeurs obtenues, on procède aux calculs de la sensibilité et de la spécificité de la méthode dans chacun des cas. La sensibilité est le pourcentage de vrai positifs détectés dans le jeu de données :

$$Sensi = \frac{TP}{TP + FN} \tag{7.8}$$

La spécificité est le pourcentage d'arêtes réellement présentes dans le modèle parmi toutes celles qui ont été détectées par la méthode :

$$Speci = \frac{TP}{TP + FP} \tag{7.9}$$

Les résultats de cette comparaison montrent une progression attendue des trois méthodes en fonction du nombre d'échantillons. On constate aussi que *GS* est dans tous les cas bien en deçà des deux autres, que ce soit en terme de sensibilité ou de

spécificité. La qualité de détection de *GES* et de *MBOR* sont déjà plus comparables.

Cette dernière bénéficie d'une spécificité « intéressante » de 75% dès 50 échantillons utilisés pour l'apprentissage (à comparer aux 55% obtenus pour *GES*). Il est intéressant de constater qu'avec seulement 50 échantillons, on peut accorder une bonne confiance à la réelle présence des arêtes détectées à l'aide de cette méthode. Avec ce nombre d'échantillons, la meilleure méthode est donc *MBOR* avec un gain de spécificité de 20 points par rapport à *GES*, malgré une perte de sensibilité de 5 points.

Avec 100 échantillons, *MBOR* surclasse de nouveau *GES* avec 2 points supplémentaire de sensibilité, et 9 points de spécificité.

S'appuyant sur 200 échantillons, les deux méthodes sont très performantes en terme de spécificité (100% avec *GES* et 94% *MBOR*). La sensibilité de *MBOR* est assez supérieure à *GES* avec 69% contre 58%.

Quelque-soit la taille de l'échantillon, on obtient globalement les meilleurs résultats avec *MBOR* qui sont bons en spécificité dès 50 échantillons, jusqu'à devenir excellents pour les tableaux de plus grande taille, et qui devient de surcroit bien plus sensible que ses concurrentes à partir de 100 échantillons. Le fait qu'*MBOR* ait ces résultats est une bonne nouvelle, car elle a en plus l'avantage d'être la plus rapide, et la seule à proposer directement la recherche de la couverture de Markov d'une variable cible sans passer par l'apprentissage de la structure entière au préalable.

Dans ce qui suit, le but est dans un premier temps de reconstruire le réseau de l'apoptose à partir de données de puces appliquées au cancer du sein, en extrayant auparavant les gènes correspondant à ce réseau cellulaire (les trois méthodes étaient donc applicables). Dans un deuxième temps, l'objectif est de rechercher un réseau au voisinage d'un gène d'intérêt (ce qui implique obligatoirement d'après l'état de nos connaissances le choix d'une méthode de type recherche de couverture de Markov).

7.3 Recherche du réseau apoptose à partir de données de biopuces de cancer du sein

Les données utilisées dans ce cas d'étude sont une nouvelle fois celles qui ont été publiées par [Chanrion *et al.*, 2008]. Afin de réaliser une étude « raisonnable », les gènes ciblés par la biopuce faisant partie du réseau cellulaire de l'apoptose selon *KEGG* ont

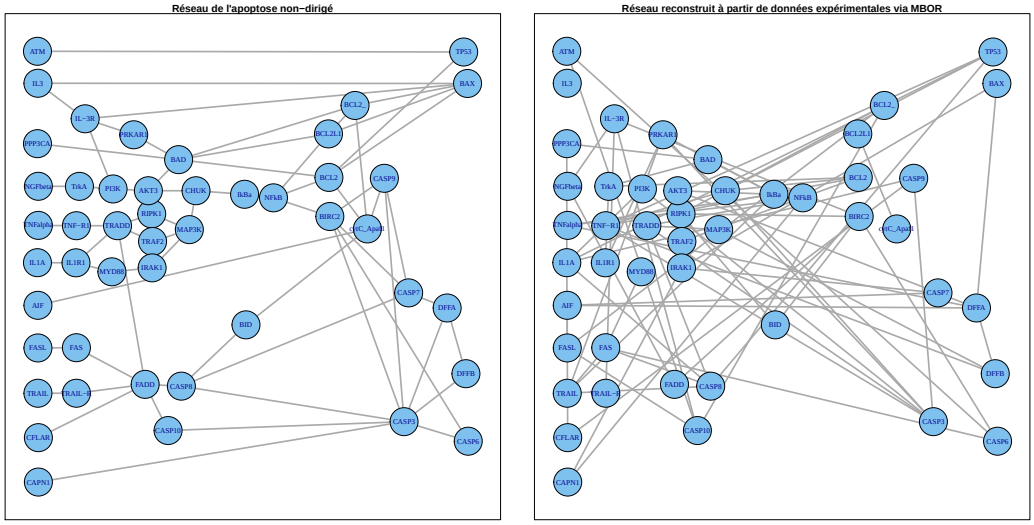


FIGURE 7.3 – **Graphes désorientés de l'apoptose.** A gauche : apoptose selon KEGG, après désorientation du graphe. A droite : résultat de la reconstruction du graphe à l'aide de l'algorithme MBOR.

été préalablement extraits des données. La démarche est ici de reconstruire le graphe de l'apoptose à partir des gènes du réseau biologique, de le comparer au modèle de référence (selon *KEGG*), et de confronter cette structure à des réseaux générés au hasard. Lorsque plusieurs sondes sont annotées comme ciblant le même gène, on considère la moyenne de leur intensité dans le tableau proposé en instance de l'apprentissage de structure (2 gènes parmi les 47 sont chacun ciblés par 2 sondes).

7.3.1 Reconstruction du graphe

Avant d'appliquer la procédure de reconstruction du graphe, les données ont été binarisées avec trois seuils différents, correspondants aux trois percentiles suivants : 75%, 80% et 85%. *MBOR* a ensuite été exécutée 100 fois par seuil, et par cible, à chaque fois en conservant aléatoirement 90% des biopuces. Il en résulte un tableau symétrique de 47×47 gènes, ou chaque cellule correspond au pourcentage de présence de l'arête dans la reconstruction. La comparaison topologique entre ces deux graphes peut être visualisée à travers la figure 7.3. Dans la reconstruction, les arêtes présentes dans plus de 50% des cas sont représentées (graphe de droite). Par coïncidence, l'application de ce seuil entraine la génération du même nombre d'arêtes dans le graphe reconstruit que dans le modèle.

Il est difficile de constater visuellement des similarités entre le graphe de référence, et le graphe reconstruit. Dans le but de proposer un critère de comparaison, on décide de

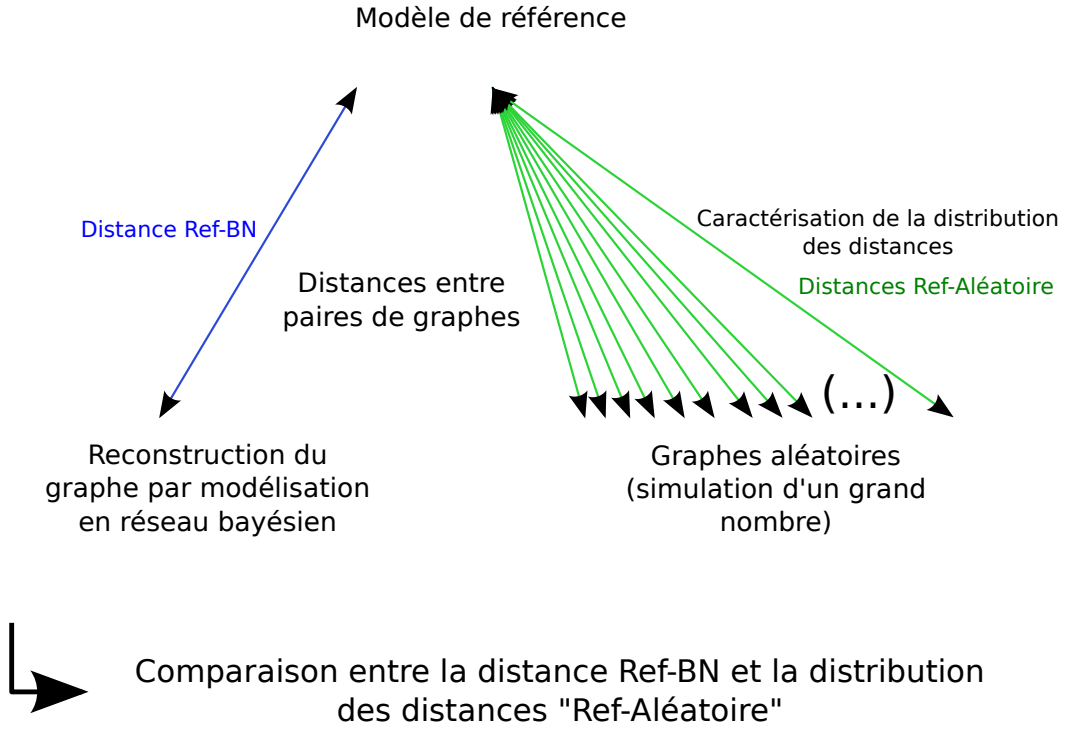


FIGURE 7.4 – **Stratégie d'évaluation de la reconstruction du graphe.** Une fois la reconstruction automatique effectuée, la distance avec le graphe de référence est calculée. Puis cette distance est comparée avec les distances obtenues à l'aide de graphes aléatoires.

générer des graphes aléatoires (avec plus ou moins de contraintes), puis de comparer la distance qui sépare la reconstruction via les Réseaux Bayésiens et le modèle de référence, de celle qui sépare les graphes aléatoires du modèle de référence. La stratégie de d'évaluation de la méthode développée est résumée dans la figure 7.4.

Pour suivre cette stratégie, on a donc besoin de définir un critère générant des graphes aléatoires (en particulier les contraintes de façon à rester dans un cadre réaliste), et un critère de distance (ou similarité).

7.3.2 Reconstruction du graphe de l'apoptose : Réseaux Bayésiens *VS* hasard

a. Critères de comparaison entre les graphes

La comparaison nécessite un ou plusieurs critères. Les critères retenus peuvent être comme dans le chapitre précédent la sensibilité et la spécificité. Ces paramètres

manquent cependant de « souplesse » : en pratique lorsque trois gènes A, B et C sont reliés en chaîne, et que sur la cinquantaine de gènes la méthode relie directement A et C, on est généralement assez satisfait du résultat. C'est un problème que l'on retrouve aussi lorsque que l'on veut utiliser la distance d'édition : nombre minimal d'opérations d'ajouts ou de suppressions d'arêtes pour convertir le graphe 1 en graphe 2. C'est pourquoi on introduit ici un nouveau critère de comparaison basé sur la taille des plus courts chemins. Un plus court chemin entre deux sommets A et B dans un graphe, est un chemin comptabilisant le nombre d'arêtes minimum reliant A et B.

En pratique pour calculer une distance basée sur les plus courts chemins, on calcule pour chaque graphe à comparer une matrice 47×47 (gènes) des plus courts chemins : chaque cellule de la matrice contient ce nombre minimal d'arêtes entre les deux nœuds considérés.

Le graphe généré aléatoirement n'est pas toujours connexe, cela signifie qu'il n'y a pas de chemin (et *a fortiori* de plus court chemin) reliant tout couple de sommets, cela se traduit par des cellules vides dans la matrice. Le calcul de la distance imposant des données complètes, les cellules vides sont remplacées par la valeur la plus importante possible à l'échelle du graphe : le diamètre +1. Le diamètre d'un graphe est le plus long des plus courts chemins de celui-ci. La distance est ensuite calculée de la manière suivante, soient M_1 et M_2 les matrices de plus courts chemins des graphe G_1 et G_2 (ayant le même nombre de nœuds) que l'on cherche à comparer :

$$D = |M_1 - M_2| \tag{7.10}$$

la matrice différentielle est ensuite sommée (puis divisée par 2 car elle est symétrique). Cela nous amène à la distance d :

$$d = \sum_{i=1}^{n} \sum_{j=1}^{n} D_{ij} \tag{7.11}$$

b. Génération des graphes aléatoires

Il y a plusieurs façons de générer des graphes aléatoires, selon les contraintes que l'on veut appliquer. Les graphes simulés ont tous le même nombre de nœuds que dans le graphe de référence. Dans un premier temps, le choix a été de générer des graphes

ayant pour contrainte la conservation de la distribution des degrés[1] par rapport au graphe de référence.

Les générateurs de graphes utilisés proviennent de la librairie *iGraph* développée pour *R*. Les résultats avec ce type de génération de graphe sont représentés dans la figure 7.5.

Ces histogrammes montrent que la reconstruction ne se distingue aucunement de ce que l'on peut obtenir aléatoirement avec la contrainte de distribution des degrés. En effet si elle avait été meilleure, on aurait obtenu des sensibilités et spécificités plus importantes (traits rouges décalés à droite dans les graphes du haut), ou des distances moins importantes (traits rouges plus à gauche dans le graphe du bas). Dans l'hypothèse où l'imposition du degré moyen est trop avantageuse dans la génération de graphes aléatoires, on décide de relâcher cette contrainte au profit de simulations plus basiques à l'aide de graphes de Erdös-Rényi (équiprobabilité de la présence des arêtes dans la reconstruction) comportant un nombre d'arêtes identique à celui du modèle.

Les histogrammes obtenus via cette génération de graphes aléatoires sont dessinés dans la figure 7.6

Outre une amélioration de la qualité relative de la reconstruction du graphe de l'apoptose à l'aide des Réseaux Bayésiens, on ne peut pas toujours pas distinguer celle-ci de ce que l'on peut obtenir aléatoirement (avec faibles contraintes) avec le modèle de génération aléatoire de Erdös-Rényi.

7.3.3 Conclusion

Malgré le succès démontré en terme de sélection de variables dans le cadre de la recherche de signatures de gènes ou de classification (*cf. 6*), l'utilisation des Réseaux Bayésiens dans la recherche d'un réseau d'interaction bien précis et caractérisé se solde ici par un échec. La technologie (puces à ADN) n'est pas à remettre en cause, puisqu'avec les mêmes données on est arrivé à extraire des gènes caractérisant ces malades en terme de transcriptome. On peut néanmoins formuler une autocritique en ce qui concerne le choix des données : on effet on est parti d'une hypothèse assez forte considérant le réseau de l'apoptose observable, étant donné que les données sont issues de tumeurs cancéreuses. Il est vrai que les cellules cancéreuses ne s'autodétruisent pas (ce qui explique en partie leur prolifération) traduisant un dérèglement de l'apoptose. Néanmoins l'expression combinée des gènes concernés ne se distingue pas forcément dans l'ensemble des 21329 gènes étudiés pour les 132 échantillons pris en compte. Bien-

1. le degré d'un nœud correspond au nombre d'arêtes connectées à celui-ci

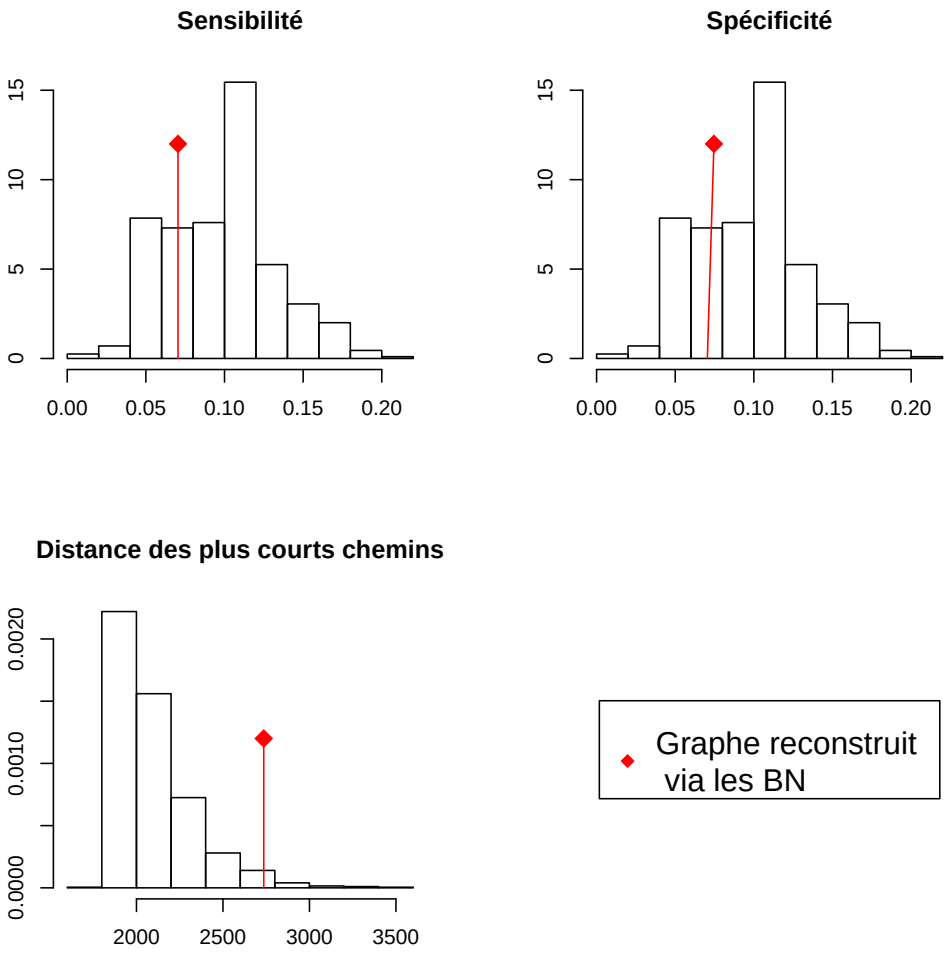


FIGURE 7.5 – **Comparaison de la reconstruction avec une distribution de graphes aléatoires obtenus en respectant la même distribution des degrés du modèle d'apoptose.** En haut : résultats de sensibilité et de spécificité. En bas : distances obtenues à l'aide de la comparaison des plus courts chemins au modèle. Le trait rouge pointe dans les histogrammes le résultat issu de la comparaison du graphe reconstruit avec le modèle.

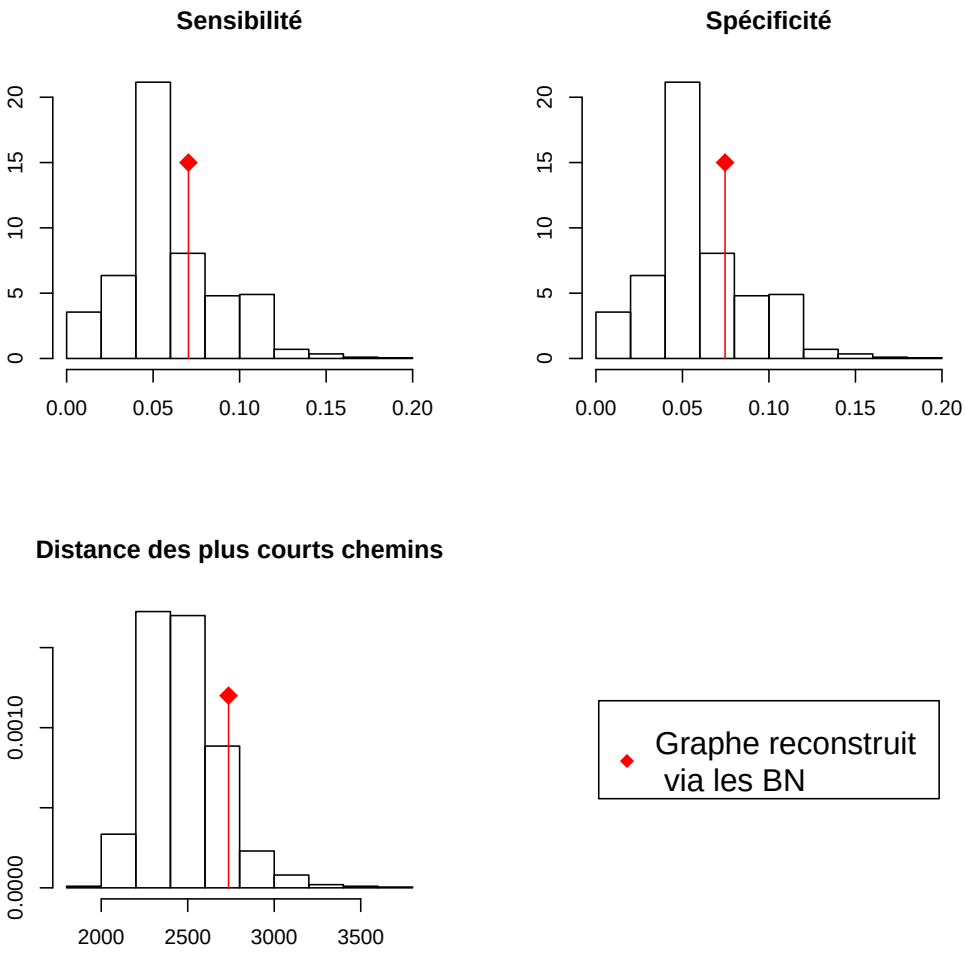


FIGURE 7.6 – **Comparaison de la reconstruction avec une distribution de graphes aléatoires obtenus en respectant un nombre d'arêtes identique au modèle d'apoptose.** En haut : résultats de sensibilité et de spécificité. En bas : distances obtenues à l'aide de la comparaison des plus courts chemins au modèle. Le trait rouge pointe dans les histogrammes le résultat issu de la comparaison du graphe reconstruit avec le modèle.

sûr, on peut rejeter le fait que l'application des Réseaux Bayésiens (du moins dans cette réalisation) soit conseillée dans cet exercice. Cependant, la méthode est peut-être adaptée à la recherche de réseaux se caractérisant plus fortement dans le transcriptome, impliquant des gènes peut-être plus spécifiques (ou sensibles) à l'affection (en l'occurrence le cancer du sein). Le problème, comme ici, survient quand on ne connait pas de tels réseaux validés et consensuels. Une idée serait de partir d'un gène qui nous intéresse, jouant *a priori* un rôle dans le cancer du sein, puis de rechercher dans son voisinage des gènes, voire une structure suscitant l'intérêt d'un spécialiste. Cette approche fait l'objet du paragraphe suivant.

7.4 Recherche d'un réseau local autour d'un gène d'intérêt

La protéine *ZNF217* suscite l'intérêt de nos collaborateurs de la plateforme de pharmacogénomique Lyon-Est pour son implication dans le cancer du sein [Nonet *et al.*, 2001]. La protéine est nommée ainsi car sa structure est adaptée au transport d'ions de zinc. D'après les annotations de la base de données *RefSeq*, celle-ci jouerait un rôle dans la régulation de la transcription. Pour toutes ces raisons, le gène éponyme codant pour cette protéine est celui qui a été choisi dans la recherche d'une structure avoisinante un objet biologique d'intérêt dans un tel jeu données. Ce gène mesure environ 30 Kb, et se situe sur le chromosome 20.

La procédure de recherche de structure est similaire à celle qui a été décrite dans le paragraphe 7.3.1. Simplement, un seul gène est considéré comme cible de la couverture de Markov. On limite la recherche du voisinage a des gènes dont le plus court chemin avec la cible ne dépasse pas 3 arêtes. Comme précédemment, à l'aide du partitionnement aléatoire, et des différences dans le seuil binarisation on obtient une matrice d'adjacence avec non-pas des valeurs binaires, mais des pourcentages de présence dans les graphes obtenus. Plus ce pourcentage est important, plus on peut avoir confiance dans l'existence réelle d'une relation entre les deux gènes concernés. Le graphe 7.7 expose les résultats de cette analyse (par souci de lisibilité, les relations dont la fréquence d'apparition est inférieure à 30% ne sont pas représentées).

Les gènes présents dans la première couche de cette couverture (directement reliés

à ZNF217) sont :

- CNOT2 : dont la protéine est la deuxième sous-unité du complexe de transcription CCR4-NOT. Ce complexe a été décrit comme ayant un rôle dans la régulation de la transcription, et fait partie du même processus biologique GO (GO :0006355) que ZNF217.
- PFDN4 : sous-unité d'une protéine permettant la stabilisation de polypeptides nouvellement synthétisés, en leur permettant de se replier correctement. Il est à noter que ce gène se situe au même emplacement que ZNF217, c'est à dire à l'emplacement 20q13.2.
- VIT1 (ou FBXO11) : protéine F-box 11. Cette protéine est l'une des quatre sous-unités constituant une ligase ubiquitaire : le complexe SCFs. Celle-ci jouerait un rôle dans l'inhibition de la protéine p53 phosphorylée. C'est un résultat intéressant dans le contexte de cancer, car p53 est bien connue, dans sa relation avec ces maladies (elle est d'ailleurs présente dans le réseau de l'apoptose selon KEGG).
- PRO1866 (ou EPS15L2) : récepteur de facteur de croissance de l'épiderme.
- LASS2 : ceramide synthase 2, homologue de LAG1, qui chez la levure joue un rôle dans la longévité de la cellule. La fonction du gène humain, moins connue, est soupçonnée d'être impliquée dans la régulation de la croissance cellulaire.

Pour les cinq gènes directement reliés au gène visé, on retrouve donc une information cohérente avec ce que l'on connait de ZNF217, ou des processus impliqués dans le cancer en général. Ces résultats sont très satisfaisants, car on aurait eu du mal à obtenir des gènes aussi étroitement liés au contexte biologique par le simple fait du hasard.

Il n'en est pas moins difficile de juger de l'aspect structurel de celui-ci. Il est en effet difficile de chercher à valider le graphe expérimentalement, et le champ de connaissance dans les interactions géniques humaines ne sont pas encore assez mâtures pour pouvoir se contenter des annotations dans les banques de données disponibles actuellement.

La procédure montre néanmoins une différence notable avec la recherche de signature génique précédemment effectuée : la cible est ici un gène parmi les autres, et non une condition biologique. On ne s'attend donc pas à ce qu'il y ait une différence de comportement entre plusieurs conditions données.

7.5 Conclusion

Dans ce chapitre, les Réseaux Bayésiens ont été utilisés de plusieurs manières dans le cadre de la recherche de réseaux d'interactions locaux (c'est-à-dire, ne cherchant pas à mettre en relation tous les gènes présents dans les données). Dans un premier temps, afin de maitriser parfaitement le jeu de données, celles-ci ont été simulées à partir d'un réseau de référence : la voie de l'apoptose, telle qu'elle annotée dans la base de connaissances KEGG. On a choisi celui-ci car sa dimension est raisonnable sans être minimaliste ou simpliste, et que l'apoptose est un processus cellulaire central dans les problématiques de cancer.

La première étape a été de développer une stratégie de passage de ce réseau biologique a un Réseau Bayésien (graphe bayésien et paramètres) de manière à ce qu'il retranscrive au mieux les annotations (en particulier le rôle activateur ou inhibiteur dans la régulation). Une fois ce Réseau Bayésien de l'apoptose obtenu, un jeu de données de 200 échantillon a été généré en suivant le modèle. Grâce à cela, il a été possible de comparer plusieurs stratégies d'apprentissage de structure, et de choisir la plus appropriée pour la suite, la méthode MBOR.

Pour appliquer cette méthode dans un contexte expérimental et non de simulation, les mêmes données de biopuces ont été utilisées (132 patientes atteintes d'un cancer du sein traitées en adjuvant par du tamoxifène) pour rechercher les interactions entre tous les gènes présents dans le réseau de l'apoptose. Les résultats de cette reconstruction ont pu être comparés au modèle grâce à des critères de similarité ou de distance (sensibilité, spécificité, et les plus courts chemin dans un graphe). Cette démarche n'a pas aboutir à des résultats satisfaisants lorsqu'on les a comparés à des reconstructions aléatoires (en conservant soit la distribution des degrés, soit simplement le nombre d'arêtes du graphe bayésien de l'apoptose de référence).

Cependant, dans le cas où cette méthode a été appliquée à la recherche d'un voisinage de gène spécifique à ce jeu de données (dont l'implication dans les cancers du sein semble importante), on obtient des résultats probants, au moins en ce qui concerne la nature des gènes identifiés comme voisins. La qualité de la recherche des relations entre ces gènes (la connectivité) reste encore à être validée, nécessitant cependant une résolution plus fine de ces réseaux à l'aide de méthodes expérimentales.

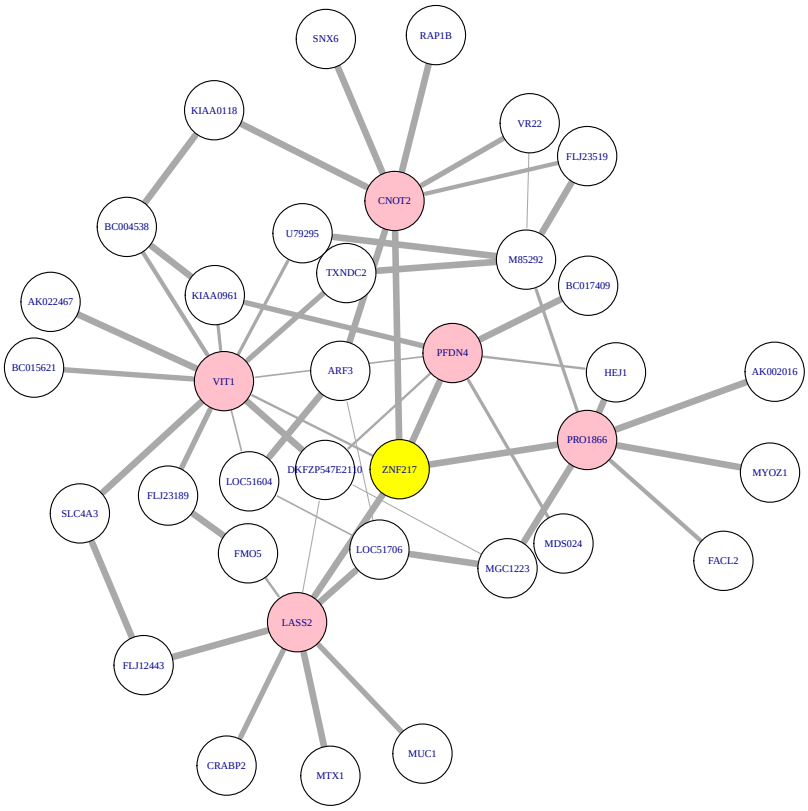


FIGURE 7.7 – **Graphe local autour du gène d'intérêt ZNF217**. Les gènes directement connectés à ZNF217 sont colorés en rose. L'épaisseur des relations est fonction de la « robustesse » de celle-ci : plus le trait est épais, plus le niveau de confiance est important.

Conclusion et perspectives

Dans ce projet, on s'est intéressé à la modélisation de réseaux de régulation de gènes, ainsi qu'à l'utilisation dans plusieurs problématiques de tels modèles. Les Réseaux Bayésiens ont été choisis à ces fins car ils présentent certaines qualités. Ils permettent de modéliser des données statiques, composées uniquement de répétitions biologiques et non spatiales ou temporelles. Ce sont des modèles graphiques, particulièrement adaptés à l'esprit humain. Ils formalisent des indépendances et des dépendances conditionnelles, ce qui semble parmi les modèles graphiques connus être un avantage. Ce sont des modèles probabilistes, facilitant une utilisation directe dans l'inférence des distributions lorsque de nouvelles informations sont disponibles. Cependant ils ont quelques défauts : les Réseaux Bayésiens ont des graphes sans circuit, alors que l'on sait que dans les réseaux cellulaires que l'on cherche à modéliser on retrouve des cycles.

Cette limite à la modélisation n'est néanmoins pas aussi gênante qu'elle n'y parait : en effet l'information dans un graphe bayésien ne circule pas toujours dans le sens des flèches (due à l'inversion bayésienne). Un autre défaut (le prix à payer pour leur finesse) est la mise en œuvre. Il est nécessaire de connaître la structure et les paramètres si on veut utiliser un Réseau Bayésien pour faire de l'inférence. La structure est, on l'a vu, très difficile à apprendre. Pourtant, dans ce travail où la priorité est la détection d'une structure d'interactions entre gènes à partir de données d'expression, on se focalise particulièrement sur cet exercice d'apprentissage.

La première difficulté de l'apprentissage de structure est due à la taille des jeux de données à analyser. Pour résister à ce problème, des stratégies de parcours de l'espace de recherche doivent être élaborées en fonction des algorithmes d'apprentissage, qu'ils soient sous contraintes ou basés sur un score. Quelques uns de ces aspects ont été approfondis dans ce travail. Dans le cas où on s'intéresse à tous les gènes du système (ici le génome humain), une stratégie d'accélération de la procédure basée sur un score, à l'aide d'un graphe précalculé de façon simple et en pratique (avec quelques milliers de variables) instantané, a été proposée ici. Elle ne semble pas suffire à l'obtention d'un réseau global (de plusieurs milliers de variables) satisfaisant ou validé, mais elle accélère

significativement la procédure, laissant à penser qu'elle est adaptable et serait utile à plus petite échelle. C'est dans ce sens, que la suite de ce projet de thèse se focalise sur l'étude de plus petits réseaux, soit au travers de travaux de classification ou de sélection de variable, soit la recherche d'un réseau local au voisinage d'un gène d'intérêt. On note et apprécie la grande souplesse de ce modèle graphique dont les potentialités sont nombreuses. La possibilité de mettre à jour la distribution de probabilité liée à n'importe-qu'elle variable du réseau est exploitée ici pour classer des tumeurs selon le type de leucémie à partir de données de biopuces. Les résultats sont probants et la méthode répond donc aux attentes que peuvent avoir les médecins dans cette récente technologie que constitue les biopuces.

Une deuxième difficulté réside dans la validation. En effet, dans un cadre de classification, si on estime que la réalité est connue, la validation est faisable. Néanmoins lorsqu'on cherche à mettre en évidence un réseau cellulaire, où les connaissances sont incomplètes et très compartimentées (en réseaux de gènes, de protéines, métabolique : les résultats des interactions inter-réseaux ayant forcément des répercussions dans les données mesurées), il n'est pas aisé de comparer les résultats obtenus avec confiance. C'est pour cela que l'on s'est intéressé à la simulation de jeux de données. A partir d'un réseau de signalisation répertorié comme ayant une importance capitale en cancérologie, celui de l'apoptose, un Réseau Bayésien a été construit : c'est-à-dire que le graphe bayésien est adapté à partir du réseau cellulaire obtenu dans une base de connaissances, et que les paramètres ont été simulés en accord avec la typologie du réseau (en particulier les flèches d'inhibition et d'activation), et avec la biologie impliquant obligatoirement une variation aléatoire. Un tel modèle, adapté à un réseau cellulaire connu et à jour n'avait en effet pas été décelé dans la littérature scientifique. Une fois le modèle obtenu, il est très facile de simuler des jeux de données à partir desquels on peut faire des comparaisons. Trois stratégies d'apprentissage ont donc été mise au banc d'essai, et contrairement à nos attentes, c'est la stratégie locale (qui n'a alors pas été utilisée localement dans ce contexte) de recherche basée sur la couverture de Markov d'une variable qui obtient les meilleurs résultats en sensibilité et spécificité, de plus ces indicateurs sont déjà encourageants avec 50 échantillons. Par ambition, on a testé la méthode pour retrouver les interactions entre gènes de l'apoptose à travers leur expression dans des données de cancer du sein. La comparaison avec des graphes aléatoires n'a pas validé la combinaison méthode et données. Cependant lorsque la stratégie d'apprentissage du graphe basée sur la couverture de Markov est appliquée à des problèmes de classification, on a pu obtenir des résultats satisfaisants en terme

de recherche de signature génique d'une condition particulière (correspondante à la rechute ou non de cancers du sein traités en adjuvant au tamoxifène). Cette signature permet avec seulement 4 gènes d'obtenir une classification bien meilleure que ce qui avait été précédemment proposé en terme de précision (gain de 13 points). Un autre résultat encourageant est obtenu lorsqu'on recherche le voisinage d'un gène duquel les *a priori* sur son implication dans le contexte biologique capturé par la biopuce sont très forts d'après les experts du domaine. Les gènes obtenus dans ce voisinage ont des annotations cohérentes avec le contexte, et la méthode fournit de fait des indices d'investigations futures. On est cependant confronté, pour ce qui concerne la méthode, au problème de la validation de la structure obtenue.

Trois perspectives peuvent découler de ce travail. La première est l'amélioration de l'algorithme de recherche de structures locales, de façon à ce qu'il soit encore plus rapide, robuste, et adaptable à des problèmes de réseaux cellulaires. Une idée, liée à l'essor des calculateurs en grilles ou clusters, est de proposer une version parallélisée de l'algorithme, ce qui d'après ses fondements semble assez simple.

Une deuxième perspective est l'adaptation de la méthode à des données d'une autre nature, prometteuses dans la recherche de réseaux de régulation : les données de séquençage à haut-débit. En effet, le coût de cette technologie baisse aussi vite que la qualité augmente, elle est déjà plus rentable que les puces dans certains domaines (les études de métagénomes par exemple). Elle permet à la fois des études génomiques et transcriptomiques et est donc vouée à remplacer les biopuces.

Enfin, une troisième perspective est de trouver une manière de profiter de toutes les données publiées (actuelles et futures) pour effectuer des méta-analyses. Une approche serait de se reposer sur une (ou quelques) voie(s) de signalisation d'intérêt, de taille moyenne, et de mettre à jour le réseau régulièrement lorsque des données pertinentes pour ces voies sont disponibles. Ainsi, on raisonnerait sur un ordre de grandeur permettant de devenir expert, et on mettrait à profit une production scientifique de plus en plus importante et accessible.

Parait-il qu'on ne voit jamais aussi bien que depuis les épaules d'un géant, la mise à profit des outils bioinformatiques permet de s'y élever.

Quatrième partie

Bibliographie et annexes

Références bibliographiques

[Adams *et al.*, 1991] Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., Merril, C. R., Wu, A., Olde, B. et Moreno, R. F. (1991). Complementary dna sequencing : expressed sequence tags and human genome project. *Science*, 252(5013):1651–6.

[Akaike, 1973] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. pages 267–281.

[Akaike, 1974] Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

[Antoniou *et al.*, 2003] Antoniou, A., Pharoah, P. D. P., Narod, S., Risch, H. A., Eyfjord, J. E., Hopper, J. L., Loman, N., Olsson, H., Johannsson, O., Borg, A., Pasini, B., Radice, P., Manoukian, S., Eccles, D. M., Tang, N., Olah, E., Anton-Culver, H., Warner, E., Lubinski, J., Gronwald, J., Gorski, B., Tulinius, H., Thorlacius, S., Eerola, H., Nevanlinna, H., Syrjäkoski, K., Kallioniemi, O.-P., Thompson, D., Evans, C., Peto, J., Lalloo, F., Evans, D. G. et Easton, D. F. (2003). Average risks of breast and ovarian cancer associated with brca1 or brca2 mutations detected in case series unselected for family history : a combined analysis of 22 studies. *Am J Hum Genet*, 72(5):1117–30.

[Aouba *et al.*, 2007] Aouba, A., Péquignot, F., Toullec, A. L. et Jougla, E. (2007). Les causes médicales de décès en france en 2004 et leur évolution 1980-2004. *Bulletin épidémiologique hebdomadaire*, pages 308–314.

[Benjamini et Hochberg, 1995] Benjamini, Y. et Hochberg, Y. (1995). Controlling the false discovery rate : A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

[Bolstad *et al.*, 2003] Bolstad, B. M., Irizarry, R. A., Astrand, M. et Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–93.

[Buffon *et al.*, 1749] Buffon, G.-L. L., Daubenton, L. J.-M., de Montbéliard, P. G. et de La Cépède, B. G. É. (1749). Histoire naturelle, générale et particulière : avec la description du cabinet du roi. *Imprimerie Royale*.

[Butte et Kohane, 2000] Butte, A. J. et Kohane, I. S. (2000). Mutual information relevance networks : functional genomic clustering using pairwise entropy measurements. *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing*, pages 418–29.

[Calzolari *et al.*, 2007] Calzolari, D., Paternostro, G., Harrington, P. L., Piermarocchi, C. et Duxbury, P. M. (2007). Selective control of the apoptosis signaling network in heterogeneous cell populations. *PLoS ONE*, 2(6):e547.

[Chanrion *et al.*, 2008] Chanrion, M., Negre, V., Fontaine, H., Salvetat, N., Bibeau, F., Grogan, G. M., Mauriac, L., Katsaros, D., Molina, F., Theillet, C. et Darbon, J.-M. (2008). A gene expression signature that can predict the recurrence of tamoxifen-treated primary breast cancer. *Clin Cancer Res*, 14(6):1744–1752.

[Cheng *et al.*, 1997] Cheng, J., Bell, D. et Liu, W. (1997). Learning belief networks from data : an information theory based approach. *CIKM '97 : Proceedings of the sixth international conference on Information and knowledge management*.

[Chickering, 1995] Chickering, D. M. (1995). A transformational characterization of equivalent bayesian network structures. *UAI'95*, pages 87–98.

[Chickering, 2002] Chickering, D. M. (2002). Learning equivalence classes of bayesian-network structures. *The Journal of Machine Learning Research*, 2.

[Cooper et Herskovits, 1992] Cooper, G. et Herskovits, E. (1992). A bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9(4):309–347.

[de Chassey *et al.*, 2008] de Chassey, B., Navratil, V., Tafforeau, L., Hiet, M. S., Aublin-Gex, A., Agaugué, S., Meiffren, G., Pradezynski, F., Faria, B. F., Chantier, T., Breton, M. L., Pellet, J., Davoust, N., Mangeot, P. E., Chaboud, A., Penin, F., Jacob, Y., Vidalain, P. O., Vidal, M., André, P., Rabourdin-Combe, C. et Lotteau, V. (2008). Hepatitis c virus infection protein network. *Mol Syst Biol*, 4:230.

[de Jong, 2002] de Jong, H. (2002). Modeling and simulation of genetic regulatory systems : a literature review. *J Comput Biol*, 9(1):67–103.

[de Morais et Aussem, 2008] de Morais, S. R. et Aussem, A. (2008). A novel scalable and data efficient feature subset selection algorithm. *Machine Learning and Knowledge Discovery in Databases*, pages 1–15.

[de Morais et Aussem, 2010] de Morais, S. R. et Aussem, A. (2010). A novel markov boundary based feature subset selection algorithm. *Neurocomputing*, 73(4-6):578–584.

[Dean et Kanazawa, 1989] Dean, T. et Kanazawa, K. (1989). A model for reasoning about persistence and causation. *Computational Intelligence*.

[Dean et Wellman, 1991] Dean, T. L. et Wellman, M. P. (1991). Planning and control. *M. Kaufmann Publishers*.

[Dray *et al.*, 2003] Dray, S., Chessel, D. et Thioulouse, J. (2003). Co-inertia analysis and the linking of ecological data tables. *Ecology*, 84(11):3078–3089.

[Edwards, 2000] Edwards, D. (2000). Introduction to graphical modelling. *Springer texts in statistics*.

[Friedman *et al.*, 2000] Friedman, N., Linial, M., Nachman, I. et Pe'er, D. (2000). Using bayesian networks to analyze expression data. *RECOMB '00 : Proceedings of the fourth annual international conference on Computational molecular biology*.

[Fung et Mangasarian, 2004] Fung, G. et Mangasarian, O. (2004). A feature selection newton method for support vector machine classification. *Comput Optim Appl*, 28(2):185–202.

[Glymour *et al.*, 1991] Glymour, C., Spirtes, P. et Scheines, R. (1991). Causal inference. *Erkenntnis*, 35(1):151–189.

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

[Guyon *et al.*, 2006] Guyon, I., Gunn, S., Nikravesh, M. et Zadeh, L. A. (2006). Feature extraction : Foundations and applications. *Springer*.

[Guyon *et al.*, 2002] Guyon, I., Weston, J., Barnhill, S. et Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422.

[Hall *et al.*, 2009] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. et Witten, I. (2009). The weka data mining software : an update. *SIGKDD Explorations Newsletter*, 11(1).

[Hanahan et Weinberg, 2000] Hanahan, D. et Weinberg, R. A. (2000). The hallmarks of cancer. *Cell*, 100(1):57–70.

[Hanczar et Dougherty, 2008] Hanczar, B. et Dougherty, E. R. (2008). Classification with reject option in gene expression data. *Bioinformatics*, 24(17):1889–95.

[Hastings, 1970] Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109.

[Havukkala et Vanderlooy, 2007] Havukkala, I. et Vanderlooy, S. (2007). On the reliable identification of plant sequences containing a polyadenylation site. *Journal of Computational Biology*, 14(9):1229–1245.

[Heckerman, 1997] Heckerman, D. (1997). Bayesian networks for data mining. *Data Min. Knowl. Discov.*, 1(1):79–119.

[Heckerman, 1998] Heckerman, D. (1998). A tutorial on learning with bayesian networks. pages 301–354.

[Heckerman *et al.*, 1995] Heckerman, D., Geiger, D. et Chickering, D. M. (1995). Learning bayesian networks : The combination of knowledge and statistical data. *Machine Learning*, 20(3):197–243.

[Hua *et al.*, 2009] Hua, J., Tembe, W. D. et Dougherty, E. R. (2009). Performance of feature-selection methods in the classification of high-dimension data. *Pattern Recognition*, 42(3):409–424.

[Irizarry *et al.*, 2003] Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B. et Speed, T. P. (2003). Summaries of affymetrix genechip probe level data. *Nucleic Acids Research*, 31(4):e15.

[Ishak, 2007] Ishak, A. (2007). Sélection de variables par les machines à vescteurs de supports pour la discrimination binaire et multiclasse en grande dimension. *Thèse de doctorat, Université Aix-Marseille II.*

[Jarnagin, 1961] Jarnagin, M. P. (1961). *Automatic Machine Methods of Testing PERT Networks for Consistency.*.

[Jensen, 1996] Jensen, F. V. (1996). An introduction to bayesian networks.

[Jensen et Nielsen, 2007] Jensen, F. V. et Nielsen, T. D. (2007). Bayesian networks and decision graphs. page 447.

[Kauffman, 1969] Kauffman, S. (1969). Homeostasis and differentiation in random genetic control networks. *Nature*, 224(5215):177–8.

[Kurgan et Cios, 2004] Kurgan, L. et Cios, K. (2004). Caim discretization algorithm. *Ieee T Knowl Data En*, 16(2):145–153.

[Liang et Pardee, 1992] Liang, P. et Pardee, A. B. (1992). Differential display of eukaryotic messenger rna by means of the polymerase chain reaction. *Science*, 257(5072):967–71.

[Ma et Huang, 2008] Ma, S. et Huang, J. (2008). Penalized feature selection and classification in bioinformatics. *Brief Bioinformatics*, 9(5):392–403.

[Maxam et Gilbert, 1977] Maxam, A. M. et Gilbert, W. (1977). A new method for sequencing dna. *Proc Natl Acad Sci USA*, 74(2):560–4.

[Metropolis *et al.*, 1953] Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. et Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087.

[Miki *et al.*, 1994] Miki, Y., Swensen, J., Shattuck-Eidens, D., Futreal, P. A., Harshman, K., Tavtigian, S., Liu, Q., Cochran, C., Bennett, L. M. et Ding, W. (1994). A strong candidate for the breast and ovarian cancer susceptibility gene brca1. *Science*, 266(5182):66–71.

[Murphy, 2001] Murphy, K. P. (2001). The bayes net toolbox for matlab. *Computing Science and Statistics*, 33.

[Nagalakshmi *et al.*, 2008] Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M. et Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by rna sequencing. *Science*, 320(5881):1344–9.

[Nilsson *et al.*, 2006] Nilsson, R., Peña, J. M., Bjöerkegren, J. et Tegner, J. (2006). Evaluating feature selection for svms in high dimensions.

[Nilsson *et al.*, 2007] Nilsson, R., Peña, J. M., Bjöerkegren, J. et Tegner, J. (2007). Consistent feature selection for pattern recognition in polynomial time. *The Journal of Machine Learning Research*, 8:612.

[Nonet *et al.*, 2001] Nonet, G. H., Stampfer, M. R., Chin, K., Gray, J. W., Collins, C. C. et Yaswen, P. (2001). The znf217 gene amplified in breast cancers promotes immortalization of human mammary epithelial cells. *Cancer Res*, 61(4):1250–4.

[Ogata *et al.*, 1999] Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. et Kanehisa, M. (1999). Kegg : Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 27(1):29–34.

[Pearl, 1988] Pearl, J. (1988). Probabilistic reasoning in intelligent systems : Networks of plausible inference. *Morgan Kaufmann Ed.*

[Polanski *et al.*, 2007] Polanski, A., Polanska, J., Jarzab, M., Wiench, M. et Jarzab, B. (2007). Application of bayesian networks for inferring cause–effect relations from gene expression profiles of cancer versus normal cells. *Mathematical Biosciences*, 209(2):528–546.

[Rakotomamonjy, 2003] Rakotomamonjy, A. (2003). Variable selection using svm based criteria. *Journal of Machine Learning Research*, 3:1357–1370.

[Ren *et al.*, 2000] Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T. L., Wilson, C. J., Bell, S. P. et Young, R. A. (2000). Genome-wide location and function of dna binding proteins. *Science*, 290(5500):2306–9.

[Robinson, 1973] Robinson, R. W. (1973). Counting labeled acyclic digraphs.

[Robinson, 1977] Robinson, R. W. (1977). Counting unlabeled acyclic digraphs. pages 28–43.

[Sackmann *et al.*, 2006] Sackmann, A., Heiner, M. et Koch, I. (2006). Application of petri net based analysis techniques to signal transduction pathways. *BMC Bioinformatics*, 7:482.

[Saeys *et al.*, 2008] Saeys, Y., Abeel, T. et Peer, Y. (2008). Robust feature selection using ensemble feature selection techniques. *Proceedings of the European conference on Machine Learning and Knowledge Discovery in Databases-Part II*, pages 313–325.

[Sanger *et al.*, 1977] Sanger, F., Nicklen, S. et Coulson, A. R. (1977). Dna sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA*, 74(12):5463–5467.

[Schafer et Strimmer, 2005] Schafer, J. et Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, 4:Article32.

[Schena *et al.*, 1995] Schena, M., Shalon, D., Davis, R. W. et Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467–70.

[Schwarz, 1978] Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

[Simes, 1986] Simes, R. (1986). An improved bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754.

[Smith *et al.*, 2006] Smith, V. A., Yu, J., Smulders, T. V., Hartemink, A. J. et Jarvis, E. D. (2006). Computational inference of neural information flow networks. *PLoS Comput Biol*, 2(11):e161.

[Smyth *et al.*, 2003] Smyth, G. K., Yang, Y. H. et Speed, T. P. (2003). Statistical issues in cdna microarray data analysis. *Methods Mol Biol*, 224:111–36.

[Spirtes *et al.*, 2000] Spirtes, P., Glymour, C. et Scheines, R. (2000). Causation, prediction, and search. *MIT Press*.

[Statnikov *et al.*, 2008] Statnikov, A., Wang, L. et Aliferis, C. (2008). A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinformatics*, 9(1):319.

[Tamada *et al.*, 2003] Tamada, Y., Kim, S., Bannai, H., Imoto, S., Tashiro, K., Kuhara, S. et Miyano, S. (2003). Estimating gene networks from gene expression data by combining bayesian network model with promoter element detection. *Bioinformatics*, 19(90002):227–236.

[Tang *et al.*, 2007] Tang, Y., Zhang, Y.-Q. et Huang, Z. (2007). Development of two-stage svm-rfe gene selection strategy for microarray expression data analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB*, 4(3).

[Thomas, 1973] Thomas, R. (1973). Boolean formalization of genetic control circuits. *Journal of Theoretical Biology*, 42(3):563–585.

[Tibshirani *et al.*, 2002] Tibshirani, R., Hastie, T., Narasimhan, B. et Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci USA*, 99(10):6567–72.

[Tuleau, 2005] Tuleau, C. (2005). Sélection de variables pour la discrimination en grande dimension et classification de données fonctionnelles. *Thèse de doctorat, UNIVERSITÉ PARIS XI ORSAY*.

[van't Veer *et al.*, 2002] van't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R. et Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–6.

[Velculescu *et al.*, 1995] Velculescu, V. E., Zhang, L., Vogelstein, B. et Kinzler, K. W. (1995). Serial analysis of gene expression. *Science*, 270(5235):484–7.

[Verma et Pearl, 1990] Verma, T. et Pearl, J. (1990). Equivalence and synthesis of causal models. *Proceedings 6th Conference on Uncertainty in AI*, pages 220–227.

[Verma et Pearl, 1991] Verma, T. et Pearl, J. (1991). A theory of inferred causation.

[Watson et Crick, 1953] Watson, J. D. et Crick, F. H. (1953). Molecular structure of nucleic acids ; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–8.

[Weston *et al.*, 2003] Weston, J., Elisseeff, A., Schölkopf, B. et Tipping, M. (2003). Use of the zero-norm with linear models and kernel methods. *The Journal of Machine Learning Research*, 3:1439–1461.

[Whittaker, 1990] Whittaker, J. (1990). Graphical models in applied multivariate statistics. *Wiley series in probability and mathematical statistics*.

[Wooster *et al.*, 1995] Wooster, R., Bignell, G., Lancaster, J., Swift, S., Seal, S., Mangion, J., Collins, N., Gregory, S., Gumbs, C. et Micklem, G. (1995).

Identification of the breast cancer susceptibility gene brca2. *Nature*, 378(6559):789–92.

[Wu *et al.*, 2004] Wu, Z., Irizarry, R. A., Gentleman, R., Martinez-Murillo, F. et Spencer, F. (2004). A model-based background adjustment for oligonucleotide expression arrays. *Journal of the American Statistical Association*, 99(468):909+.

[Yang *et al.*, 2002] Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J. et Speed, T. P. (2002). Normalization for cdna microarray data : a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, 30(4):e15.

[Yang et NP, 2003] Yang, Y. H. et NP, T. (2003). Goldstein dr, editor. normalization for two-color cdna microarray data. science and statistics : A festschrift for terry speed. *IMS Lecture Notes–Monograph Series*, 40:403–418.

Annexes

Sommaire

7.6 Algorithmes de recherche de couverture de Marcov

Algorithm 7.6.1

Require: T : target ; **U** : variables
Ensure: **MB** : Markov boundary of T
Phase I : *Remove X if $T \perp X$*
PCS = **U** $\setminus T$
for all $X \in$ **PCS do**
  if $(T \perp X)$ **then**
    PCS = **PCS** $\setminus X$
    dSep(X) = $\varnothing$
  end if
end for
Phase II :*Remove X if $T \perp X|Y$*
for all $X \in$ **PCS do**
  for all $Y \in$ **PCS** $\setminus X$ **do**
    if $(T \perp X \mid Y)$ **then**
      PCS = **PCS** $\setminus X$
      dSep(X) = Y ; **go to 15**
    end if
  end for
end for
Phase III : *Find super set for SP*
SPS = $\varnothing$

17: **for all** $X \in \mathbf{PCS}$ **do**
18: $\quad \mathbf{SPS}_X = \varnothing$
19: $\quad$ **for all** $Y \in \mathbf{U} \setminus \{T \cup \mathbf{PCS}\}$ **do**
20: $\quad\quad$ **if** $(T \not\perp Y | \mathbf{dSep}(Y) \cup X)$ **then**
21: $\quad\quad\quad \mathbf{SPS}_X = \mathbf{SPS}_X \cup Y$
22: $\quad\quad$ **end if**
23: $\quad$ **end for**
24: $\quad$ **for all** $Y \in \mathbf{SPS}_X$ **do**
25: $\quad\quad$ **for all** $Z \in \mathbf{SPS}_X \setminus Y$ **do**
26: $\quad\quad\quad$ **if** $(T \perp Y | X \cup Z)$ **then**
27: $\quad\quad\quad\quad \mathbf{SPS}_X = \mathbf{SPS}_X \setminus Y$; **go to 30**
28: $\quad\quad\quad$ **end if**
29: $\quad\quad$ **end for**
30: $\quad$ **end for**
31: $\quad \mathbf{SPS} = \mathbf{SPS} \cup \mathbf{SPS}_X$
32: **end for**
Phase IV : *Find PC of* T
33: $\mathbf{PC} =$ *Inter-IAPC*$(T, \mathcal{D}(\mathbf{PCS} \cup \mathbf{SPS})$
34: **for all** $X \in \mathbf{PCS} \setminus \mathbf{PC}$ **do**
35: $\quad$ **if** $T \in$ *Inter-IAPC*$(X, \mathcal{D})$ **then**
36: $\quad\quad \mathbf{PC} = \mathbf{PC} \cup X$
37: $\quad$ **end if**
38: **end for**
Phase V : *Find spouses of* T
39: $\mathbf{SP} = \varnothing$
40: **for all** $X \in \mathbf{PC}$ **do**
41: $\quad$ **for all** $Y \in$ *Inter-IAPC*$(X, D) \setminus \{\mathbf{PC} \cup T\}$ **do**
42: $\quad\quad$ Find minimal $\mathbf{Z} \subset \mathbf{PCS} \cup \mathbf{SPS} \backslash \{T \cup Y\}$ such that $T \perp Y | \mathbf{Z}$
43: $\quad\quad$ **if** $(T \not\perp Y | \mathbf{Z} \cup X)$ **then**
44: $\quad\quad\quad \mathbf{SP} = \mathbf{SP} \cup Y$
45: $\quad\quad$ **end if**
46: $\quad$ **end for**
47: **end for**

Algorithm 7.6.2

Require: T : target ; D : data set ; $\mathbf{V}$ set of variables
Ensure: $\mathbf{PC}$: Parents and children of T ;
1: $\mathbf{MB} = \varnothing$
2: **repeat**
3: $\quad$ *Add true positives to* $\mathbf{MB}$
4: $\quad Y = \mathbf{argmax}_{X \in (\mathbf{V} \setminus \mathbf{MB} \setminus \{T\})}$
5: $\quad\quad$ *AssocMeasure*$(T, X | \mathbf{MB})$
6: $\quad$ **if** $T \not\perp Y | \mathbf{MB}$ **then**
7: $\quad\quad \mathbf{MB} = \mathbf{MB} \cup Y$
8: $\quad$ **end if**

Remove false positives from $\mathbf{MB}$

```
 9:    for all X ∈ MB do
10:      if T ⊥ X|(MB ∖ X) then
11:        MB = MB ∖ X
12:      end if
13:    end for
14: until MB has not changed

    Remove parents of children from MB
15: PC = MB
16: for all X ∈ MB do
17:    if ∃Z ⊂ (MB ∖ X)
18:    such that T ⊥ X | Z then
19:      PC = PC ∖ X
20:    end if
21: end for
```

7.7 Programme en langage R générant des tables de probabilités conditionnelles aléatoires

```
########## CPT() constructs the conditional probability table frame
########## for a node according to the number of parents,
########## the number of states and their labels
CPT<- function(nparents, nstates, lstates){
nc<-nparents+1
config<-matrix(nrow=nstates^nc,ncol=nc)
for (i in 0:nparents){
config[,i+1]<-as.character(gl(nstates, nstates^i, nstates^nc,
labels=lstates))
}
config<-as.data.frame(config)

return(config)
}

########## GenerateProbas() simulates probabilities of X according
########## to the kind of the relationship (activator or repressor)
########## between X and its parents
GenerateProbas <- function(CPT, ParentsWeight=rep(1,dim(CPT)[2]-1)){
if (dim(CPT)[2]-1!=length(ParentsWeight)) stop("Error : number of
parents in CPT and number of weights are not the same !!")
```

```
nstates<-exp(log(dim(CPT)[1])/dim(CPT)[2])

### if X has no parent, the probability value
### is simulated by the uniform law
###  between 0.4 and 0.6
if (length(ParentsWeight)==0){
tab<-runif(dim(CPT)[1],0.4,0.6)
}

else {
# to inverse effect for negative ParentsWeight values
neg<-which(ParentsWeight<0)
if (length(neg)>0) CPT[,neg]<-rev(CPT[,neg])

# distance between state and averaged (through the
# ParentsWeigh coefficients)
# states of the parents
pw<-abs(ParentsWeight)
dist<-abs(apply(CPT,1,function(x) (sum(as.numeric(x)
[-length(x)]*pw)/sum(pw)) -(mean(pw)
*as.numeric(x)[length(x)])))

sim <- (max(dist)-dist)^3

# perturbation : adds random residues
tab <- sim +rnorm(length(sim),sd=max(sim)/10)
tabraw<-tab
}

### scaling similarity values between 0 and 1
if (min(tab)<0) tab<-tab-min(tab)
for (i in 1:nstates^(dim(CPT)[2]-1)){
elements<-seq(i,nstates*nstates^(dim(CPT)[2]-1),
nstates^(dim(CPT)[2]-1))
if(sum(tab[elements]!=0)) tab[elements]
<-tab[elements]/sum(tab[elements])
else stop()
}

return(tab)

}
```

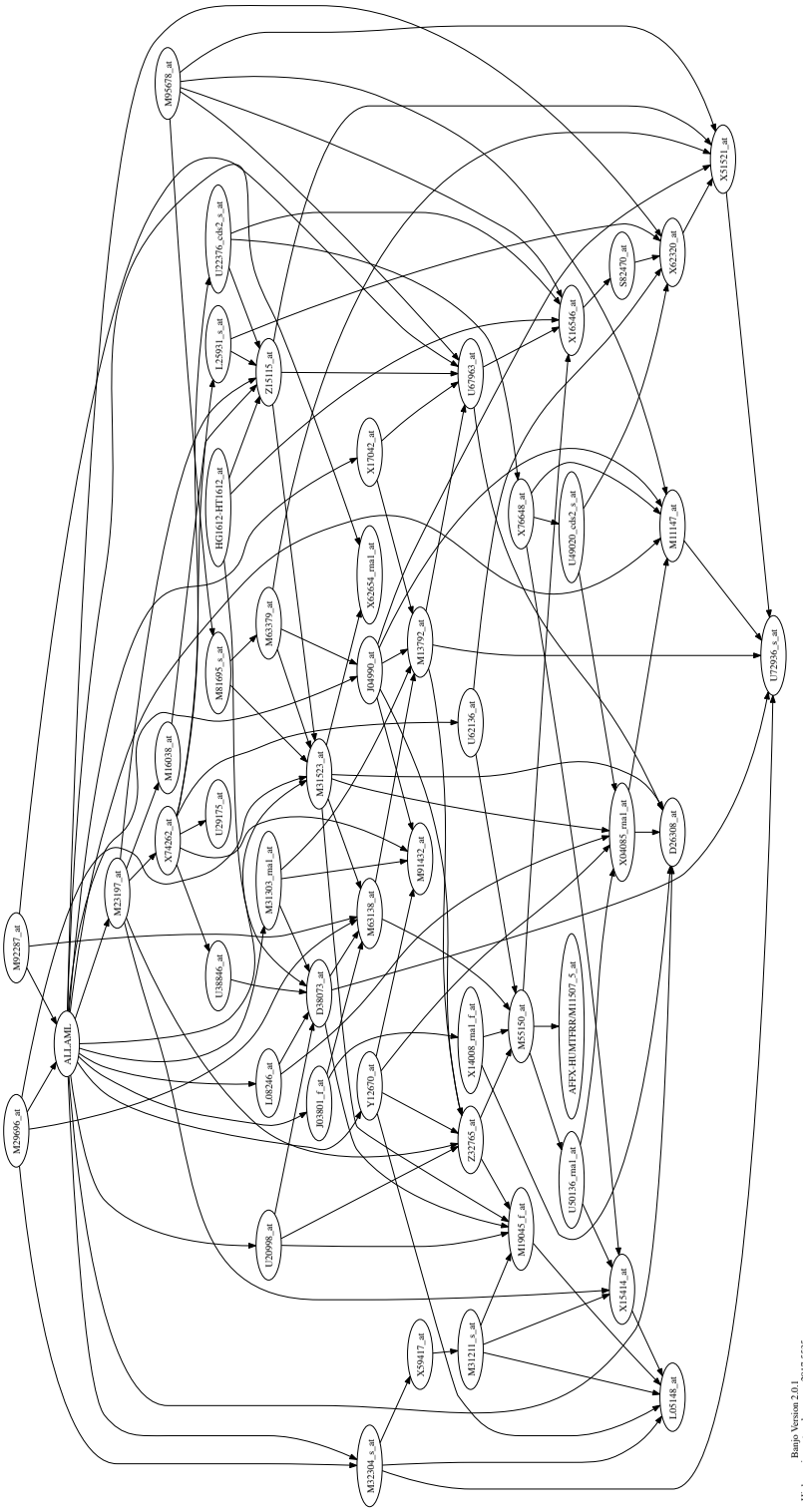


FIGURE 7.8 – Graphe ayant obtenu le meilleur score BDe.

Glossaire

ADN	Acide DesoxyibuNucléique. Molécule polymère support de l'information génétique constituée de nucléotides, eux-mêmes composés d'un phosphate, d'un sucre (desoxyribose), et de l'une des quatre base azotée suivantes : adénine, cytosine, guanine et thymine. L'ADN est bicaténaire, c'est-à-dire qu'elle est formée de l'accolement de deux brins. Ces deux brins sont liés par complémentarité des bases azotées. L'adénine est complémentaire à la thymine, et la guanine est complémentaire à la cytosine. Les deux brins de la molécule sont donc semblables mais inversés. Elle est utilisée par tout le Vivant, on la trouve dans les chromosomes, les plasmides, les mitochondries, les chloroplastes. Seuls certains virus ont un génome qui n'est pas constitué d'ADN mais d'ARN., 11
ADNc	ADN complémentaire : ADN issu d'une rétrotranscription, c'est-à-dire synthétisé à partir d'une molécule d'ARN. Cette synthèse peut être soit naturelle (comme chez les retrovirus), soit artificielle (en utilisant une enzyme de retrovirus), 37, 39
agoniste	Un agoniste d'une molécule M est une autre molécule se plaçant dans les mêmes récepteurs que M, engendrant la même action que M, 107
antagoniste	Un antagoniste d'une molécule M est une autre molécule se plaçant dans les mêmes récepteurs que M mais qui n'engendre pas d'action, 107
arête	Objet reliant deux sommets dans un graphe., 52

arc	Arête dirigée, partant d'un sommet parent, et pointant un sommet enfant., 52
ARN	Transcrit dans sa forme utilisable par la cellule. Il peut être messager lui permettant d'être traduit en protéine, ribosomique (constitue alors les ribosomes qui servent à la traduction des messagers), de transfert (participe à la polymérisation des acides aminés), ou d'autres types participant à la régulation de la transcription., 32
cytoplasme	Le cytoplasme est l'intérieur d'une cellule. Chez les procaryotes, le cytoplasme est composé de tout les constituants internes de la cellule, chez les eucaryotes, il est délimité par la membrane externe du noyau., 36
enzymes	protéines catalysant une réaction enzymatique, 27
eucaryotes	organismes vivants possédant un noyau (plantes, animaux, champignons. . ., 36
gène	une définition opérationnelle (d'autres existent) : séquence nucléique se situant sur un chromosome pouvant être transcrite en ARN par l'ARN polymérase, 11
graphe	Un graphe est un objet mathématique constitué de sommets et d'arêtes reliant des sommets., 26
méta-analyse	Analyse utilisant d'autres analyses dans un but de synthèse des informations, ou même de proposer de nouvelles conclusions., 13
métabolisme	ensemble des réactions de synthèse (anabolisme) et de dégradation (catabolisme) de macromolécules à l'échelle d'un organisme, 27

PCR	« Polymerase Chain Reaction », outil de biologie moléculaire permettant l'amplification de fragments d'ADN. Le principe est basé sur des cycles de température, permettant la séparation des double-brins et l'utilisation d'une ADN polymerase particulière capable de résister aux hautes températures., 37
phénotype	Un phénotype est un caractère visible chez un être vivant. C'est la résultante de l'interaction entre le génotype et l'environnement., 24
protéines	Une protéine est formée à partir d'un ou plusieurs polypeptides qui sont des polymères d'acides aminés. Ils sont issus de la traduction d'un ARN messager dont chaque triplet nucléotidique est associé à un acide aminé en respectant le code génétique : un ensemble de règles d'association entre les 64 triplets possibles et les 20 acides aminés connus comme existant dans la nature., 11
SNP	Single Nucleotide Polymorphism. A l'échelle d'une population, un SNP est dans un génome une position unique (une base azotée) qui varie., 13
sommet	Objet d'un graphe pouvant être relié à d'autres sommets du même graphe par des arêtes. Les sommets sont aussi nommés « nœuds »., 27
traits d'histoire de vie	Un trait d'histoire de vie est un caractère influant la reproduction ou la survie d'un individu., 12
transcription	La transcription est le processus produisant une molécule d'ARN à partir d'un gène., 32
transcrit	Produit de la transcription d'une séquence d'ADN. Un transcrit n'est pas forcément mâture. On parle alors de « transcrit primaire »., 32

TITRE en français

Les Réseaux Bayésiens : classification et recherche de réseaux locaux en cancérologie

RÉSUMÉ en français

En cancérologie, les puces à ADN mesurant le transcriptome sont devenues un outil commun pour chercher à caractériser plus finement les pathologies, dans l'espoir de trouver au travers des expressions géniques : des mécanismes, des classes, des associations entre molécules, des réseaux d'interactions cellulaires. Ces réseaux d'interactions sont très intéressants d'un point de vue biologique car ils concentrent un grand nombre de connaissances sur le fonctionnement cellulaire. Ce travail de thèse a pour but, à partir de ces mêmes données d'expression, d'extraire des structures pouvant s'apparenter à des réseaux d'interactions génétiques. Le cadre méthodologique choisi pour appréhender cette problématique est les « Réseaux Bayésiens », c'est-à-dire une méthode à la fois graphique et probabiliste permettant de modéliser des systèmes pourtant statiques (ici le réseau d'expression génétique) à l'aide d'indépendances conditionnelles sous forme d'un réseau. L'adaptation de cette méthode à des données dont la dimension des variables (ici l'expression des gènes, dont l'ordre de grandeur est 10^5) est très supérieure à la dimension des échantillons (ordre 10^2 en cancérologie) pose des problèmes statistiques (de faux positifs et négatifs) et combinatoires (avec seulement 10 gènes on a 4×10^{18} graphes orientés sans circuit possibles). A partir de plusieurs problématiques de cancers (leucémies et cancers du sein), ce projet propose une stratégie d'accélération de recherche de réseaux d'expression à l'aide de Réseaux Bayésiens, ainsi que des mises en œuvre de cette méthode pour classer des tumeurs, sélectionner un ensemble de gènes d'intérêt reliés à une condition biologique particulière, rechercher des réseaux locaux autour d'un gène d'intérêt. On propose parallèlement de modéliser un Réseau Bayésien à partir d'un réseau biologique connu, utile pour simuler des échantillons et tester des méthodes de reconstruction de graphes à partir de données contrôlées.

MOTS-CLEFS en français

réseaux cellulaires ; transcriptome ; Réseaux Bayésiens ; classification ; sélection de variables ; cancer

TITRE en anglais

Classification and capture of regulation networks with Bayesian Networks in oncology

RÉSUMÉ en anglais

In oncology, microarrays have become a classical tool to search and characterize pathologies at a deeper level than previous methods, using genetic expression to find the mechanisms, classes, molecular associations, and cellular interaction networks of different cancers. From a biological point of view, these cellular networks are interesting because they concentrate a large amount of knowledge about cellular processes. The goal of this PhD thesis project is to extract structures that could correspond to genetic interaction networks from the expression data. "Bayesian Networks", *i.e.* a graphic and probabilistic method that models even static systems (like the expression network) with conditional independences, are used as the framework to investigate this problem. The adaptation of this method to data where the dimension of the variables (about 10^5 for gene expression) is much greater than the dimension of the samples (about 10^2 in oncology) aggravates some statistical and combinatorial problems. For several cancer problematics, this project proposes an acceleration strategy for capturing expression networks with Bayesian Networks and some methods to classify tumors, finding gene signatures of particular biological conditions by searching for local networks in the neighborhood of a gene of interest. In parallel, we propose to model a Bayesian Network from a known biological network, which is useful to simulate samples and to test these methods to reconstruct graphs from controlled data.

MOTS-CLEFS en anglais

cellular networks ; transcriptome ; Bayesian Networks ; classification ; gene selection ; cancer

DISCIPLINE : Bioinformatique

INTITULE ET ADRESSE DE L'U.F.R. OU DU LABORATOIRE :

Laboratoire de Biométrie et Biologie Évolutive - UMR 5558 CNRS
Bâtiment Gregor Mendel - Université Claude Bernard Lyon1
43, bd du 11 novembre 1918 - 69622 Villeurbanne cedex