Contributions à la fouille de données complexes

Claude Pasquier

UNIVERSITÉ CÔTE D'AZUR

École doctorale STIC
Sciences et Technologies de l'Information et de la Communication

Mémoire présenté en vue de l'obtention du

DIPLÔME D'HABILITATION À DIRIGER DES RECHERCHES

CONTRIBUTIONS À LA FOUILLE DE DONNÉES COMPLEXES

---

par

Claude Pasquier

Date de soutenance : 20 septembre 2018

JURY

Richard Christen	Directeur de Recherche	Examinateur
Amedeo Napoli	Directeur de Recherche	Rapporteur
Frédéric Précioso	Professeur des Universités	Examinateur
Céline Robardet	Professeur des Universités	Rapporteur
Alain Robichon	Directeur de Recherche	Examinateur
Alexandre Termier	Professeur des Universités	Rapporteur

I3S : Laboratoire d'Informatique, Signaux et Systèmes de Sophia-Antipolis
UMR7271 - UNS CNRS - 06900 Sophia Antipolis - France

Remerciements

Les travaux de recherche présentés dans ce mémoire ont été réalisés au sein de différents laboratoires. Ils n'ont pas été accomplis ex nihilo mais sont le fruit de collaborations nombreuses. Je tiens donc à remercier ici toutes les personnes qui ont concouru à la réalisation de ces projets en me donnant la chance de travailler avec elles : Paul Franchi Zannettacci, mon directeur de thèse, qui m'a initié au travail de recherche, Stavros Hamodrakas à l'Université d'Athènes, qui m'a permis d'aborder le domaine de la bioinformatique, Didier Parigot, de l'INRIA Méditerranée, qui m'a beaucoup appris des divers paradigmes de programmation, Richard Christen, qui a fortement influencé ma manière d'aborder la recherche scientifique et Nazha Selmaoui-Folcher, de l'Université de la Nouvelle Calédonie, qui m'a fait découvrir la fouille de données spatio-temporelles. Je veux également remercier tous les collègues, jeunes chercheurs ou étudiants que j'ai croisé ces dernières années, pour toutes les discussions et moments de réflexion que nous avons partagés.

Je veux aussi remercier chaleureusement, les membres du jury Richard Christen, Amedeo Napoli, Frédéric Précioso, Céline Robardet Alain Robichon et Alexandre Termier pour m'avoir fait l'honneur d'examiner et de rapporter mon travail.

Table des matières

Liste des figures

Liste des tableaux

Chapitre 1

Aperçu des travaux, résultats et perspectives

Sommaire

1.1 Modélisation des documents par langage à prototype

J'ai effectué un DEA puis une thèse dans le domaine du génie logiciel sous la supervision de Paul Franchi-Zannettacci, professeur à l'Université de Nice - Sophia Antipolis. Mes travaux portaient sur la modélisation et la manipulation de documents structurés. Ils visaient à généraliser les techniques du génie logiciel au domaine de la gestion des documents. Bien que peu de recherches étaient consacrées à ce thème, cette approche paraissait naturelle si l'on considère que les programmes informatiques constituent une forme simplifiée de documents pour lesquels se posent déjà des problèmes d'organisation, de partage, d'interprétation, de génération ou de modification dynamique. Comme pour les programmes, le traitement de documents structurés nécessite l'utilisation de grammaires abstraites pour définir les règles syntaxiques et sémantiques de leur composition. Par ailleurs, l'introduction de documents structurés actifs (i.e. dont la représentation abstraite peut évoluer dynamiquement) nécessite la mise en œuvre de mécanismes permettant d'une part d'organiser les différents modèles de documents et d'autre part, de décrire l'évolution des différentes versions d'un même document. Les problèmes de taxinomie et d'héritage sont très proches des mécanismes d'héritage et de création dynamique développés par les langages à objets. Se posent également des questions cruciales de la réutilisation des documents et de leur inter-opérabilité dans le cas de systèmes hétérogènes [Pasquier, 1991].

> *Le système s'inspire de la programmation par prototypes pour proposer un partage dynamique des documents reposant sur la délégation et le clonage*

Mes travaux de thèse ont permis le développement du système BIBLE dont la principale contribution résidait dans une organisation intelligente des différentes représentations logiques et physiques associées aux documents manipulés [Pasquier, 1992, 1994]. Le système que j'ai élaboré mettait en œuvre une forme de partage dynamique reposant sur la délégation et le clonage de prototypes. Le mécanisme sous-jacent était inspiré de la programmation orientée prototype qui était un nouvel axe de recherche prometteur au début des années 1990.

A l'issue de ma thèse, persuadé que les documents structurés allaient très rapidement supplanter les documents formatés, j'ai créé l'entreprise « PARADOCS » pour valoriser mes recherches. Après deux années de développements, les perspectives de succès s'amenuisant, j'ai décidé de changer de thématique en effectuant un post doctorat en bioinformatique.

1.2 Prédiction de la structure des protéines

Au sein du **département de biologie moléculaire de l'université d'Athènes**, j'ai travaillé sur la prédiction de la structure secondaire des protéines sous la supervision du professeur Stavros Hamodrakas. Plusieurs approches, combinant des connaissances biologiques connues (influence de la charge et de l'hydrophobicité des acides aminés dans la structure des protéines), des motifs récurrents identifiés sur les structures primaires et l'apprentissage automatique (réseau de neurones type perceptron) ont permis de mettre au point des logiciels performants qui ont rapidement été adoptés par les biologistes (200 000 prédictions effectuées durant mon post-doctorat et les deux années suivantes) et sont toujours utilisés actuellement (60 références à l'utilisation des logiciels au cours des 5 dernières années ; plus d'un million de requêtes effectuées au total au 20 décembre 2017). Mes travaux s'inscrivaient dans le cadre du **projet Européen GeneQuiz** dont le but était de concevoir un système logiciel permettant d'annoter automatiquement les gènes. Grâce à ce séjour post-doctoral, j'ai découvert la recherche en biologie et les vertus de l'interdisciplinarité. Les résultats de GeneQuiz n'auraient,

en effet, pas pu être obtenus sans la collaboration étroite entre chercheurs en informatique et en biologie. Cette expérience a été décisive pour la suite de ma carrière.

Dans mes recherches, je me suis plus particulièrement intéressé à une catégorie spécifique de protéines : les protéines membranaires. La plupart des cibles des médicaments correspondent en effet soit à des protéines insérées dans la membrane (protéines membranaires), soit à des facteurs secrétés. Malgré leur importance physiologique, les informations sur la structure tridimensionnelle des protéines membranaires sont peu abondantes car plus difficiles à obtenir que celles des protéines de type globulaire.

> *La méthode de détection des régions transmembranaires combine une analyse d'hydrophobicité et une détection de motifs types caractérisant les extrémités de segments*

Il est bien connu que les régions transmembranaires sont des régions hydrophobes par excellence. Malheureusement, les méthodes de prédiction basées sur une analyse des variations d'hydrophobicité d'une séquence donnée peinent à distinguer les segments correspondant à de réelles parties transmembranaires des régions ayant simplement une forte proportion de résidus hydrophobes. L'idée que j'ai développée consistait à raffiner une analyse d'hydrophobicité standard par une détection de motifs types signalant les extrémités potentielles (début et fin) des régions transmembranaires. Ainsi, les régions hydrophobes qui ne sont pas délimitées par des motifs clairs de début et de fin pouvaient être éliminées. A l'opposé, la mise en évidence de motifs favorables pouvait permettre d'isoler des régions transmembranaires qui ne sont pas détectables par une simple analyse d'hydrophobicité. Les résultats obtenus avec l'algorithme soutenaient favorablement la comparaison avec les meilleures méthodes développées jusqu'alors. Les résultats ont été publiés dans la revue Protein Engineering [Pasquier et collab., 1999a]. La méthode développée, appelée PRED-TMR, est utilisable en ligne sur le site du laboratoire de biophysique de l'Université d'Athènes.

En utilisant la même stratégie, j'ai travaillé avec Thodoris Liakopoulos, un étudiant en licence, au raffinement des motifs caractérisant une séquence protéique lors des traversées de la membrane pour élaborer une méthode de prédiction de l'orientation des segments transmembranaires. La qualité des prédictions, qui a été testée sur plusieurs ensembles de test, avoisinait les 90 %. Ces résultats ont été présentés à la 21e conférence HSBS (21st conference of the Hellenic Society for Biological Sciences) [Liakopoulos et collab., 1999], puis publiés dans le journal Protein Engineering [Liakopoulos et collab., 2001]. L'implémentation de l'algorithme, OrienTM, peut être exécuté en ligne.

Avec Vasilis Promponas, doctorant dans l'équipe, nous avons travaillé sur une méthode ensembliste combinant les résultats de plusieurs algorithmes de prédictions. Les résultats ont été présentés à la 20e conférence HSBS (20st conference of the Hellenic Society for Biological Sciences) [Promponas et collab., 1998], puis publiés dans le journal « In Silico Biology » [Promponas et collab., 1999]. Une implémentation de la méthode CoPreTHi est exécutable en ligne.

> *Associés à des classifieurs basés sur des réseaux de neurones, les méthodes développées sont adaptées à l'annotation sans supervision de génomes complets*

Partant du constat que les algorithmes de prédiction les plus performants identifiaient, dans 10 à 30 % des cas, des segments transmembranaires dans des protéines de type globulaire ou fibreux, j'ai développé un réseau de neurones de type perceptron capable de prédire si une séquence inconnue contient ou non des segments transmembranaires. L'idée de base est d'utiliser, en entrée du réseau de neurones, une représentation d'une partie de la séquence analysée de manière à obtenir un résultat unique indiquant si la région considérée est représentative d'un segment transmembranaire ou non. Le réseau de neurones que j'ai proposé

n'utilisait pas une codification discrète des acides aminés (ce qui était la pratique standard à l'époque) mais les valeurs de certaines caractéristiques identifiées précédemment. La méthode a été présentée à la 21e conférence HSBS (21st conference of the Hellenic Society for Biological Sciences) [Pasquier et collab., 1999b] et détaillée dans un article publié dans la revue Protein Engineering [Pasquier et Hamodrakas, 1999]. Une implémentation PRED-TMR2, est utilisable en ligne.

Encouragés par les bonnes performances de la méthode, nous avons ensuite développé, Vasilis Promponas et moi-même, deux autres réseaux de neurones. L'un pour identifier les protéines globulaires et l'autre, les protéines fibreuses. Nous avons utilisé, pour chacun de ces classifieurs 60 paramètres d'entrée. Parmi ceux-ci, 20 valeurs correspondent à la fréquence constatée de chacun des 20 acides aminés dans la séquence, 10 correspondent à la fréquence d'apparition des acides aminés représentatifs d'une classe particulière (composants d'hélices alpha, résidus hydrophobes, etc.). Les 30 autres valeurs indiquent la périodicité de chaque acide aminé ou groupe de résidus à l'intérieur de la séquence. Ces périodicités sont mises en évidence à l'aide d'une méthode baptisée FT qui est basée sur la transformée de Fourier et sur laquelle j'ai travaillé au début de mon post-doctorat [Pasquier et collab., 1998a,b]. En combinant les différents classifieurs, nous avons obtenu un système permettant d'annoter des séquences protéiques selon quatre catégories : transmembranaires, globulaires, fibreuses et mixtes (pour les protéines ne rentrant dans aucune autre catégorie). Le système mis en place fut l'un des premiers autorisant le traitement de génomes complets dans des temps raisonnables. Trente génomes ont ainsi été complètement analysés par le système. La méthode ainsi que les résultats sont détaillés dans un article publiés dans le journal Proteins [Pasquier et collab., 2001]. PRED-CLASS, une implémentation de la méthode, est utilisable en ligne. Ce système a permis d'obtenir rapidement une vision synthétique de la composition de nouveaux génomes. Il a également permis d'apporter des éléments de réponses aux questions suivantes : quelle est la fréquence des protéines membranaires, globulaires, fibreuses ou mixtes à l'intérieur d'un génome complet ? Est-ce que les répartitions entre les différentes classes sont liées au type d'organisme étudié ?

Les travaux effectués durant mon post-doctorat participaient à l'objectif d'annotation automatique et à grande échelle de gènes ou produits de gènes poursuivi par le projet GeneQuiz. Les différentes méthodes développées ont été intégrées à un outil, utilisable en ligne, dédié à l'analyse des séquences et des structures des protéines [Liakopoulos et collab., 2000, 2001]. Un article de synthèse, écrit par le consortium a été publié en 2003 dans la revue Bioinformatics [Iliopoulos et collab., 2003].

1.3 Contributions à la programmation générative

A mon retour en France, j'ai travaillé deux années à l'**INRIA méditerranée**, au sein des équipes Lemme puis Oasis, et une année à **Schlumberger** dans le domaine du génie logiciel, sur une thématique s'inscrivant dans la continuité de mon travail de thèse. Mes recherches portaient sur la manipulation des documents structurés et sur la programmation générative, c'est-à-dire la production automatique d'outils logiciels à partir de spécifications. Les déclarations et transformations des spécifications s'appuyaient sur les standard du Web en émergence, c'est à dire principalement XML, XSLT et les services Web. J'ai contribué à l'élaboration de la fabrique logicielle SmartTools et implémenté des solutions pour des langages dédiés. Je me suis également intéressé à la distribution des opérations d'édition sur un réseau.

> *Les recherches effectuées ont abouti à l'élaboration d'un générateur d'outils d'édition distribués par assemblage de composants graphiques*

KeTuk illustre une manière de définir des éditeurs par assemblage de composants graphiques. Un affichage particulier est obtenu par une transformation du document à éditer en une hiérarchie de composants java. Le processus de transformation est rendu inversible de manière à assurer que toute modification des objets java puisse être répercutée sur le document XML. L'architecture qui a été mise en place est suffisamment ouverte pour permettre l'intégration de n'importe quel composant Java. Les travaux portant sur KeTuK n'ont pas été publiés. Cependant, le système est consultable et téléchargeable à partir de ma page Web.

Capitalisant les travaux effectués sur KeTuK, j'ai travaillé avec Laurent Théry à des solutions permettant de distribuer les opérations d'édition sur un réseau. Le principe de fonctionnement retenu consiste à regrouper sur un serveur les documents XML à éditer ainsi que toutes les applications logicielles assurant l'édition de données. Les opérations d'édition réalisées par les clients (qui interagissent sur une vue personnalisée des données) sont transmises au serveur qui effectue les opérations adéquates sur le document XML central et informe chaque client des modifications à répercuter sur les vues. Les données qui transitent sur le réseau ne concernent pas des documents complets mais des codifications en XML des opérations effectuées ou à effectuer. Le système, qui limite le code nécessaire chez le client au strict minimum, est ainsi particulièrement adapté à l'édition de documents sur des matériels ayant des ressources limitées (téléphones portables, organiseurs électroniques). Le système a été présenté à un workshop de la conférence ECOOP (European Conference on Object-Oriented Programming) [Pasquier et Théry, 2000].

> *Les recherches sur la génération automatique d'outils logiciels à partir de spécifications ont été utilisé dans un cadre industriel pour développer un environnement de développement JavaCard*

Le principe d'édition par assemblage de composants et la manière de distribuer les vues d'édition étaient similaire à l'approche adoptée dans SmartTools, un projet dirigé par Didier Parigot. Après avoir collaboré de manière informelle avec Didier Parigot et son équipe durant près de deux années, j'ai rejoint le projet SmartTools en 2001.

SmartTools est un projet ambitieux qui visait à la génération automatique d'environnements de développements à partir de spécifications. Le projet est fortement basé sur XML et les technologies objets et se veut générique et hautement paramétrable. Outre les possibilités d'édition graphique sur des vues pouvant être réparties sur un réseau, SmartTools permet de générer automatiquement des outils (parseur, afficheurs, visiteurs sur la structure du document) à partir d'une spécification d'un langage. J'ai participé au développement de SmartTools et assuré le transfert de compétences entre l'équipe Oasis de l'INRIA et l'entreprise Schlumberger. J'ai en particulier utilisé dans un cadre industriel les outils fournis par l'atelier logiciel SmartTools pour développer un application dédiée à JavaCard œuvrant au niveau des sources et du bytecode. SmartTools est distribué en licence libre auprès des utilisateurs non industriels depuis novembre 2001. Le système a été présenté à des workshops de conférences de premier plan : International Conference on Software Engineering (ICSE) en 2001 [Attali et collab., 2001a], European Joint Conferences on Theory and Practice of Software (ETAPS) en 2001 et 2002 [Attali et collab., 2001b; Parigot et collab., 2002].

1.4 Intégration et analyse de données transcriptomiques

À partir de 2002, date à laquelle j'ai été recruté par le CNRS, j'ai orienté très nettement mes activités vers la bioinformatique. Je me suis efforcé durant toutes les années qui ont suivi, de poursuivre des recherches aux frontières de la biologie moléculaire et de l'informatique. Mon activité de recherche, à partir de cette date, tout en étant diversifiée quant aux méthodes

et aux applications, se focalise sur le thème de **l'analyse des données**.

À l'**iBV (Institut de Biologie Valrose - UMR7277)**, j'ai mené des recherches portant sur l'intégration des données biologiques et l'analyse des données d'expression des gènes. Au début des années 2000, il existait de nombreuses sources stockant des données de biologie moléculaire, mais toutes ces sources étaient complètement indépendantes les unes des autres. J'ai proposé une méthodologie facilitant l'intégration des données en utilisant les technologies du Web Sémantique dont les préceptes étaient très proches des règles qui régissent actuellement le Web des Données (officiellement né en 2007). Ces travaux ont été publiés dans la revue « Biochimie » [Pasquier, 2008] et on constitués un chapitre du livre « Data Management in the Semantic Web » [Pasquier, 2011].

Concernant la fouille de données transcriptomiques proprement dite, nous pouvons distinguer trois axes de recherche : les approches guidées par les données numériques, les approches guidées par les connaissances et les approches visant à utiliser parallèlement les deux aspects (une analyse détaillée de ces différentes approches a été publiée dans « International Journal of Software and Informatics » [Pasquier et collab., 2008]). J'ai travaillé avec des collègues de l'iBV (Ronnie Alves, Richard Christen, Karim De Fombelle, Fabrice Girardot, Saïd El Kasmi) et des membres de l'I3S (Martine Collard, Ricardo Martinez, Nicolas Pasquier) sur chacun de ces trois axes.

> *Les recherches ont permis l'élaboration d'un système d'annotation automatique des gènes en utilisant les informations biologiques provenant d'une base de connaissances*

L'approche guidées par les données consiste à identifier des groupes de gènes dont l'expression varie de manière similaire pour ensuite intégrer les connaissances sur les gènes. Mes recherches sur ce thème se sont concrétisées sous la forme d'un système intégré appelé THEA (Tools for High-throughput Experiments Analysis). Le logiciel intègre plusieurs algorithmes de fouille de données permettant d'annoter automatiquement les groupes de gènes partageant le même profil d'expression et de mettre en évidence les corrélations existantes entre les niveaux d'expression de certains gènes et leur localisation sur le génome. Les expérimentations montrent que l'utilisation de THEA permet, non seulement d'obtenir facilement et rapidement tous les résultats mis en évidence de manière manuelle mais que cela permet aussi d'identifier de nouvelles informations. THEA est décrit en détail dans un article publié dans « Bioinformatics » [Pasquier et collab., 2004].

L'approche guidée par la connaissance consiste à trouver d'abord des groupes de gènes co-annotés puis, dans un second temps, à intégrer les données concernant les profils d'expression. La méthode AGGC (Analyse des Groupes de Gènes Co-exprimés), que Ricardo Martinez, Nicolas Pasquier et moi même avons développée, s'inscrit dans cette approche.

Les tests que nous avons effectués montrent que les annotations fonctionnelles fournies par AGGC constituent un outil efficace et rapide permettant de réduire la complexité du problème de l'analyse de données transcriptomiques en intégrant des informations de diverses natures sur les gènes. Les résultats expérimentaux ont montré l'intérêt de l'approche et ont permis d'identifier des informations pertinentes sur les processus biologiques étudiés [Martinez et collab., 2005].

Les travaux s'inscrivant dans l'approche guidée par les connaissances ont été publiés dans « Journal of Integrative Bioinformatics » [Martinez et collab., 2006a], dans la « Revue des Nouvelles Technologies de l'Information » [Martinez et collab., 2008c] et ont été présentés à la conférences ICDS (International Conference on Discovery Science) [Martinez et collab., 2006c] ainsi qu'à l'occasion des rencontres de la société francophone de classification (SFC) [Martinez et collab., 2006b]. L'implémentation de AGGC est distribuée sous licence Open Source.

> *Notre approche, permet d'extraire à partir de grandes quantités de données, un ensemble minimal de règles associant les informations connues sur les gènes et leurs niveaux d'expression*

Les deux approches décrites précédemment, confèrent, selon le cas, une importance prépondérante soit aux données d'expression, soit aux connaissances biologiques. Ceci présente un certain nombre d'inconvénients. Pour pallier ceux-ci, j'ai travaillé, avec Nicolas Pasquier et Ricardo Martinez à une solution n'imposant aucune priorité sur le traitement des différentes sources de données. Nous avons proposé d'utiliser la méthode de découverte des règles d'association (Association Rule Discovery (ARD)) qui est une technique d'exploration de données non supervisée permettant de découvrir les liens entre des ensembles de valeurs variables (items).

Nous avons développé une application appelée GenMiner pour résoudre les faiblesses des approches existantes et exploiter complètement les capacités de l'ARD dans le cadre d'une fouille de données biologiques. GenMiner permet l'utilisation conjointe des informations connues sur les gènes et de leur niveau d'expression dans certaines conditions de manière à découvrir les relations qui existent entre connaissance a priori et mesures expérimentales. Notre méthode inclut un nouvel algorithme, appelé NorDI (Normal Discretization algorithm) pour discrétiser les mesures d'expression des gènes et générer des profils d'expression (présenté à la conférence BIBM (International Conference on Bioinformatics and Biomedicine)) [Martinez et collab., 2007]).

Les expérimentations, que nous avons effectuées, ont confirmé les avantages de GenMiner par rapport aux approches connues. GenMiner permet de rechercher des règles d'association en utilisant un support minimum bien plus petit que ce qui est possible avec les approches classiques. De plus, GenMiner réduit considérablement le nombre de règles extraites, ce qui facilite énormément le travail d'exploration et d'interprétation par l'utilisateur final. La méthode proposée ainsi que des analyses détaillées des résultats obtenus ont été publiées dans le journal « Bioinformatics » [Martinez et collab., 2008a] et présentées à la conférence CIBB (international conference on computational intelligence methods for bioinformatics and biostatistics) [Martinez et collab., 2008b]).

1.5 Fouille de graphes attribués

Durant mon séjour au **PPME (Pôle Pluridisciplinaire de la Matière et de l'Environnement - EA3325)**, un laboratoire regroupant des géologues, des physiciens et des informaticiens autour d'une thématique de gestion des connaissances des écosystèmes littoraux, j'ai développé des recherches portant sur l'élaboration de nouvelles méthodes de fouille de données intégrant conjointement la composante spatiale et temporelle. Plus précisément, je m'intéressais à la définition et l'extraction de nouveaux motifs spatio-temporels susceptibles de décrire des phénomènes dans le but de les prédire.

À mon arrivée, les chercheurs du PPME développaient déjà des méthodes de fouille de données portant sur les graphes attribués. Les graphes attribués sont tout simplement des graphes dans lesquels les sommets ou les arêtes sont étiquetés par plusieurs attributs. Pour représenter des phénomènes spatio-temporels, ils utilisaient des graphes attribués orientés dans lesquels les sommets représentent des objets spatiaux, caractérisés par un ensemble d'attributs, tandis que les arcs expriment leur proximité spatio-temporelle (par exemple des objets voisins dans des temps consécutifs).

Avec deux collègues, Nazha Selmaoui-Folcher et Frédéric Flouvat, enseignants chercheurs à l'Université de la Nouvelle Calédonie et Jérémy Sanhes, étudiant en thèse, nous nous sommes basés sur cette structure de données pour mener des recherches portant sur la fouille de motifs fréquents, sur la définition et l'extraction de motifs condensés et sur une approche permettant d'intégrer les connaissances des experts dans le processus de fouille de données.

> *Nos algorithmes permettent la recherche de motifs fermés fréquents ou de chemins pondérés dans différentes structures attribuées en combinant fouille d'itemsets et parcours structurel*

La stratégie de recherche de motifs fréquents que nous avons proposée est de partir d'un ensemble de motifs initiaux composés d'un nœud associé avec un unique item et de construire l'arbre de recherche par un parcours de l'arbre attribué couvrant en utilisant un ordre basé sur un code de représentation des graphes acycliques attribués que nous avons défini. Nous avons proposé une méthode complète de parcours de l'espace de recherche qui combine deux types d'extensions : l'extension d'itemset et l'extension de structure.

Nous avons défini empiriquement la notion de fermeture sur les graphes attribués en adoptant la définition suivante : un graphe attribué est un graphe attribué fermé s'il n'est inclus dans aucun autre graphe attribué qui possède le même support que lui. Dans le cadre de motifs ayant une structure de graphe, le calcul des fermés nécessite d'effectuer plusieurs tests d'isomorphisme de graphes qui sont des opérations coûteuses. Nous avons proposé deux autres représentations résumées des motifs qui sont définies soit en fonction de l'inclusion portant sur les itemsets (les motifs c-fermés), soit en fonction de l'inclusion portant sur la structure (les motifs s-fermés). Nous avons montré que l'énumération des motifs c-fermés permet de réduire drastiquement à la fois le nombre de motifs retournés et le temps d'exécution. Des tests ont montré que cette représentation condensée offrait un bon compromis entre vitesse d'exécution et concision des résultats. Ces travaux ont tout d'abord été présentés dans les conférences PAKDD (Pacific-Asia Conference on Knowledge Discovery and Data Mining) [Pasquier et collab., 2013b] et EGC (conférence francophone sur l'Extraction et la Gestion des Connaissances) [Pasquier et collab., 2013a] où nous avons obtenu le prix du meilleur article académique. Une version étendue a par la suite été publiée dans la revue « Knowledge and Information Systems » [Pasquier et collab., 2016].

Après avoir travaillé sur les arbres attribués, nous avons étendu notre méthode pour traiter les DAG attribués (résultats non publiés) et les graphes attribués orientés. Si le passage de la fouille d'arbres attribués à celle des DAG attribués a été relativement aisée, la prise en compte des cycles a nécessité de modifier le code des arbres attribués couvrants et à éliminer les motifs isomorphes qui sont fatalement générés pour toutes les explorations qui débutent sur un autre nœud faisant partie du cycle. Les motifs qui possèdent de nombreux isomorphismes de sous-graphes avec le graphe analysé posent des difficultés à tous les algorithmes existants, car le problème d'isomorphismes de sous-graphes en lui même est NP-complet.

Nous avons proposé deux optimisations qui permettent d'une part d'élaguer l'arbre de recherche généré à partir d'un motif automorphe et d'autre part, de supprimer certaines manières d'obtenir des motifs automorphes qui ne permettent pas de générer de nouveaux motifs canoniques. Notre démarche est expliquée dans un article publié dans « Knowledge and Information Systems » [Pasquier et collab., 2017b]. Ces travaux ont également été présentés à la conférence francophone sur l'Extraction et la Gestion des Connaissances (EGC) [Pasquier et collab., 2014b].

Nous avons abordé le problème de l'extraction de chemins pondérés fréquents dans un unique graphe attribué sans cycle (DAG) où chaque poids exprime la fréquence d'une transition. Les chemins fréquents sont utilisés pour analyser la relation causale entre séquences d'événements et/ou attributs. Comme le nombre de motifs peut être très grand, nous avons conçu une représentation condensée pour de telles collections. Ces travaux se situent à l'intersection de plusieurs sujets importants tels que l'extraction de séquences et la fouille de graphes. À partir d'une mesure de support anti monotone des chemins que nous avons définie, nos travaux ont abouti à la première définition d'une représentation condensée (exacte) dans le cas d'un unique graphe attribué. Sur la base de cette représentation condensée, nous avons proposé un algorithme efficace pour extraire des motifs fréquents (les chemins pondérés) dans un unique DAG attribué. Ces travaux ont été présentés aux conférences IJCAi (International Joint Conference on Artificial Intelligence) [Sanhes et collab., 2013b] et EGC

(conférence francophone sur l'Extraction et la Gestion des Connaissances) [Sanhes et collab., 2013a]. Une version étendue est en préparation pour le journal « Knowledge and Information Systems ».

> *Nous avons exploité les modèles mathématiques définis par les experts d'un domaine pour dériver de nouvelles contraintes pouvant être intégrées à la fouille de motifs*

En constatant que, dans de nombreux contextes de science des données, les experts ont souvent capitalisé une partie de leur connaissance dans des modèles mathématiques nous avons proposé d'exploiter ces modèles pour dériver de nouvelles contraintes pouvant être intégrées aux processus de fouille et ainsi améliorer la pertinence des extractions. Nous avons défini une méthode permettant d'effectuer la recherche des motifs sous contrainte d'un modèle. Nous avons également étudié certaines propriétés des prédicats et contraintes afin de les exploiter pour optimiser les calculs de motifs. Nous avons montré que la prise en compte de contraintes provenant de modèles mathématiques permet de mieux cibler l'analyse, tout en améliorant les performances grâce à des propriétés des modèles. Nous avons ainsi obtenu des motifs plus pertinents, venant compléter ou contredire la connaissance experte sur les phénomènes étudiés. Ces travaux ont été présentés aux conférences ECAI (European Conference on Artificial Intelligence) [Flouvat et collab., 2014a] et RFIA (Reconnaissance de Formes et l'Intelligence Artificielle) [Flouvat et collab., 2014b].

1.6 Approches informatiques pour l'étude fonctionnelle des ARN

Au laboratoire **I3S (Laboratoire d'Informatique, Signaux et Systèmes de Sophia Antipolis - UMR7271)**, j'initie une nouvelle thématique de recherche en bio-informatique centrée sur le développement de méthodes de fouille de données/apprentissage automatique dédiés à l'étude fonctionnelle des ARN. Je m'efforce de ne pas perdre de vue l'aspect applicatif de mes recherche en entretenant des liens avec des équipes de recherche en biologie [Pasquier et collab., 2014a]. J'essaie également, dans tous les projets collaboratifs multidisciplinaires auxquels je participe, de trouver un équilibre entre recherche en biologie et recherche en informatique. L'idée de base de mon travail est de traiter les données disponibles de manière computationnelle pour en faire émerger de nouvelles connaissances. Plusieurs axes de recherches sont explorés : la prédiction des liens entre micro ARN (miARN) et maladies et l'étude des interactions des ARN dans la cellule.

> *Nos recherches utilisent la sémantique distributionnelle pour prédire les associations entre micro ARN et maladies*

Les miARN sont une classe de petits ARN qui peuvent réguler l'expression des gènes au niveau post-transcriptionnel et qui jouent un rôle critique dans de nombreux processus physiologiques tels que le développement, l'apoptose ou la différenciation cellulaire. Les dérégulations des miARN sont également étroitement liées au développement et à la progression de diverses maladies humaines, y compris le cancer.

En collaboration avec la startup Biomanda, j'ai travaillé à l'élaboration d'une nouvelle méthode permettant de prédire les associations entre miARN et maladies à partir de multiples sources de données. Nous regroupons toutes les données disponibles sur les miARN dans un tableau dans lequel chaque ligne représente un miARN, et chaque colonne, un type de données associée. Nous pouvons ainsi représenter les associations connues entre miARN et maladies, les gènes ciblés par les miARN, les miARN physiquement proches sur le génome, les familles d'appartenance, etc. Nous pouvons même faire figurer dans la même matrice les termes utilisés

dans les articles scientifiques traitant du miARN (pas forcément de l'association entre le miARN et une maladie).

Dans cette recherche, nous avons étudié l'utilisation de la sémantique distributionnelle pour analyser les données associées aux miARN. Nous avons émis l'hypothèse que la sémantique distributionnelle peut être utilisée pour révéler de nouvelles informations attachées aux miARN. L'idée de base de la sémantique distributionnelle, utilisée en linguistique, est que la signification d'un mot peut être déduite uniquement du contexte dans lequel ce mot est utilisé. Dans notre étude, on peut considérer que les miARN représentent des mots et que, par conséquent, les données associées aux miARN jouent le rôle du contexte. En gardant l'analogie, l'objectif est alors d'utiliser les informations disponibles sur les miARN dans leur ensemble (contexte) pour en déduire une nouvelle connaissance sur les miARN.

Les performances de notre algorithme sont caractéristiques d'un excellent classifier et correspondent à l'état de l'art dans le domaine. Nous avons publié une première version de la méthode en 2016 dans le journal « Scientific Reports [Pasquier et Gardès, 2016]. Dans une seconde phase, nous avons travaillé avec Denis Pallez à l'identification des paramètres permettant d'obtenir les meilleurs résultats. Cette paramétrisation, qui a permis d'améliorer sensiblement les performances de la méthode, a été effectuée à l'aide d'algorithmes évolutionnaires. Les résultats ont été également publiés dans le journal « Scientific Reports [Pallez et collab., 2017]. Le logiciel MiRAI a été déposé à l'Agence pour la Protection des Programmes.

> *Nos analyses suggèrent fortement que l'appariement ADN :ARN sous forme de triplex pourrait constituer une couche inattendue de régulation en biologie du développement*

Il est connu, depuis les années soixante, que certaines courtes séquences d'ARN sont susceptibles de s'apparier avec des zones particulières de d'ADN pour former des structures triple brins appelées ADN triplex. Bien que les structures en triple hélice aient été observées expérimentalement, on connaît encore très peu de chose sur leurs fonctions.

En collaboration avec Alain Robichon (équipe Génétique, Environnement et Plasticité de l'INRA), nous avons entrepris une étude consistant à localiser, quantifier et analyser les ADN triplex sur un génome dans le but d'accroître nos connaissances sur ces structures. Toute l'étude se fait in silico, de l'identification sur le génome des zones susceptibles de former des structures triplex à l'analyse des données obtenues, en passant par le calcul des appariements. Nous avons choisi de baser notre étude sur Drosophila melanogaster dont le génome est très annoté.

Nos analyses ont permis d'identifier de nombreux sites potentiels de triplex localisés à l'intérieur des gènes, principalement dans les zones introniques. Nous avons mis en évidence que les gènes concernés sont principalement liés à la biologie du développement, et à la biochimie des acides nucléiques et qu'ils possèdent de nombreuses interactions génétiques entre eux. Nos analyses suggèrent fortement que certains fragments d'ARN, codants ou non codants, pourraient avoir une action importante sur de nombreux loci des chromosomes pour effectuer des contrôles génétiques ou épigénétique à grande échelle. Nos données ouvrent la voie à une possible nouvelle voie de régulation génétique par l'entremise de fragments d'ARN. Notre étude a été publiée dans le journal « G3 : Genes | Genomes | Genetics » [Pasquier et collab., 2017a].

1.7 Synthèse des activités

Une vue synthétique des mes activités de recherche est présentée dans le tableau 1.1.

Thème de recherche Période Contrat [1] – Porteur du contrat	Partenaires industriels *Partenaires académiques*	Productions
Modélisation des documents structurés 1991-1994 Thèse CIFRE – ANRT/CDC	Caisse des Dépôts (CDC) *I3S*	1 conf. internat. 1 logiciel
Prédiction de la stucture des protéines 1997-199 GENEQUIZ [2] – C. Sanders	IBM Research Center *EMBL, Max-Delbrück Centre, CNB-CSIC, Univ. of Athens, Whitehead Institute, IRBM, Univ. of Oslo*	7 journaux internat. 6 conf. nationales 9 logiciels
Programmation générative 1999-2002 Contrat Schulmberger – I. Attali	Schulmberger *INRIA méditerranée*	4 conf. nationales 2 logiciels
Intégration/analyse de données transcriptomiques 2002-2010 Prog. Inter-EPST [3] – C. Pasquier CNRS Bio-STIC – F. Chevenet Post-doc CNRS – C. Pasquier	*iBV, I3S, IRD*	6 journaux internat. 6 conf. internat. 1 conf. nationale 7 logiciels
Fouille de graphes attribués 2013-2014 ANR FOSTER [4] – N. Selmaoui-Folcher	BlueCham *LIRIS UMR5205, Icube UMR7005, LISTIC EA3703*	2 journaux internat. 3 conf. internat. 4 conf. nationales 2 logiciels
Étude fonctionnelle des ARN 2014-present ANR METH [5] – A. Robichon	Biomanda *Institut Sophia Agrobiotech, I3S*	3 journaux internat. 2 logiciels
Sélection de caractéristiques 2017-present fondation ARC – O. Soriani UCA – O. Soriani & C. Pasquier INCa – M. Gabut	*iBV, CRCM, I3S, CRCL, Université de Tours*	

Tableau 1.1 – Synthèse des activités

Dans la suite de ce manuscrit je présenterai plus en détails mes recherches portant sur l'analyse des données et la bioinformatique. Je traiterai successivement les problématiques de l'intégration et de l'analyse des données biologiques, la fouille des graphes attribués et les

1. Contrat finançant les recherches
2. EEC-TMR 'GENEQUIZ', grant ERBFMRXCT960019, Integrated Software system for molecular Biologists
3. programme Inter-EPST pour la bioinformatique regroupant le CNRS, l'INSERM, l'INRA, l'INRIA et le ministère de la recherche
4. ANR-10-COSI-012 FOSTER, Spatio-temporal data mining : application to the understanding and monitoring of soil erosion, 2011-2014
5. ANR-12-BSV6-0006, Methylclonome, Analyse de l'héritabilité des traces épigénétiques dans la reproduction clonale, 2013-2015

recherches portant sur l'étude fonctionnelle des ARN.

Chapitre 2

Intégration et analyses de données biologiques

Sommaire

2.1 Introduction

Au cours des dernières décennies, les génomes de nombreux organismes ont été séquencés. La connaissance de la séquence des génomes est un pas important vers la compréhension de la composition moléculaire des organismes et l'assignation de fonctions aux gènes. Toutefois, cela ne fournit que des vues statiques. De multiples questions portant sur les aspects dynamiques demeurent. Dans quels processus cellulaires les gènes sont-ils impliqués ? Comment sont-ils régulés ? Dans quels types de cellules et selon quelles conditions sont-ils activés ?

Les mesures du transcriptome (l'ensemble des ARNs issus de la transcription du génome) à un moment donné et dans des conditions données fournissent des données dynamiques qui peuvent être utilisées pour répondre à ces questions. La quantité d'ARN mesurée est en effet directement liée à l'activité des gènes. En comparant l'expression d'un gène dans une condition donnée par rapport à son expression dans une condition normale (témoin), il est possible d'identifier les gènes qui sont sur-exprimés (plus actifs que dans la condition témoin) et ceux qui sont sous-exprimés (moins actifs que dans la condition témoin). A partir de ces mesures effectuées, à l'échelle d'un génome, dans différentes conditions expérimentales, les chercheurs peuvent déduire les fonctions des gènes et les voies impliquées dans les processus biologiques.

Quelque soit la technologie utilisée, le pipeline d'analyse des données peut être décomposé en trois étapes :

1. identification et mesure des ARN,
2. nettoyage des données, normalisation des mesures et traitements statistiques permettant de mettre en évidence les variations d'expression significatives,
3. analyse des données dont le but est de transformer les données brutes en savoir.

Les deux premières étapes regroupent des traitements essentiellement numériques sur lesquels, historiquement, la majorité des recherches se sont concentrées. La troisième étape est celle vers laquelle j'ai orienté mes recherches. Elle consiste à combiner les données numériques obtenues dans l'étape précédente avec des connaissances biologiques disponibles pour générer de nouvelles connaissances. Ce travail peut être effectué de manière manuelle par des experts. Les données étaient d'ailleurs interprétées de cette manière à la fin des années 90, lorsque les premiers résultats produits par des expériences de puces à ADN ont été analysés. Cependant, il a vite été considéré évident que l'adoption des nouvelles technologies de séquençage rendait utopique le traitement manuel des données générées. Il est devenu nécessaire : i) d'automatiser le processus permettant la recherche et l'intégration des nombreuses sources de données biologiques disponibles, ii) de représenter cette connaissance dans un système formel de manière à ce qu'elle soit manipulable par des systèmes informatiques et iii) de combiner cette connaissance avec les données pour élaborer de nouvelles connaissances.

Intégration sémantique des données biologiques, **gestion des connaissances** et **analyse des données transcriptomiques** sont trois axes de recherche relevant des sciences et technologies de l'information. Mes activités de recherches, s'inscrivant dans la thématique de l'intégration et de l'analyse des données biologiques, s'articulent autour de ces trois axes en les abordant dans le contexte des sciences du vivant.

2.2 Intégration sémantique des données biologiques

La biologie est maintenant une science dans laquelle l'information joue un rôle de plus en plus important. Les recherches en génomique, transcriptomique ou protéomique dépendent fortement de la disponibilité et de l'utilisation efficace de l'information. Lorsque les données étaient structurées et organisées en une collection d'enregistrements dans des bases de données dédiées et autosuffisantes, les informations pouvaient être extraites à l'aide de langages de requêtes spécialisés ; par exemple Structured Query Language (SQL) pour les bases de

données relationnelles ou Object Query Language (OQL) pour les bases de données à objets. En biologie, il est difficile d'exploiter les différents types d'information disponibles car les données sont disséminées sur le World Wide Web (WWW), éclatées en un grand nombre de ressources indépendantes, hétérogènes et hautement spécialisées.

Le WWW est un système de documents interconnectés distribués sur Internet. Il permet d'accéder à un grand nombre de ressources, principalement conçues pour l'usage et la compréhension humaine. En fait, les liens hypertextes peuvent être utilisés pour lier tout à n'importe quoi. En cliquant sur l'hyperlien d'une page Web, on obtient une autre ressource liée à l'élément cliqué (texte, image, son, clip vidéo, etc.). La relation entre la source et la cible d'un lien peut avoir de multiples significations : une explication, une traduction, une localisation, un ordre de vente ou d'achat, etc. Les utilisateurs humains sont capables de déduire le rôle des liens et d'utiliser le Web pour effectuer des tâches complexes. Cependant, un ordinateur ne peut pas accomplir les mêmes tâches sans supervision humaine parce que les pages Web sont conçues pour être lues par des personnes et non par des machines.

Le traitement automatique des données nécessite de passer d'un Web de documents, compréhensible uniquement par les humains, à un Web de données dans lequel l'information est exprimée non seulement en langage naturel, mais aussi dans un format lisible et utilisable par des agents logiciels, leur permettant ainsi de trouver, de partager et d'intégrer plus facilement l'information [Berners-Lee et Hendler, 2001]. Parallèlement au Web des données, des efforts internationaux considérables sont en cours pour développer l'interopérabilité programmatique sur le Web dans le but de permettre la mise en place d'un Web des programmes [Bratt, 2005]. Ici, les descriptions sémantiques s'appliquent aux processus, par exemple représentés sous la forme de Services Web [Kiryakov, 2006]. L'extension de la partie statique et dynamique du Web actuel s'appelle le Web sémantique.

Dans le cadre du Web Sémantique, de nouveaux langages ont été développés pour permettre aux contenus d'être compréhensibles par des machines [Berners-Lee et Hendler, 2001]. Resource Description Framework (RDF) permet d'exprimer des assertions sur des ressources. Une déclaration de base en RDF, également appelée triplet ou assertion, consiste en un sujet (une ressource identifiée par un Uniform Resource Identifier (URI)), un prédicat (une propriété de la ressource) et un objet (la valeur de cette propriété qui peut être un littéral ou une autre ressource). Une collection de triplets peut être représentée comme un graphe orienté (appelé graphe RDF) dans lequel, chaque nœud représente un sujet ou un objet, et chaque arc, représente un prédicat. Une assertion RDF peut porter sur d'autres assertions ; par exemple pour noter la date à laquelle l'assertion a été effectuée, l'auteur de cette assertion ou d'autres informations contextuelles. Une assertion qui porte sur une autre assertion est appelée réification. RDF Schema (RDFS) et Ontology Web Language (OWL) sont utilisés pour représenter explicitement la signification des ressources utilisées sur le Web et leurs relations. Ces spécifications, appelées ontologies, décrivent la sémantique des classes et propriétés utilisées dans les documents Web. SPARQL Protocol and RDF Query Language (SPARQL) est un langage de requête et un protocole qui permet de rechercher, d'ajouter, de modifier ou de supprimer des données RDF disponibles à travers Internet. SPARQL est capable de rechercher des motifs de graphe (appelés « graph pattern ») dans un graphe RDF. Les motifs de graphes sont formés de motifs de triplets, qui ressemblent à des triplets RDF avec la possibilité d'utiliser des variables en lieu et place des sujets, prédicats ou objets. Dans la suite de ce document, l'ensemble des langages utilisés par le Web Sémantique sera désigné par le terme générique Semantic Web Language (SWL).

Du point de vue de l'intégration des données, l'application de ces technologies se heurte à des difficultés qui sont amplifiées en raison de certaines spécificités des connaissances biologiques :

— les données biologiques représentent un volume énorme,

— l'hétérogénéité des informations biologiques rend difficile l'intégration des données,

— l'isolement des ontologies biologiques complique l'intégration des données,

— une grande proportion des connaissances en biologie est dépendante du contexte,

— la prise en compte de la provenance des connaissances biologiques est importante.

2.2.1 Collecte des données

Nous avons collecté diverses sources de données concernant les gènes et produits des gènes. Ce choix est arbitraire et est seulement destiné à illustrer la variété de données disponibles et la manière dont ces données sont traitées. Les données utilisées sont, soit directement exprimées en SWL, soit représentées sous un format tabulé ou stockées dans les tables d'une base de données relationnelle.

Les informations exprimées en SWL concernent les données protéiques de la base de données UniProt [UniProt, 2017], les interactions protéine-protéine de IntAct [Kerrien et collab., 2007] et la structure de Gene Ontology (GO) [GeneOntology, 2008]. Ces données sont décrites selon deux ontologies différentes. Les données d'UniProt et de IntAct sont décrites dans une ontologie nommée « core.owl » qui est disponible sur le site d'UniProt. GO est un cas spécial dans le sens où il ne représente pas la définition d'instances d'une ontologie existante, mais est une ontologie en elle-même dans laquelle les termes sont représentés par des classes.

Les données en format tabulé concernent les interactions protéine-protéine connues et prédites qui sont répertoriées dans la base de données STRING [Szklarczyk et collab., 2015], les interactions moléculaires et les réseaux de régulation stockés dans KEGG [Kanehisa et Goto, 2000], les annotations fonctionnelles des gènes de GeneRIFs [Mitchell et collab., 2003], les annotations de Gene Ontology Annotation (GOA) [Camon et collab., 2004], les données concernant la littérature et divers fichiers de mapping figurant dans la base de données Entrez du NCBI [Maglott et collab., 2007].

Les données stockées dans des bases de données relationnelles sont extraites à l'aide de requêtes SQL. Ce type d'information concerne les données d'Ensembl [Birney et collab., 2004] qui sont récupérées sur un serveur MySQL situé à l'adresse « ensembldb.ensembl.org ».

2.2.2 Conversion des données

Les données exprimées en SWL, sont directement importées dans la base de connaissances. Toutes les données qui ne sont pas exprimées en SWL doivent être converties. Les données tabulées sont converties en RDF à l'aide d'une procédure simple qui est similaire à celle utilisée dans YeastHub [Cheung et collab., 2005]. Chaque colonne devant être convertie en RDF est associée avec un espace de nom (namespace) qui est utilisé pour construire les URI identifiant les valeurs des colonnes (voir plus loin la section « principe d'encodage des URI »). Les relations entre le contenu de deux colonnes sont exprimées en RDF par un triplet ayant le contenu de la première colonne comme sujet, le contenu de la seconde colonne comme objet et la propriété spécifiée. Les résultats obtenus par les requêtes SQL, qui sont composés d'un ensemble d'enregistrements, ont été traités de la même manière que les données en format tableau.

2.2.3 Construction de l'ontologie unifiant les données

Le vocabulaire utilisé dans les descriptions RDF générées a été défini dans une nouvelle ontologie appelée Biowl. Les classes (par exemple « Gene », « Transcript », « Translation ») et les propriétés (par exemple « interacts_with », « has_score », « annotated_with ») sont définies dans cette ontologie en utilisant l'espace de nom « http ://www.unice.fr/bioinfo/biowl# ».

2.2.4 Principe d'encodage des URI

Dans les spécifications RDF générées à partir des fichiers tabulés ou des résultats retournés par des requêtes SQL, les ressources sont identifiées par des URI. Par exemple,

le peptide « ENSP00000046967 » de la base de données Ensembl est identifié par l'URI « http ://www.ensembl.org#ENSP00000046967 » alors que le gene « 672 » de NCBI Entrez est identifié par l'URI « http ://www.ncbi.nlm.nih.gov/EntrezGene#672 ».

2.2.5 Unification des ressources

Plusieurs listes de correspondance entre les identifiants utilisés dans différentes bases de données sont accessibles sur le Web. Nous avons utilisé les informations existantes sur les sites Ensembl, KEGG et le NCBI pour générer des descriptions OWL spécifiant les relations entre différentes ressources. Lorsque plusieurs ressources identifient le même objet, nous avons spécifié les équivalences avec la propriété OWL « sameAs », dans le cas contraire, nous avons utilisé la propriété appropriée définie dans l'ontologie Biowl.

Dans les cas où l'équivalence entre des ressources n'est pas spécifiée dans un fichier de mapping existant, l'identification des variations de nommage d'une même ressource a été effectuée manuellement. En examinant les différentes URI utilisées pour identifier une même ressource, il a été possible, par exemple, d'identifier le fait que le processus biologique « cell proliferation » est identifié par l'URI « http ://purl.org/obo/owl/GO#GO_0008283 » dans GO et par l'URI « urn :lsid :uniprot.org :go :0008283 » dans UniProt. A partir de cette connaissance, une règle a été établie pour indiquer qu'une ressource identifiée avec l'URI « http ://purl.org/obo/owl/GO#GO_$id » dans GO est équivalente à la ressource identifiée avec l'URI « urn :lsid :uniprot.org :go :$id » dans Uniprot ($id est une variable qui correspond à la même sous-chaîne). Les déclarations GO et Uniprot sont ensuite traitées par un programme qui utilise les règles précédemment définies pour générer un fichier de déclarations OWL exprimant les équivalences entre les ressources.

2.2.6 Fusion des ontologies

Comme spécifié précédemment, en plus de Biowl, nous avons utilisé deux autres ontologies définies par UniProt (core.owl) et GO (go_daily-termdb.owl). Ces ontologies modélisent différents sous-ensembles de la connaissance biologique mais possèdent néanmoins des parties qui se chevauchent. Pour pouvoir être utilisables, ces trois différentes ontologies doivent être unifiées. Il existe de nombreux outils permettant de fusionner ou d'établir des correspondances entre des ontologies [Do et Rahm, 2007; Kalfoglou et Schorlemmer, 2003] mais ils sont assez difficiles à utiliser et nécessitent une édition manuelle du résultat produit si l'on veut obtenir un résultat fiable (voir sur ce sujet l'évaluation de ces outils dans le cadre de la bio-informatique effectué par Lambrix et Edberg [2003]). Avec l'aide de l'outil de fusion d'ontologie PROMPT [Noy et Musen, 2003] et de l'éditeur d'ontologie Protégé [Rubin et collab., 2007], nous avons créé une ontologie unifiée décrivant les équivalences entre les classes et les propriétés définies dans les trois différentes ontologies. Par exemple, le concept de protéine est défini par la classe « urn :lsid :uniprot.org :ontology :Protein » dans Uniprot et par la classe « http ://www.unice.fr/bioinfo/owl/biowl#Translation » dans Biowl. L'unification de ces deux classes est déclarée dans une ontologie séparée où est définie une nouvelle classe dédiée à la représentation du concept unifié de protéine à laquelle est assigné l'URI « http ://www.unice.fr/bioinfo/owl/unification#Protein ». Chaque représentation du concept de protéine dans les autres ontologies est déclarée comme étant une sous-classe du concept unifié.

Le même principe est appliqué à l'unification des propriétés en spécifiant que plusieurs propriétés équivalentes sont des sous-propriétés de la propriété unifiée. Par exemple, le concept de nom, défini par la propriété « urn :lsid :uniprot.org :ontology :name » dans UniProt et par la propriété « http ://www.unice.fr/bioinfo/owl/biowl#denomination » dans Biowl est unifié avec la propriété « http ://www.unice.fr/bioinfo/owl/unification#name ».

L'ontologie unifiée permet à un système capable d'effectuer des inférences basées sur la hiérarchie des classes ou des propriétés d'effectuer des requêtes de manière unifiée sur plusieurs

spécifications définies de manière indépendante par plusieurs ontologies différentes.

2.3 Gestion des connaissances

Nous avons choisi d'utiliser une architecture d'entrepôt de données pour stocker les données contenues dans les différents documents que nous allons collecter. De manière à pouvoir exploiter complètement l'information exprimée dans l'un des langages du Web Sémantique, il est nécessaire d'utiliser un système à base de connaissances (Knowledge-Base System (KBS)) capable de stocker et d'interroger un ensemble important de spécifications RDF/OWL. Le KBS doit inclure des capacités de raisonnement basées sur l'inférence de type, la transitivité et au moins deux propriétés OWL qui sont « sameAs » et « inverseOf ». De plus, il doit être capable de stocker la provenance de l'information et les réifications.

En 2003, lors de l'initiation de ce projet, aucun KBS existant ne possédait toutes les caractéristiques requises. La quantité maximale d'information pouvant être stockée dans les systèmes les plus performants et leur capacité à traiter des requêtes complexes étaient bien en dessous de nos besoins (voir à ce sujet les tests effectués sur plusieurs bases de connaissances RDF par Guo et collab. [2004]). Pour cette raison, nous avons développé, en Java, un KBS appelé AllOnto, qui a été spécialement conçu pour répondre aux besoins identifiés.

En plus des fonctionnalités de base de stockage et d'interrogation de données RDF, AllOnto possède deux caractéristiques innovantes qui le différencie des autres KBS.

2.3.1 Prise en compte du contexte des assertions

Dans AllOnto, chacun des triplets est associé à une ressource représentant le contexte dans lequel la relation décrite est vraie. Les chercheurs en Intelligence Artificielle utilisent la notation $ist(C, a)$ introduite par Guha [1992] pour exprimer le fait que l'assertion a est vraie dans le contexte C. *ist* est l'abréviation de l'expression « is true in ». De nombreux travaux, effectués par Guha et d'autres, ont été réalisés dans le but d'élaborer une théorie définissant la manière d'effectuer des raisonnements utilisant le contexte. Dans AllOnto, aucun traitement ou raisonnement n'est effectué sur le contexte. Celui-ci est uniquement utilisé pour associer un ensemble de données à certains triplets. C'est un moyen commode de représenter les réifications exprimées en RDF. Dans le codage en RDF d'UniProt, la réification est intensivement utilisée pour représenter les codes évidence associés aux annotations GO (une indication du type d'expérience qui est à la source de l'annotation), des informations supplémentaires portant sur les citations (le type de tissu par exemple) ou la position des « features ». Ainsi, le fait que « 14-3-3 protein beta/alpha », identifiée par « P31946 » dans UniProt est une protéine localisée dans le cytoplasme (terme GO correspondant à « 5737 ») et que cette affirmation est étayée par une expérience biologique dédiée (code évidence « IDA » signifiant « Inferred from Direct Assay ») est exprimé en RDF dans UniProt par :

```
<rdf:Description rdf:about="urn:lsid:uniprot.org:uniprot:P31946">
    <classifiedWith rdf:resource="urn:lsid:uniprot.org:go:5737" rdf:ID="_55"/>
</rdf:Description>
<rdf:Description rdf:about="#_55">
    <evidence rdf:resource="urn:lsid:uniprot.org:ontology:IDA"/>
</rdf:Description>
```

Dans AllOnto, ces informations sont encodées en utilisant le contexte. La description précédente est codée en utilisant une ressource c pour le contexte et une ressource T, « True », pour identifier un contexte global qui est toujours vrai, par :

```
ist(c, (urn:lsid:uniprot.org:uniprot:P31946,
    classifiedWith,
    urn:lsid:uniprot.org:go:5737))
ist(T, (c, evidence, urn:lsid:uniprot.org:ontology:IDA))
```

Concrètement, cette information est stockée dans AllOnto sous forme de deux quadruplets :

```
(urn:lsid:uniprot.org:uniprot:P31946, classifiedWith,
 urn:lsid:uniprot.org:go:5737, c)
(c, evidence, urn:lsid:uniprot.org:ontology:IDA, T)
```

Cela permet d'utiliser le contexte dans les requêtes. Par exemple, retrouver la liste de toutes les protéines de la base ainsi que leurs annotations GO effectuées avec le code évidence « Traceable Author Statement (TAS) » se traduit par :

```
(?P, classifiedWith, ?X, ?C)
(?C, evidence, urn:lsid:uniprot.org:ontology:TAS)
```

2.3.2 Modélisation de la provenance des assertions

En plus des triplets et des contextes, nous avons choisi de représenter de manière séparée la provenance des assertions. La provenance est une ressource associée à chaque quadruplet qui identifie l'origine. Chaque morceau d'information est donc finalement défini par un quintuplet de la forme (sujet, prédicat, objet, contexte, origine). L'extrait RDF défini précédemment est codé en utilisant une ressource « up », qui identifie la base de données UniProt, par :

```
(urn:lsid:uniprot.org:uniprot:P31946, classifiedWith,
urn:lsid:uniprot.org:go:5737, c, up)
(c, evidence, urn:lsid:uniprot.org:ontology:IDA, T, up)
```

La ressource « up » peut être définie par ailleurs pour préciser les caractéristiques de la source de données (nom, url, indice de fiabilité par exemple). Cela permet d'utiliser le contexte et l'origine dans les requêtes. Par exemple, retrouver les annotations GO les plus fiables (avec un code évidence « TAS ») concernant la protéine « P31946 » stockées dans la base UniProt s'exprime par :

```
(urn:lsid:uniprot.org:uniprot:P31946, classifiedWith, ?X, ?C, ?O)
(?C, evidence, urn:lsid:uniprot.org:ontology:TAS)
(?O, name, 'UNIPROT')
```

2.3.3 Interrogation de la base de connaissances

Une fois que les données sont nettoyées, fusionnées et encodées dans l'un des langages du Web Sémantique, il ne reste plus qu'à les importer dans notre base de connaissances. Des requêtes sont effectuées sur le KBS pour retrouver les informations d'intérêt. A ce point, nous sommes confrontés à un nouveau problème qui est celui de l'abondance des résultats, parfois similaires. En effet, comme les sources de données utilisées sont redondantes, la même pièce d'information peut être décrite de nombreuses fois et souvent sous des formes différentes. D'où la nécessité de procéder à une fusion des résultats.

Les triplets stockés dans la base de connaissances, les données encodées en utilisant le principe de réification et la provenance des assertions peuvent être retrouvées avec des requêtes SPARQL. Un exemple de requête permettant de retrouver toutes les annotations de la protéine identifiée par « P38398 » avec leurs indices de fiabilité et leurs provenances est donné dans la figure 2.1.

Le moteur d'interrogation de la base de connaissances comprend également un module de désambiguïsation et un module de fusion des résultats qui est notamment utilisé dans l'interface de recherche de l'application « Thea-online ».

2.3.4 Caractéristiques et performances de la base de connaissances

AllOnto est un système de gestion des connaissances capable de stocker de grandes quantités de données OWL contenant des instructions réifiées Le système comprend des capacités de raisonnement qui vont au-delà de ce qui peut être exécuté dans les descriptions existantes. Ces caractéristiques seront nécessaires pour être capable de manipuler les futures version d'OWL qui utiliseront des définitions plus complexes (comme l'illustrent J Wroe et collab.

```
PREFIX     up: <urn:lsid:uniprot.org:uniprot:>
PREFIX   unif: <http://www.unice.fr/bioinfo/owl/unification#>
PREFIX    rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT
    ?annotation ?reliability ?source
WHERE {
    GRAPH ?source {
        ?r rdf:subject up:P38398 .
        ?r rdf:predicate unif:annotated_by .
        ?r rdf:object ?annot .
        ?r unif:reliability ?reliability
    }
}
```

Figure 2.1 – Exemple de requête SPARQL
Cette requête affiche les données décrivant les annotations, la fiabilité des annotations et la provenance de l'information pour la protéine « P38398 ». La base de connaissances est recherchée pour trouver un triplet « r » ayant la ressource « urn :lsid :uniprot.org :UniProt :P38398 » comme sujet, la ressource « http ://www.unice.fr/bioinfo/owl/unification annotated_by# » comme propriété et la variable « annot » comme objet. La valeur de la propriété « http ://www.unice.fr/bioinfo/owl/unification#reliability » des triplets trouvés est stockée dans la variable « reliability ». La provenance de l'information est obtenue en récupérant le « named graph » qui contient ces spécifications. Le KBS effectue des inférences basées sur la propriété « owl :sameAs » pour unifier la protéine UniProt « P38398 » avec des ressources équivalentes définies dans d'autres bases de données. Il utilise également l'ontologie unifiée pour rechercher des données exprimées en utilisant des sous-propriétés de « annotated_by » et « reliability ».

[2003]). AllOnto est développé entièrement dans le langage Java. Il est construit au dessus d'une base de données relationnelle (MySQL et HSQLDB ont notamment été utilisées) en utilisant Hibernate comme moteur de persistance. Les fonctionnalités de raisonnement sont fournies par Pellet [Sirin et collab., 2007] et la boîte à outils Jena [Wilkinson et collab., 2003] est utilisée pour importer les spécifications OWL 2.2.

Le Lehigh University BenchMark (LUBM), mis au point par Guo et collab. [2004] a été utilisé pour évaluer les performances d'AllOnto et les comparer avec celles des KBS existants. Le benchmark est composé d'une ontologie et d'un ensemble d'instances générées automatiquement dont la taille peut être paramétrée. L'ontologie a pour thème la modélisation d'universités et définit des classes telles que « Professor », « Student », « Department », « Publication » et des propriétés comme « degreeFrom », « memberOf », « teacherOf », etc. Les capacités d'interrogation des bases de connaissances sont évaluées à l'aide de 14 requêtes qui permettent de mesurer les vitesses de réponse, mais aussi les capacités d'inférence des KBS. Les requêtes de référence ainsi que leurs caractéristiques sont listées dans l'article de Guo et collab. [2005].

Prenons pas exemple la requête numéro 13 qui est définie avec les deux triplets suivants :

```
(?X, type, Person)
(http://www.University0.edu, hasAlumnus, ?X)
```

Le moteur de recherche de la base de connaissances doit rechercher parmi les données, les triplets qui satisfont la requête. « ?X » permet de déclarer une variable X pouvant prendre n'importe quelle valeur et qui contiendra le résultat de la requête. La première expression recherche toutes les ressources X tel que X soit du type « Person ». La seconde expression recherche toutes les ressources X pour lesquelles X possède une propriété « hasAlumnus » qui a pour valeur « http ://www.University0.edu ». En clair, on recherche dans la base de connaissances, toutes les personnes diplômées d'une université donnée.

Une recherche naïve des triplets correspondants dans la base ne retourne aucun résultat. Pourquoi ? Parce que les informations importées dans la base ne sont pas définies avec des triplets de cette forme. Il est nécessaire d'inférer de la connaissance à partir des données de

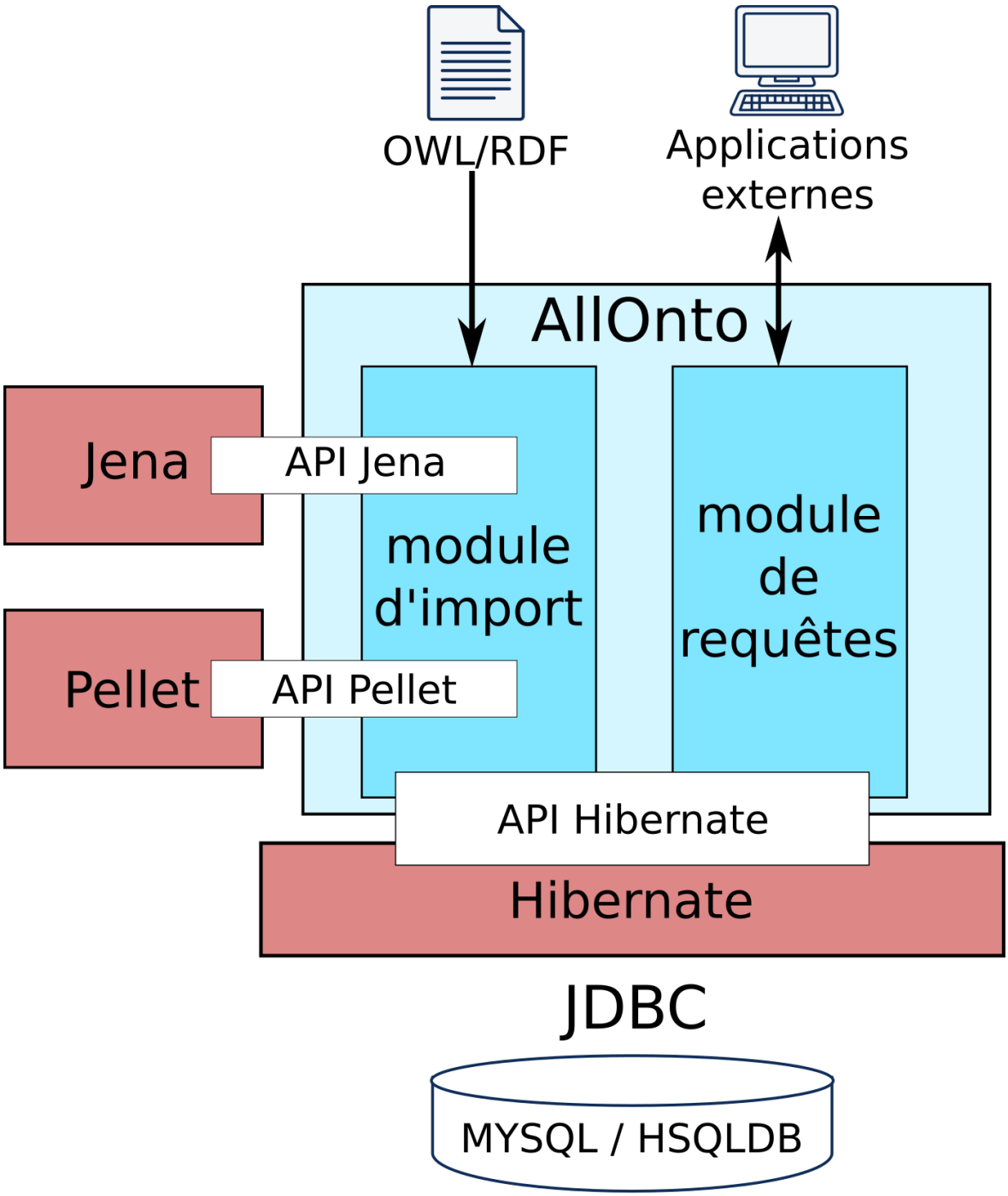

Figure 2.2 – Vue générale du KBS AllOnto et de ses modules
Allonto contient un module d'import des données qui fait appel à Jena pour l'extractions des graphes RDF et à Pellet pour calculer les inférences. Il contient également un module accessible par des applications externes qui traduit les requêtes SPARQL en appels à l'API d'Hibernate et retourne le résultat au format XML.

base. Plusieurs mécanismes d'inférence sont mis en jeu avec cette simple requête :

1. **raisonnement utilisant la hiérarchie des classes**. Les individus ne sont pas définis en tant qu'instances de la classe « Person » mais comme des instances de « Professor », « GraduateStudent », etc. Le moteur d'inférence doit utiliser l'ontologie pour déduire que les instances d'une sous-classe de « Person » sont également des instances de la classe « Person ».
2. **raisonnement sur les propriétés inverses**. Aucune université n'est définie avec une propriété « hasAlumnus » qui pointe sur les étudiants diplômés. Le moteur d'inférence doit utiliser l'ontologie pour découvrir que « hasAlumnus » est défini comme propriété inverse de « degreeFrom » et faire une recherche des personnes possédant une propriété « degreeFrom » ayant comme valeur l'université donnée.
3. **raisonnement utilisant la hiérarchie des propriétés**. Malheureusement, aucune personne ne possède de propriété « degreeFrom ». Le moteur d'inférence doit utiliser l'ontologie pour chercher les individus possédant des sous-propriétés de « degreeFrom » comme « DoctoralDegreeFrom ».

Sont également testées, les capacités des KBS à utiliser les propriétés transitives, à construire une hiérarchie des classes lorsque celle-ci n'est pas définie explicitement et à attribuer automatiquement une classe à une instance en fonction de ses caractéristiques (requête numéro 12, qui n'est résolue par aucun KBS).

dataset	nom #instances taille (MB)	LUBM5.0 645 649 51		LUBM50.0 6 888 642 540		LUBM500.0 69 099 760 5454
nom du KBS		DLDB	AllOnto	DLDB	Allonto	Allonto
temps d'import (h :m :s)		00 :51 :57	00 :02 :16	12 :37 :57	00 :24 :19	07 :57 :54
taille sur le disque (MB)		91	97	958	1 100	10 343
temps de réponse aux requêtes (mm :ss,00)	Requête 1	00 :00,22	00 :00,88	00 :02,21	00 :01,76	00 :01,36
	Requête 2	00 :02,32	00 :03,73	Echec	00 :09,53	3h 45mn
	Requête 3	00 :02,54	00 :00,12	00 :36,16	00 :00,17	00 :01,10
	Requête 4	00 :02,49	00 :00,13	00 :10,11	00 :00,14	00 :02,19
	Requête 5	00 :04,64	00 :00,07	02 :15,05	00 :00,07	00 :01,89
	Requête 6	00 :04,36	00 :00,94	02 :31,90	00 :05,87	04 :43,68
	Requête 7	00 :02,63	00 :00,07	02 :01,67	00 :00,17	00 :01,76
	Requête 8	00 :03,00	00 :00,19	00 :39,84	00 :00,22	00 :16,21
	Requête 9	00 :07,75	00 :01,79	05 :23,57	00 :16,61	1h 09mn
	Requête 10	00 :01,05	00 :00,01	00 :05,83	00 :00,01	00 :00,24
	Requête 11	Incomplet	00 :00,01	Incomplet	00 :00,01	00 :00,07
	Requête 12	Incomplet	Incomplet	Incomplet	Incomplet	Incomplet
	Requête 13	Incomplet	00 :00,05	Incomplet	00 :00,08	00 :17,58
	Requête 14	00 :02,93	00 :00,33	01 :46,76	00 :03,59	04 :34,60

Tableau 2.1 – Comparaison entre AllOnto et DLDB-OWL
Les 14 requêtes ainsi que les jeux de données LUBM5.0 et LUBM50.0 sont définis dans l'article de Guo et collab. [2004]. Le jeu de données LUBM500.0, qui contient dix fois plus d'instances que LUBM50.0 a été crée avec le générateur fourni par Guo et collab. [2004]. Il n'a pas été testé par Guo et collab. [2004] à cause des médiocres performances des KBS. Les performances de DLDB-OWL (identifiées par DLDB dans le tableau) sont telles que reportées par les auteurs. Les performances d'AllOnto ont été mesurées sur une machine de bureau avec un processeur AMD à 2.8GHz et possédant 1.5GB de RAM.

Comme le montre le tableau 2.1, les performances d'AllOnto surclassent DBLB-OWL, le meilleur KBS évalué en 2004 [Guo et collab., 2004]. Le point faible est la place occupée sur

le disque. En effectuant des tests avec un jeu de données dix fois plus important que le plus grand testé par Guo et collab. (LUBM500.0), on constate un ralentissement de la vitesse d'import et on voit apparaître des problèmes de performances importants avec certaines requêtes (numéro 2 et 9). Il est à noter cependant que AllOnto est handicapé par le fait qu'il traite le contexte et la provenance des assertions ; informations qui ne sont pas prises en compte par les autres KBS. Le système AllOnto est à la base de tous les développements effectués concernant l'analyse des données du transcriptome.

2.4 Analyse des données transcriptomiques

Comme cela a été présenté au début de ce chapitre, les techniques de quantification à haut débit permettent de mesurer des milliers de variables et de comparer leurs valeurs dans des centaines de conditions. Une fois que les données sont quantifiées, normalisées et classifiées, la phase d'analyse est essentiellement une tâche manuelle et subjective basée sur l'inspection visuelle des résultats à la lumière de la grande quantité d'information disponible. L'interprétation des données constitue clairement le goulot d'étranglement de ces analyses et il existe un besoin évident d'outils capables d'analyser les données de manière automatique en utilisant les connaissances biologiques existantes.

Il existe plusieurs stratégies permettant d'élaborer de nouvelles connaissances à partir des données numériques et des connaissances. Trois axes de recherche principaux peuvent être distingués : les approches guidées par les connaissances, les approches guidées par les données numériques et les approches visant à utiliser parallèlement les deux aspects.

L'approche guidée par les données consiste à identifier des groupes de gènes dont l'expression varie de manière similaire pour ensuite intégrer les connaissances sur les gènes. **L'approche guidée par la connaissance** consiste à analyser les variations d'expressions de gènes qui possèdent des caractéristiques similaires. Enfin, **l'approche basée sur le co-partitionnement** permet d'analyser simultanément les données d'expression et les connaissances.

2.4.1 Approche guidée par les données

THEA (Tools for High-throughput Experiments Analysis) est un système intégré de traitement de l'information permettant une manipulation aisée des données. Il permet d'annoter automatiquement des groupes de gènes partageant le même profil d'expression avec des informations biologiques provenant d'une base de connaissances. Le logiciel permet de rechercher et de parcourir manuellement les annotations ou de générer automatiquement des généralisations significatives selon des critères statistiques. Fondamentalement, THEA prend comme entrée le résultat d'un clustering effectué sur les données du transcriptome et propose des annotations pertinentes.

L'interface utilisateur graphique de THEA permet aux utilisateurs d'explorer facilement les données biologiques. Il est possible de parcourir les ontologies, de rechercher un domaine de connaissance particulier et de visualiser les éléments associés avec des marqueurs de couleur. L'emploi successif de plusieurs termes révèle immédiatement quels groupes de gènes se rapportent simultanément à différents domaines de connaissance. Cela permet également de visualiser rapidement les zones enrichies par tel ou tel terme. Cette exploration manuelle est toutefois biaisée car elle est guidée par les connaissances de l'utilisateur et son champ d'intérêt. Pour pallier ce biais, THEA intègre plusieurs algorithmes de fouille de données permettant d'annoter automatiquement une classification complète.

Identification des annotations enrichies

Pour chaque groupement de gènes, THEA effectue une analyse statistique selon l'hypothèse nulle d'une distribution uniforme des annotations. Un groupe donné est considéré

comme enrichi de façon significative avec une annotation si le nombre de gènes possédant cette annotation dépasse le nombre que l'on devrait théoriquement obtenir par chance. La distribution hypergéométrique est celle qui doit être utilisée mais son calcul devient coûteux lorsque l'on doit traiter plusieurs milliers de groupes de gènes. La loi binomiale, qui est moins coûteuse en temps de calcul et qui constitue une bonne approximation de la loi hypergéométrique lorsque la population est importante, peut la plupart du temps être utilisée à sa place. THEA offre la possibilité de choisir entre ces deux lois statistiques. Pour un terme donné présent dans la population totale avec une fréquence p, les lois calculent la probabilité (appelée *valeur-p*) d'observer par hasard, avec l'hypothèse nulle, au moins k génes annotées avec ce terme dans un groupe de taille n. Les termes enrichis sont identifiés en sélectionnant ceux associés à une *valeur-p* sous un seuil donné. Le même principe est utilisé pour mettre en évidence les termes sous-représentés en calculant la probabilité d'observer un maximum de k gènes annotées avec un terme dans un groupe.

Le calcul d'enrichissement peut être relatif, soit à tous les gènes associés aux termes d'une ontologie, soit à la liste des gènes inclus dans l'expérimentation, soit à une liste spécifique définie par l'utilisateur. L'utilisation de toutes les annotations GO est une méthode couramment utilisée [Drăghici et collab., 2003] mais l'annotation peut être biaisée lorsque les gènes analysés ne représentent qu'un sous-ensemble du génome. Si l'on analyse, par exemple, les résultats d'une expérience ne mesurant qu'une catégorie de gènes donnés, on obtiendra pour tous les groupes de gènes, de manière logique, une surreprésentation artefactuelle de la catégorie étudiée. Dans ce cas, il vaut mieux effectuer le calcul avec l'hypothèse nulle d'une distribution uniforme des annotations restreinte aux gènes analysés. Une autre option consiste à utiliser comme référence l'ensemble des gènes inclus dans la classification ou toute autre liste définie par l'utilisateur.

Identification de co-localisations

Un second type de recherche possible avec THEA consiste à mettre en évidence les corrélations existantes entre les niveaux d'expression de certains gènes et leur localisation sur le génome. Une technique similaire à celle utilisée pour l'identification d'une surreprésentation des termes dans les clusters est utilisée. Sous l'hypothèse nulle d'une distribution uniforme des gènes sur chaque chromosome, on peut calculer une *valeur-p* qui représente la probabilité pour que des gènes co-exprimés soient également co-localisés. L'utilisateur doit d'abord spécifier la distance maximale d entre deux gènes pour les considérer comme co-localisés. La mesure de la distance que nous avons choisie d'utiliser est le nombre de régions intergéniques séparant deux gènes : si l'on attribue un indice pour chaque gène sur un chromosome (le gène le plus proche de l'extrémité 5' se voit attribuer l'indice 1, le deuxième le plus proche de l'extrémité, l'indice 2, etc), alors la distance entre deux gènes est la différence de leurs indices. Pour un gène donné et une distance donnée d, il existe, dans l'ensemble du génome, un maximum de $2d$ autres gènes à une distance inférieure ou égale à d (ces gènes voisins sont ceux qui sont considérés comme co-localisés). Pour chaque gène d'un groupe, nous déterminons le nombre c de gènes co-localisés présents dans le même groupe et nous utilisons la loi binomiale pour calculer la probabilité d'observer au moins c gènes dans un groupe de taille $(n-1)$ par hasard, en considérant une fréquence de $2d/G$ (G représentant le nombre de gènes dans le génome). Cette procédure ne permet de détecter que des groupes d'au moins trois gènes (le gène pris comme référence plus deux gènes présents dans son voisinage). Les gènes qui sont considérés comme co-localisés sont mis en évidence par un marqueur similaire à celui utilisé pour étiqueter les gènes sur la classification (Fig. 2.3).

Applications

THEA a été utilisé, à des fins de démonstration, pour réanalyser plusieurs résultats d'expériences transcriptomiques publiés. Les expérimentations montrent que l'utilisation de THEA

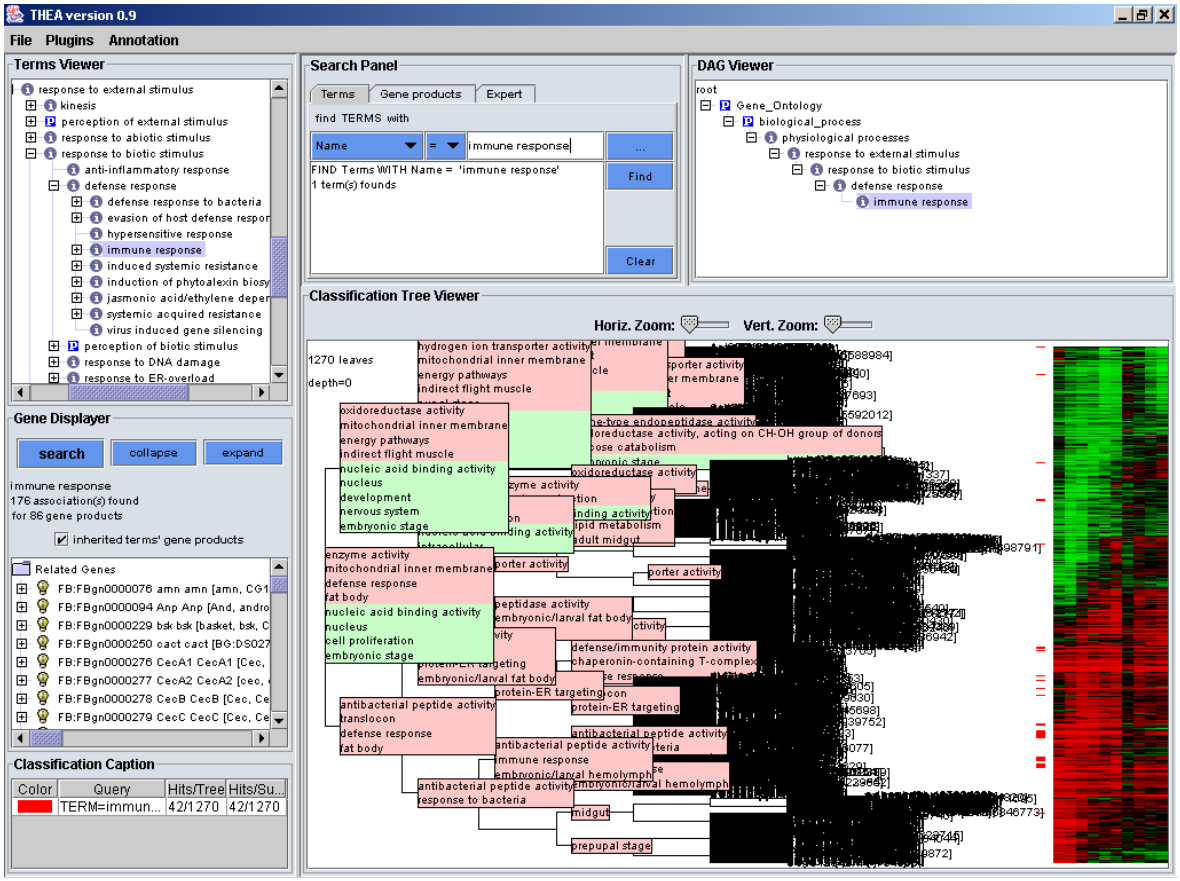

Figure 2.3 – Vue générale de l'interface utilisateur de THEA

Dans cette vue, THEA affiche une classification annotée de 1270 gènes (données extraites d'un article de De Gregorio et collab. [2001]. La configuration de THEA illustrée contient six modules activés. Le visualisateur de termes (en haut, à gauche), permet de naviguer graphiquement dans les ontologies et de sélectionner des termes. Le visualisateur DAG (en haut, à droite), affiche le chemin dans l'ontologie jusqu'au terme sélectionné. Le visualisateur de gènes (au milieu, à gauche) affiche la liste des gènes associés au terme sélectionné en incluant éventuellement ses sous-termes (ici, les 86 genes associés au terme « immune response » et à ses sous termes sont affichés). Le panneau de recherche (en haut, au centre) permet de rechercher des termes particuliers ou des chaînes de caractères parmi les gènes ou les termes présents dans AllOnto. Des requètes plus élaborées sont également possibles. Ici, le terme « immune response » a été sélectionné. Le visualisateur de classification (en bas, à droite) affiche le résultat d'une classification hiérarchique effectuée sur le profil d'expression des gènes. Il est possible de naviguer dans l'arbre en zoomant sur les nœuds Les niveaux d'expression sont affichés sur la droite sous une forme désormais classique. Les gènes associés au terme sélectionné et à ses fils sont identifiés par des barres de couleur rouge. Les noms automatiquement assignés à chaque cluster sont également affichés par ce module. Le paneau de légendes (en bas, à gauche) affiche la légende des codes de couleur utilisés ainsi que quelques informations statistiques.

permet, non seulement d'obtenir facilement et rapidement tous les résultats mis en évidence de manière manuelle mais que cela permet aussi d'identifier de nouvelles informations. Dans l'article présentant THEA [Pasquier et collab., 2004]), le lecteur intéressé pourra trouver l'analyse, avec le logiciel, des expériences effectuées par De Gregorio et collab. [2001] portant sur des mouches atteintes d'infections bactériennes ou fongiques.

THEA a également été utilisé par Rousset et collab. lors d'une études concernant la fermeture dorsale de l'embryon de drosophile [Rousset et collab., 2017] ainsi que par Girardot et collab. dans son étude sur les gènes liés au vieillissement chez la drosophile [Girardot et collab., 2006].

2.4.2 Approche guidée par la connaissance

L'approche guidée par la connaissance consiste tout d'abord à trouver des groupes de gènes co-annotés (Groupe Fonctionnellement Enrichi (GFE)). Dans un second temps, les données concernant les profils d'expression sont intégrées. La signification statistique de chaque GFE est ensuite validée en utilisant un test basé sur un score enrichi [Mootha et collab., 2003], un test issu d'un *z-score* [Kim et Volsky, 2005] ou un test basé sur une *pc-value* avec une distribution hypergéométrique [Breitling et collab., 2004].

Notre méthode, appelée AGGC (Analyse des Groupes de Gènes Co-exprimés), s'inspire de cette approche : les GFE sont d'abord formés à partir de GO, puis, une fonction qui synthétise l'information contenue dans les données d'expression est appliquée afin d'obtenir une liste ordonnée de gènes [Breitling et collab., 2004]. Dans cette liste, les gènes sont triés par variabilité d'expression décroissante. La significativité statistique des GFE obtenus est alors testée à l'aide d'une preuve d'hypothèse. Finalement, nous obtenons la liste des GFE co-exprimés et statistiquement significatifs. La méthode AGGC est une extension de la méthode IGA [Breitling et collab., 2004] permettant d'obtenir tous les sous-ensembles possibles de GFE de gènes co-exprimés, sans se limiter aux GFE constitués des gènes les plus exprimés.

L'algorithme CGGA

Afin d'incorporer les profils d'expression des gènes, nous avons utilisé une mesure de la variation d'expression, le *f-score*, qui est plus robuste que d'autres mesures telles que l'anova, le fold change ou le test de Student [Riva et collab., 2005]. Cette mesure permet d'établir une liste des gènes, *g-rank*, ordonnés selon ordre décroissant de leur variation d'expression. Nous avons utilisé le programme SAM [Tusher et collab., 2001] pour calculer le *f-score* associé à chaque gène.

AGGC est basé sur l'idée que tout changement affine (co-expression) d'un sous-ensemble de gènes appartenant à un GFE est physiologiquement important. Nous disons que deux gènes sont co-exprimés s'ils sont proches par rapport à la métrique de variabilité d'expression (*f-score*). L'algorithme AGGC permet de déterminer pour chaque GFE la *pc-value* qui estime sa cohérence (à partir du *g-rank*) et donc de détecter les groupes statistiquement significatifs.

L'algorithme AGGC commence par construire la liste *g-rank* à partir des niveaux d'expression et les GFE à partir des données GO. Pour chaque GFE constitué de n gènes, l'algorithme détermine les $n(n+1)/2$ sous-ensembles de gènes dont nous voulons tester la co-expression. Pour chacun de ces sous-ensembles nous calculons sa *p–value*.

Soit H_0 la probabilité que les x gènes d'un de ces sous-ensembles aient été associés par hasard. Cette probabilité correspond à la distribution hyper-géométrique suivante :

$$p(X = x|N, R_{g(x)}, n) = \frac{\binom{R_{g(x)}}{x}\binom{N-R_{g(x)}}{n-x}}{\binom{N}{n}} \qquad \text{où} \qquad p(X = 0|N, R_{g(x)}, n) = 0$$

avec :

— N : nombre total de gènes dans le jeu de données,

— n : nombre de gènes dans le GFE,

— x : position (n° d'ordre) du gène dans le GFE,

— $r_{g(x)}$: rang absolu du gène de position x dans g-rank,

— $R_{g(x)}$: nombre de rangs dans g-rank qui séparent le gène de position x de son prédécesseur dans le GFE.

L'intervalle entre rang, $R_{g(x)}$, est calculé a partir des rangs absolus $r_{g(x)}$ selon la formule :

$$R_{g(x)} = r_{g(x)} - r_{g(x-1)} + 1 \qquad o\grave{u} R_{g(0)} = R_{g(0)} = 1$$

La *pc–value* correspondant à cette preuve d'hypothèse est [Drăghici et collab., 2003] :

$$pc{-}value(x) = 1 - \sum_{k=1}^{x} p(X = k|N, R_{g(k)}, n)$$

Afin d'accepter ou rejeter l'hypothèse H_0, nous utilisons comme seuil de significativité :

$$p{-}value = Min\left\{\frac{1}{N}, \frac{1}{|\Omega|}\right\}$$

où $|\Omega|$ est la cardinalité de l'ensemble de toute les annotations fonctionnelles.

Ainsi, pour chaque GFE, si $pc{-}value(x) < p{-}value$ alors on rejette H_0 , i.e. le GFE est statistiquement significatif. Le pseudo-code de l'algorithme AGGC est présenté dans la figure 2.4. L'algorithme a été implémenté en langage Perl. Il reçoit en entrée les listes d'annotations de chaque gène et la liste ordonnée g-rank des N gènes. Il renvoie la liste des groupes de gènes co-exprimés significatifs.

Input: Liste des annotations de chaque gene G : $annotations(G)$
Input: Liste ordonnée des N gènes : g-rank
Result: Ensemble résultat des GFE de gènes co-exprimés

```
1  déterminer p-value
2  foreach annotation A de GO do
3  |  foreach gène G do
4  |  |  if A ∈ annotations(G) then
5  |  |  |  GFE_A ← GFE_A ∪ G
6  |  |  end
7  |  end
8  end
9  foreach GFE_A do
10 |  foreach sous-ensemble S de GFE_A do
11 |  |  calculer pc-value(S)
12 |  |  if pc-value(S) < p-value then
13 |  |  |  résultat(GFE_A) ← résultat(GFE_A) ∪ S
14 |  |  end
15 |  end
16 |  supprimer de résultat(GFE_A) les S non maximaux vis-à-vis de l'inclusion
17 end
18 résultat ← ⋃_{i=A} résultat(GFE_i)
```

Figure 2.4 – L'algorithme AGGC

L'algorithme commence par déterminer la valeur de la *p-value* (ligne 1), puis, il génère les GFE à partir des annotations GO (lignes 2 à 8). Il considère ensuite successivement chaque GFE (lignes 9 à 17). Pour chaque GFE, il détermine tous ses sous-ensembles non vides et

calcule la *pc-value* de chacun (lignes 10 à 15). Si la *pc-value* calculée est inférieure à la *p-value*, le sous-ensemble est inséré dans le résultat du GFE (lignes 12 à 14). Les sous-ensembles insérés dans le résultat du GFE qui ne sont pas maximaux vis-à-vis de l'inclusion sont ensuite supprimés (ligne 16).

Résultats

Afin d'évaluer notre méthode, nous avons comparé, sur le même jeu de données, les résultats obtenus manuellement par DeRisi et collab. [1997] avec ceux générés par IGA [Breitling et collab., 2004] et AGGC. Les résultats obtenus avec AGGC pour les gènes sur-exprimés sont présentés dans le tableau 2.2 et ceux obtenus pour les gènes sous-exprimés sont présentés dans tableau 2.3. Les tableaux présentent pour chaque GFE, le nombre de gènes dans le groupe concerné, le nombre de gènes différentiellement exprimés, la *pc-value* obtenue et la méthode d'identification. Dans cette dernière colonne, la lettre 'A' correspond à AGGC, la lettre 'D' à DeRisi et la lettre 'I' à IGA.

Groupe GO fonctionnellement riche	nb. gènes	nb. gènes sur-exp.	pc-value	meth. ident.
proton-transporting ATP synthase complex	2	2	4.38E-06	A
invasive growth (sensu Saccharomyces)	5	3	6.13E-06	A
signal transduction during filamentous growth	2	2	8.77E-06	A
respiratory chain complex II	4	4	3.75E-05	A D
succinate dehydrogenase activity	4	4	3.75E-05	A D
mitochondrial electron transport	4	4	3.75E-05	A D
aerobic respiration	36	10	3.30E-05	A
tricarboxylic acid cycle	14	5	6.54E-05	A D
gluconeogenesis	12	2	9.64E-05	A
response to oxidative stress	10	3	1.55E-06	A
filamentous growth	8	4	9.06E-05	A
vacuolar protein catabolism	4	2	2.63E-05	A I
respiratory chain complex IV	8	2	4.05E-04	A D
cytochrome-c oxidase activity	8	2	4.05E-04	A D
glycogen metabolism				D
glycogen synthase				D

Tableau 2.2 – Comparaison entre les GFE sur-exprimés obtenus avec trois méthodes
Pour les trois méthodes, les groupes sont considérés fonctionnellement enrichis lorsqu'ils sont associés avec une *p-value* inférieure à 6×10^{-3}

Dans le cas de gènes sur-exprimés (tableau 2.2), AGGC a permis de retrouver sept des neuf groupes de gènes obtenus manuellement par DeRisi et collab.. Les deux groupes annotés « glycogen metabolism » et « glycogen synthase » n'ont pas été identifiés par AGGC car ils s'expriment uniquement dans la phase initiale du processus. Toutefois AGGC a identifié huit autres groupes statistiquement significatifs et cohérents vis-à-vis du processus étudié. Un seul de ces huit autres groupes avait été identifié par IGA et aucun lors de l'étude de DeRisi et collab. [1997].

Pour le cas de gènes sous-exprimés, AGGC a retrouvé sept des huit groupes de gènes obtenus manuellement par DeRisi et collab. Comme pour les gènes sur-exprimés, un groupe, annoté « ribosome biogenesis », n'a pas été identifié par AGGC car s'exprimant seulement durant la phase finale du processus. AGGC a également identifié sept autres groupes statistiquement significatifs et cohérents vis-à-vis du processus étudié qui n'ont été identifiés ni lors de l'analyse de DeRisi et collab. ni par IGA.

Groupe GO fonctionnellement riche	nb. gènes	nb. gènes sous-exp.	pc-value	meth. ident.	
chromatin modification	6	5	2.35E-06	A	
mitochondrial inner memb. prot. inser. complex	3	2	3.60E-06	A	
regulation of nitrogen utilization	4	2	7.20E-06	A	
acid phosphatase activity	4	2	7.20E-06	A	
histone acetylation	4	4	7.95E-06	A	
nucleolus	52	10	3.41E-05	A	D
rRNA modification	10	3	2.75E-05	A	D
transcription initiation from RNA poly. II prom.	14	3	1.00E-05	A	
mitochondrial matrix	15	3	1.25E-05	A	
processing of 20S pre-rRNA	11	2	1.97E-04	A	D
ribosomal large subunit biogenesis	9	4	3.17E-04	A	D
small nucleolar ribonucleoprotein complex	20	3	2.52E-04	A	D
cytosolic large ribosomal subunit	69	13	2.87E-04	A	D
ribosomal large subunit assembly and maint.	21	2	2.52E-04	A	D
ribosome biogenesis					D

Tableau 2.3 – Comparaison entre les GFE sous-exprimés obtenus avec trois méthodes
Pour les trois méthodes, les groupes sont considérés fonctionnellement enrichis lorsqu'ils sont associés avec une *p-value* inférieure à 6×10^{-3}

Les trois groupes identifiés par DeRisi et collab. qui ne sont pas identifiés par AGGC, à savoir les groupes sur-exprimés « glycogen metabolism » et « glycogen synthase » et le groupe sous-exprimé « ribosome biogenesis » partagent deux propriétés importantes. Premièrement, ils contiennent des gènes appartenant à une structure hétérogène, i.e. des gènes appartenant à plusieurs groupes fonctionnels. Deuxièmement ces GFE s'expriment uniquement durant une phase spécifique du processus étudié et non pas tout au long de celui-ci.

Ces résultats montrent que l'algorithme AGGC permet d'identifier automatiquement les groupes de gènes co-exprimés significatifs et fonctionnellement riches sans avoir de connaissance a priori des résultats. Il est extensible aux annotations biologiques de toutes natures et aux diverses mesures de variabilité. AGGC analyse tous les sous-ensembles possibles de chaque GFE, accroissant ainsi la sensibilité de la détection des groupes de gènes co-exprimés, même en présence de données très bruitées. Il est également robuste contre les mauvaises assignations lors de la création des groupes fonctionnels à partir des sources publiques (annotations erronées) ou bien de processus automatiques (erreurs de nommage, fautes d'orthographe, etc.). AGGC peut être utilisé pour valider et comparer les résultats d'expériences avec ceux stockés dans les bases de données et les bases documentaires publiques en identifiant les groupes de gènes d'intérêt pour l'expérience en question.

2.4.3 Approche basée sur le co-partitionnement

Des deux approches décrites précédemment, confèrent, selon le cas, une importance prépondérante soit aux données d'expression, soit aux connaissances biologiques. Ceci présente un certain nombre d'inconvénients. Pour l'approche guidée par les données, celle qui est la plus fréquemment utilisée, on peut évoquer :

— le fait que les gènes soient regroupés en fonction de leur similarité d'expression sur toutes les conditions biologiques alors que des gènes impliqués dans un processus biologique particulier peuvent ne s'exprimer de la même manière que pour un sous-ensemble des conditions [Altman et Raychaudhuri, 2001].

— le fait que la plupart des méthodes de partitionnement placent les gènes dans un groupe

de manière exclusive, ce qui a pour effet de ne mettre en évidence que l'un des multiples rôles qu'un gène peut jouer et de masquer les relations complexes qui peuvent exister entre différents groupes de gènes [Gasch et Eisen, 2002].

— le fait que même lorsque des profils d'expression similaires correspondent à des rôles biologiquement proches, déterminer les connexions qui existent entre les gènes d'un même groupe n'est pas une tâche triviale et demande beaucoup de travail d'analyse supplémentaire [Shatkay et collab., 2000].

— le fait que les gènes qui sont fonctionnellement apparentés peuvent démontrer une forte anti-corrélation dans leurs niveaux d'expression (un gène peut être fortement réprimé pour permettre à un autre d'être exprimé) et donc, être regroupés en groupes distincts, brouillant ainsi la relation entre eux.

— le fait que les gènes exprimés simultanément n'ont pas toujours la même fonction, de même que des gènes qui sont exprimés à des moments différents peuvent assurer les rôles complémentaires d'une fonction commune.

Pour pallier ces inconvénients, nous avons travaillé sur une solution n'imposant aucune priorisation sur le traitement des différentes sources de données. Les méthodes basées sur le co-partitionnement permettent de traiter de manière simultanée les informations numériques et les connaissances. Parmi les méthodes existantes, nous avons proposé d'utiliser la méthode de découverte des règles d'association (ARD).

L'ARD est une technique d'exploration de données non supervisée qui permet de découvrir les liens entre des ensembles de valeurs variables (*items*), comme les profils d'expression des gènes ou leurs annotations génétiques, à partir de grandes quantités de données. Les règles d'association identifient les groupes d'éléments qui ont une forte co-occurence en établissant des relations de la forme : $A \Rightarrow B$, qui signifie que lorsque A se produit, il est probable que B se produise également. Pour chaque règle, des mesures de support et de confiance indiquent respectivement la portée et la précision de la règle. Un exemple de règle d'association extraite d'une base de données de ventes de supermarché pourrait être : « céréales $\wedge$ sucre $\Rightarrow$ lait (support 7%, confiance 50%) ». Cette règle indique que les clients qui achètent des céréales et du sucre ont également tendance à acheter du lait. La mesure de support définit la portée de la règle, c'est-à-dire la proportion de clients qui ont acheté les trois articles, et la mesure de confiance définit la précision de la règle, c'est-à-dire la proportion de clients qui ont acheté du lait parmi ceux qui ont acheté des céréales et du sucre. L'extraction de règles d'association consiste à extraire les règles dont le support et la confiance sont au moins égaux à des seuils minimaux de support et de confiance définis par l'utilisateur. Les règles d'association ont été utilisées avec succès dans de nombreux domaines, parmi lesquels l'aide à la planification commerciale, l'aide au diagnostic, l'amélioration des processus de télécommunications, l'organisation et l'accès aux sites Internet ou encore l'analyse d'images. L'extraction de règles d'association est un processus itératif et interactif constitué de plusieurs phases allant de la sélection et la préparation des données jusqu'à l'interprétation des résultats, en passant par la phase de recherche des connaissances.

L'utilisation d'ARD pour analyser les données transcriptomiques a plusieurs avantages :

1. l'ARD peut générer des règles contenant des gènes qui ne sont co-exprimés que dans un sous-ensemble des conditions biologiques ; cette méthode peut donc être considérée comme une méthode de *co-clustering* de ce point de vue,
2. avec l'ARD, tout gène peut être assigné à n'importe quel nombre de règles tant que son profil d'expression remplit les critères d'assignation ; cela signifie qu'un gène impliqué dans de nombreux groupes co-exprimés apparaîtra dans chacun de ces groupes, sans limitation,
3. l'ARD génère des motifs de connaissance orientés « **si** *condition(s)* **alors** *conséquence(s)* », ainsi, tout type de relation entre les mesures d'expression et les annotations biologiques peut être découvert.

4. cette méthode facilite l'intégration de diverses sources d'informations biologiques hétérogènes.

Dans le domaine biologique, ARD a été utilisé pour analyser l'expression des gènes dans le but de découvrir des motifs de variation communs à un sous-ensemble de conditions biologiques [Creighton et Hanansh, 2003; Georgi et collab., 2005; Tuzhilin et Adomavicius, 2002]. Les règles d'association générées par cette approche sont de la forme :

$$\textit{gène } g1\downarrow \quad \Rightarrow \quad \textit{gène } g2\uparrow,\ \textit{gène } g3\downarrow$$

qui signifie que, dans un nombre significatif de conditions biologiques, lorsque *gène* $g1$ est sous-exprimé, il est probable d'observer une sur-expression de *gène* $g2$ et une sous-expression de *gène* $g3$. Cette technique a été appliquée avec succès pour grouper les gènes en fonction de leur niveau d'expression en évitant les inconvénients des techniques de clustering standard [Georgi et collab., 2005]. Cependant, ces algorithmes utilisent exclusivement les mesures d'expression des gènes sans tenir compte de la connaissance biologique. La tâche consistant à découvrir et interpréter les similarités biologiques cachées à l'intérieur des groupes de gènes est laissée à l'expert.

L'ARD a également été utilisé [Carmona-Saez et collab., 2006] pour intégrer les profils d'expression et les informations disponibles sur les gènes dans le but d'extraire des règles de la forme :

$$\textit{annotation } a1 \quad \Rightarrow \quad \textit{condition } c1\downarrow,\ \textit{condition } c2\uparrow$$

qui signifie que les gènes annotés avec *annotation* $a1$ sont susceptibles d'être sur-exprimés dans la *condition* $c1$ et sous-exprimés dans la *condition* $c2$. Cette approche présente plusieurs faiblesses qui ont été présentées par Martinez et Collard [2007].

Premièrement, toutes les méthodes proposées sont basées sur l'utilisation de l'algorithme Apriori [Agrawal et Srikant, 1994] qui est très coûteux en temps et en mémoire dans le cas des données corrélées. De plus, elles génèrent toutes un grand nombre de règles dont beaucoup sont redondantes, ce qui complique l'interprétation des résultats. Il s'agit d'une limitation majeure bien connue de l'algorithme Apriori pour les données corrélées [Tuzhilin et Adomavicius, 2002]. Deuxièmement, les règles extraites sont limitées à quelques formes prédéfinies ; par exemple les annotations du côté gauche et les profils d'expression du côté droit. Or, toutes les règles contenant des annotations ou des profils d'expression, qu'ils interviennent dans les conditions ou les conséquences, apportent des informations importantes pour le biologiste. Troisièmement, elles sont basées sur des méthodes de discrétisation des variations d'expression simplistes (un simple seuil de variation au delà duquel les gènes sont considérés sur- ou sous-exprimés) dont les inconvénients sont connus [Pan et collab., 2005].

Nous avons développé une application appelée GenMiner pour résoudre ces faiblesses et exploiter complètement les capacités de l'ARD dans le cadre d'une fouille de données biologiques. GenMiner permet l'utilisation conjointe des informations connues sur les gènes et de leur niveau d'expression dans certaines conditions de manière à découvrir les relations qui existent entre connaissance a priori et mesures expérimentales. Notre méthode inclut un nouvel algorithme, appelé NorDI (Normal Discretization algorithm) pour discrétiser les mesures d'expression des gènes et générer des profils d'expression. GenMiner utilise l'algorithme Close [Pasquier et collab., 2005], basé sur la fermeture de la connexion du treillis de Galois pour générer, de manière efficiente des règles d'association. Lorsque les données sont denses ou corrélées, telles que les données génomiques, Close réduit à la fois les temps d'exécution et l'espace mémoire par rapport à Apriori, ce qui permet l'analyse d'énormes ensembles de données. De plus, il améliore la pertinence du résultat en extrayant un ensemble minimal de règles ne contenant que des règles non redondantes, réduisant ainsi le nombre de règles et facilitant leur interprétation par les biologistes. Avec ces caractéristiques, GenMiner est une approche ARD adéquate pour répondre aux exigences de l'analyse des données génomiques.

L'extraction des règles d'association est un problème difficile étant donné que l'espace de recherche, c'est-à-dire le nombre de règles potentielles, est exponentiel par rapport à la

taille de l'ensemble d'items et que plusieurs balayages coûteux de données sont nécessaires. Il a été démontré que l'extraction des règles d'association est un problème NP-complet et qu'une approche triviale n'est pas réalisable pour les grands ensembles de données. La première approche efficace proposée pour extraire les règles d'association est l'algorithme Apriori [Agrawal et Srikant, 1994]. Plusieurs optimisations de cette approche ont été proposées depuis, mais les temps de réponse restent du même ordre de grandeur. En effet, cette approche est efficace lorsque les données sont faiblement corrélées et peu abondantes mais la performance diminue considérablement lorsque les données sont corrélées ou denses [Brin et collab., 1997]. De plus, avec de telles données, un grand nombre de règles sont extraites et la plupart de ces règles sont redondantes, c'est-à-dire qu'elles couvrent les mêmes informations. Considérons par exemple, les cinq règles suivantes qui ont toutes le même support la même confiance et le même antécédent :

1. *annotation* ⇒ *gene1*↑
2. *annotation* ⇒ *gene2*↑
3. *annotation* ⇒ *gene1*↑, *gene2*↑
4. *annotation*, *gene1*↑ ⇒ *gene2*↑
5. *annotation*, *gene2*↑ ⇒ *gene1*↑

La règle la plus pertinente du point de vue de l'utilisateur est la règle 3 puisque toutes les autres règles peuvent être déduites par inférence de celle-ci, y compris le support et la confiance. Comme les informations apportées par toutes les autres règles sont résumées dans la règle 3, celle-ci est une règle d'association non redondante avec un antécédent minimal et un conséquent maximal, ou une régle d'association minimale non redondante. Cette situation est fréquente lors de l'extraction de données corrélées ou denses, telles que des données statistiques ou génomiques, et pour résoudre ce problème, l'approche ARD de GenMiner utilise l'algorithme Close pour extraire uniquement des itemsets minimaux non redondants.

La plupart des algorithmes ARD (y compris Close), ne travaillent que sur des valeurs discrètes. Pour pouvoir analyser les variations d'expression qui sont des valeurs continues, il est nécessaire d'effectuer une discrétisation des données pour les transformer en valeurs discrètes. Dans le cas des valeurs d'expression, nous transformons des mesures en étiquettes telles que « sur-exprimé », « sous-exprimé », « fortement sur-exprimé », etc. L'identification des gènes différentiellement exprimés dans une expérience de microarray est un processus très délicat qui n'a encore aujourd'hui, pas de solution satisfaisante. Dans NorDi, nous avons opté pour une méthode simple, mais statistiquement solide. Très brièvement, voici le processus : les valeurs d'expression correspondant à un ensemble de mesures dans une condition donnée sont tout d'abord normalisées. Ensuite, la méthode de Grubb [Grubbs, 1969] est utilisée pour enlever temporairement une valeur aberrante (« outlier »). Chaque fois qu'un outlier est supprimé, un test de normalité, selon la méthode de Jaque-Bera [Bera et Jarque, 1981] est effectué. Si la mesure de normalité est meilleure que la mesure initiale, on essaye de supprimer un autre outlier et on teste à nouveau l'accroissement ou le décroissement de la normalité de l'échantillon. Lorsque la normalité n'est plus améliorée, on détermine un seuil de sur-expression et sous-expression avec un calcul de *z-score*, selon la proportion des gènes que l'on veut identifier comme étant sur ou sous exprimés. On utilise le seuil calculé pour découper l'échantillon initial en classes discrètes.

Pour valider notre approche, nous avons utilisé GenMiner pour fouiller les résultats de puces à ADN utilisés dans un article de Eisen et collab. [1998]. Cet ensemble de données contient des mesures d'expression de 2465 gènes de *Saccharomyces cerevisiae* effectués par quatre études indépendantes : des expériences sur le cycle cellulaire [Spellman et collab., 1998], des expériences de sporulation [Chu et collab., 1998], des expérience impliquant des chocs thermiques [Eisen et collab., 1998] et des mesures durant le phénomène de diauxie [DeRisi et collab., 1997]. Au total, cela représente 79 mesures d'expression. A ces données numériques (que nous avons discrétisées avec NorDi), nous avons ajouté pour chaque gène

Règle	Antécédent	Conséquent	Sup. (#)	Conf. (%)
1	go:0006412 (translation) kegg:sce03010 (ribosome pathway)	heat3↓	95	74
2	go:0005198 (struct. molecule activity) go:0005840 (ribosome)	heat3↓	96	56
3	heat3↓ heat4↓ heat5↓	go:0006412 (translation)	35	88
4	heat2↓	go:0006996 (organelle organization)	41	69
5	heat2↓	go:0042254 (ribosome biogenesis)	39	66
6	heat2↓ heat3↑ heat4↑	go:0006950 (response to stress)	15	52
7	cold4↓	heat3↓	66	68
8	kegg:sce00190 (purine metabolism)	go:0005737 (cytoplasm)	52	96
9	pubmed:16155567	phenotype:inviable	168	93
10	regulator:FHL1	regulator:RAP1	114	86
11	regulator:RAP1	regulator:FHL1	114	61

Tableau 2.4 – Exemples de règles d'associations générées par GenMiner

Dans ce tableau, heat1 à heat6 et cold1 à cold4 font référence aux différentes mesures effectuées lors des expériences de choc thermique. ↑ représente une sur-expression alors que ↓ représente une sous-expression. Les préfixes « go », « kegg », « pubmed », « phenotype » et « regulator » identifient des annotations avec, respectivement des gènes avec les termes GO, les réseaux KEGG, les identifiants pubmed, les descriptions de phénotypes et les nom de facteurs de transcription. Les supports sont donnés en nombre de gènes et les indices de confiance sont des pourcentages.

jusqu'à 658 annotations (24 annotations GO, 14 annotations KEGG, 25 régulateurs transcriptionnels, 14 phenotypes et 581 mots-clés Pubmed). Au final, la matrice que nous avons traité comporte 2465 lignes (une par gènes) et 737 colonnes. Nous avons appliqué GenMiner sur ce jeu de données en utilisant un support minimum de 0.003 (ce qui représente 7 lignes) et un indice de confiance minimum de 30%. Nous avons considéré tous les types de règles, c'est à dire, toutes celles contenant des annotations ou des niveaux d'expression dans l'antécédent ou dans le conséquent.

Le tableau 2.4 montre quelques exemple de règles extraites par GenMiner (une analyse plus détaillée a été publiée dans deux articles [Martinez et collab., 2007, 2008b]. Les règles 1 et 2 illustrent une réduction générale de la synthèse des protéines, de l'organisation du ribosome et de la maintenance des cellules à la suite d'un choc thermique. Les règles 3 à 6 montrent que les gènes qui sont sous exprimés durant un choc de chaleur sont impliquées dans la synthèse des protéines, l'organisation cellulaire et l'organisation ribosomale alors que les gènes qui sont sur exprimés sont impliqués dans la réponse au stress. La règle 7 identifie un groupe de gènes qui sont sous exprimés à la suite des chocs thermiques chaud et froid. Les règles 8 et 9 révèlent de possibles liens entre les annotations provenant de différentes sources comme la relation entre le pathway « purine metabolism » et le terme GO « cytoplasm » ou la forte association entre les gènes cités dans un article présentant une revue des gènes essentiels de la levure et le phénotype « inviable ». Les règles 10 et 11 permettent d'établir une relation forte entre les régulateurs de transcription FHL1 et RAP1. On remarque dans la règle 10, que les gènes régulés par FHL1 sont aussi régulés par RAP1. L'inverse est moins vrai puisque l'on peut remarquer qu'il y a une proportion non négligeable de gènes qui sont régulés par RAP1 mais pas par FHL1. Ces deux règles reflètent le fait, déjà bien connu, que

la reconnaissance de RAP1 est essentielle au recrutement de FHL1.

Ces expérimentations confirment les avantages de GenMiner par rapport aux approches connues. GenMiner permet de rechercher des règles d'association en utilisant un support minimum bien plus petit que ce qui est possible avec les approches basées sur l'algorithme Apriori. De plus, GenMiner réduit considérablement le nombre de règles extraites, ce qui facilite énormément le travail d'exploration et d'interprétation par l'utilisateur final. Nous avons également comparé les règles générées par GenMiner avec les connaissances extraites manuellement des données par plusieurs laboratoires de recherche. Cela nous a permis de mettre en évidence des règles qui ont été corroborées ultérieurement par des articles de recherche.

Chapitre 3

Fouille de graphes attribués

Sommaire

3.1 Introduction

De nombreux domaines de recherche doivent faire face à une abondance de données qui, d'une part rend inefficaces toutes les méthodes d'interprétation manuelles, et d'autre part, induit une spécialisation de la connaissance puisque plus personne n'est en mesure d'avoir une vision globale et exhaustive d'un domaine donné. Cela est particulièrement vrai pour les sciences du vivant et de l'environnement où l'on est maintenant capable, grâce à l'utilisation de nouvelles technologies (télédétections, capteurs, pyroséquençage de transcriptomes, de génomes ou de métagénomes), de générer des quantités faramineuses de nouvelles données.

Une caractéristique essentielle de ces données est le fait qu'elles sont éminemment dynamiques, c'est-à-dire qu'elles évoluent dans l'espace et dans le temps. Dans de nombreux domaines, la compréhension et la modélisation de la dynamique spatio-temporelle sont essentielles. On peut prendre comme exemple les études visant à comprendre la propagation d'une maladie, l'adaptation d'organismes particuliers aux changements environnementaux, la croissance de tumeurs ou l'érosion des sols.

Dans les sciences du vivant, les séries temporelles sont très communes. Elles sont générées lorsque l'on étudie l'évolution d'un phénomène dans le temps. L'aspect spatial est également présent mais il est beaucoup moins étudié. Certaines études, comme celles visant à comprendre les processus de croissance d'un organe [Silk, 1984], analysent les données biologiques en prenant en compte leur aspect spatial et temporel. Dans un tissu en croissance, le temps et l'espace sont en effet intimement, mais non linéairement liés. Les premiers travaux dans ce domaine, jusqu'au début des années 2000, ont traité les dimensions spatiale et temporelle séparément [Yao, 2003]. Dans un second temps, ces travaux ont été étendus pour intégrer simultanément les deux dimensions. Le problème a été abordé de deux manières : par la détection de trajectoires [Giannotti et collab., 2007] ou par l'identification de séquences d'événements localisés [Mohan et collab., 2010]. Cependant, dans ces travaux, les domaines de motifs utilisés et donc les motifs extraits ne sont pas à la hauteur de la complexité spatiale des objets à étudier.

Très récemment, les séquences et plus généralement les graphes, ont été utilisés et étendus au spatio-temporel pour représenter la propagation de phénomènes dans l'espace et dans le temps [Alatrista-Salas et collab., 2012]. Des modèles de graphes plus riches ont été considérés où les sommets et les arêtes sont étiquetés par plusieurs attributs ou propriétés, au lieu d'un unique label. Un réseau social peut, par exemple, être représenté par un grand graphe où chaque sommet désigne une personne ainsi que ses domaines d'intérêt. Ces graphes sont nommés graphes attribués et les attributs associés aux sommets sont appelés itemsets. Quand la notion de temps est exprimée, les arêtes sont le plus souvent orientées (on parle alors d'arcs). Dans un graphe attribué, les sommets peuvent représenter des objets spatiaux, caractérisés par un ensemble d'attributs ou événements, tandis que les arcs peuvent exprimer leur proximité spatio-temporelle (c-à-d, les objets voisins dans des temps consécutifs).

Le problème qui est donc posé revient à rechercher dans un ou plusieurs graphes attribués, des motifs dont la fréquence (également appelé support) est supérieure à un seuil donné. Cela permet en effet d'identifier des combinaisons et successions d'événements qui sont susceptibles d'aider à la compréhension de certains phénomènes, voire de les prédire.

La plupart des algorithmes d'extraction de graphes s'appliquent sur des graphes étiquetés, où chaque sommet ou arc n'a qu'une seule étiquette associée. Dans le cas des graphes attribués, l'espace de recherche porte à la fois sur la structure des graphes et sur la combinaison des attributs ce qui provoque une explosion combinatoire. Peu d'études ont considéré les graphes attribués. Miyoshi et collab. [2009] ont proposé une solution pour les graphes étiquetés attribués, c'est à dire des graphes où chaque sommet est associé à une étiquette et à des attributs quantitatifs. En gardant les étiquettes et en leur attribuant une place prépondérante, la recherche de motifs fréquents est simplifiée. Il suffit de combiner un algorithme « classique » d'exploration de sous-graphe pour le graphe étiqueté, et un algorithme existant

de fouille d'itemsets pour les attributs quantitatifs. Certains auteurs (par exemple Fukuzaki et collab. [2010]) se concentrent sur la fouille de motifs cohésifs et de motifs partageant des itemsets, c'est-à-dire des motifs représentant des sous-graphes avec des attributs partagés.

L'approche de Miyoshi et collab. [2009] ne peut pas être utilisée si les attributs associés aux sommets ont tous la même importance. Les travaux portant sur les motifs cohésifs ne répondent pas à notre problématique dans le sens où nous recherchons des motifs dans lesquels les itemsets associés aux sommets ne sont pas forcément identiques. De plus, les approches existantes ne traitent que l'extraction de motifs dans dans un ensemble de graphes ; la recherche des motifs fréquents dans un unique graphe attribué n'a pas encore été étudiée.

Après quelques préliminaires, je décrirai plusieurs de nos travaux qui ont permis de faire avancer l'état de l'art dans le domaine de la fouille de graphes attribués.

3.2 Préliminaires

La figure 3.1 est un exemple de graphe attribué. Sur chaque sommet, les chiffres en gras représentent les identifiants des sommets. Ils sont uniquement affichés pour la clarté de l'illustration et ne font pas partie des données. Les lettres a,b,c,d,e,f,g,h,i représentent des attributs des sommets. Chaque sommet est associé à un ensemble d'attributs (un itemset).

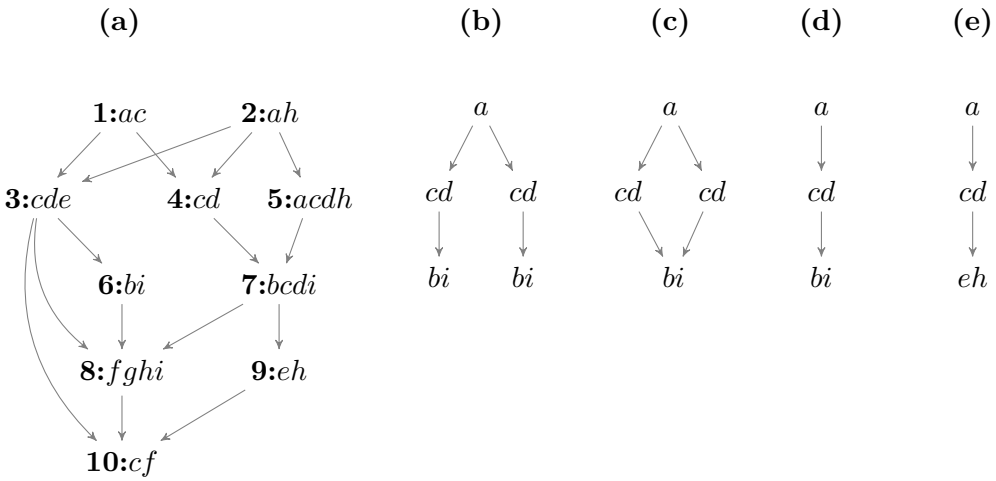

Figure 3.1 – Exemple d'un graphe attribué (a) et de quelques uns de ses sous graphes attribués
Dans (a), les chiffres en gras représentent les numéros de sommets mais ne font pas partie des données du graphe. Les sous-graphes (b) et (d) sont présents 2 fois, (c) n'apparait qu'une seule fois et (e) ne peut être trouvé (2 fois) qu'avec une recherche de motifs enchâssés (voir détails dans le texte).

Soit $\mathcal{I} = \{i_1, i_2, .., i_n\}$ un ensemble d'items. Un **itemset** est un ensemble $\mathcal{P} \subseteq \mathcal{I}$. Les items appartenant à un itemset sont triés selon l'ordre lexicographique et peuvent être accédés par leur index (par exemple $\mathcal{P}_2$ pour référencer le deuxième item de $\mathcal{P}$).

Un **graphe attribué orienté** $G = (V, E, \lambda)$ sur un ensemble d'items $\mathcal{I}$ consiste en un ensemble de sommets V, un ensemble (orienté) d'arcs $E \subseteq \{(u, v) \in V^2 / u \neq v\}$ et une fonction $\lambda(v) : V \rightarrow pow(\mathcal{I})$ qui associe un ensemble d'items $i \in \mathcal{I}$ à chaque sommet $v \in V$. La notation $pow(\mathcal{I})$ représente l'ensemble des parties de $\mathcal{I}$. Pour un arc $(u, v) \in E$, u est le parent de v et v est le fils de u. Il existe un cycle dans le graphe orienté si un chemin peut être trouvé entre un sommet et lui-même. Un graphe attribué orienté est **enraciné** si l'on peut trouver un sommet v tel qu'il existe un chemin entre v et n'importe quel autre sommet du graphe. Un **DAG attribué** est un graphe attribué acyclique orienté. Un **arbre attribué** est un DAG attribué possédant une unique racine et tel que tous les sommets (sauf la racine) ont

un unique parent. L'**arbre attribué couvrant** d'un graphe attribué connexe est un arbre attribué inclus dans ce graphe et qui connecte tous les sommets du graphe.

Tous les algorithmes de fouille de graphe travaillant avec des graphes non-ordonnés doivent prendre en compte le problème d'isomorphisme. Deux graphes attribués $G_1 = (V_1, E_1, \lambda_1)$ et $G_2 = (V_2, E_2, \lambda_2)$ sont **isomorphes** *ssi* il existe une bijection $\varphi : V_1 \to V_2$ telle que : (i) $\forall v \in V_1 : \lambda_1(v) = \lambda_2(\varphi(v))$, (ii) $\forall (u, v) \in E_1 \Leftrightarrow (\varphi(u), \varphi(v)) \in E_2$. Un **automorphisme** est un isomorphisme pour lequel $G_1 = G_2$. Un graphe attribué $G_1 = (V_1, E_1, \lambda_1)$ est un **sous graphe attribué** de $G = (V, E, \lambda)$ *ssi* il existe une correspondance $\varphi : V_1 \to V$ telle que : (i) $\forall v \in V_1 : \varphi(v) \in V$, (ii) $\forall v \in V_1 : \lambda_1(v) \subseteq \lambda(\varphi(v))$, et (iii) $\forall (u, v) \in E_1 : (\varphi(u), \varphi(v)) \in E$.

A partir des définitions précédentes, nous généralisons la notion de sous-graphe attribué de la façon suivante. $G_1 = (V_1, E_1, \lambda_1)$ est un sous-graphe attribué d'un graphe attribué $G_2 = (V_2, E_2, \lambda_1)$ et on note $G_1 \sqsubset G_2$ si G_1 est un sous-graphe attribué isomorphe de G_2. Si G_1 est un sous-graphe attribué de G_2, on dit que G_2 est un super-graphe attribué de G_1. G_1 est un sous-graphe attribué induit de G_2 *ssi* G_1 est un sous-graphe attribué isomorphe de G_2 et φ préserve la relation directe des arcs. G_1 est un sous-graphe attribué enchâssé de G_2 *ssi* G_1 est un sous-graphe attribué isomorphe de G_2 et φ représente une relation pouvant être transitive (i.e. : dans la figure 3.1, le sommet 1 est relié au sommet 6 de manière indirecte via le sommet 3). $G_1 = (V_1, E_1, \lambda_1)$ est appelé un sous-graphe attribué d'intervalle i de $G_2 = (V_2, E_2, \lambda_2)$ *ssi* G_1 est un sous-graphe attribué enchâssé de G_2 et φ satisfait à la contrainte suivante : $\forall u \forall v \in E_1$ tel que v est accessible à partir de u et $d(\varphi(u), \varphi(v)) = 1, d(u, v) \leq i$ où $d(x, y)$ représente le nombre d'arcs entre x et y dans le graphe attribué.

Les données à analyser sont disponibles soit sous la forme d'une base de données $\mathcal{B}$ de graphes orientés attribués, soit sous la forme d'un unique graphe orienté attribué G, non nécessairement connexe. Dans le premier cas, il est possible de calculer, pour chaque motif identifié G, une fréquence par transaction qui est donnée par le nombre de graphes dans $\mathcal{B}$ pour lesquels G est un sous-graphe attribué. Dans le second cas, nous utilisons une mesure de support anti-monotone proche de celle définie par Bringmann et Nijssen [2008] qui est égale au nombre minimal d'occurrences dans G de chacun des sommets du motif. Dans la figure 3.1, par exemple, le motif (b) a une occurrence de 2 car il y a deux manières distinctes de choisir l'étiquette a alors que pour obtenir le motif (c), l'étiquette a doit forcément être issue du sommet 2. Nous pouvons avec cette mesure calculer une fréquence relative qui est donnée par la fréquence absolue divisée par le nombre de sommets du graphe. Nous considérons qu'un motif est fréquent si sa fréquence est supérieure ou égale à un seuil minimum.

3.3 Recherche de motifs fréquents combinant fouille d'itemsets et parcours structurel

Les graphes attribués peuvent être vus comme des itemsets organisés selon une structure de graphe. L'inclusion sur les graphes attribués peut donc être définie en considérant soit l'inclusion d'itemsets, soit l'inclusion structurelle. Pour l'inclusion portant sur les itemsets, on considère que le graphe attribué G_1 est contenu dans un autre graphe attribué G_2 si les deux graphes attribués ont la même structure et que, pour chaque sommet de G_1, l'itemset qui lui est associé est inclus dans l'itemset du sommet correspondant de G_2. Plus formellement, $G_1 = (V_1, E_1, \lambda_1)$ est contenu dans $G_2 = (V_2, E_2, \lambda_2)$, et on note $G_1 \sqsubset_I G_2$, si $V_1 = V_2$ et $E_1 = E_2$ et $\forall x \in V_1, \lambda_1(x) \subseteq \lambda_2(x)$. L'inclusion structurelle est quant à elle représentée par le concept classique de sous-graphe. Elle est notée $\sqsubset_S$.

La stratégie proposée est de partir d'un ensemble de motifs initiaux composés d'un sommet associé avec un unique item. L'arbre de recherche est construit en effectuant un parcours de l'arbre attribué couvrant et en utilisant un ordre basé sur un code de représentation des graphes acycliques attribués que nous avons défini. Le code utilisé permet de déterminer rapidement si le motif est sous une forme canonique sans avoir à effectuer des calculs coûteux

d'isomorphisme de graphes. Pour l'extension de la structure arborescente attribuée, nous utilisons une variante de la technique connue sous le nom d'extension du chemin droit (rightmost path extension [Asai et collab., 2002]).

Le code que nous avons proposé possède la propriété suivante : chaque préfixe d'un code canonique est lui-même canonique. L'extension d'un motif ne change pas l'ordre des éléments du code lors de la génération d'un motif canonique (si un changement intervient, alors, on sait que le motif n'est pas canonique). Chaque motif canonique ne peut être obtenu qu'en ajoutant un nouvel élément à la fin. De plus, en utilisant la stratégie d'extension du chemin droit, il est possible de vérifier la canonicité d'un motif en ne testant que les sommets appartenant au chemin droit. Soit les fonctions l et p qui, appliquées à un sommet retournent respectivement son dernier et son avant-dernier fils. Il est possible de déterminer la canonicité d'un motif de la manière suivante : pour tous les sommets n composant le chemin droit, si $\exists p(n)$ et $code(p(n)) \leq code(l(n))$, alors, le motif est sous forme canonique.

La génération de nouveaux motifs est effectuée par extension des sommets composant le motif. Nous avons proposé une méthode complète de parcours de l'espace de recherche qui combine deux types d'extensions :

— **l'extension d'itemset**, qui permet d'ajouter un nouvel item à l'itemset associé au sommet terminal le plus à droite,

— **l'extension de structure**, qui consiste à ajouter un fils à l'un des sommets composant le chemin droit.

Comme nous utilisons une mesure de support anti-monotone (l'extension d'un motif ne peut pas donner un nouveau motif avec une fréquence supérieure), il est possible d'arrêter l'exploration d'une branche lorsque la fréquence d'un candidat est inférieure au support minimum. Par exemple, dans la figure 3.1, si l'on détermine que la fréquence du motif (d) est inférieure au support minimum, il n'est pas nécessaire de générer des extensions de ce motif, et ainsi éviter l'exploration des motifs (b) et (c) dont on sait qu'ils ne répondent pas au critère minimum de fréquence. L'extension est également stoppée si le candidat examiné n'est pas sous forme canonique.

Une première méthode a été élaborée pour traiter les arbres attribués. Par la suite, la méthode a été étendue pour traiter les DAG attribués (résultats non publiés) et les graphes attribués orientés. Si le passage de la fouille d'arbres attribués à celle des DAGs attribués a été relativement aisée, la prise en compte des cycles a nécessité de modifier le code des arbres attribués couvrants et à éliminer les motifs isomorphes qui sont fatalement générés pour toutes les explorations qui débutent sur un autre sommet faisant partie du cycle.

3.4 Fermeture portant sur les graphes attribués

Même en utilisant les stratégies d'élagage de l'espace de recherche définies plus haut, la méthode de fouille identifie de nombreux motifs non pertinents. C'est le cas, par exemple, du motif (d) de la figure 3.1 qui est inclus dans (b) et apparaît avec la même fréquence que ce dernier dans le graphe initial. On voit bien que le motif (d) n'apporte aucune information supplémentaire car il peut être déduit de la fréquence de (b).

Dans des applications réelles, générer toutes les solutions peut s'avérer très coûteux ou même impossible. Depuis la proposition de Mannila et Toivonen [1996] d'importants efforts on été consacrés à l'élaboration de représentations condensées qui résument toutes les solutions dans un ensemble restreint. L'ensemble des motifs fermés est un exemple d'une telle représentation condensée [Pasquier et collab., 1999c].

La notion de fermeture peut être prise en compte dans la fouille de graphes attribués en adoptant la définition empirique suivante : un graphe attribué G est un graphe attribué fermé si aucun de ses super-graphes attribués ne possède le même support que lui. Dans le cadre

de motifs ayant une structure de graphe, le calcul des fermés nécessite d'effectuer plusieurs tests d'isomorphisme de graphes qui sont des opérations coûteuses.

Nous avons proposé deux autres représentations résumées des motifs qui sont définie soit en fonction de l'inclusion portant sur les itemsets (les motifs c-fermés), soit en fonction de l'inclusion portant sur la structure (les motifs s-fermés).

Soit G un graphe attribué candidat, $\mathcal{H}$ l'ensemble de tous les sous-graphes attribués fréquents identifiés précédemment et $\mathcal{X}$ l'ensemble de tous les candidats générés par l'extension de G. Nous distinguons deux sous ensembles de $\mathcal{X}$: $\mathcal{X}_I$, l'ensemble des motifs générés par extension d'itemset de G et $\mathcal{X}_S$, l'ensemble des motifs obtenus par extension de structure. Nous définissons deux fonctions : fs qui donne le support d'un sous-graphe attribué et fo qui retourne son nombre total d'occurrences. G est un graphe attribué c-fermé si $\nexists G' \in \mathcal{H} \cup \mathcal{X}_I$ tel que $G \sqsubset_I G'$ et $fs(G') = fs(G)$. Cependant, identifier une extension d'itemset de G avec la même fréquence par transaction que G ne permet pas de stopper l'exploration des autres candidats de $\mathcal{X}$. La condition supplémentaire suivante doit également être satisfaite : $\exists G' \in \mathcal{H} \cup \mathcal{X}_I : G \sqsubset_I G'$ et $fo(G') = fo(G)$.

Les motifs s-fermés sont définis de la même manière en remplaçant, dans les formules du paragraphe précédent $\sqsubset_I$ par $\sqsubset_S$. Pour l'énumération des motifs fermés ou s-fermés, il faut également supprimer les arbres attribués non fermés stockés dans $\mathcal{H}$, i.e. tous les motifs qui sont des sous-arbres attribués de $\mathcal{H}$ avec la même fréquence.

Le test de fermeture pour les motifs fermés ou s-fermés nécessite d'effectuer des tests d'isomorphisme. Il en va autrement pour l'énumération des motifs c-fermés pour lesquels le test de fermeture se limite à un test d'inclusion d'itemsets. L'énumération des motifs c-fermés permet de réduire drastiquement à la fois le nombre de motifs retournés et le temps d'exécution. Des tests ont montré que cette représentation condensée offrait un bon compromis entre vitesse d'exécution et concision des résultats.

3.5 Réduction de l'espace de recherche par élagage des motifs automorphes

Deux raisons principales rendent la fouille des graphes attribués très difficile. En premier lieu, il est nécessaire de combiner de manière non triviale l'exploration de la structure du graphe avec l'identification d'itemsets fréquents associés aux sommets. La seconde raison est que, comme pour la fouille des graphes étiquetés, les performances sont très impactées par la présence de sous-graphes isomorphes [Yan et Han, 2002].

L'une des méthodes privilégiées pour éviter d'explorer plusieurs fois une même partie de l'espace de recherche est d'ajouter des arcs et des sommets aux motifs candidats de manière ordonnée et d'utiliser une représentation canonique des sous-graphes. L'idée sous-jacente à l'utilisation des formes canoniques est de n'étendre qu'une seule des solutions isomorphes. Pour éviter d'effectuer des tests coûteux d'isomorphisme complets, une représentation canonique adaptée à la manière d'étendre les sous-graphes est choisie. Cette forme canonique permet de décider rapidement si un motif a déjà été examiné et donc, ne doit pas être étendu. La solution la plus utilisée est de créer une séquence de tous les arcs contenus dans le sous-graphe dans l'ordre ou les arcs ont été ajoutés [Borgelt, 2007; Yan et Han, 2002].

L'utilisation des formes canoniques, si elle permet d'éviter d'explorer plusieurs fois le même sous-graphe, ne dispense pas de la nécessité d'identifier toutes les manières de générer les motifs. Ainsi, si l'on ordonne les sommets dans l'ordre lexicographique de leurs étiquettes, le motif (b) de la figure 3.2, est canonique, mais il est présent 4 fois dans le graphe initial (a). Les motifs (c) et (d), qui sont eux aussi canoniques, peuvent être générés de 12 façons différentes (dans le graphe initial, il y a 12 manières possibles de choisir 2 sommets b parmi les 4 fils étiquetés b du sommet a). Il est nécessaire de tenir compte de toutes ces formes car la configuration choisie conditionne les extensions futures du motif ; par exemple, la possibilité

ou non d'étendre chacun des sommets b avec un sommet étiqueté c. Les motifs qui possèdent de nombreux isomorphismes de sous-graphes avec le graphe analysé posent des difficultés à tous les algorithmes existants, car le problème d'isomorphismes de sous-graphes en lui-même est NP-complet. Les motifs (c) et (d), qui possèdent des automorphismes posent le plus de problèmes et contribuent de manière importante à la dégradation des performances des algorithmes, car le nombre d'isomorphismes de sous-graphes avec le graphe initial est augmenté. Comme indiqué par Yan et Han [2002], dans le cas de graphes étiquetés creux avec des labels diversifiés, le nombre de sous-graphes isomorphes est la plupart du temps faible. Au contraire, dans les graphes attribués, la combinatoire sur les itemsets augmente grandement la probabilité d'observer des sous-graphes partageant un ou plusieurs attributs identiques. La conséquence est un nombre potentiellement très important d'isomorphismes de sous-graphe.

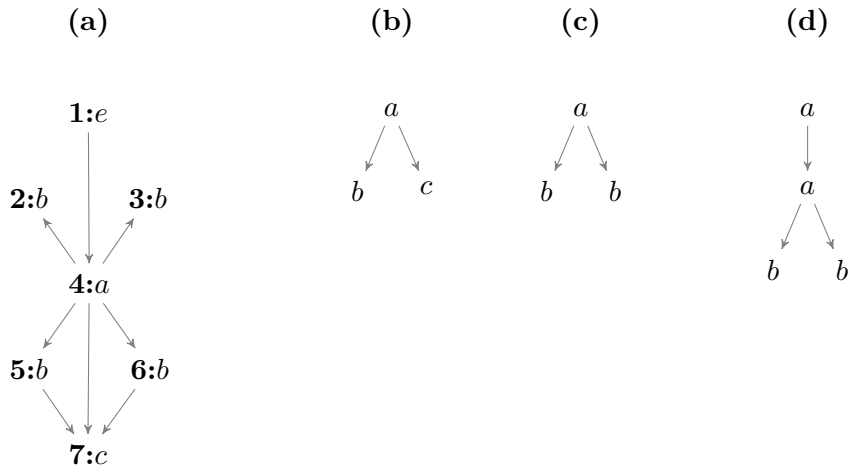

Figure 3.2 – Exemple d'un graphe étiqueté orienté (a) et de 3 sous motifs (b, c et d)

Dans une section précédente, nous avons décrit la manière de déterminer la canonicité d'un motif en utilisant les fonctions l et p qui, appliquées à un sommet retournent respectivement son dernier et son avant-dernier fils. Il est également assez aisé, en utilisant les deux mêmes fonctions, d'identifier les extensions qui produisent des motifs automorphes. S'il existe un sommet n dans le chemin droit du motif tel que n a plus d'un fils et $code(p(n)) = code(l(n))$, alors le motif possède des automorphismes. Sur un tel motif, on appelle sommet de séparation le sommet du chemin droit, le plus proche de la racine, qui possède plus d'un fils. Ainsi, les motifs (c) et (d) de la figure 3.2 possèdent des automorphismes. Dans ces deux figures, le sommet de séparation est le sommet a.

L'utilisation de formes canoniques permet d'élaguer l'espace de recherche et cette optimisation est utilisée dans notre algorithme AADAGE. Cependant, cela ne permet pas de résoudre le problème posé par la présence de motifs automorphes.

Prenons par exemple le graphe orienté attribué représenté dans la figure 3.3(a). Ce graphe contient plusieurs sous-structures automorphes qui rendent l'élagage basé sur l'utilisation des formes canoniques moins opérant. Lorsque le motif illustré dans la figure 3.3(b) est analysé, il est conservé et étendu puisqu'il est sous forme canonique. Or, ce motif peut être obtenu de douze façons différentes. L'algorithme naïf consiste, pour chacune de ces formes, à étendre le motif en utilisant comme défini précédemment, l'extension d'itemset ou l'extension de structure.

L'extension d'itemset génère le motif illustré dans la figure 3.3(c). Ce motif, qui n'est pas sous forme canonique, sera écarté. L'extension de structure à partir du sommet b génère l'un des motifs décrits dans la figure 3.3(e). Dans tous les cas, le motif obtenu n'est pas sous forme canonique donc il sera écarté. La seule opération qui permet de générer un motif canonique est l'extension de structure à partir du sommet de séparation (figure 3.3(d)).

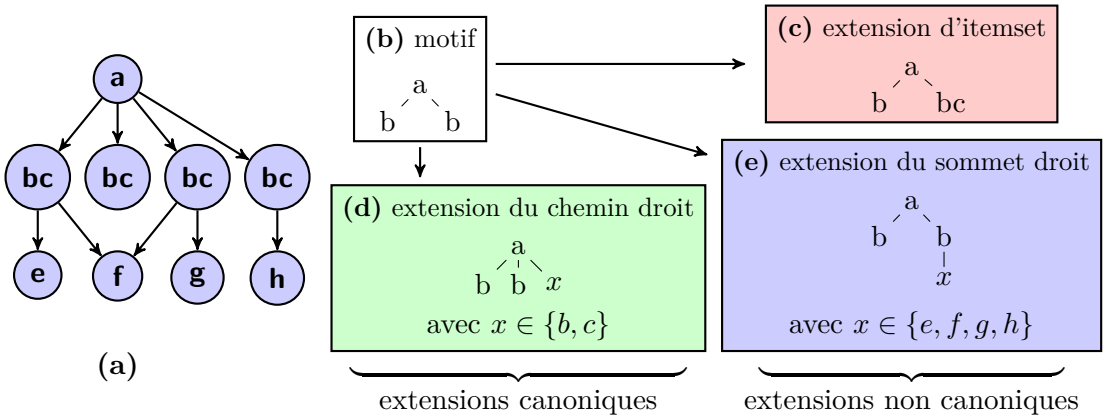

Figure 3.3 – Illustration des extensions d'un motif canonique possédant des automorphismes
Le graphe orienté attribué en (a) contient plusieurs sous-graphes automorphes comme celui choisi comme motif (b). Les extensions de sommet droit (e) ou d'itemset (c) effectuées sur le motif génèrent obligatoirement des motifs non canoniques. Seule l'extension du chemin droit (d) peut produire des motifs canoniques.

Lorsque l'on identifie un nouveau motif possédant des automorphismes, on sait que l'on ne pourra pas faire d'extension sur le dernier sommet, ni sur tous les sommets situés après le sommet de séparation. En effet, pour un sommet de séparation n, on a $code(p(n)) = code(l(n))$ donc, n'importe quelle extension effectuée sur le motif de racine $l(n)$ rendra son code inférieur à $p(n)$, donc le motif ne sera plus canonique. Nous pouvons utiliser ce fait pour éviter de générer inutilement des motifs. Ceci est une première optimisation, mais son impact est limité car elle ne modifie pas le nombre d'instances du motif isomorphe qui constitue le cœur du problème.

La seconde optimisation consiste à supprimer certaines manières d'obtenir des motifs automorphes qui ne permettent pas de générer de nouveaux motifs canoniques. Dans la figure 3.2, par exemple, le motif (d), qui est obtenu de 12 façons différentes, ne peut être étendu que par l'ajout d'un troisième fils (qui peut être étiqueté b ou c) au sommet de séparation a. Quelle que soit la manière dont le motif (d) a été obtenu, les possibilités d'extension sont inchangées. Il est donc possible, dans ce cas, sans omettre de solution, de ne conserver qu'une seule mise en correspondance du motif ; n'importe laquelle.

Avec le graphe tel que présenté dans la figure 3.3(a), la situation est plus compliquée car il est possible, à partir du motif, d'ajouter un fils c au sommet a. Dans ce cas, il sera possible lors de la génération suivante, d'appliquer n'importe quel type d'extension sur tous les sommets du chemin droit. Dans l'exemple, on pourrait en particulier ajouter e, f, g ou h en tant que fils du sommet c. Pour que tous les motifs possibles puissent être générés, il faut conserver les possibilités d'étendre le motif avec 2 sommets bc parmi les 3 sommets bc ayant un fils, soit $\binom{3}{2} = 3$ formes différentes.

Dans un graphe réel, la situation est souvent beaucoup moins simple et déterminer quelles formes il est possible de supprimer n'est pas une tâche facile. Pour éviter d'effectuer des calculs portant sur la structure du graphe qui nuiraient aux performances de l'algorithme, on choisit de ne conserver qu'une seule forme parmi celles utilisant les mêmes sommets. On passe ainsi à un nombre de formes déterminé par un nombre d'arrangements possibles à un nombre de formes basé sur le nombre de combinaisons. Cette solution n'est pas optimale mais elle est suffisante pour limiter l'explosion combinatoire. Sur une configuration en étoile dans laquelle un sommet possède n fils avec une même étiquette, il existe $\sum_{i=1}^{n} {}^{n}P_k$ manières différentes de générer des motifs inclus. Après optimisation, ce nombre chute à $\sum_{i=0}^{n-1} (n-i)\binom{n}{i}$.

3.6 Représentation condensée exacte des chemins fréquents

Nous avons abordé le problème de l'extraction de chemins pondérés fréquents dans un unique graphe attribué sans cycle (DAG) où chaque poids exprime la fréquence d'une transition. Les chemins fréquents sont utilisés pour analyser la relation causale entre séquences d'événements et/ou attributs. Comme le nombre de motifs peut être très grand, nous avons conçu une représentation condensée pour de telles collections. Ces travaux se situent à l'intersection de plusieurs sujets importants tels que l'extraction de séquences et la fouille de graphes.

Plusieurs algorithmes ont été proposés pour l'extraction de séquences fréquentes. La fermeture, qui a été intensivement étudiée pour la fouille d'itemsets [Pasquier et collab., 1999c] a été étendue à la fouille de séquences [Yan et collab., 2003]. Les représentations condensées de sous-graphes fréquents ont également été étudiées. Yan et Han [2003] ont proposé d'extraire les sous-graphes fermés dans une base de données transactionnelle et Termier et collab. [2007] ont adapté cette approche pour les DAGs. Ces études se concentrent sur la recherche de motifs fréquents dans des transactions. Dans nos travaux, nous avons considéré la fouille dans un unique DAG qui soulève différents problèmes.

L'un des premiers problèmes était de définir la mesure de fréquence des motifs. Le second problème consistait à définir une représentation condensée appropriée.

3.6.1 Chemin et occurrence de chemin

Soit P une suite d'itemsets $I_i \in \mathcal{P}(\mathcal{I})$ notée $P = I_1 \rightarrowtail I_2 \rightarrowtail \cdots \rightarrowtail I_{|P|}$. P est appelé **chemin** ssi il existe une suite de sommets $v_1, v_2, \ldots, v_{|P|} \in V_G$ qui satisfont $\forall I_i \in P$, $I_i \subseteq \lambda_G(v_i)$, et chaque sommet v_i est un père de v_{i+1} dans G. La succession de sommets $O = v_1 \rightarrowtail v_2 \rightarrowtail \cdots \rightarrowtail v_{|P|}$ est une **occurrence** de P. Par exemple, dans le DAG attribué donné en figure 3.1, les occurrences du chemin de **taille 3** $ah \rightarrowtail cd \rightarrowtail i$ sont $2 \rightarrowtail 3 \rightarrowtail 6$, $2 \rightarrowtail 3 \rightarrowtail 8$, $2 \rightarrowtail 4 \rightarrowtail 7$, $2 \rightarrowtail 5 \rightarrowtail 7$ et $5 \rightarrowtail 7 \rightarrowtail 8$. L'occurrence $2 \rightarrowtail 3 \rightarrowtail 6$ quant à elle **supporte** les chemins $ah \rightarrowtail cde \rightarrowtail bi$, $a \rightarrowtail cde \rightarrowtail bi$, $h \rightarrowtail cde \rightarrowtail bi$, $h \rightarrowtail c \rightarrowtail bi$, etc. De plus, nous posons que P_i représente le $i^{\text{ème}}$ itemset de P, que O_i représente le $i^{\text{ème}}$ sommet de O, et que $occur_G(P)$ représente l'ensemble des occurrences de P dans G.

Nous définissons les chemins pondérés comme étant des chemins avec un poids sur chaque arc qui représente le nombre d'occurrences différentes de cet arc parmi toutes les occurrences dans le chemin. Dans les données de l'exemple de la figure 3.1, le chemin $P = ah \rightarrowtail cd \rightarrowtail i$, dont les occurrences ont été énumérées précédemment, nous donne le motif :

$$ah \overset{4}{\rightarrowtail} cd \overset{5}{\rightarrowtail} i.$$

En effet, les différentes occurrences de `ahcd` dans $occur_G(P)$ sont au nombre de 4, et les différentes occurrences de `cdi` dans $occur_G(P)$ sont au nombre de 5. Une telle représentation permet de voir que l'itemset ah apparaît 4 fois avant l'apparition du chemin `cdi`, et que l'itemset i apparaît 5 fois après l'apparition du chemin `ahcd`. Dorénavant, $\omega_G(P_i \rightarrowtail P_{i+1})$ désignera le poids de l'arc entre les itemsets P_i et P_{i+1}.

3.6.2 Relation d'inclusion

Nous avons défini l'opérateur $\sqsubseteq$ sur un couple de chemins pondérés de la manière suivante : $P \sqsubseteq P'$ ssi $|P| \leq |P'|$ et $\exists k \in [0, |P'|[$ tel que

$$\begin{cases} \forall i \in [1, |P|], \quad P_i \subseteq {P'}_{k+i} & \text{(inclusion d'itemsets)} \\ \forall j \in [1, |P|[, \quad \omega_G(P_j \rightarrowtail P_{j+1}) = \omega_G({P'}_{k+j} \rightarrowtail {P'}_{k+j+1}) & \text{(égalité des poids)} \end{cases}$$

Un chemin pondéré P' absorbe un autre chemin pondéré P si nous pouvons trouver P dans une sous-séquence de P'. Nous disons ainsi que P' est un **super-chemin-pondéré** de P, ou que P' **contient** P.

A partir de la mesure proposée par Bringmann et Nijssen [2008] [1], nous définissons une mesure de support anti-monotone pour un motif P dans un a-dag G, notée $\sigma_G(P)$:

$$\begin{aligned}\sigma_G(P) &= \min_{1\leq i<|P|} |occur(P_i \rightarrowtail P_{i+1}) \text{ dans } occur_G(P)| \\ &= \min_{1\leq i<|P|} \omega_G(P_i \rightarrowtail P_{i+1})\end{aligned}$$

La collection de motifs fréquents dans G est l'ensemble des motifs P tels que $\sigma_G(P) > minsup$.

Dans un DAG attribué G donné, nous avons cherché une représentation condensée notée $cond(G)$ de tous les chemins pondérés. Notre objectif était que chaque chemin pondéré fréquent *ainsi que son support* puisse être déductible de $cond(G)$. De plus, aucun élément de $cond(G)$ ne doit pouvoir être déductible à partir d'un autre élément de $cond(G)$. En d'autres termes, nous avions besoin à la fois de la maximalité, de l'unicité et de la complétude des solutions dans la collection $cond(G)$.

Comme il est possible de lire directement le support d'un chemin pondéré à partir de celui-ci (en prenant le poids minimum parmi tous ceux du motif), nous avons pu définir une représentation condensée comme suit :

$$cond(G) = \{P \ / \ \nexists P' \text{ t.q. } P \sqsubseteq P'\}$$

Les chemins pondérés de taille 2 se représentent naturellement par des triplets de la forme $\{\omega(P_{orig} \rightarrowtail P_{dest}), I_{orig}, I_{dest}\}$. Ces triplets correspondent à des concepts triadiques fréquents tels que définis par Cerf et collab. [2008]. Nous avons donc utilisé Data-Peeler ; l'algorithme développé par Cerf et collab. pour calculer les chemins pondérés de taille 2.

Soient $L2W$ l'ensemble des chemins pondérés de taille 2, ainsi que la relation ternaire suivante :

1. La première dimension est E_G, la seconde et la troisième sont $\mathcal{I}$,
2. Pour un arc $v_1 \rightarrowtail v_2 \in E_G$ donné, le tuple correspondant $T \in (E_G, \mathcal{I}, \mathcal{I})$ est $T = (v_1 \rightarrowtail v_2, \lambda_G(v_1), \lambda_G(v_2))$,

Les motifs fermés de dimension 3 (c-à-d, les chemins de taille 2) **sont équivalents aux** chemins pondérés condensés par rapport à $L2W$.

Par exemple, à partir du motif de dimension 3 $S = \{\{1 \rightarrowtail 3, 1 \rightarrowtail 4\}, \{b, c\}\{c, d, e\}\}$, on obtient le chemin pondéré $P = bc \overset{2}{\rightarrowtail} cde$. Avoir un tel ensemble de chemins pondérés de taille 2 nous permet de procéder à une extension via une recherche en profondeur, à partir de n'importe quel chemin pondéré de taille 2 trouvé.

Le but de la deuxième étape était d'étendre les chemins pondérés jusqu'à ce qu'ils deviennent des condensés. Pour ce faire, nous avons exploité les chemins pondérés précédemment extraits qui sont assurés d'être maximaux au niveau des itemsets des sommets d'origine et de destination. Nous avons proposé d'effectuer des extensions à la fois vers le bas (en ajoutant des fils) et vers le haut (en ajoutant des pères).

Pour un motif P, nous avons défini $V_{dest} = \{v_{|P|} \in occur_G(P)\}$ et $Li = \{i \in \mathcal{I}$, tel que $\forall v \in V_{dest}, v$ a au moins un fils u t.q. $i \in \lambda_G(u)\}$. La méthode d'extension est fondée sur la proposition suivante :

Soit P un chemin pondéré à étendre (vers le bas). Soit $DB|_P$, la relation binaire projetée par rapport à P, définie comme suit :

1. La première dimension est E_G, la seconde est $\mathcal{I}$,
2. Pour un arc $v_1 \rightarrowtail v_2 \in E_G$ t.q. $v_1 \in V_{dest}$, la transaction correspondante $T \in (E_G, \mathcal{I})$ est $T = (v_1 \rightarrowtail v_2, \lambda_G(v_2) \cap Li)$

1. Cette mesure peut tout aussi bien s'appliquer sur des sommets ou des arcs

Les chemins pondérés étendus $\{P \rightarrowtail u \ / \ u \text{ est fermé dans } DB|_P\}$ sont des chemins pondérés condensés, ou leurs prochaines extensions le seront.

A partir d'une mesure de support anti-monotone des chemins que nous avons définie, nos travaux ont abouti à la première définition d'une représentation condensée (exacte) dans le cas d'un unique graphe. Sur la base de cette représentation condensée, nous avons proposé un algorithme efficace pour extraire des motifs fréquents (les chemins pondérés) dans un unique DAG attribué [Sanhes et collab., 2013a].

3.7 Recherche de motifs contraints par un modèle

Pour assister la découverte de connaissances à partir de données, de nombreuses techniques de calcul de motifs ont été proposées. L'un des verrous à leurs dissémination est que nombre des motifs extraits apparaissent triviaux et/ou inintéressants au regard de la connaissance du domaine et des experts.

Travailler à l'intégration de la connaissance des experts dans les processus de fouille n'est pas nouveau. Cette connaissance est souvent exprimée sous la forme de règles/contraintes expertes (e.g., de la forme si ... alors ...) définies manuellement. Ce type d'explicitation est difficile à obtenir et ne représente que très partiellement la connaissance du domaine ; en pratique, elle reste limitée à quelques règles basiques. Très tôt, des algorithmes de fouille ont exploité des taxinomies ou hiérarchies explicitées (par exemple, sur les attributs) pour fournir des motifs plus pertinents. En constatant que dans de nombreux contextes de science des données, les experts ont souvent capitalisé une partie de leur connaissance dans des modèles mathématiques, l'originalité de notre approche a consisté à exploiter de tels modèles pour dériver de nouvelles contraintes pouvant être intégrées aux calculs des motifs pour améliorer la pertinence des extractions. Plus précisément, nous avons ciblé un domaine de motif de type « itemsets » et nous nous sommes intéressé aux nombreux modèles mathématiques qui prennent la forme de fonctions à plusieurs variables.

Une approche simple permettant d'exploiter la représentation des connaissances d'un domaine est d'en dériver des contraintes primitives à prendre en compte lors des extractions. Nous avons proposé de définir des contraintes dérivées des conditions des modèles à appliquer aux motifs plutôt que des contraintes définies « à la main ». Ces contraintes représentent bien plus que de simples règles « si ... alors ... » car un seul modèle expert peut condenser l'information sur un nombre considérable de telles règles. Rappelons que les contraintes dérivables des modèles dont nous parlons seront utilisées conjointement avec celles qui auront été spécifiées par les analystes lorsqu'ils définissent l'intérêt a priori des motifs.

Différents types de contraintes peuvent être dérivés en fonction des données, des modèles utilisés, mais aussi du problème étudié. Dans le cadre de notre étude, nous nous sommes focalisés plus particulièrement sur une contrainte qui apparaît similaire à une contrainte de fréquence minimale même si ses propriétés sont très différentes. Le cadre d'application de notre étude concernait l'érosion des sols. Nous avons pris en compte des modèles quantitatifs fondés sur des propriétés physiques et calibrés à partir des données expérimentales. Par exemple, les modèles WEPP (Water Erosion Prediction Project) et RMMF (Revised Morgan-Morgan Finney) qui sont basés sur plusieurs modèles physiques non linéaires et non polynomiaux.

Tout comme la contrainte de fréquence minimale, nous avons pu définir une contrainte filtrant les motifs dont la valeur par le modèle f est supérieure ou égale à un seuil. En fonction de ce que modélise f, cette contrainte a différentes sémantiques. Par exemple, si f estime la perte en sol (en kg/m^2 par an) comme c'est le cas dans le modèle RMMF, la contrainte permet de ne conserver que les motifs susceptibles de représenter une perte en sol supérieure à une certaine quantité. En l'absence de « vérité terrain » (i.e., de mesures sur la perte en sol), cette contrainte permet de mettre en avant si de telles pertes risquent d'être fréquentes dans la zone d'étude et dans quelles situations (grâce aux valeurs des autres paramètres

environnementaux décrits par les motif extraits). Elle peut également permettre d'identifier (si les motifs sont fréquents) à quels autres facteurs, non couverts par le modèle, ces pertes en sols sont fréquemment liées dans les données. En présence de « vérité terrain », cette contrainte permet de comparer la prévision du modèle des experts à la réalité des données collectées. Les motifs confirmant le modèle sont intéressants car ils sont doublement validés par la vérité terrain et la connaissance du domaine (i.e., le modèle expert). De plus, les items additionnels du motif peuvent venir compléter les explications du modèle. Les motifs contredisant le modèle des experts sont tout aussi intéressants car ils permettent de mettre en avant certaines spécificités non prises en compte dans les modèles experts utilisés, déterminant donc des possibilités de révision.

Nous avons proposé d'utiliser et d'exploiter de tels modèles pour dériver des contraintes utilisables au cours des processus de fouille et ainsi améliorer la pertinence des motifs calculés tout en gagnant en performances. Cela a nécessité de définir des méthodes permettant de calculer la valeur des itemsets dans un modèle. Dans le cas général, cette opération n'est pas triviale est soulève de nombreux problèmes.

Nous avons défini une méthode permettant d'effectuer la recherche des motifs sous contrainte d'un modèle. Nous avons également étudié certaines propriétés des prédicats et contraintes afin de les exploiter pour optimiser les calculs de motifs. Nous avons proposé des solutions dans les cas où des variables du modèle ne sont pas présentes dans les données et dans les cas où les domaines de valeurs sont différents dans les données et dans le modèle. Un effort particulier a été consacré aux itemsets différents qui sont associés à une même règle du modèle car cela permet d'apporter des compléments à la règle initiale.

Nous avons montré que la prise en compte de contraintes provenant de modèles mathématique permet de mieux cibler l'analyse, tout en améliorant les performances grâce à des propriétés des modèles. Nous avons ainsi obtenu des motifs plus pertinents, venant compléter ou contredire la connaissance experte sur les phénomènes étudiés [Flouvat et collab., 2014a].

3.8 Applications

3.8.1 Analyse de l'érosion des sols

Dans le cadre du projet FOSTER (contrat ANR-2010-COSI-012 FOSTER), qui a pour objectif de proposer aux géologues un processus complet permettant de les aider à modéliser et à prédire les phénomènes érosifs, nous avons proposé une approche générique consistant à représenter les données disponibles par des graphes (dynamiques) attribués et à les analyser grâce aux algorithmes que nous avons développés.

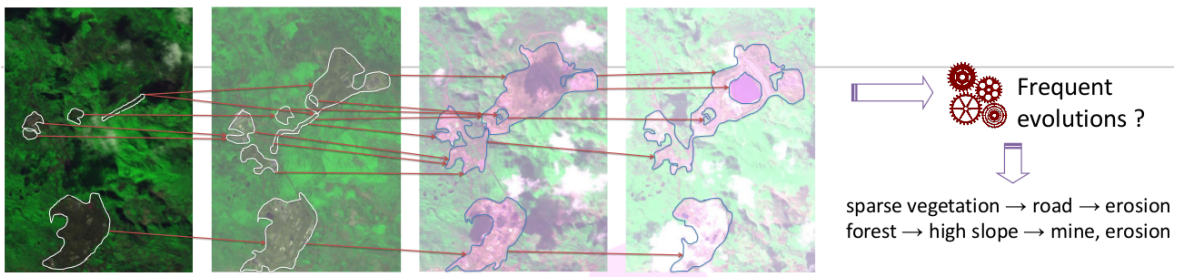

Figure 3.4 – Exemple de fouille de motifs sur des images satellite
Le graphe attribué qui est la source de la fouille de données est construit à partir des évolutions de terrain identifiées sur des images satellite représentant un même espace géographique à des temps différents. Dans ces graphes, chaque sommet représente un objet géologique avec ses caractéristiques et les arcs reliant deux sommets représentent l'évolution d'un objet dans le temps.

Les graphes attribués ont été construits à partir de l'analyse des évolutions de terrain identifiées en analysant des successions d'images satellites. Dans ces graphes, chaque sommet

représente un objet géologique avec ses caractéristiques et les arcs reliant deux sommets représentent l'évolution d'un objet dans le temps (figure 3.4). Le graphe obtenu, qui est orienté, acyclique et attribué, représente ainsi l'évolution des objets spatiaux dans le temps.

Nous avons utilisé les algorithmes de fouille de sous-graphes attribués fréquents et d'extraction de chemins condensés pour fouiller ces structures. Nous avons identifié des motifs topologiques qui sont des ensembles d'attributs des sommets fortement corrélés dans le graphe. Nous avons également proposé plusieurs mesures d'intérêt pour évaluer la pertinence de ces motifs. Nous avons notamment défini, en collaboration avec des géologues, une mesure d'intérêt qui calcule la corrélation entre la structure du graphe et le motif. Ces développements ont été utilisés dans la suite de logiciels de suivi environnemental développés par la société Blue Cham (http ://www.bluecham.net/).

3.8.2 Étude de la propagation de la dengue

L'étude avait pour but d'analyser un ensemble de données et d'identifier des motifs caractéristiques d'un pic épidémiologique de dengue. Le jeu de données étudié représentait l'évolution hebdomadaire d'un ensemble de données entomologiques, météorologiques, médicales et d'urbanisme mesurées sur la ville de Nouméa et sa banlieue proche de février 1998 à septembre 2004 [2]. L'espace géographique étudié a été divisé en 32 zones et les données (qui sont toutes de type numérique) ont été discrétisées (figure 3.5(a)).

À partir de ces données, nous avons construit, pour chaque point temporel, un graphe attribué dans lequel chaque sommet représente un quartier avec ses caractéristiques et chaque arc identifie une relation de voisinage (figure 3.5(b)). L'ensemble des données est représenté par une succession de graphes attribués qui évoluent dans le temps (figure 3.5(c)). Comme nous nous intéressions aux facteurs qui sont susceptibles d'indiquer le début d'une phase épidémique, nous avons construit un arbre attribué à partir de chaque zone dans laquelle au moins 4 cas de dengue ont été reportés pour une semaine donnée. Cette zone constitue la racine d'un arbre attribué et les caractéristiques de la zone représentent les attributs associés à la racine (figure 3.5(d)). Ces embryons d'arbres, composés d'un sommet unique, ont ensuite été étendus pour refléter l'évolution des zones voisines au temps précédent (figure 3.5(e)). Récursivement, nous avons généré, pour chacune des feuilles d'un arbre qui contient les caractéristiques d'une zone à un temps t, plusieurs branches, qui lient cette feuille aux sommets représentant les caractéristiques des zones précédentes au temps $t-1$ (figure 3.5(f)).

Les arbres de notre jeu de données sont de profondeur 5, c'est à dire qu'ils représentent l'évolution des caractéristiques des zones voisines d'un quartier infecté sur une période de 4 semaines. Au final, nous avons obtenu un jeu de données composé de 181 arbres attribués dont chaque sommet est associé à un ensemble de 15 attributs parmi 46.

Afin de mettre en évidence des motifs qui apparaissent souvent avant un pic de dengue, nous avons utilisé notre programme de recherche de motifs fermés fréquents dans les arbres attribués pour analyser le jeu de données avec un support de 70 %. Nous avons effectué la fouille en mode enchâssé avec un intervalle de 3 ce qui a permis d'identifier des successions d'événements n'ayant pas exactement la même durée. Nous avons extrait du jeu de données « dengue » 9265 motifs fermés présents dans plus de 70 % des arbres.

Pour évaluer la pertinence des motifs extraits, nous avons construit un autre jeu de données, selon la même procédure, pour représenter l'évolution précédent un point spatio-temporel pour lequel aucun cas de dengue n'a été reporté. Ce jeu de données, que nous avons appelé « no-dengue », contient 9087 arbres attribués. En calculant la fréquence des motifs survenant fréquemment avant un pic de dengue dans le jeu de données « no-dengue », nous avons identifié 8 motifs qui sont présents dans moins de 30 % des arbres « no-dengue » et qui pourraient donc être caractéristiques des pics de dengue.

2. Les données nous ont été fournies par la ville de Nouméa, l'institut Pasteur, la DDASS et l'antenne de Météo France en Nouvelle Calédonie

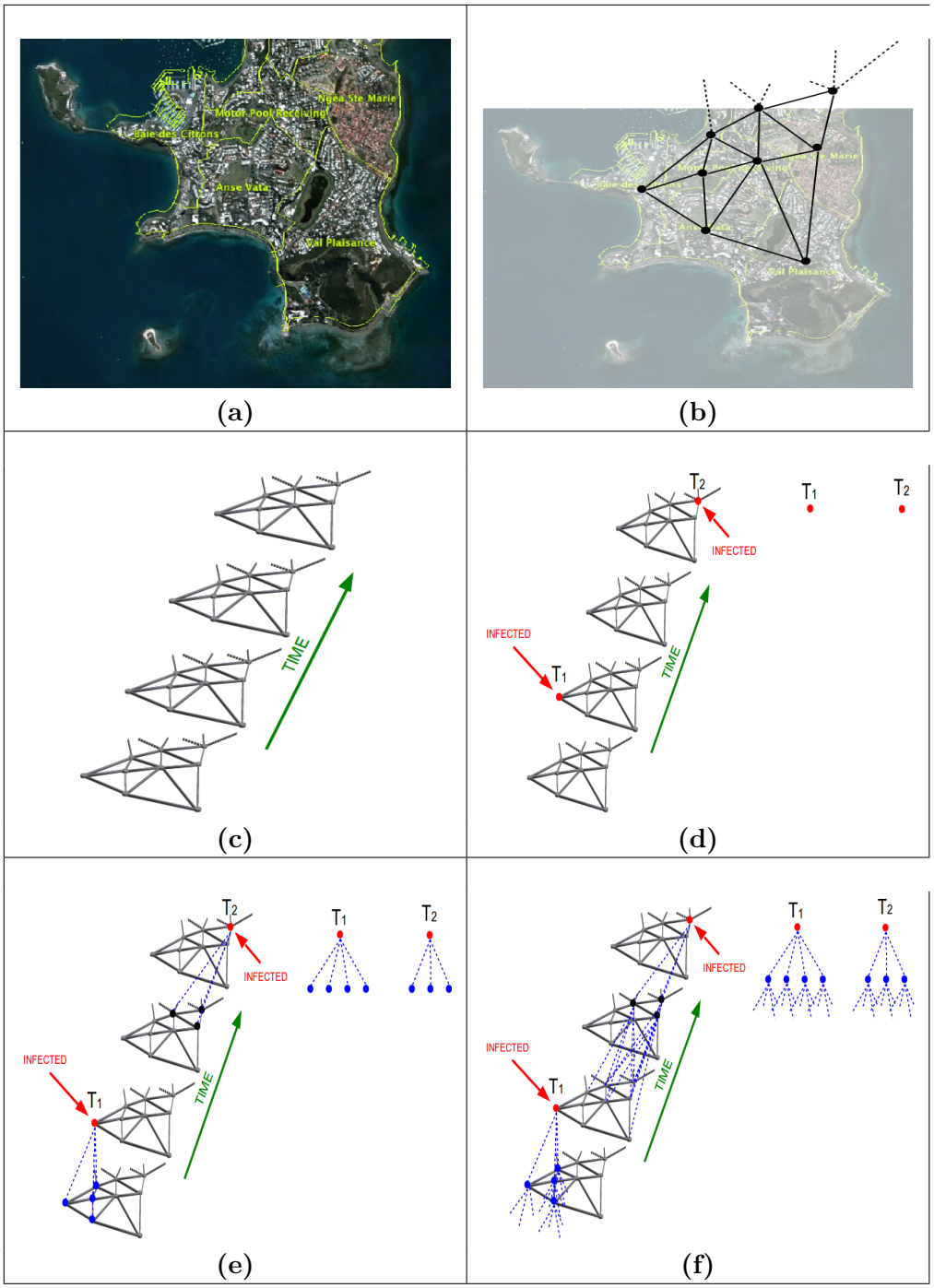

Figure 3.5 – Principe de création d'une forêt d'arbres attribués à partir des données dengue
(a) la zone géographique est divisée en quartiers pour lesquels sont disponibles des données concernant la démographie, l'entomologie, la météorologie, l'urbanisme et la santé. (b) la représentation des données à un temps t s'effectue sous la forme d'un graphe attribué. (c) l'ensemble des données est représenté sous la forme d'une succession de graphes attribués. (d) les racines des arbres attribuées correspondent aux quartiers infectés. (e) les sommets situés juste sous la racine représentent les quartiers voisins au temps $t-1$. (f) les sommets situés au $n^{ème}$ niveau représentent les quartiers voisins au temps $t-n$.

Malheureusement, les motifs trouvés ne révèlent pas de nouvelles informations concernant les signes avant coureurs d'une épidémie. Cependant, ils confirment ce qui est déjà connu dans ce domaine. Par exemple, une température élevée associée avec un indice de risque entomologique élevé suivi par une forte humidité, également, un indice de risque entomologique élevé associé à une température élevée suivi par un vent d'Est, ou encore un faible rayonnement global suivi d'une grande humidité. Une combinaison des facteurs suivants est également annonciatrice de cas de dengue : indice de risque entomologique élevé, faibles précipitations, températures élevées et ensoleillement élevé de 2 heures à 14 heures. Le détail des résultats obtenus sur ce jeu de données a été publié en 2015 [Pasquier et collab., 2016]. La recherche de chemins condensés fréquents a également été appliquée sur ces données et les résultats obtenus reportés par Sanhes et collab. [2013a].

Chapitre 4

Approches informatiques pour l'étude fonctionnelle des ARN

Sommaire

4.1 Introduction

Depuis une vingtaine d'années, le dogme central de la biologie moléculaire (ADN ⇒ ARN ⇒ protéine) est secoué par les découvertes de plus en plus nombreuses de brins d'ARN qui ne codent pas pour des protéines (ARN non codant (ARNnc)) mais semblent jouer un rôle critique dans de nombreux processus physiologiques. Les dérégulations de ces ARNnc semblent aussi être étroitement liées au développement et la progression de diverses maladies humaines, y compris le cancer. D'autres molécules d'ARN sont encore moins étudiées que les ARNnc. Je pense aux brins d'ARN qui sont produits lors de la dégradation des ARNm. Ces bouts d'ARNm, qui ne codent plus pour des protéines pourraient néanmoins avoir des fonctions auxiliaires qui seraient mise en œuvre lorsque la production des protéines est terminée.

Depuis mon arrivée à l'I3S, j'ai initié des recherches centrée sur le développement ou l'utilisation de méthodes de data mining/apprentissage automatique pour étudier les fonctions des ARN non codants. Sous le terme d'ARN non codant, je regroupe les ARNnc, les miARN mais aussi les molécules résultant de la dégradation des ARNm. Je travaille sur plusieurs axes : la prédiction des associations entre micro ARN (miARN) et maladies et l'étude des ARN (codants ou non) qui sont susceptibles de s'apparier avec des zones particulières de d'ADN pour former des structures triple brins.

4.2 Prédiction de l'implication des microARN dans les maladies

Les miARN forment une classe d'ARN de longueur comprise, en général, entre 19 et 24 nucléotides, qui peuvent réguler l'expression des gènes au niveau post-transcriptionnel par appariement avec les régions non traduites des ARN messagers (ARNm) cibles. Les miARN pourraient réguler entre une dizaine et plusieurs milliers de gènes et chaque gène cible peut aussi être régulé par des centaines de miARN. Ces interactions miARN-ARNm représentent un important réseau de régulation post-transcriptionnelle qui joue un rôle critique dans de nombreux processus physiologiques tels que le développement, l'apoptose ou la différenciation cellulaire. Les dérégulations des miARN sont également étroitement liées au développement et à la progression de diverses maladies humaines, y compris le cancer. L'identification de nouveaux miARN associés à des maladies contribue, par conséquent, à une meilleure connaissance de la pathogenèse des maladies. Ces nouvelles associations entre miRNA et maladies pourraient être utilisées pour le diagnostic précoce de maladies ou même la thérapie. Ces toutes dernières années, plusieurs méthodes bio-informatiques ont été présentées pour prédire de nouvelles associations entre miARN et maladies. Les approches proposées exploitent plusieurs hypothèses reposant sur des connaissances disponibles actuellement dans le domaine. Nous pouvons distinguer les méthodes qui utilisent les similarités fonctionnelles entre les ARNm cibles, les méthodes qui utilisent les similarités entre les maladies, les méthodes qui utilisent des systèmes de recommandation et les méthodes qui utilisent l'apprentissage automatique. Au vu des travaux existants nous pouvons faire plusieurs constatations :

— le point faible des méthodes qui reposent sur les similarités des ARNm cibles vient du peu de données fiables disponibles et des faibles performances des algorithmes de prédiction,

— plusieurs méthodes utilisent les similarités connues entre maladies en les mettant en relation avec les similarités fonctionnelles des ARNm cibles qui elles, sont calculées en fonction des similarités entre maladies,

— les meilleurs méthodes testées nécessitent de connaître un nombre relativement important de miARN associés à une maladie (au moins 100) pour pouvoir en prédire, de manière fiable, de nouveaux,

— les maladies peu connues, qui n'ont peu ou aucun miARN associé ne sont pas traités par les méthodes,

— beaucoup d'informations, connues des biologistes, ne sont pas du tout utilisées dans les programmes existants. On pense bien sûr à toutes les données concernant les miARN qui sont contenues dans la littérature scientifique mais aussi le fait, par exemple, que les miARN soient organisés en clusters qui agissent de concert sur le réseau de régulation des gènes ou qu'ils puissent être catégorisés en familles ayant des fonctionnalités très proches.

Les remarques évoquées ci-dessus permettent de proposer les bases d'une méthode plus efficace :

— combiner plusieurs sources différentes de données devrait permettre d'améliorer les résultats en étant moins sensible aux biais ou erreurs provenant d'une source de données unique,

— diverses sources de données, qui ne sont pas utilisées dans les méthodes existantes devraient pouvoir être utilisées,

— la méthode doit tirer profit des similarités contenues dans les données.

En collaboration avec la startup Biomanda, j'ai travaillé à l'élaboration d'une nouvelle méthode permettant de prédire les associations entre miARN et maladies à partir de multiples sources de données.

4.2.1 Idée de base

Nous regroupons toutes les données disponibles sur les miARN dans un tableau dans lequel chaque ligne représente un miARN, et chaque colonne, un type de données associé. Nous pouvons ainsi représenter les associations connues entre miARN et maladies, les gènes ciblés par les miARN, les miARN physiquement proches sur le génome, les familles d'appartenance, etc. Nous pouvons même faire figurer dans la même matrice les termes utilisés dans les articles scientifiques traitant du miRNA (pas forcément de l'association entre le miARN et une maladie). Les données de ce tableau peuvent être analysées avec des algorithmes classiques du domaine de l'apprentissage automatique (SVM, chaines de markov, règles d'associations, etc). Nous pouvons aussi remarquer que le tableau regroupant toutes les données ressemble fort à une matrice termes-documents utilisée en traitement statistique de la langue naturelle. Ici, les lignes ne représenteraient pas un document, mais un miARN, et les colonnes contiendraient diverses données concernant les miARNs.

Dans cette recherche, nous avons étudié l'utilisation de la sémantique distributionnelle pour analyser les données associées aux miARN. Nous avons émis l'hypothèse que la sémantique distributionnelle peut être utilisée pour révéler de nouvelles informations attachées aux miARN. L'idée de base de la sémantique distributionnelle, utilisée en linguistique, est que la signification d'un mot peut être déduite uniquement du contexte dans lequel ce mot est utilisé. Dans notre étude, on peut considérer que les miARN représentent des mots et que, par conséquent, les données associées aux miARN jouent le rôle du contexte. En gardant l'analogie, l'objectif est alors d'utiliser les informations disponibles sur les miARN dans leur ensemble (contexte) pour en déduire une nouvelle connaissance sur les miARN (mots).

Un modèle vectoriel est un modèle algébrique permettant de représenter des objets sous la forme de vecteurs. Le principe est que chaque composant d'un vecteur est représenté par une valeur (poids) qui doit quantifier l'importance d'une caractéristique de l'objet modélisé. Les études portant sur l'analyse de textes utilisent des techniques telles que l'analyse sémantique latente (Latent Semantic Analysis (LSA)) pour traiter des vecteurs d'un espace vectoriel. Pour les données textuelles, il existe de nombreuses façons de calculer les poids. Parmi ces méthodes, des variations de la méthode Term Frequency-Inverse Document Frequency (TF-IDF) [Turney et Pantel, 2010], qui consiste à diviser la fréquence d'un terme dans une ligne par la fréquence totale de ce terme dans la matrice, sont fréquemment utilisées.

Dans notre scénario, les vecteurs représentant les objets à analyser (miARN) ne sont pas homogènes. En plus des associations entre les miARN et les mots extraits des documents texte, nous devons aussi représenter les relations entre les miARN et les maladies, les cibles, les familles ou les miARN voisins. Les liens entre les miARN et les maladies, les cibles ou les familles se composent d'informations binaires qui dépendent de l'existence ou non d'une association. Les relations entre miARN voisins sont exprimés par des nombres qui représentent une distance sur un chromosome en paires de bases.

Notre approche consiste à représenter les informations sur les miARN, les maladies et les facteurs environnementaux dans un espace vectoriel de grande dimension et de définir les associations entre miARN et maladie ou environnement en termes de similarités de vecteurs.

4.2.2 Collecte des données

Nous avons téléchargé les associations miRNA-maladie vérifiés expérimentalement à partir de la base de données miRNA-maladie HMDD v2.0 de Juin, 14 2014 25. La plupart des miARN sont associés à des maladies répertoriées dans le vocabulaire controlé Medical Subject Headings (MeSH). Pour les maladies qui ne sont pas incluses dans MeSH, nous avons associé manuellement leurs noms aux termes MeSH les plus pertinents. Après cette étape de prétraitement, l'ensemble des données représente 10737 associations entre 559 miARN et 369 maladies. Les cibles des miARN ont été téléchargées à partir de miRTarBase, une base de données d'interactions miARN-ARNm validés expérimentalement. L'ensemble contient 48030 associations entre 1886 miARN et 11985 cibles. Pour chaque miARN, nous avons également récupéré, dans la base de données miRBase, sa famille, sa localisation génomique et les résumés des études de recherche qui lui sont associés.

4.2.3 Création des matrices

Les données collectées représentent, pour chaque miARN, les maladies connues qui lui sont associées, les mARN qu'il cible, sa famille, la distance avec les miARN voisins et des informations complémentaires au format texte. Ces données peuvent être représentées par des matrices distinctes, chacune contenant un type d'association spécifique (figure 4.1, partie a). Dans toutes les matrices, les lignes représentent les miARN et les colonnes représentent des caractéristiques.

Les associations miRNA-cible sont stockées dans une matrice $MT = [x_{ij}]$ de taille $m \times t$, où m est le nombre de miARN, t est le nombre de différents gènes cibles, et x_{ij} est égal à 1 si le miARN à la position i cible le gène à la position j et 0 sinon.

Les associations miRNA-maladie sont stockées dans une matrice $MD = [x_{ij}]$ de dimension $m \times d$, où m est le nombre de miARN, d est le nombre de maladies différentes, et x_{ij} est égal à 1 si une association entre le miRNA à la position i et la maladie à la position j est présente dans les données, ou 0 si l'association n'est pas connue.

Dans la matrice d'associations miRNA-famille (matrice $MF = [x_{ij}]$), les colonnes représentent les familles de miRNA et MF_{ij} est égal à 1 si le miRNA à la position i appartient à famille à la position j, ou 0 sinon.

Dans la matrice d'associations miRNA-voisin (matrice $MN = [x_{ij}]$), les colonnes représentent des miRNA et mn_{ij} est égal à la distance génomique, en nombre de paires de bases séparant le miRNA à la position de ligne i et le miRNA à la position de colonne j.

Les associations entre les miARN et les mots utilisés dans les textes qui leurs sont associés sont stockées dans une matrice $MW = [x_{ij}]$ de dimension $m \times w$, où m est le nombre de miARN, w est le nombre de termes différents qui sont présents dans les documents associés, et x_{ij} est égal au nombre de fois que le mot à la position j est utilisé dans les descriptions textuelles du miRNA à la position i. En ce qui concerne les sources des descriptions en texte brut, nous pouvons choisir d'utiliser le résumé des articles associés aux miRNA dans PubMed (en effectuant des recherches avec le nom des miRNA), les résumés des articles référencés

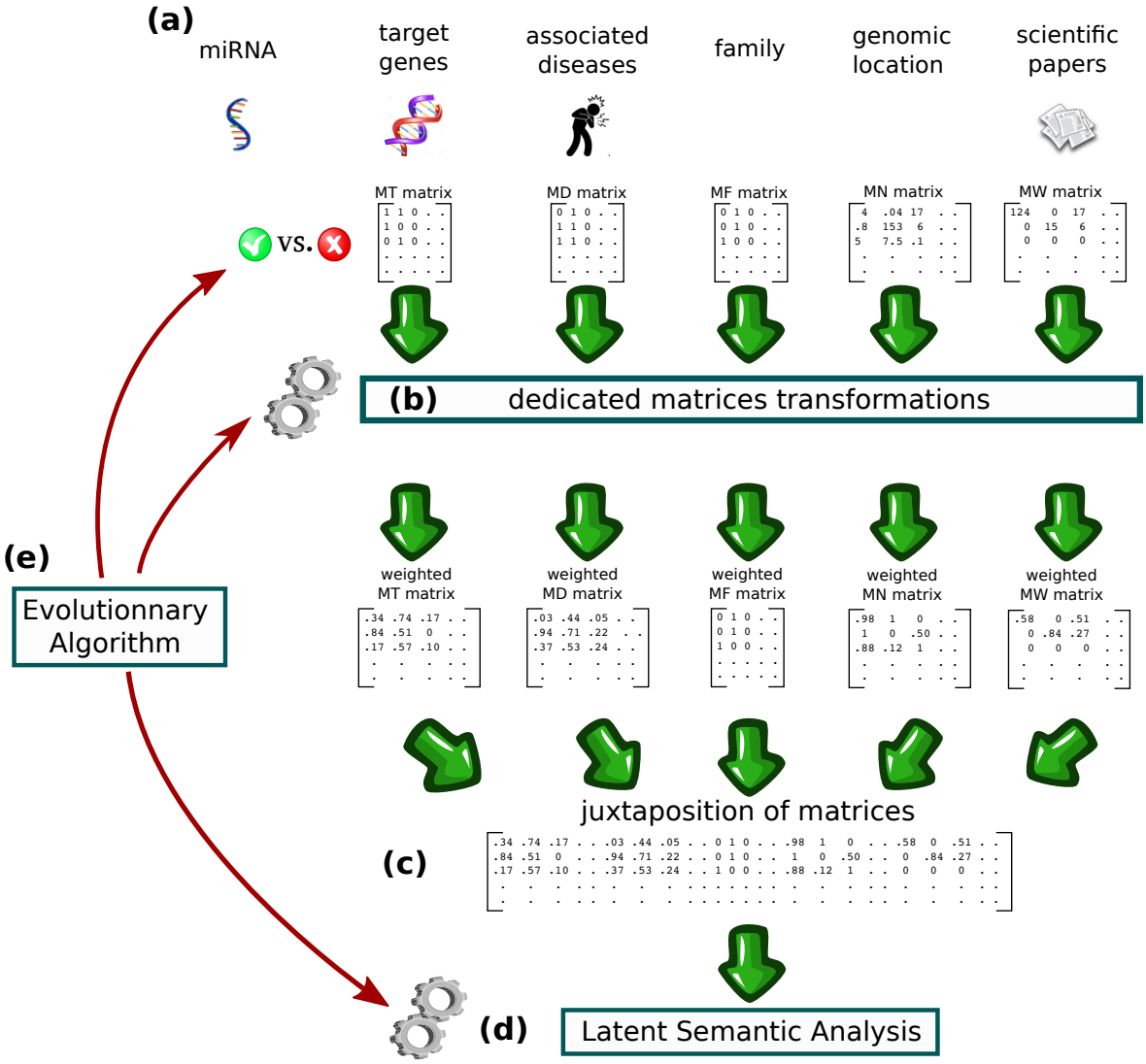

Figure 4.1 – Vue d'ensemble de la méthode MiRAI
a) les miARN sont caractérisés par plusieurs types de données qui sont stockées dans des matrices distinctes. b) chaque matrice est traitée par une méthode dédiée pour la transformer en une matrice pondérée dans laquelle l'importance d'une association entre un miRNA et une caractéristique est représentée par un nombre décimal. c) concatenation des matrices d) les similitudes et les dissemblances entre les miARN et les maladies sont mises en évidence par la méthode LSA. e) Des algorithmes évolutionnaires sont utilisées pour sélectionner les sources de données à utiliser, déterminer les transformations à appliquer sur les matrices, ajuster la taille de l'espace latent et choisir d'élargir ou non les termes utilisés pour les requêtes LSA.

dans la base de données MiRBase, les champs de description miRNA dans MiRBase ou toute combinaison de ces sources.

4.2.4 Pondération des associations

Lorsqu'une source de données est sélectionnée, nous avons la possibilité d'appliquer différentes stratégies de transformation des données figurant dans la matrice correspondante pour pondérer le poids des associations

Associations miRNA-maladies

Le lien entre un miARN et une maladie va être d'autant plus probable que le miARN en question est connu pour être impliqué dans des maladies proches. Pour tenir compte de cela, les poids des associations peuvent être mis à jour par la fonction suivante : $x_{ij} = max(simil(j,k) \times x_{ik})_{k=1}^{d}$ où $simil$ est une mesure de similarité qui représente la distance entre deux maladies dans l'ontologie MeSH. Elle peut être calculée avec la formule proposée par Lin [1998].

$$simil(x,y) = \frac{2 \times logP(D_0)}{logP(x) + logP(y)}$$

où $P(d)$ est la probabilité qu'une maladie choisie au hasard soit égale à d, et D_0 est la maladie la plus spécifique qui subsume à la fois x et y.

Au lieu d'utiliser une mesure unique $m \in [0,1]$ reflétant le poids de l'association entre un miARN et une maladie, nous pouvons aussi discrétiser cette valeur pour obtenir plusieurs indicateurs, chacun d'eux indiquant si le poids de l'association est au-dessus d'un seuil donné ou non. Par exemple, en utilisant des seuils de $1/3$ et $2/3$, on obtient trois plages de valeurs chevauchantes qui associent $m \in [0,1/3]$ à l'indicateur no_assoc, $m \in]1/3,1]$ à l'indicateur significant_assoc et $m \in]2/3,1]$ à l'indicateur high_assoc. Avec cette représentation, une mesure de similarité de $0,8$ sera associée aux annotations significant_assoc et high_assoc.

Associations avec les miARN voisins

Bandyopadhyay et collab. [2010] et Baskerville et Bartel [2005] ont mis en évidence une coexpression importante des miARN colocalisés sur le génome. Ils ont noté qu'au-delà d'une distance de 50k paires de base, la corrélation chute brusquement. Pour imiter cette corrélation dans l'ensemble de données, nous transformons une distance génomique en un poids, exprimant la corrélation entre deux miARN avec une fonction sigmoïde de la forme suivante :

$$w(dist) = 1 - \frac{1}{1 + e^{-k(dist - dist_0)}}$$

où $dist$ est la distance génomique entre deux miARN, $dist_0$ est le milieu du sigmoïde qui est fixé à 6×10^5 (une distance de 60kb est associé à une corrélation de $0,5$), et k est la pente de la courbe qui est fixée à 2×10^{-4}.

Associations miARN-cibles

La matrice MT peut être représentée par un graphe biparti qui relie les éléments d'un ensemble de miARN aux éléments d'un ensemble de cibles. Déduire les relations entre les éléments du même ensemble est généralement effectué à l'aide d'une opération dite de projection. Celle-ci consiste à se ramener à un graphe ne comportant qu'un seul des deux ensembles de sommets du graphe initial. Pour cela, on conserve tous les sommets de l'ensemble en question, puis on décide qu'un lien entre deux sommets existe dans le graphe projeté selon une métrique relative au graphe biparti ; par exemple le fait qu'ils partagent plus de k voisins. L'opération de projection est fréquemment utilisée pour construire des graphes de similarités

[Bhajun et collab., 2015; Le, 2015; Lu et collab., 2008]. Cependant, cette transformation est quantitativement biaisée et produit une perte d'information [Li et collab., 2005].

Pour surmonter ce problème, Zhou et collab. [2007] ont conçu une méthode de pondération qui combine des informations partielles données par les projections pour extraire des informations cachées dans le réseau. La méthode permet au final de moduler les poids des liens du graphe biparti en tenant compte des proximités entre les sommets composant chaque partition. Zhou et collab. [2007] ont appelé cette méthode Network-Based Inference (BNI) et nous avons le choix de l'utiliser ou non. Des explications plus détaillées du fonctionnement de la méthode dans le cadre de l'association entre les miARN et leurs cibles sont données par Pasquier et Gardès [2016].

Associations miRNA-termes

Parmi les nombreux schémas de pondération existants pour ternir compte de l'importance des termes dans des phrases, nous nous en tenons à la mesure populaire TF-IDF, qui consiste à diviser la fréquence du terme dans un corpus par la fréquence de ce terme dans le document considéré [Turney et Pantel, 2010].

4.2.5 Application de l'Analyse Sémantique Latente

Combinaison des matrices

Les matrices contiennent différents types de données qui sont représentés par des valeurs décimales qui vont de 0 à 1. Une grande matrice X qui rassemble toutes les données disponibles concernant les miARN est tout simplement construite par concaténation des matrices existantes. $X = [MD_{(m\times d)}MN_{(m\times n)}MT_{(m\times t)}MW_{(m\times w)}MF_{(m\times f)}]$

Réduction de dimensionnalité

L'objectif de la réduction de dimensionnalité est de convertir les données de haute dimensionnalité en données de dimensionnalité inférieure, ce qui a pour effet de rapprocher les éléments semblables et d'éloigner ceux qui sont les plus dissemblables.

La décomposition en valeurs singulières (Singular Value Decomposition (SVD)), qui est la méthode de réduction de dimensionnalité utilisée dans LSA, est une technique de décomposition de la matrice qui peut être utilisée pour transformer les données dans un espace de grande dimension en un espace avec moins de dimensions. Avec SVD, la matrice d'origine X est décomposé comme un facteur de trois autres matrices, U, Σ and V, comme suit :

$$X = U\Sigma V^T$$

où U est une matrice $m \times m$, Σ est une matrice diagonale $m \times n$ avec des nombres réels non négatifs sur la diagonale, et V^T, une matrice $n \times n$, désigne la transposée de V.

Il est souvent utile d'approximer X en utilisant uniquement un nombre fixé, r, de valeurs singulières (avec $r < min(m, n)$), de manière à obtenir $X = U_r\Sigma_r V_r^T + E$, où E est une erreur ou matrice résiduelle, U_r est une matrice $m \times r$, Σ_r est une matrice diagonale $r \times r$ et V_r est une matrice $n \times r$. La matrice $X_r = U_r\Sigma_r V_r^T$ est la matrice de rang r qui se rapproche le plus la matrice originale X. Le bon choix de la dimensionnalité est important. Cependant, il n'y a aucune méthode générale pour trouver la dimension optimale.

Recherche sur la matrice

Dans X_r, les maladies et les miARN sont représentés par des vecteurs. En appliquant l'approche classique qui est utilisée pour l'indexation des documents, la proximité de deux vecteurs est mesurée par la distance cosinus. Pour obtenir une liste ordonnée des miARN qui sont liés à une maladie d, les requêtes sont effectuées dans l'espace latent en priorisant les vecteurs de miARN selon leur distance par rapport au vecteur de d.

Évaluation de la performances des prédictions

Pour évaluer la capacité de notre méthode à prédire les associations miARN-maladies, nous avons effectué une validation croisée cinq fois (five-fold cross validation). Pour une maladie spécifique d, l'ensemble de données est divisé aléatoirement en cinq sous-ensembles de taille égale. Quatre des cinq sous-ensembles sont utilisés pour créer l'espace latent tandis que le dernier sous-ensemble est utilisé pour l'interroger et tester le modèle. Dans le sous-ensemble de test, toutes les associations entre miARN et d sont supprimées. En outre, pour être certain que les liens entre miARN et maladies sont mis en évidence grâce de leur liens sous-jacents et non pas à cause de certaines informations trouvées dans la littérature, toutes les données provenant de la littérature sont également supprimées. Le processus de validation croisée est ensuite répété cinq fois, avec chacun des cinq sous-ensembles utilisé exactement une fois en tant que données de validation.

L'espace latent est alors interrogé avec la maladie d pour obtenir une liste ordonnée des miARN les plus liés à cette maladie. Plus les miARN qui sont effectivement associés à d sont trouvés en tête de liste, meilleure est la performance.

Si un miARN associé à d apparaît dans la liste à un rang supérieur à un seuil θ donné, la prédiction est correcte, tout comme un miARN qui est pas associé à d et qui apparaît à un rang inférieur à θ. Cependant, comme le choix de θ a un impact significatif sur la mesure de la performance, nous avons utilisé la courbe Receiver Operating Characteristic (ROC), qui combine en un seul graphique, le compromis entre le taux de faux positifs (False Positive Rate (FPR)) et le taux de vrais positifs (True Positive Rate (TPR)) sur toute la plage de θ, où :

$$TPR = \frac{TP}{(TP + FN)} \qquad FPR = \frac{FP}{(FP + TN)}$$

avec TP, TN, FP, et FN étant respectivement le total des vrais positifs, vrais négatifs, faux positifs et faux négatifs. Une unique mesure de performance est obtenue en calculant l'aire sous la courbe ROC (Area Under the ROC Curve (AUC)) (une AUC de 1 reflète une classification parfaite alors qu'une AUC de $0,5$ indique un classement aléatoire [Lasko et collab., 2005]).

4.2.6 Paramétrisation de MiRAI

La méthode MiRAI dépend de nombreux paramètres. Tout d'abord au niveau des données à partir desquelles est construit le modèle. L'utilisation des informations concernant les liens connus entre les miARN et les maladies est obligatoire, mais les autres données peuvent être utilisées ou non. En considérant que les données textuelles peuvent être choisies parmi 3 sources différentes, nous avons un total de 6 choix individuels, selon que les données correspondantes sont utilisées ou non. D'autres facteurs influent également sur l'efficacité de la méthode comme le choix des différentes techniques utilisées pour pondérer les valeurs des matrices ou l'utilisation ou non de la subsomption entre maladies. Tenir compte de la similarité entre les maladies semble être judicieux. Cependant, il existe de nombreuses manières de calculer les similarités et il se pose la question de l'utilisation concrète que l'on fait de la mesure obtenue. Faut-il utiliser directement la mesure dans la matrice miARN-maladie ou faut-il fixer un seul à partir duquel deux maladies sont considérées comme similaires. Nous pouvons également décider d'utiliser plusieurs bornes fixant les seuils à partir desquels deux maladies sont considérées très similaires, moyennement similaires, etc.

La prise en compte de tous ces facteurs représente un nombre énorme de combinaisons. Il n'est pas possible de les tester toutes, même en automatisant le processus car un seul cycle, comprenant la construction du modèle puis l'évaluation, peut nécessiter jusqu'à 4 heures de calcul.

Dans un premier temps, nous avons choisi de manière manuelle et très arbitraire les différents paramètres. Dans une seconde étude, nous avons utilisé une stratégie d'optimisa-

tion stochastique basée sur les algorithmes évolutionnaires pour déterminer de manière plus rationnelle les paramétrages de MiRAI.

Paramétrisation manuelle de MiRAI

Lors de nos premiers essais avec différentes combinaisons de paramètres, nous avons rapidement constaté qu'utiliser toutes les données disponibles n'était pas systématiquement corrélé avec la qualité des prédictions obtenues. De même qu'utiliser un espace latent de grande dimension ne permettait pas forcément d'obtenir les meilleurs résultats. L'utilisation de la méthode NBI sur la matrice miRNA-cibles et de la méthode TF-IDF sur les matrices à contenu textuel s'est révélé la plupart du temps avoir un effet positif sur la qualité des prédictions. Il en était de même pour ce qui concerne la prise en compte de la similarité entre les maladies.

Nous avons déterminé manuellement que les meilleures performances de la méthode étaient obtenues en combinant les matrices MT, MD, MF, MN et MW lorsque MW est construit avec les termes collectés dans les résumés des articles retrouvés sur PubMed. Nos expérimentations nous ont également montré que l'utilisation de la méthode NBI sur la matrice MT et de la méthode TF-IDF sur la matrice MW étaient bénéfiques. Nous avons également trouvé que la discrétisation des similarités entre maladies en utilisant les deux plages de valeurs suivantes donnaient de bons résultats : $[0,2-1]$ et $[0,9-1]$ Pour déterminer la taille optimale de l'espace latent, nous avons testé les performances des prédictions pour toutes les dimensions comprises entre 50 et 500 avec un pas de 50. Nous avons constaté que l'utilisation d'une dimension de 400 est le choix qui permettait d'obtenir de meilleurs résultats. Selon ce résultat, la décomposition en valeurs singulières est appliquée sur la matrice combinée, et seuls les 400 premiers vecteurs sont retenus ($r = 400$).

MiRAI a été testé sur les 83 maladies humaines de la base de données HMDD version 2.0 [Li et collab., 2014] qui sont associées à au moins 20 miARN.

La valeur AUC moyenne obtenue avec MiRAI est 0,867 avec un minimum de 0,624 pour Lupus vulgaris et un maximum de 0,987 pour la cardiomyopathie hypertrophique. Dans la figure 4.2, qui illustre la distribution univariée des scores AUC, on remarque que peu de prédictions (11 sur 82, 13%) sont inférieures à 0,8. Seuls deux maladies (Lupus vulgaris et Infection à adénovirus) obtiennent un score inférieur à 0,7. Les scores obtenus par notre méthode sont bons dans l'ensemble. De plus, il est probable que les performances réelles de notre méthode soient plus élevées car nous avons détecté un certain nombre de fausses associations dans la base de données HMDD que nous utilisons comme référence.

La méthode est décrite en détail dans l'article de Pasquier et Gardès [2016]. Une comparaison de cette première version de MiRAI avec d'autres méthodes existantes est présentée dans le tableau 4.1.

Paramétrisation de MiRAI par un algorithme évolutionnaire

Dans la première version de MiRAI, les valeurs des paramètres ont été trouvées manuellement par une méthode essai-erreur. Nous nous sommes ensuite intéressés à l'élaboration d'un algorithme évolutionnaire (EA) susceptible de déterminer, dans un délai raisonnable, un réglage satisfaisant sans avoir à évaluer toutes les configurations possibles. Afin de réduire le temps de calcul, les configurations ont été évaluées en parallèle sur une grille de calcul. Pour minimiser le nombre d'évaluations réelles, nous avons utilisé un modèle de substitution permettant de prédire l'intérêt de chaque configuration.

Comme résumé dans le Tableau 4.2, la méthode MiRAI est contrôlée par 12 paramètres binaires, un ensemble de nombres décimaux (les seuils de discrétisation) et un entier (la dimension de l'espace latent). L'objectif étant de trouver la meilleure combinaison de valeurs, tous les paramètres sont regroupés dans un vecteur unique. De plus, afin d'éviter de traiter différents types de paramètres (binaires, réels ou entiers) et de simplifier l'adaptation des

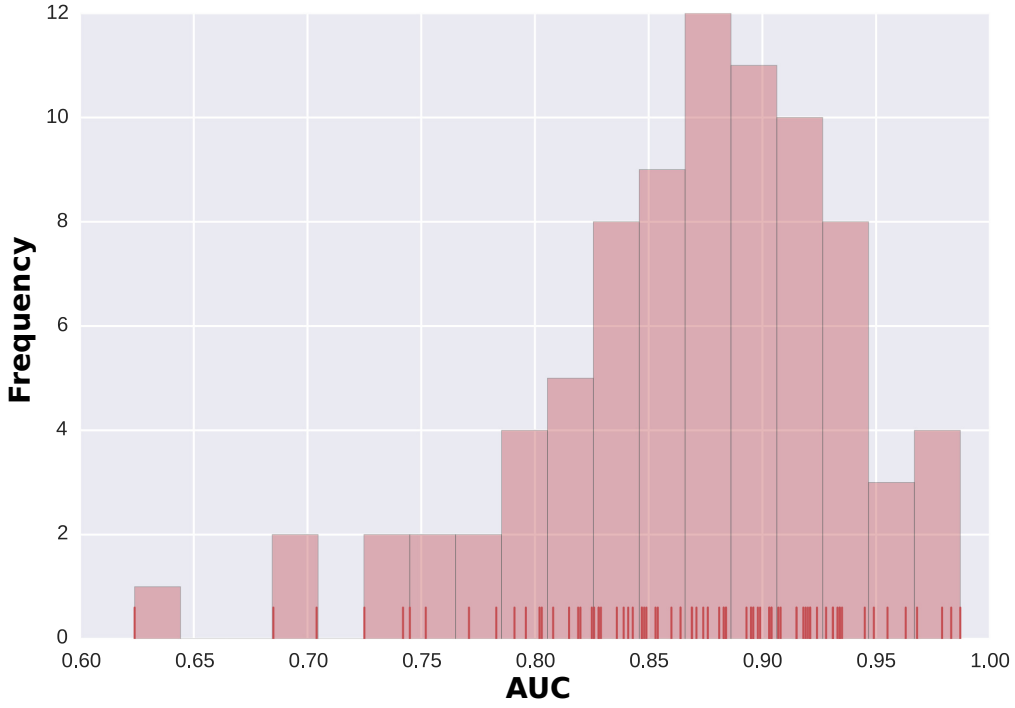

Figure 4.2 – Distribution des scores de MiRAI
Les petites lignes verticales affichées sur l'axe horizontal identifient les valeurs d'AUC obtenues pour chacune des 83 maladies. Les valeurs d'AUC ont été divisés en partitions de taille égale. Les rectangles verticaux représentent le nombre de cas pour lesquels la valeur AUC appartient à chaque partition.

techniques d'optimisation, nous avons décidé de les convertir tous en valeurs binaires. Pour ce faire, les paramètres entiers et réels ont été convertis en utilisant le codage Gray. En ce qui concerne les seuils de discrétisation utilisés pour quantifier les similarités entre maladies, nous avons décidé d'autoriser un maximum de 19 seuils et de prédéterminer leur valeur de 0,05 à 0,95 en utilisant un pas de 0,05. Les seuils sont donc représentés par 19 valeurs binaires, chacun correspondant à une valeur de seuil. La valeur binaire est 1 lorsque le seuil est utilisé pour la discrétisation, 0 sinon. Nous avons encodé la dimension de l'espace latent sur quatre bits pour englober toutes les dimensions entre 50 et 800 avec un pas de 50. Le nombre d stocké dans le vecteur est encodé avec le code de Gray. La dimension correspondante est obtenue avec $dim = 50(d+1)$. Au final, l'ensemble des paramètres de MiRAI est encodé par un vecteur binaire de 35 bits comme détaillé dans le Tableau 4.2 qui représentent un total de 2^{35} ou 5×10^{16} configurations différentes.

Les Algorithmes évolutionnaires (EA) sont des algorithmes stochastiques, inspirés de la théorie de l'évolution de Darwin, utilisés pour l'optimisation des problèmes [Siddique et Adeli, 2015]. Soit un problème P à maximiser (ou à minimiser), l'objectif est de trouver une solution x^* telle que $f(x^*) = max$ (ou min) $\{f(y)/y \in S\}$ avec S, l'ensemble de toutes les solutions possibles de P, $f : S \to \mathbb{R}$ est appelée *fonction objectif* ou *fitness* et indique la qualité de la solution y pour le problème P.

De nombreuses variantes des algorithmes évolutionnaires ont été proposées dans la littérature (algorithme génétique, stratégies d'évolution, programmation génétique, évolution différentielle...) mais elles partagent toutes une structure commune : un ensemble de solutions candidates, appelées *population*, est créé et le plus souvent initialisé de façon aléatoire. Ensuite, chaque solution de la population, appelée *individu*, est évaluée à l'aide de la fonction fitness f. En fonction de la valeur de fitness de chaque individu, un opérateur de sélection

Nom de la maladie	Chen 2013	HDMP 2013	RLSMDA 2014	MIDP 2015	Liu 2016	MiRAI 2016	MiRAI+EA 2017
Acute myeloid leukemia	0.716	0.858	0.853	**0.913**	0.871	0.895	0.906
Breast neoplasms	0.653	0.801	0.832	0.838	0.826	**0.864**	0.858
Colorectal neoplasms	0.662	0.802	0.831	0.845	0.833	0.864	**0.868**
Glioblastoma	0.607	0.700	0.714	0.786	0.839	**0.898**	0.872
Heart failure	0.761	0.77	0.738	0.821	0.812	0.796	**0.847**
Liver carcinoma	0.613	0.759	0.794	0.807	0.802	0.808	**0.825**
Lung neoplasms	0.606	0.835	0.855	0.876	0.925	0.904	**0.926**
Melanoma	0.642	0.790	0.807	0.837	0.834	0.849	**0.875**
Ovarian neoplasms	0.644	0.884	**0.909**	0.923	0.896	0.874	0.906
Pancreatic neoplasms	0.684	0.895	0.887	**0.945**	0.901	0.928	0.925
Prostatic neoplasms	0.629	0.854	0.841	**0.882**	0.842	0.871	0.872
Renal cell carcinoma	0.627	0.833	0.839	0.862	0.815	0.869	**0.882**
Squamous carcinoma	0.676	0.820	0.849	0.870	0.872	0.883	**0.888**
Stomach neoplasms	0.628	0.787	0.797	0.821	0.798	0.815	**0.848**
Bladder neoplasms	0.632	0.85	0.845	0.897	0.851	0.884	**0.900**
AUC MOYEN	0.652	0.816	0.826	0.862	0.848	0.867	**0.880**

Tableau 4.1 – Comparaison de 7 méthodes de prédiction d'associations miARN-maladies Les scores AUC de MiRAI configuré manuellement et à l'aide d'un algorithme évolutionnaire (MiRAI+EA) sont comparés avec les scores de 5 autres méthodes publiées récemment. Chen est la méthode décrite par Chen et Zhang [2013], HDMP est la méthode proposée par Xuan et collab. [2013], RLSMDA est présenté par Chen et Yan [2014], MIDP est détaillé par Xuan et collab. [2015], Liu est la méthode décrite par Liu et collab. [2016].

est appliqué pour choisir les individus qui sont autorisés à créer des descendants. Ceux-ci sont générés en utilisant des opérateurs génétiques comme le croisement ou la mutation, chacun appliqué avec une probabilité donnée. Ces opérateurs introduisent une variabilité dans l'espace de solutions permettant d'échapper à un optimum local. Chaque descendant est à son tour évalué en utilisant la fonction fitness f. Une itération de ce processus est appelée *génération* et est répétée jusqu'à ce qu'un critère donné soit atteint. Une très bonne description d'ensemble des algorithmes évolutionnaires a été proposée par Bartz-Beielstein et collab. [2014] où sont introduits de nombreux sous-domaines fondamentaux tels que les objectifs multiples, dynamiques, bruités ou comprenant des évaluations coûteuses comme celui auquel nous sommes confrontés. Les algorithmes évolutionnaires sont préférés aux optimisations déterministes lorsque $|S|$ est très grand, dans la mesure où ils permettent de trouver des solutions prometteuses dans un délai raisonnable.

L'évolution différentielle, proposée par Storn et Price [1997], s'est révélée plus efficace que d'autres types d'algorithmes évolutionnaires. Elle fonctionne de la manière suivante : pour chaque individu de la population, appelé *vecteur cible* et formalisé par x_i^t, un *vecteur mutant* μ_i^t associé à x_i^t est tout d'abord généré en additionnant la différence pondérée entre deux vecteurs choisis aléatoirement (*vecteurs paramètres* $x_{i_2}^t$ et $x_{i_3}^t$) à un troisième vecteur (*vecteur base* $x_{i_1}^t$) en utilisant l'équation suivante :

$$\mu_{ij}^t = x_{i_1 j}^t + F \cdot \left(x_{i_2 j}^t - x_{i_3 j}^t \right) \tag{4.1}$$

où $i \neq i_1 \neq i_2 \neq i_3$; i_1, i_2 et i_3 sont choisis de manière uniforme et aléatoire entre 1 et la taille de la population λ ; $F \in \mathbb{R}^+$ est un facteur d'échelle, contrôlant l'amplification de la variation différentielle et x_{ij}^t représente le $j-$ème gène du $i-$ème individu dans la population à la génération t. Dans une seconde étape, un descendant, appelé *vecteur d'essai* x_i^{t+1}, est

obtenu par croisement entre le vecteur mutant μ_i^t et le vecteur cible x_i^t en utilisant :

$$x_{ij}^{t+1} = \begin{cases} \mu_{ij}^t & \text{if } (rand \leq CR) \text{ or } j = rand(i) \\ x_{ij}^t & \text{otherwise} \end{cases} \tag{4.2}$$

où CR représente la probabilité de croisement dans l'intervalle $([0;1]$. $rand$ est une valeur aléatoire uniformément distribuée dans l'intervalle $[0;1]$; $rand(i)$ est un entier aléatoire dans l'intervalle $[1;N]$ où N est le nombre d'individus. Pour finir, le vecteur cible est remplacé par le meilleur entre le vecteur test et le vecteur cible.

À l'origine, l'évolution différentielle a été conçue pour opérer dans un espace continu, ce qui signifie que les x_i^t sont des vecteurs de valeurs décimales. En 2010, Wang et collab. [2010] ont proposé une version modifiée de la méthode, appelée MBDE, pour permettre de traiter les problèmes d'optimisation dans un espace binaire. MBDE est basé sur la même stratégie que l'algorithme initial mais introduit un opérateur d'estimation de probabilité $P(x_{ij}^t)$ pour exprimer la probabilité d'un vecteur x_{ij}^t :

$$P(x_{ij}^t) = \frac{1}{1 + e^{-\frac{2 \cdot b \cdot \left[x_{i_1 j}^t + F \cdot (x_{i_2 j}^t - x_{i_3 j}^t) - 0.5 \right]}{1 + 2 \cdot F}}} \tag{4.3}$$

où $b \in \mathbb{R}^+$ (les auteurs suggèrent d'utiliser $b = 6$). Cette probabilité est utilisée pour définir le vecteur mutant dans l'espace binaire en utilisant l'équation ci-dessous à la place de l'équation 4.1 :

$$\mu_{ij}^t = \begin{cases} 1 & \text{if } rand \leq P(x_{ij}^t) \\ 0 & \text{otherwise} \end{cases} \tag{4.4}$$

Cette adaptation de l'évolution différentielle dans un espace binaire est bien adaptée à notre problème. Cependant, elle nécessite toujours, à chaque génération, d'effectuer de nombreuses évaluations par MiRAI qui est une procédure coûteuse.

Différentes approches [Emmerich et collab., 2006; Fonseca et collab., 2010; Jin, 2005] ont été proposées pour réduire le coût de calcul en exploitant la connaissance des évaluations passées. Elles sont principalement basées sur la méta-modélisation. L'idée est d'apprendre un nouveau modèle, appelé *modèle de substitution*, qui approxime au mieux la fonction fitness réelle qui est coûteuse à calculer.

Rasmussen [2004], puis Buche et collab. [2005] suggèrent d'utiliser une régression gaussienne (où *Krigeage*) comme modèle de substitution à cause des propriété suivantes : (1) elle peut approximer n'importe quelle fonction comme un réseau de neurones artificiel (ANN), (2) elle peut prédire la moyenne et l'écart-type de la fitness de n'importe quel nouvel individu, (3) elle dépend d'un très petit nombre d'hyperparamètres. L'inconvénient d'un tel modèle est son coût en temps de calcul qui est en $\mathcal{O}(n^3)$ pour l'apprentissage, en $\mathcal{O}(n^2)$ pour la prédiction de l'écart-type et en $\mathcal{O}(n)$ pour la prédiction de la valeur moyenne. Cependant, le temps nécessaire pour apprendre le modèle de substitution peut être considéré comme insignifiant par rapport au temps requis pour évaluer réellement les configurations de MiRAI.

La recherche de la paramétrisation optimale de MiRAI s'est effectué avec une méthode d'évolution différentielle, en exécutant, avant chaque évaluation réelle des paramètres, un modèle de substitution basé sur le Krigeage qui permet d'identifier les zones prometteuses de l'espace de recherche en réduisant considérablement les coûts de calcul.

Une analyse du meilleur individu, c'est-à-dire de la meilleure paramétrisation de MiRAI, trouvée par l'algorithme révèle qu'en plus des associations miRNA-maladie, seules les données relatives à la famille, aux miRNA voisins et aux gènes cibles sont utilisées (bits 1,2,3 du vecteur de codage décrit dans le Tableau 4.2). Fait intéressant, aucune des sources de données textuelles n'a été utilisée. Le vecteur d'encodage indique que les données relatives aux associations avec les gènes cibles doivent être pondérées par la méthode NBI (bit 7 décrit dans le Tableau 4.2). Les maladies associées aux miARN sont d'abord déduites des données

(bit 11 décrit dans le Tableau 4.2). Ensuite, une mesure de similarité est effectuée entre les maladies et 3 seuils sont utilisés pour discrétiser la valeur obtenue. Les seuils de coupure optimaux sont les suivants : 0.05,0.25 et 0.65. En ce qui concerne la dimension de l'espace latent, la méthode évolutionnaire a confirmé la valeur optimale de 400 qui avait été trouvée manuellement. Enfin, la maladie recherchée est étendue à toutes les maladies subsumées avant d'interroger l'espace latent (bit 35 décrit dans le Tableau 4.2).

La valeur AUC moyenne obtenue est 0,897 avec un minimum de 0,713 pour Lupus vulgaris et un maximum de 0,986 pour la cardiomyopathie hypertrophique. Cette performance est meilleure que la version de MiRAI configurée manuellement qui obtenait une AUC moyenne de 0,867. Une comparaison avec la version 1 de MiRAI ainsi qu'avec d'autres méthodes existantes sur les 15 maladies les plus annotées est présentée dans le tableau 4.1. Comme ce tableau le montre, les performances de MiRAI paramétrisé par un algorithme évolutionnaire sont caractéristiques d'un excellent classifieur.

Nous devons relativiser ces résultats en signalant que peu après la date à laquelle nous avons effectué les évaluations, deux méthodes très performantes ont été proposées par deux équipes de recherches distinctes : NCPMDA par Gu et collab. [2016] et PBMDA par You et collab. [2017]. Ces méthodes surclassent sans aucune ambiguïté MiRAI puisqu'elles revendiquent toutes les deux des scores AUC de 0.917 sur le même jeu de test composé de 15 maladies.

Malgré ces résultats un peu en dessous de l'état de l'art en ce qui concerne la qualité des prédictions, MiRAI se distingue des autres méthodes par la possibilité d'invalider des associations présentes dans les données sources.

La méthode est décrite en détail dans l'article de Pallez et collab. [2017].

Id	Description des paramètres	type	encodage binaire
a	Inclusion des sources de données		
	Gènes cibles ?	1 binaire	bit 1
	Familles de miRNA ?	1 binaire	bit 2
	Distance entre miRNA ?	1 binaire	bit 3
	Résumés des articles associés dans pubMed ?	1 binaire	bit 4
	Résumés des articles associés par MiRBase ?	1 binaire	bit 5
	Description des miRNA dans MiRBase ?	1 binaire	bit 6
b	Transformation des sources de données		
	NBI sur la matrice miRNA-cible	1 binaire	bit 7
	TF-IDF sur les résumés PubMed ?	1 binaire	bit 8
	TF-IDF sur les résumés MiRBase ?	1 binaire	bit 9
	TF-IDF sur les descriptions de MiRBase ?	1 binaire	bit 10
	Subsomption entre maladies dans la matrice MD ?	1 binaire	bit 11
	Similarités entre maladies ? / param. de discrétisation	19 décimaux	bits [12-30]
c	Dimension de l'espace latent	1 entier	bits [31-34]
d	Subsomption entre maladies lors de la requête ?	1 binaire	bit 35

Tableau 4.2 – Description et encodage des paramètres de MiRAI

Identification des associations potentiellement erronées

Notre méthode génère, pour chaque maladie, une liste ordonnée de miARN qui lui sont potentiellement associés. Les miARN qui ont une forte probabilité d'être associés à une maladie donnée sont situés au début de la liste, tandis que les miARN qui ne sont pas liés à la maladie sont placés à l'autre extrémité. En regardant les listes ordonnées obtenues pour chaque ensemble de tests, nous avons constaté que certaines des associations connues se trouvent à la fin de la liste. Ce phénomène est illustré par des courbes ROC qui atteignent 100% pour

le faux positif avant que tous les vrais positifs soient confirmés. Cette découverte suggère que les associations positives qui se trouvent à la fin de la liste des prédictions pourraient être potentiellement fausses. Nous considérons qu'une association miARN-maladie est erronée si elle est reportée par notre méthode avec un taux de faux positifs supérieur à 0,85. L'application systématique de ce critère pour toutes les maladies conduit à l'identification de 86 associations miARN-maladies potentiellement erronées.

Nous avons manuellement vérifié chaque association en examinant en détail les données fournies dans les publications associées de chaque association. Normalement, nous nous attendons à ce que les associations miARN-maladie spécifiées dans la base de données de référence correspondent à des marqueurs des maladies, c'est à dire, des associations qui reflètent une différence d'expression des miARN entre des échantillons sains et des échantillons malades. Cependant, parmi les associations douteuses mises en évidence par MiRAI, plus de la moitié (57%) proviennent de travaux qui étudient l'effet de molécules ou de traitements sur les cellules cancéreuses [Findlay et collab., 2008; Tang et collab., 2013; Yamamoto et collab., 2011] ou l'évolution de l'expression des miARN dans la progression du cancer [Kumar et collab., 2011; Pizzimenti et collab., 2009]. De plus, nous avons noté que 19% des associations infirmées concernent des recherche portant sur les miARN circulants.

Bien que les miARN contenus dans le sérum ou le sang représentent potentiellement des biomarqueurs avec un énorme potentiel, la pertinence de leur utilisation est actuellement contestée par une partie de la communauté scientifique [Jarry et collab., 2014].

Le reste des associations marquées comme erronées correspondent à des problèmes de contrôle (comme l'utilisation de cellules d'un patient malade pour le contrôle) [Sato et collab., 2011; Sun et collab., 2013], à l'extraction automatisée de données à partir de tableaux sans tenir compte des scores statistiques [Thum et collab., 2007] et à des prédictions bioinformatiques [Zhou et collab., 2012].

Seules 8 invalidations effectuées par MiRAI n'ont pas pu être expliquées à la lecture des publications. Nous considérons que ces 8 associations sont valides et qu'elles ont été invalidées par MiRAI de manière erronée. Pour ce qui concerne la tâche d'identification des fausses associations contenues dans la base de données de référence, MiRAI affiche donc un taux de faux positif légèrement inférieur à 10% ($8/86 = 0.093$).

Prédiction de nouvelles associations miARN-maladie

En suivant le même type de raisonnement que celui présenté dans le paragraphe précédent, nous suggérons que les miARN qui ne sont pas associés à une maladie et qui apparaissent à proximité du début de la liste ordonnée des candidats pourraient être proposés en tant que nouvelles associations. Nous considérons que de nouveaux miARN sont potentiellement associés à une maladie si l'association est rapportée avec un taux de vrais positifs supérieur à $0,85$. L'application systématique de ce critère pour toutes les maladies a permis la génération d'une liste de 126 nouvelles associations. Tous les résultats sont détaillés par Pasquier et Gardès [2016] et Pallez et collab. [2017].

La dernière version de la base de données HMDD date de juin 2014. C'est donc avec les données contenues dans les articles publiés jusqu'à cette date que les prédictions ont été effectuées. Pour évaluer la qualité des prédictions de MiRAI, nous avons procédé à une vérification manuelle en recherchant dans les articles publiés après 2014 ceux qui décrivent des associations prédites par notre méthode.

Il s'avère que 54 associations miARN-maladie (parmi les 126 qui ont été prédites) ont été découvertes au cours des deux dernières années [Pallez et collab., 2017]. Parmi les 72 prédictions restantes, 53 concernent l'implication de mir-188 et mir-765 dans de multiples maladies.

Prédiction de l'implication de mir-188 et mir-765 dans de multiples maladies

Mir-188 est un miARN situé dans le génome humain sur le chromosome X (50003503-50003588 +). Il a été isolé pour la première fois en 2003 d'un rein de souris par l'équipe de Tuschl en Allemagne [Lagos-Quintana et collab., 2003].

Mir-188 est connu pour être dérégulé dans diverses maladies comme la maladie d'Alzheimer [Lee et collab., 2016; Zhang et collab., 2014], l'azoospermie [Song et collab., 2017], le cancer du sein [Hamam et collab., 2016], quelques carcinomes [Fang et collab., 2015; Pichler et collab., 2016; Wang et Liu, 2016], des leucémies [Chakraborty et collab., 2016; Jinlong et collab., 2015], l'infarctus [Wang et collab., 2015], le myélome [Bi et collab., 2015], la pré-éclampsie [Yang et collab., 2015] et le cancer de la prostate [Zhang et collab., 2015b]. Mir-188 participe au maintien du cycle cellulaire (prolifération cellulaire, transition du cycle cellulaire G1/S, formation de colonies tumorales) en ciblant les gènes impliqués dans les points de contrôle du cycle cellulaire (CCND1, CCND3, CCNE1, CCNA2, CDK4 et CDK2) [Wu et collab., 2014]. Son rôle a également été démontré dans les processus épigénétiques cellulaires [Song et collab., 2017] et dans l'activation de la synthèse des protéines [Wu et collab., 2014].

Notre analyse révèle l'implication potentielle de mir-188 dans au moins 32 autres maladies, y compris les maladies cardiovasculaires, les troubles neuronaux et les maladies auto-immunes. Son action sur les fonctions cellulaires fondamentales clés peut expliquer son implication potentielle dans un si large éventail de maladies.

Mir-765 est un miARN situé dans le génome humain sur le premier chromosome (156936131-156936244 -). Il a été isolé pour la première fois en 2006 à partir de cellules embryonnaires et primaires humaines par l'équipe de Cuppen aux Pays-Bas [Berezikov et collab., 2006]. Mir-765 inhibe la phosphorylation de la signalisation eNOS [Ho et collab., 2013] et ERK/Akt/AMPK en ciblant l'apelin [Liao et collab., 2015], un ligand endogène de la protéine G. Mir-765 interfère également avec la voie MAPK en réprimant le récepteur neurotrophique tyrosine kinase [Muiños-Gimeno et collab., 2009]. Ces voies sont impliquées dans la régulation du cycle cellulaire induit par un stimulus externe (blocage de la progression du cycle cellulaire à la transition G2/M, migration et invasion cellulaire). Mir-765 diminue le niveau d'HMGA1 [Leung et collab., 2014], une protéine chromatine non-histone impliquée dans la régulation des processus 3R dépendants de l'ADN (réplication, recombinaison et réparation). Il est impliqué dans de nombreuses maladies comme l'oligoasthénozoospermie [Abu-Halima et collab., 2016], le carcinome hépatocellulaire [Xie et collab., 2016], l'insuffisance cardiaque [Cai et collab., 2015], ainsi que dans le cancer du sein [Lv et collab., 2014], de la prostate [Leung et collab., 2014] et du rectum [Della Vittoria Scarpati et collab., 2012].

Nos prédictions font état d'un lien potentiel entre le mir-765 et 21 maladies, dont les troubles neuronaux, les maladies cardiovasculaires, les rhumatoïdes, divers lymphomes et carcinomes, la leucémie, la cirrhose du foie et les cancers de l'appareil reproducteur féminin.

4.2.7 Etude de cas sur le cancer du sein

hsa-mir-1323	hsa-mir-1469	hsa-mir-215	hsa-mir-2355
hsa-mir-3130-1	hsa-mir-3130-2	hsa-mir-3186	hsa-mir-411
hsa-mir-4257	hsa-mir-4306	hsa-mir-718	

Tableau 4.3 – Liste des miARN faussement associés au cancer du sein d'après notre méthode

Pour confirmer la robustesse de notre méthode et sa capacité à découvrir de nouvelles informations sur les principales pathologies humaines, nous présentons une étude de cas pour l'une des maladies les plus étudiées : le cancer du sein.

Le tableau 4.3 présente une liste de 11 miARN qui sont décrits comme étant associés au cancer du sein mais qui sont identifiés par notre méthode comme ayant peu de liens avec cette maladie. Parmi eux, 9 miARN proviennent d'une table figurant dans une publication de Schrauder et collab. [2012] (article déjà évoqué auparavant) qui liste des miARN différentiellement exprimés dans le cancer du sein. Les 2 miARN restants sont proposés dans une étude réalisée par van Schooneveld et collab. [2012]. Les chercheurs ont travaillé sur les miARN circulants (à partir de sang ou de sérum) qui pourraient être utilisés comme biomarqueurs pour diagnostiquer le cancer du sein au stade précoce. Cependant, selon une étude de Jarry et collab. [2014] publiée plus récemment, les conclusions reposant sur la base de miARN circulants doivent être effectuées avec prudence. L'hypothèse de base dans les études sur les ARN circulants est que la gravité de la maladie d'un patient va influencer de manière significative la qualité et la quantité des miRNA trouvés dans le sang. Plusieurs causes sont susceptibles de libérer des miARN dans les vaisseaux : sécrétions de miRNA [Nishida-Aoki et Ochiya, 2015], réactions inflammatoires [Su et collab., 2016], destruction des cellules [Jarry et collab., 2014] et d'autres maladies mineures, entre autres.

Les réactions inflammatoires sont un processus habituel des réponses immunitaires. Au cours du développement du cancer, des activations chroniques de la réponse inflammatoire ont été observées [Pandey et collab., 2015]. Ces phénomènes influencent le modèle des ARN circulants. Plusieurs articles [Peck et collab., 2015; Tsuchiya et collab., 2013; Zhang et collab., 2015a] montrent un lien entre miR-215 (un faux positif) et le processus inflammatoire. Ces observations tendent à confirmer que les miARN énumérés dans le tableau 4.3 sont plus impliqués dans les voies générales de la réponse immunitaire que dans des phénomènes spécifiques à la cancérogenèse des cancers du sein.

Le sang est un milieu multifactoriel et il est impossible pour un médecin de diagnostiquer toutes les maladies présentes chez un patient à un moment donné. Dans ces conditions, seuls des miARN spécifiques à une maladie peuvent être utilisés comme biomarqueurs. Les travaux de Jarry et collab. [2014] n'ont pas réussi à montrer un motif spécifique au cancer du sein. Les miARN mis en exergue par notre méthode sont probablement représentatifs d'artefacts dans l'expérience. L'association de ces miARN avec le cancer du sein doivent être considérées avec prudence.

miARN	**PubMed ID de l'article reportant l'association**	**Date publication**
hsa-mir-106a	19706389	Sep. 2009
hsa-mir-130a	23528537	Jun. 2013
Hsa-mir-138-1	23300839	Dec. 2012
Hsa-mir-138-2	23300839	Dec. 2012
Hsa-mir-142	25406066	Nov. 2014
Hsa-mir-144	25465851	Dec. 2014
Hsa-mir-150	24312495	Dec. 2013
hsa-mir-15b	22908280	Sep. 2012
hsa-mir-181c	23524334	Jul. 2013
hsa-mir-19b-2	21059650	Jan. 2011
hsa-mir-208a		
hsa-mir-30e	19432961	May 2009
hsa-mir-378a	20889127	Oct. 2010
hsa-mir-99a	21575166	May 2011

Tableau 4.4 – Liste des nouveaux miRNA prédits pour être associés au cancer du sein

Le tableau 4.4 présente une liste de 14 miARN qui sont prédits par notre méthode comme étant associés au cancer du sein alors que leurs associations ne sont pas reportées dans HMDD v2.0. Nous avons trouvé la confirmation de la plupart des associations dans la littérature

récente. Nous avons également mis en évidence le fait que 11 associations sont reportées dans des études publiées antérieurement à la distribution de HMMM v2.0. Ceci met en évidence les limites concernant l'exhaustivité des bases de données miARN dont le remplissage est toujours manuel. Au final, nous pouvons proposer miR-208a, dont l'association avec le cancer du sein n'est reporté dans aucune étude, comme un nouveau miRNA associé à cette maladie.

Une description beaucoup plus détaillés du travail figure dans un article que nous avons publié en 2016 [Pasquier et Gardès, 2016].

4.3 Étude des structures ADN triplex

Certaines courtes séquences d'ARN sont susceptibles de s'apparier avec des zones particulières de d'ADN pour former des structures triple brins appelées ADN triplex. Ces structures particulières obéissent à des règles d'appariement qui ont été définies par Karst Hoogsteen [Hoogsteen, 1963] et qui portent maintenant le nom de règles d'Hoogsteen. Bien que les structures ADN triplex aient été observées expérimentalement, elles sont très peu étudiées. Si quelques études ont montré l'impact inhibiteur de la structure en triple hélice, on sait encore très peu de choses sur leur rôle dans la distribution euchromatine/hétérochromatine, le processus d'expression des gènes, la biologie du développement ou la régulation épigénétique des organismes.

En collaboration avec Alain Robichon (équipe Génétique, Environnement et Plasticité de l'INRA), nous avons entrepris une étude consistant à localiser, quantifier et analyser les ADN triplex sur un génome dans le but d'accroitre nos connaissances sur ces structures. Nous avons choisi de baser notre étude sur Drosophila melanogaster dont le génome est très annoté.

Identification, sur les ARN, des motifs susceptibles de former des structures triplex

Nous avons identifié de manière systématique, sur les ARNnc mais aussi sur les ARN messagers ARNm, les sites putatifs de formation triplex d'au moins 20 ribonucléotides qui respectent strictement les règles d'Hoogsteen. Sur les 2910 séquences ARNnc et les 30440 séquences ARNm stockées dans FlyBase, environ 10% (respectivement 235 et 3047) contiennent des motifs susceptibles de former des triplex (Triplex-Forming Oligonucleotides (TFO)). Les ARNlnc contiennent 270 TFO de longueur allant de 20 à 48 ribonucléotides avec une moyenne de 24. Seulement 8 séquences avec une longueur de plus de 40 bases ont été identifiées dans 5 ARNnc. Tous les gènes codant pour ces ARNnc ont des fonctions inconnues. 3874 TFO ont été identifiés sur les ARNm. La longueur moyenne est également de 24 mais des motifs plus longs, jusqu'à 220 ribonucléotides de longueur, ont été identifiés. Les motifs de plus de 100 bases sont tous situés dans le transcript Fbtr0077999 qui contient un domaine de lectine de type C pour lesquel les fonctions biologiques sont encore peu documentées.

4.3.1 Identification, sur l'ADN, des motifs pouvant être des sites de formation de triplex

Nous avons identifié, sur le génome de D. melanogaster, 9702 séquences pouvant s'apparier avec un troisième brin (Triplex-Target Sites (TTS)). Certains sites sont très étendus, avec une longueur allant jusqu'à 490 nucléotides. Une analyse détaillée montre que 7486 TSS (77% du nombre total de sites) sont situés à l'intérieur de cadres de lecture ouverts (open reading frame ou ORF en anglais) avec 5801 d'entre eux (77,5%) situés à l'intérieur d'introns. Beaucoup de TTS sont constitués de séquences répétées le long des chromosomes, ce qui suggère fortement qu'un TFO particulier pourrait cibler de nombreux sites bien au-delà du gène dans lequel il se trouve. Le nombre de TTS par gène va de 1 à 93. Sur l'ensemble de l'organisme, 2471

différents gènes contiennent des TTS. Leur nombre augmente à 4817 si l'on inclut une région de 2kb avant et après les gènes.

4.3.2 Analyse des sites potentiel de triplex

L'évaluation des sites potentiels de triplex a été réalisée en regardant les correspondances parfaites entre les TFO et les TTS. En raison de leurs séquences composées de motifs répétés, chaque TFO peut souvent cibler un TTS à de nombreux endroits. Afin de réduire ce biais d'énumération, nous avons considéré qu'un ensemble de triplex potentiels qui se chevauchent sur plus de 10 bases constituent en fait le même site. Malgré cela, la combinatoire entre TFO et TTS produit encore un effet multiplicateur qui est difficile à estimer. Par exemple, les 3874 TFO trouvés dans les ARNm peuvent se lier parfaitement avec les 9702 TTS identifiés sur le génome de plus de 3 millions de façons différentes. Enlever les appariements qui se chevauchent sur plus de 10 bases permet de réduire ce nombre par un facteur 5. Cela vaut également pour les ARNnc. Le dénombrement des sites potentiels de triplex révèle que des centaines de milliers de triplex peuvent être formés sur le génome. Fait intéressant, malgré le fait que nous avons identifié des TFO allant jusqu'à 220 bases et des TTS d'une longueur allant jusqu'à 490 bases, les correspondances entre les motifs ne produisent pas des triplex potentiels de plus de 49 bases. Pour les ARNnc, toutes les séquences de plus de 40 bases sont originaires des deux transcrits d'un unique gène, CR43650, qui a des fonctions inconnues. Les séquences provenant d'ARNm d'au moins 40 bases de longueur sont toutes originaires des transcrits de deux gènes : CG3078 et CANA-14F.

En résumé, 51 781 et 655 705 sites potentiels de triplex, formés respectivement par les ARNnc et les ARNm, ont été trouvés dans le génome complet ; les quatre cinquièmes d'entre eux étant localisés à l'intérieur des gènes, principalement dans les zones introniques. Nous observons aussi que la distribution/localisation des triplex sur les chromosomes présente de grandes disparités. Le chromosome X, par exemple, contient de nombreux triplex potentiels ; plus du double que les autres chromosomes de même taille. Sur chaque chromosome, nous pouvons voir quelques zones particulières qui regroupent un grand nombre de triplex. La concentration la plus remarquable de triplex se produit autour de la position 21,5 millions de paires de base sur le chromosome L. Cette zone contient des répétitions d'un groupe de gènes codant des histones (His1, His2B, His2A, HIS4 et His3). Les triplex sont principalement situés entre His4 et His3.

	TTS		ARNnc		ARNm	
	nombre	%	nombre	%	nombre	%
génome	9702	100,00 %	51781	100,00 %	655705	100,00 %
gènes	7486	77,16 %	41520	80,18 %	505815	77,14 %
introns	5801	59,79 %	34203	66,05 %	416424	63,51 %
exons	1685	17,37 %	7317	14,13 %	89391	13,63 %

Tableau 4.5 – Distribution des triplex potentiels sur le génome de D. melanogaster

Une analyse d'enrichissement avec GO des transcripts qui contribuent principalement à la formation de triplex met en évidence une surreprésentation des annotations relatives à la modulation/régulation de processus biologiques. L'analyse d'enrichissement des gènes les plus ciblés par les TFO révèle que les TFO provenant des ARNnc et des ARNm ciblent le même type de gènes dont les fonctions sont essentiellement liées à la biologie du développement, et à la biochimie des acides nucléiques. Étonnamment, les cibles des TFO provenant des ARNnc et des ARNm semblent se chevaucher dans de nombreux cas, et avec une correspondance parfaite, sur des gènes qui jouent un role dans le processus de morphogenèse. Ce ciblage très spécifique suggère fortement que l'appariement ADN :ARN sous forme de triplex pourrait constituer une couche inattendue de régulation en biologie du développement.

L'étude des interactions génétiques entre les gènes principalement ciblés par les TFO situés dans les ARN étudiés permet identifier deux réseaux d'interactions qui contiennent de nombreux gènes en commun. Sans surprise, les analyses d'enrichissement GO effectuées sur ces réseaux mettent en évidence des annotations principalement liées au développement et à la morphogenèse. Chose intéressante, on distingue très nettement sur le graphe des gènes ciblés par les TFO des ARNnc, 5 sous-réseaux qui regroupent différents processus biologiques (figure 4.3). Dans le réseau des gènes ciblés par les TFO des ARNm, l'existence de sous-réseaux n'est pas visuellement évidente. Cependant l'utilisation d'un algorithme dédié à la détection des clusters denses dans les graphes a identifié 8 modules contenant de 7 à 37 gènes correspondant à des fonctions distinctes. Plus de détails peuvent être trouvés dans Pasquier et collab. [2017b].

Nos analyses suggèrent fortement que certains fragments d'ARN, codants ou non codants, pourraient avoir une action importante sur de nombreux loci des chromosomes pour effectuer des contrôles génétiques ou épigénétique à grande échelle. Nos données conduisent à une possible nouvelle voie de régulation génétique par l'entremise de fragment d'ARN.

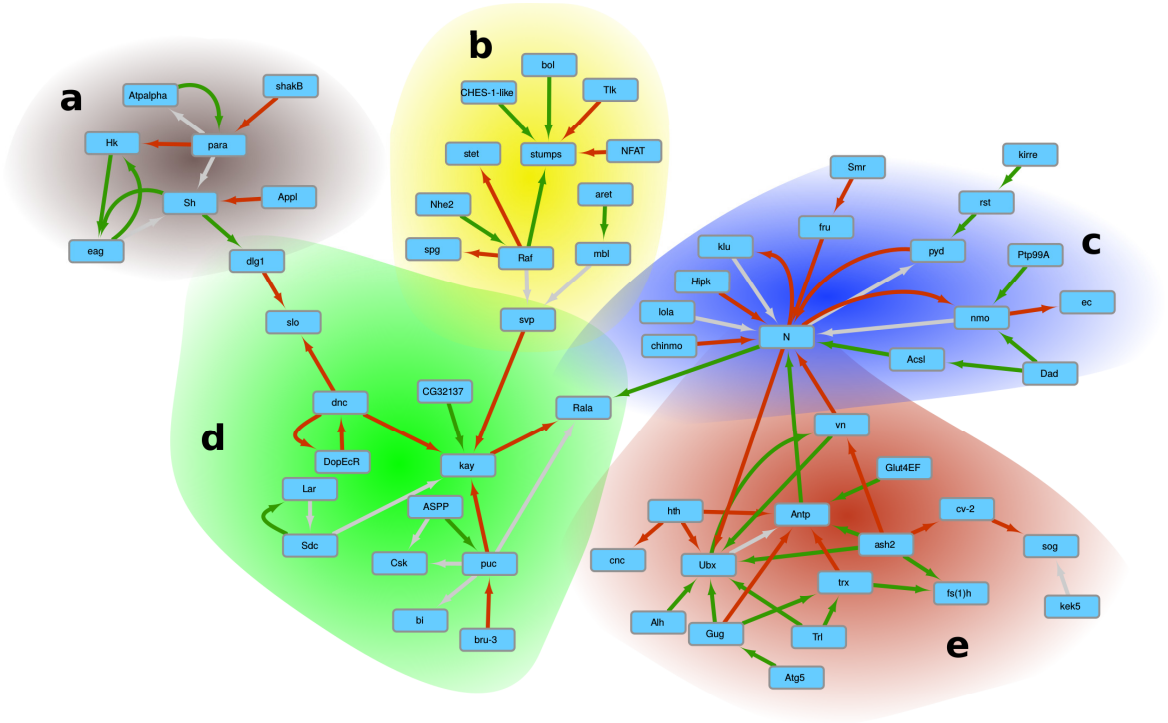

Figure 4.3 – Réseau des interactions entre les gènes les plus ciblés par les TFO des ARNnc
Sur le graphe, les flèches rouges représentent des relations d'inhibition, les flèches vertes des relations de renforcement et les flèches grises des relations combinant les deux types. Le graphe est décomposé en 5 sous-réseaux. L'analyse par enrichissement GO met en évidence des annotations correspondant à différents processus. (a) Les 8 gènes de ce sous-réseau sont enrichis avec des termes liés à des processus de comportement ou de transport. (b) Ce groupe de 12 gènes se caractérise par des fonctions touchant principalement le développement des organes. (c) Un groupe de 16 gènes impliqués dans la morphogenèse organique, principalement des yeux et la génération des neurones. (d) Un groupe de 15 gènes liés à la régulation biologique, le développement et la locomotion. (e) Le plus grand groupe contient 17 gènes liés au développement de la structure anatomique.

Chapitre 5

Perspectives

Sommaire

J'ai l'objectif d'accentuer mes activités en bioinformatique au sein de l'I3S. Je pense qu'il y a une réelle opportunité de développer des collaborations entre informaticiens et biologistes. Le domaine biologique constitue en effet un champ d'application porteur pour les recherches en science des données menées dans l'équipe SPARKS car les expériences biologiques actuelles génèrent de grandes quantités de données dont l'analyse pose des problématiques intéressantes. De nombreux laboratoires de biologie moléculaire sont implantés dans la région et il semble bénéfique, dans un contexte de recherches de plus en plus multidisciplinaires, d'intensifier les collaborations avec ces instituts.

Les recherches qui sont en cours vont naturellement être poursuivies. Je pense en particulier aux deux volets de la thématique portant sur l'étude fonctionnelle des ARN.

— Avec l'équipe GEP de l'INRA, nous prévoyons de poursuivre les recherches consistant à **étudier les interactions qui se produisent dans les cellules entre les ARN et les autres molécules présentes**.

— Avec l'entreprise Biomanda, nous prévoyons d'**étendre l'utilisation de notre méthode de prédiction des liens entre miARN et maladies** à d'autres types d'ARN non codants et à d'autres facteurs (les facteurs environnementaux ou les substances chimiques, par exemple). Nous étudions également la possibilité de nous baser sur des représentation vectorielles plus riches (comme word2vect par exemple).

Je viens également de **débuter une nouvelle thématique de recherche portant sur la sélection de caractéristiques par optimisation multicritères**. Le cadre d'application concerne les données transcriptomique et des collaborations ont déjà été établies avec deux équipes de l'iBV.

J'aimerais également beaucoup **poursuivre les recherches portant sur la fouille de graphes attribués**. De nouveaux développements algorithmiques sont d'autant plus justifiés que plusieurs applications des méthodes sont envisagées. Divers algorithmes de fouille de graphes attribués ont, par exemple, été appliquée=s à des thématiques environnementales (prédiction de l'érosion, étude de la diffusion des épidémies de dengue) mais leur utilisation dans d'autres domaines est également envisageable. Dans le domaine biologique, les nouvelles techniques de séquençage single-cell dans lesquelles on obtient des mesures pouvant être placées dans un contexte spatial (en particulier la technique seqFISH - sequential Fluorescence In Situ Hybridization) pourraient constituer un champ d'application intéressant pour les méthodes que j'ai développées. J'ai comme objectif, à moyen terme, d'orienter une partie de mes recherches dans cette direction. Pour cela, il convient de monter des projets avec des équipes de biologistes qui travaillent sur cette thématique. C'est une voie sur laquelle je me suis engagé ces derniers mois. La fouille de textes pourrait être un autre type d'application possible. J'ai en effet commencé à travailler sur l'extraction d'information dans les textes scientifiques en adoptant une approche basée sur la fouille de structures attribuées.

5.1 Étude des interactions ARN-ARN

Alain Robichon et moi-même, avons déjà exploré les appariements entre ARN et ADN qui forment des structures triple brins appelées ADN triplex [Pasquier et collab., 2017a]. Maintenant, nous émettons l'hypothèse que de petites séquences résultant de la dégradation des ARN puissent être des sources potentielles de petits ARN interférents et par ce biais, influer sur l'expression des ARNm. Nous serions alors en présence d'un système de régulation dans lequel des ARN, codant ou non pour des protéines, peuvent perturber la traduction d'autres ARNm. L'idée est que certaines séquences d'ARN trouvent dans la cellule, des séquences d'ARNm qui leurs sont complémentaires et s'apparient avec elles pour former un ARN double brin. Ces ARN peuvent ensuite être coupés par la protéine Dicer pour former de petits ARN double brins qui sont traités par le complexe protéique RISC (pour RNA-induced silencing complex) qui est à la base du phénomène d'interférence par ARN.

Pour vérifier cette hypothèse, nous avons entrepris d'étudier, à l'échelle d'un génome, l'appariement des ARNm avec des brins d'ARN complémentaires provenant du transcriptome complet. Nous avons utilisé comme base de travail le génome de Drosophila *melanogaster* et avons entrepris une analyse systématique des interactions possibles entre des séquences transcrites (correspondant à des exons ou à des introns) et des ARNm. Les premiers résultats montrent, de manière théorique, que le phénomène est potentiellement très fréquent. Pour vérifier qu'il se produit effectivement dans la cellule, nous avons comparé nos séquences avec les résultats d'autres études (notamment celles de Czech et collab. [2008] et Ghildiyal et collab. [2008]) qui ont répertorié les séquences réellement traitées par le complexe RISC. Les premiers résultats montrent qu'environ trois quart des séquences réellement trouvées dans le complexe RISC ont été prédites par notre traitement in-silico. Une analyse, encore très préliminaire de nos résultats indique que les ARNm les plus ciblés par d'autres brins d'ARN participent à des fonctions très similaires à celles identifiées lors de notre travail portant sur les structures triplex, en particulier la morphogenèse, le développement et la génération des neurones.

> *Nous avons entrepris une analyse systématique des interactions possibles entre des séquences transcrites (correspondant à des exons ou à des introns) et des ARN messagers*

Tous les travaux que nous avons effectués jusqu'à présent reposent sur des analyses statiques, à partir des séquences. Nous n'utilisons pas de mesures quantitatives des ARN sur lesquels nous travaillons. Ainsi, même si l'on identifie une courte séquence d'ARN qui permet potentiellement de cibler des dizaines d'autres gènes, nous n'avons aucun moyen de savoir si cet ARN est très exprimé, ce qui impliquerait une influence importante sur les molécules ciblées, ou au contraire très peu exprimé. Cela limite grandement la portée de nos découvertes. Pour prendre un peu plus en compte l'aspect dynamique, nous prévoyons de croiser nos analyses avec les quantifications d'expression fournies par exemple par le projet modENCODE.

5.2 Prédictions des liens miARNs-maladies

Des analyses détaillées des prédictions miARN-maladie obtenues par la méthode que nous avons développée ont montré que les résultats sont dégradés par le fait que les données utilisées pour construire notre modèle contiennent des informations erronées (cela est vrai pour toutes les méthodes existantes qui utilisent toutes les mêmes sources de données en entrée). Des tests consistant à faire quelques corrections manuelles sur les données montrent une amélioration de la qualité des prédictions.

D'où notre idée de travailler à l'élaboration d'une base de données contenant des associations fiables entre miARN et maladies. À partir des données publiques, l'identification des informations potentiellement fausses peut se faire en pointant les prédictions qui sont en inadéquation avec les données en entrée (voir notre approche détaillée dans le chapitre 5).

Une seconde option, que nous étudions, avec Nicolas Pasquier, consiste à adapter les algorithmes de partitionnement conceptuels sur lesquels il travaille [Al-Najdi et collab., 2016] pour mettre en évidence les valeurs aberrantes dans les données source.

Une troisième voie consisterait à récupérer directement les bonnes informations dans les articles de recherche. Cela déborde le cadre de la simple collecte de données concernant les miARN puisqu'une méthode qui permettrait cela pourra être utilisée pour récupérer des données concernant d'autres entités.

5.3 Collecte de relations dans des textes basée sur une stratégie de fouille de graphes attribués

Partant du constat qu'une partie de l'information contenue dans beaucoup de bases de données biologiques est erronée et que cela impacte négativement les performances des algorithmes de fouille de données développés, j'ai l'ambition d'explorer une nouvelle méthode de fouille de textes permettant d'extraire automatiquement et de manière fiable les informations à leur source, c'est à dire, à partir des articles scientifiques. J'ai déjà effectué des travaux dans le domaine de la fouille de texte en proposant, en 2010, un système permettant d'extraire automatiquement les mots clés d'un texte en langage naturel [Pasquier, 2010].

D'un point de vue conceptuel, il n'y a pas de différence entre l'extraction des données concernant les liens entre miARN et maladies et, d'une manière globale, toutes les associations entre des objets biologiques (miARN, ARNnc, gène) et des concepts (maladie, substance chimique, médicament, facteurs environnementaux). Dans un premier temps, je me focaliserai sur l'extraction des données concernant les miARN pour plusieurs raisons. Les miARN sont des molécules relativement peu nombreuses qui sont stockées dans les bases de données depuis peu de temps (une dizaine d'années). Leur identification et leur nommage canonique est plus homogène que celui des gènes ou des protéines, par exemple, qui ont beaucoup évolué au cours des dernières années. La dernière raison est que nous disposons d'un outil (MiRAI) qui utilise comme source de données les liens entre ces molécules et les maladies et qui pourra être utilisé pour valider de manière automatique une partie des données collectées.

> *J'ai l'ambition d'explorer une nouvelle méthode de fouille de texte permettant d'extraire automatiquement et de manière fiable les informations à leur source, c'est à dire, à partir des articles scientifiques*

La thématique de recherche que je compte aborder n'est pas uniquement motivée par la perspective d'obtenir des annotations plus fiables. Elle a aussi pour but d'obtenir des annotations plus complètes et plus contextualisées. Par exemple, il serait intéressant, en plus des associations proprement dites entre un miARN et une maladie, de savoir si ce miARN contribue à la pathologie ou au contraire la réduit. Il serait également important de prendre en compte le contexte des associations comme les caractéristiques de la population observée ou le type de traitement administré. Ces données contextuelles ne sont actuellement pas disponibles dans les bases de données publiques. Proposer une nouvelle base de données incluant ce type d'informations serait d'une utilité certaine pour les biologistes et permettrait aux spécialistes du traitement des données de proposer des prédictions plus pertinentes.

J'envisage de réaliser l'extraction d'information dans les textes en utilisant mes travaux portant sur la fouille de graphes attribués. Cette manière de procéder n'a, à ma connaissance, jamais été explorée. Je pense d'ailleurs que le seul algorithme de fouille de graphes attribués n'imposant pas de contraintes sur la structure des données en entrée est celui que j'ai développé.

La méthode que je compte explorer est illustrée par la figure 5.1. Chaque phrase, ou éventuellement groupe de phrases, est analysée par un parseur pour en déduire un arbre de dépendance qui est transformé, grâce à l'application de règles, en une autre structure plus simple sur laquelle les liens implicites contenus dans l'arbre de dépendance sont ajoutés. Cette nouvelle représentation peut éventuellement prendre la forme d'un graphe dont les arcs et les sommets sont étiquetés. Sur cette structure, il est facile de transférer les relations de dépendance associées aux arcs vers un nouvel attribut du sommet qui est à l'origine de l'arc. On obtient ainsi un graphe attribué qui est équivalent à l'arbre de dépendance de la phrase analysée. En utilisant des sources de données externes (annotations sémantiques, propriétés lexicales, synonymes, hyperonymes), il est également possible d'ajouter des attributs aux

sommets du graphe attribué. La finalité de toutes ces opérations est de transformer chaque phrase en un graphe attribué qui représente une généralisation des informations véhiculées par la phrase. Par exemple, en utilisant les hyperonymes du verbe « inhibit », il va être possible de faire le lien avec des phrases utilisant des termes plus généraux comme « impede » ou « prevent » pour décrire l'association.

A partir de l'ensemble des graphes attribués construits de la sorte, l'idée est de faire une recherche de sous-graphes attribués fréquents afin d'identifier des motifs récurrents. Un programme de fouille de graphes attribués peut ainsi être utilisé pour extraire les motifs qui sont fréquemment utilisés pour décrire des associations miARN-maladie reportés dans les bases de données. Les motifs fréquents identifiés peuvent constituer une sorte de signature qui servirait à extraire, dans l'ensemble des textes disponibles, de nouvelles relations décrites par le même motif. Cela pourra s'effectuer par un simple filtrage par motif (pattern matching) sur la base de graphe attribués. Les structures extraites nous permettront alors d'identifier de nouvelles associations qui seront vérifiées en les comparant aux associations connues. Ces nouvelles associations seront utilisées pour ajouter de nouveaux motifs à la base de graphes attribués ce qui nous permettra d'effectuer à nouveau une recherche de motifs fréquents. Nous commencerons le processus itératif avec les associations les plus fiables qui seront vérifiées manuellement. Progressivement, les associations prises en compte et la diversité des motifs traités devraient s'étendre et permettre de traiter les nombreuses variations syntaxiques utilisées dans les textes scientifiques pour décrire des informations similaires. Un problème délicat à résoudre sera de décider du moment où les itérations devront stopper. Nous pouvons difficilement fixer à priori un taux de faux positifs maximum du fait des erreurs contenues dans les bases de données de référence. Effectuer des vérifications manuelles n'est pas non plus une option du fait de la quantité des articles à traiter.

De premières expérimentations nous ont permis d'identifier 156 motifs que nous avons vérifiés et étendus manuellement. L'extension consiste à identifier et nommer dans les motifs des informations de contexte qui ne peuvent pas être déduites automatiquement. Ces motifs ont permis d'extraire de 191773 résumés d'articles scientifiques 12548 associations. Ce chiffre est comparable au contenu des bases de données miARN-maladies actuellement disponibles mais les données extraites sont plus détaillées. Par exemple, l'association entre mir-34a et les maladies cardiovasculaires est décrite dans plusieurs bases de données. Dans la base de données HMDD on trouve simplement une association entre mir-34a et « Cardiovascular Diseases ». Dans Phenomir, on trouve que l'association avec « Cardiovascular » est associée à un contexte de surexpression de mir-34a. A partir de la phrase suivante « Silencing of miR-34a attenuates cardiac dysfunction in a setting of moderate, but not severe, hypertrophic cardiomyopathy » qui est le titre de l'article de Bernardo et collab. [2014], nous mettons en évidence les informations listées dans le tableau 5.1.

MeSH	Mir-34a
contexte associé	silencing
action	attenuate
what	dysfunction
disease	Cardiomyopathy, Hypertrophic

Tableau 5.1 – Exemple de données supplémentaires permettant de qualifier une association
Le tableau représente les données contextuelles associées au lien entre Mir-34a et « hypertrophic cardiomyopathy » qui ont été collectées en analysant la phrase « Silencing of miR-34a attenuates cardiac dysfunction in a setting of moderate, but not severe, hypertrophic cardiomyopathy ».

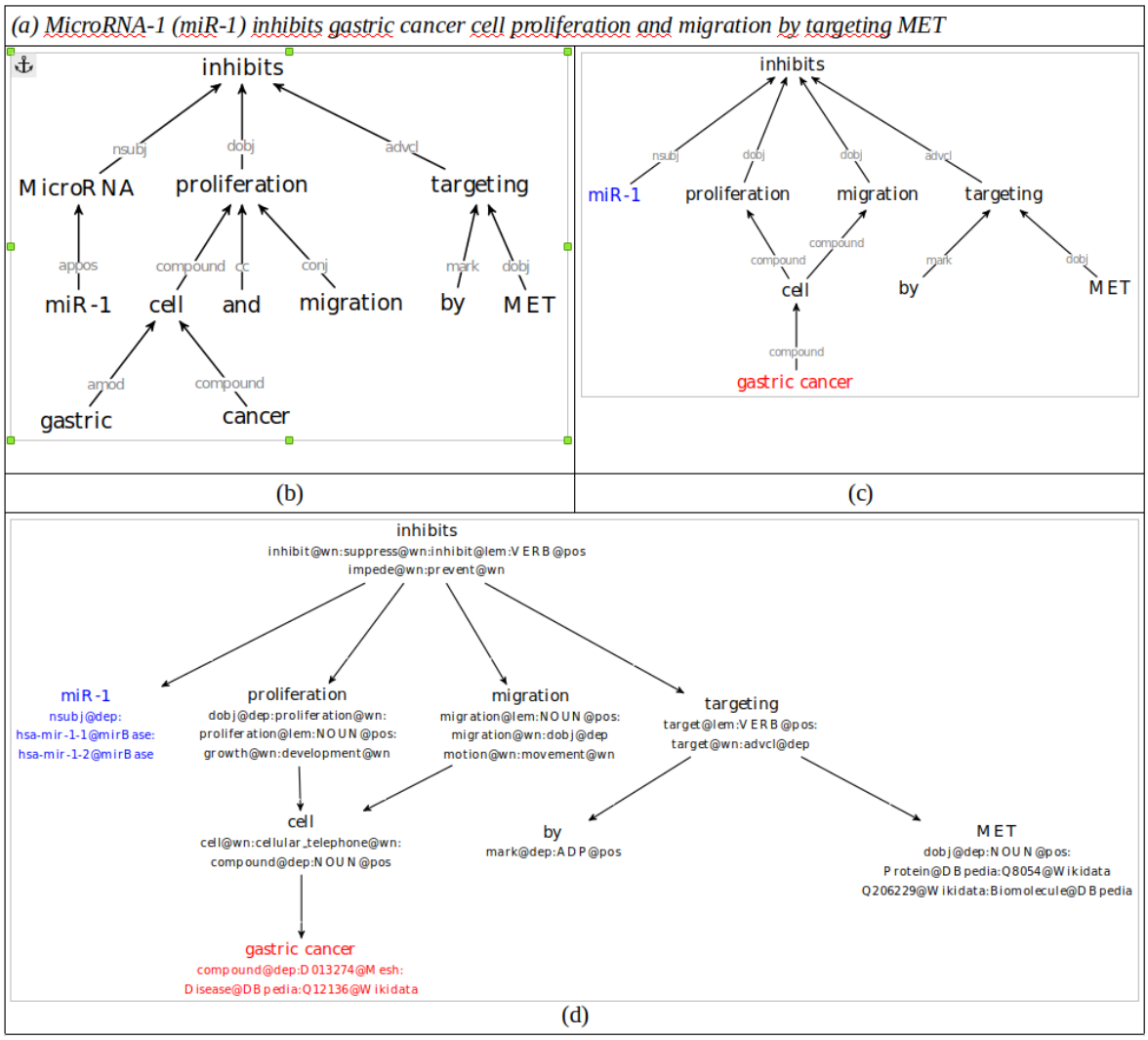

Figure 5.1 – Processus de transformation d'une phrase en un graphe attribué
(a) exemple d'une phrase extraite d'un article de Han et collab. [2015] qui contient une référence au MeSH « miR-1 » et une référence à la maladie « gastric cancer ». (b) arbre de dépendance obtenu par parsing. (c) graphe de dépendance dérivé de (b) dans lequel certaines données redondantes sont supprimées et certaines relations implicites sont décrites. (d) graphe attribué reprenant la structure du graphe (c) dans lequel les labels des nœuds sont étendus avec d'autres attributs.

5.4 Sélection de caractéristiques par optimisation multicritères

Un nouvel axe de recherche auquel je commence à m'intéresser consiste à définir une nouvelle méthode d'analyse des résultats d'expériences biologiques (notamment en transcriptomique) permettant d'identifier les facteurs clés mis en évidence par une expérience. Mon approche diffère des méthodes traditionnelles dans le sens où elle combine plusieurs sources de données et utilise l'optimisation multicritères pour identifier les gènes d'intérêt.

En biologie, une partie importante des traitements effectués sur les résultats d'expériences à haut débit consiste à identifier, le plus précisément possible, les facteurs clés qui sont à l'œuvre dans le processus étudié. Une expérimentation type consiste par exemple à mesurer les expressions des gènes dans un tissu sain et de les comparer avec des mesures effectuées dans un tissu cancéreux. Le but étant d'identifier les gènes clés qui sont à l'origine du phénotype. Suivant l'expérimentation biologique effectuée, les caractéristiques peuvent représenter différents types de molécules mais, du point de vue du traitement des données, le principe

est exactement le même. Le but est d'identifier, dans un ensemble de données, les caractéristiques (gènes, ARN, protéines) qui permettent au mieux d'expliquer un phénomène. La problématique informatique consiste en la sélection de caractéristiques (feature selection).

> *Nous envisageons d'utiliser les algorithmes évolutionnaires multi-objectifs pour sélectionner, selon plusieurs critères, les gènes clés mis en évidence par une expérience biologique*

Les données issues d'expériences biologiques à haut débit se caractérisent par un petit nombre d'échantillons (quelques dizaines dans les meilleurs de cas, mais le plus souvent moins de dix) mais un grand nombre de caractéristiques. Du fait de cette disparité, il va être possible d'identifier de nombreux ensembles de caractéristiques permettant de discriminer parfaitement les échantillons. Le problème n'est donc pas de trouver un ensemble de caractéristiques discriminantes mais d'identifier, parmi toutes les solutions possibles, celles qui sont potentiellement les plus intéressantes. Pour cela, il faut tenir compte d'autres critères basés sur des considérations biologiques. Le problème est caractérisé par un vaste espace de recherche (potentiellement 2^n solutions, n étant le nombre de mesures) et des objectifs contradictoires. Il relève de l'optimisation multi-objectif pour la sélection de caractéristiques.

Nous envisageons d'utiliser les algorithmes évolutionnaires (AE) multi-objectifs pour sélectionner de manière simultanée les groupes de caractéristiques correspondant au mieux aux différents objectifs à prendre en compte. Des travaux, portant sur utilisation des algorithmes génétiques multi-objectifs [Pighetti et collab., 2015a,b] pour la recherche d'images ont déjà été effectués dans l'équipe SPARKS et Denis Pallez, membre de l'équipe et spécialiste des AE, est impliqué dans cette démarche. Nous prévoyons de nous baser sur ces travaux antérieurs mais d'adopter une autre approche basée la programmation génétique (GP) multi-objectifs en raison de sa capacité à identifier les liens entre caractéristiques et à pouvoir entraîner parallèlement un classifieur [Xue et collab., 2016]. Nous prévoyons en particulier étudier les similitudes entre les relations inter-caractéristiques construites par la GP et les réseaux d'interactions biologique. Est-ce qu'il est possible, en utilisant un système basé sur la GP, de retrouver des interactions connues entre molécules ? Est-ce qu'un tel système pourrait être utilisé pour prédire de nouvelles interactions ? C'est un axe de recherche qui me semble prometteur et sur lequel je vais m'investir dans les prochaines années.

Nous avons reçu un financement de l'Université Côte d'Azur pour explorer cette thématique. L'étudiant sélectionné, Leandro Correa, a débuté sa thèse début janvier 2018. Il sera dirigé par Olivier Soriani et moi même.

Chapitre 6

Liste complète des références

Abu-Halima, M., N. Ludwig, M. Hart, P. Leidinger, C. Backes, A. Keller, M. Hammadeh et E. Meese. 2016, «Altered micro-ribonucleic acid expression profiles of extracellular microvesicles in the seminal plasma of patients with oligoasthenozoospermia.», *Fertility and sterility*, vol. 106, n° 5, p. 1061–1069.e3. 65

Agrawal, R. et R. Srikant. 1994, «Fast algorithms for mining association rules», dans *Proc. VLDB conf.*, p. 478–499. 31, 32

Al-Najdi, A., N. Pasquier et F. Precioso. 2016, «Using Frequent Closed Pattern Mining to Solve a Consensus Clustering Problem», dans *International Conference on Software Engineering & Knowledge Engineering*, édité par K. R. Inc., Proceedings of the SEKE'2016 International Conference. SEKE'2016, KSI Research Inc., Redwood City, United States, p. 454–461. 73

Alatrista-Salas, H., S. Bringay, F. Flouvat, N. Selmaoui et M. Teisseire. 2012, «The pattern next door : Towards spatio-sequential pattern discovery», dans *16th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, PAKDD 2012, p. 157–168. 36

Altman, R. et S. Raychaudhuri. 2001, «Whole-genome expression analysis : challenges beyond clustering», *Current Opinion Structural Biology*, vol. 11, p. 340–347. 29

Asai, T., K. Abe, S. Kawasoe, H. Arimura, H. Sakamoto et S. Arikawa. 2002, «Efficient substructure discovery from large semi-structured data», dans *the 2002 SIAM International Conference on Data Mining*, p. 158–174. 39

Attali, I., C. Courbis, P. Degenne, A. Fau, J. Fillon, D. Parigot, C. Pasquier et C. Sacerdoti Cohen. 2001a, «SmartTools : A Development Environment Generator based on XML Technologies», dans *XML Technologies and Software Engineering (XSE'01), ICSE'01, workshop proceedings*, p. 1–10. 5

Attali, I., C. Courbis, P. Degenne, A. Fau, D. Parigot et C. Pasquier. 2001b, «SmartTools : A Generator of Interactive Environments Tools», *10th International Conference on Compiler Construction (CC'01) : Held as Part of the Joint European Conferences on Theory and Practice of Software (ETAPS 2001), Electronic Notes in Theoretical Computer Science, Elsevier*, vol. 44, n° 2, p. 355–360. 5

Bandyopadhyay, S., R. Mitra, U. Maulik et M. Zhang. 2010, «Development of the human cancer microrna network», *Silence*, vol. 1, n° 1, p. 6. 56

Bartz-Beielstein, T., J. Branke, J. Mehnen et O. Mersmann. 2014, «Evolutionary algorithms», *Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery*, vol. 4, n° 3, p. 178–195. 61

Baskerville, S. et D. P. Bartel. 2005, «Microarray profiling of microrna reveals frequent coexpression with neighboring mirnas and host genes.», *RNA (New York, N.Y.)*, vol. 11, n° 3, p. 241–247. 56

Bera, A. et C. Jarque. 1981, «Efficient tests for normality, homoscedasticity and serial independence of regression residuals : Monte carlo evidence», *Economics Letters*, vol. 7, p. 313–318. 32

Berezikov, E., G. van Tetering, M. Verheul, J. van de Belt, L. van Laake, J. Vos, R. Verloop, M. van de Wetering, V. Guryev, S. Takada, A. J. van Zonneveld, H. Mano, R. Plasterk et E. Cuppen. 2006, «Many novel mammalian microrna candidates identified by extensive cloning and rake analysis.», *Genome research*, vol. 16, n° 10, p. 1289–98. 65

Bernardo, B. C., X.-M. Gao, Y. K. Tham, H. Kiriazis, C. E. Winbanks, J. Y. Y. Ooi, E. J. H. Boey, S. Obad, S. Kauppinen, P. Gregorevic, X.-J. Du, R. C. Y. Lin et J. R. McMullen. 2014, «Silencing of mir-34a attenuates cardiac dysfunction in a setting of moderate, but not severe, hypertrophic cardiomyopathy», *PLOS ONE*, vol. 9, n° 2, p. 1–12. 75

Berners-Lee, T. et J. Hendler. 2001, «Publishing on the semantic web», *Nature*, vol. 410, p. 1023–1024. 15

Bhajun, R., L. Guyon, A. Pitaval, E. Sulpice, S. Combe, P. Obeid, V. Haguet, I. Ghorbel, C. Lajaunie et X. Gidrol. 2015, «A statistically inferred microrna network identifies breast cancer target mir-940 as an actin cytoskeleton regulator», *Scientific Reports*, vol. 5, p. 8336. 57

Bi, C., T.-H. Chung, G. Huang, J. Zhou, J. Yan, G. J. Ahmann, R. Fonseca et W. J. Chng. 2015, «Genome-wide pharmacologic unmasking identifies tumor suppressive micrornas in multiple myeloma.», *Oncotarget*, vol. 6, n° 28, p. 26 508–18. 65

Birney, E., D. Andrews, P. Bevan, M. Caccamo, G. Cameron, Y. Chen, L. Clarke, G. Coates, T. Cox, J. Cuff, V. Curwen, T. Cutts, T. Down, R. Durbin, E. Eyras, X. M. Fernandez-Suarez, P. Gane, B. Gibbins, J. Gilbert, M. Hammond, H. Hotz, V. Iyer, A. Kahari, K. Jekosch, A. Kasprzyk, D. Keefe, S. Keenan, H. Lehvaslaiho, G. McVicker, C. Melsopp, P. Meidl, E. Mongin, R. Pettett, S. Potter, G. Proctor, M. Rae, S. Searle, G. Slater, D. Smedley, J. Smith, W. Spooner, A. Stabenau, J. Stalker, R. Storey, A. Ureta-Vidal, C. Woodwark, M. Clamp et T. Hubbard. 2004, «Ensembl 2004», *Nucleic Acids Research*, vol. 32, n° suppl_1, p. D468–D470. 16

Borgelt, C. 2007, *Canonical Forms for Frequent Graph Mining*, chap. 12, Springer Berlin Heidelberg, Berlin, Heidelberg, p. 337–349. 40

Bratt, S. 2005, «Toward a web of data and programs», dans *IEEE Symposium on Global Data Interoperability - Challenges and Technologies*, p. 124–128. 15

Breitling, R., A. Amtmann et P. Herzyk. 2004, «Iterative group analysis (iga) : A simple tool to enhance sensitivity and facilitate interpretation of microarray experiments», *BMC Bioinformatics*, vol. 5, n° 1, p. 34. 26, 28

Brin, S., R. Motwani, J. D. Ullman et S. Tsur. 1997, «Dynamic itemset counting and implication rules for market basket data», dans *Proc. ACM SIGMOD conf.*, p. 255–264. 32

Bringmann, B. et S. Nijssen. 2008, «What is frequent in a single graph ?», dans *12th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, édité par

T. Washio, E. Suzuki, K. M. Ting et A. Inokuchi, PAKDD 2008, Springer Berlin Heidelberg, Berlin, Heidelberg, p. 858–863. 38, 44

Buche, D., N. N. Schraudolph et P. Koumoutsakos. 2005, «Accelerating evolutionary algorithms with gaussian process fitness function models», *Trans. Sys. Man Cyber Part C*, vol. 35, n° 2, p. 183–194. 62

Cai, W.-F., G.-S. Liu, C. K. Lam, S. Florea, J. Qian, W. Zhao, T. Pritchard, K. Haghighi, D. Lebeche, L. J. Lu, J. Deng, G.-C. Fan, R. J. Hajjar et E. G. Kranias. 2015, «Up-regulation of micro-rna765 in human failing hearts is associated with post-transcriptional regulation of protein phosphatase inhibitor-1 and depressed contractility.», *European journal of heart failure*, vol. 17, n° 8, p. 782–93. 65

Camon, E., M. Magrane, D. Barrell, V. Lee, E. Dimmer, J. Maslen, D. Binns, N. Harte, R. Lopez et R. Apweiler. 2004, «The gene ontology annotation (goa) database : sharing knowledge in uniprot with gene ontology», *Nucleic acids research*, vol. 32 Database issue, p. D262–6. 16

Carmona-Saez, P., M. Chagoyen, A. Rodriguez, O. Trelles, J. Carazo et A. Pascual-Montano. 2006, «Integrated analyis of gene expression by association rules discovery», *BMC Bioinformatics*, vol. 7, n° 54. 31

Cerf, L., J. Besson, C. Robardet et J.-F. Boulicaut. 2008, «Data-peeler : Constraint-based closed pattern mining in n-ary relations», dans *SIAM International Conference on Data Mining*, SDM 2008, Atlanta, United States, p. 37–48. 44

Chakraborty, C., A. R. Sharma, B. C. Patra, M. Bhattacharya, G. Sharma et S.-S. Lee. 2016, «Micrornas mediated regulation of mapk signaling pathways in chronic myeloid leukemia.», *Oncotarget*, vol. 7, n° 27, p. 42 683–42 697. 65

Chen, H. et Z. Zhang. 2013, «Prediction of associations between OMIM diseases and MicroRNAs by random walk on OMIM disease similarity network», *The Scientific World Journal*, vol. 2013, n° December 2012. 61

Chen, X. et G.-Y. Yan. 2014, «Semi-supervised learning for potential human microRNA-disease associations inference.», *Scientific reports*, vol. 4, p. 5501. 61

Cheung, K.-H., K. Y. Yip, A. Smith, R. deKnikker, A. Masiar et M. Gerstein. 2005, «Yeasthub : a semantic web use case for integrating data in the life sciences domain», *Bioinformatics*, vol. 21, n° suppl_1, p. i85–i96. 16

Chu, S., J. DeRisi, M. Eisen, J. Mullholland, D. Botstein et P. O. e. a. Brown. 1998, «The transcriptional program of sporulation in budding yeast», *Science*, vol. 282, p. 699–705. 32

Creighton, C. et S. Hanansh. 2003, «Mining gene expression databases for association rules», *Bioinformatics*, vol. 19, p. 79–86. 31

Czech, B., C. D. Malone, R. Zhou, A. Stark, C. Schlingeheyde, M. Dus, N. Perrimon, M. Kellis, J. A. Wohlschlegel, R. Sachidanandam, G. J. Hannon et J. Brennecke. 2008, «An endogenous small interfering RNA pathway in Drosophila», *Nature*, vol. 453, n° 7196, p. 798–802. 73

De Gregorio, E., P. T. Spellman, G. M. Rubin et B. Lemaitre. 2001, «Genome-wide analysis of the drosophila immune response by using oligonucleotide microarrays», *Proceedings of the National Academy of Sciences*, vol. 98, n° 22, p. 12 590–12 595. 25, 26

Della Vittoria Scarpati, G., F. Falcetta, C. Carlomagno, P. Ubezio, S. Marchini, A. De Stefano, V. K. Singh, M. D'Incalci, S. De Placido et S. Pepe. 2012, «A specific mirna signature correlates with complete pathological response to neoadjuvant chemoradiotherapy in locally advanced rectal cancer.», *International journal of radiation oncology, biology, physics*, vol. 83, n° 4, p. 1113–9. 65

DeRisi, J., L. Iyer et V. Brown. 1997, «Exploring the metabolic and genetic control of gene expression on a genomic scale», *Science*, vol. 278, p. 680–686. 28, 29, 32

Do, H. H. et E. Rahm. 2007, «Matching large schemas : Approaches and evaluation», *Information Systems*, vol. 32, n° 6, p. 857–885. 17

Drăghici, S., P. Khatri, R. P. Martins, G. Ostermeier et S. A. Krawetz. 2003, «Global functional profiling of gene expression», *Genomics*, vol. 81, n° 2, p. 98 – 104. 24, 27

Eisen, M. B., P. T. Spellman, P. O. Brown et D. Botstein. 1998, «Cluster analysis and display of genome-wide expression patterns», *Proceedings of the National Academy of Sciences*, vol. 95, n° 25, p. 14 863–14 868. 32

Emmerich, M., K. Giannakoglou et B. Naujoks. 2006, «Single- and multiobjective evolutionary optimization assisted by gaussian random field metamodels», *IEEE Transactions on Evolutionary Computation*, vol. 10, n° 4, p. 421–439. 62

Fang, F., R.-m. Chang, L. Yu, X. Lei, S. Xiao, H. Yang et L.-Y. Yang. 2015, «Microrna-188-5p suppresses tumor cell proliferation and metastasis by directly targeting fgf5 in hepatocellular carcinoma.», *Journal of hepatology*, vol. 63, n° 4, p. 874–85. 65

Findlay, V. J., D. P. Turner, O. Moussa et D. K. Watson. 2008, «MicroRNA-mediated inhibition of prostate-derived Ets factor messenger RNA translation affects prostate-derived Ets factor regulatory networks in human breast cancer.», *Cancer research*, vol. 68, n° 20, p. 8499–506. 64

Flouvat, F., J. Sanhes, C. Pasquier, N. Selmaoui-folcher et J.-F. Boulicaut. 2014a, «Improving pattern discovery relevancy by deriving constraints from expert models», dans *European Conference on Artificial Intelligence (ECAI'14)*, p. 1–10. 9, 46

Flouvat, F., J. Sanhes, C. Pasquier, N. Selmaoui-Folcher et J.-F. Boulicaut. 2014b, «Les modèles des experts au service de l'extraction de motifs pertinents», dans *19ème congrès sur la Reconnaissance de Formes et l'Intelligence Artificielle (RFIA'14)*, p. 1–10. 9

Flouvat, F., N. Selmaoui-Folcher, J. Sanhes, C. Mu, C. Pasquier et J.-F. Boulicaut. 2020, «Mining evolutions of complex spatial objects using a single-attributed directed acyclic graph», *Knowledge and Information Systems*, vol. 62, n° 10, p. 3931–3971.

Fonseca, L. G., H. J. C. Barbosa et A. C. C. Lemonge. 2010, *On Similarity-Based Surrogate Models for Expensive Single- and Multi-objective Evolutionary Optimization*, chap. 9, Springer Berlin Heidelberg, p. 219–248. 62

Fukuzaki, M., M. Seki, H. Kashima et J. Sese. 2010, «Finding itemset-sharing patterns in a large itemset-associated graph», dans *14th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, PAKDD 2010*, vol. 6119, p. 147–159. 37

Gasch, A. et M. Eisen. 2002, «Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering», *Genome Biology*, vol. 3, p. 1–22. 30

GeneOntology. 2008, «The gene ontology project in 2008», *Nucleic Acids Research*, vol. 36, nº suppl_1, p. D440–D444. 16

Georgi, E., L. Richter, U. Ruckert et S. Kramer. 2005, «Analyzing microarray data using quantitative association rules», *Bioinformatics*, vol. 21, p. 123–129. 31

Ghildiyal, M., H. Seitz, M. D. Horwich, C. Li, T. Du, S. Lee, J. Xu, E. L. Kittler, M. L. Zapp, Z. Weng et P. D. Zamore. 2008, «Endogenous sirnas derived from transposons and mrnas in drosophila somatic cells», *Science*, vol. 320, nº 5879, p. 1077–1081. 73

Giannotti, F., M. Nanni, F. Pinelli et D. Pedreschi. 2007, «Trajectory pattern mining», dans *13th ACM International Conference on Knowledge Discovery and Data Mining*, KDD 2007, ACM, p. 330–339. 36

Girardot, F., C. Lasbleiz, V. Monnier et H. Tricoire. 2006, «Specific age related signatures in drosophila body parts transcriptome», *BMC Genomics*, vol. 7, nº 1, p. 69. 26

Grubbs, F. 1969, «Procedures for detecting outlying observations in samples», *Technometrics*, vol. 11, p. 1–21. 32

Gu, C., B. Liao, X. Li et K. Li. 2016, «Network Consistency Projection for Human miRNA-Disease Associations Inference», *Scientific Reports*, vol. 6, nº 1, p. 36 054. 63

Guha, R. 1992, *Contexts : A Formalization and Some Applications*, thèse de doctorat, Stanford, CA, USA. 18

Guo, Y., Z. Pan et J. Heflin. 2004, «An evaluation of knowledge base systems for large owl datasets», dans *Third International Semantic Web Conference (ISWC2004)*, p. 274–288. 18, 20, 22, 23

Guo, Y., Z. Pan et J. Heflin. 2005, «Lubm : A benchmark for owl knowledge base systems», *Web Semantics : Science, Services and Agents on the World Wide Web*, vol. 3, nº 2, p. 158 – 182. 20

Hamam, R., A. M. Ali, K. A. Alsaleh, M. Kassem, M. Alfayez, A. Aldahmash et N. M. Alajez. 2016, «microrna expression profiling on individual breast cancer patients identifies novel panel of circulating microrna for early detection.», *Scientific reports*, vol. 6, nº 4, p. 25 997. 65

Han, C., Y. Zhou, Q. An, F. Li, D. Li, X. Zhang, Z. Yu, L. Zheng, Z. Duan et Q. Kan. 2015, «Microrna-1 (mir-1) inhibits gastric cancer cell proliferation and migration by targeting met», *Tumor Biology*, vol. 36, nº 9, p. 6715–6723. 76

Ho, J. J. D., G. B. Robb, S. C. Tai, P. J. Turgeon, I. A. Mawji, H. S. J. Man et P. A. Marsden. 2013, «Active stabilization of human endothelial nitric oxide synthase mrna by hnrnp e1 protects against antisense rna and micrornas.», *Molecular and cellular biology*, vol. 33, nº 10, p. 2029–46. 65

Hoogsteen, K. 1963, «The crystal and molecular structure of a hydrogen-bonded complex between 1-methylthymine and 9-methyladenine», *Acta Crystallographica*, vol. 16, nº 9, p. 907–916. 67

Iliopoulos, I., S. Tsoka, M. A. Andrade, A. J. Enright, M. Carroll, P. Poullet, V. Promponas, T. Liakopoulos, G. Palaios, C. Pasquier, S. Hamodrakas, J. Tamames, A. T. Yagnik, A. Tramontano, D. Devos, C. Blaschke, A. Valencia, D. Brett, D. Martin, C. Leroy, I. Rigoutsos, C. Sander et C. A. Ouzounis.

2003, «Evaluation of annotation strategies using an entire genome sequence», *Bioinformatics (Oxford, England)*, vol. 19, n° 6, p. 717–26. 4

J Wroe, C., R. Stevens, C. Goble et M. Ashburner. 2003, «A methodology to migrate the gene ontology to a description logic environment using daml+oil», dans *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, vol. 8, p. 624–35. 19

Jarry, J., D. Schadendorf, C. Greenwood, a. Spatz et L. C. van Kempen. 2014, «The validity of circulating microRNAs in oncology : Five years of challenges and contradictions», *Molecular Oncology*, vol. 8, n° 4, p. 819–829. 64, 66

Jin, Y. 2005, «A comprehensive survey of fitness approximation in evolutionary computation», *Soft Computing*, vol. 9, n° 1, p. 3–12. 62

Jinlong, S., F. Lin, L. Yonghui, Y. Li et W. Weidong. 2015, «Identification of let-7a-2-3p or/and mir-188-5p as prognostic biomarkers in cytogenetically normal acute myeloid leukemia.», *PloS one*, vol. 10, n° 2, p. e0118 099. 65

Kalfoglou, Y. et W. M. Schorlemmer. 2003, «If-map : An ontology-mapping method based on information-flow theory», *J. Data Semantics*, vol. 1, p. 98–127. 17

Kanehisa, M. et S. Goto. 2000, «Kegg : Kyoto encyclopedia of genes and genomes», *Nucleic Acids Research*, vol. 28, n° 1, p. 27–30. 16

Kerrien, S., Y. Alam-Faruque, B. Aranda, I. Bancarz, A. Bridge, C. Derow, E. Dimmer, M. Feuermann, A. Friedrichsen, R. Huntley, C. Kohler, J. Khadake, C. Leroy, A. Liban, C. Lieftink, L. Montecchi-Palazzi, S. Orchard, J. Risse, K. Robbe, B. Roechert, D. Thorneycroft, Y. Zhang, R. Apweiler et H. Hermjakob. 2007, «Intact : open source resource for molecular interaction data», *Nucleic Acids Research*, vol. 35, n° suppl_1, p. D561–D565. 16

Kim, S.-Y. et D. J. Volsky. 2005, «Page : Parametric analysis of gene set enrichment», *BMC Bioinformatics*, vol. 6, n° 1, p. 144. 26

Kiryakov, A. 2006, *Semantic Web Technologies : Trends and Research in Ontology-based Systems*, chap. 7, Wiley, p. 115–138. 15

Kumar, S., A. Kumar, P. P. Shah, S. N. Rai, S. K. Panguluri et S. S. Kakar. 2011, «MicroRNA signature of cis-platin resistant vs. cis-platin sensitive ovarian cancer cell lines.», *Journal of ovarian research*, vol. 4, n° 1, p. 17. 64

Lagos-Quintana, M., R. Rauhut, J. Meyer, A. Borkhardt et T. Tuschl. 2003, «New microrna from mouse and human.», *RNA*, vol. 9, n° 2, p. 175–9. 65

Lambrix, P. et A. Edberg. 2003, «Evaluation of ontology merging tools in bioinformatics», dans *Pacific Symposium on Biocomputing*, p. 589–600. 17

Lasko, T. a., J. G. Bhagwat, K. H. Zou et L. Ohno-Machado. 2005, «The use of receiver operating characteristic curves in biomedical informatics», *Journal of Biomedical Informatics*, vol. 38, n° 5, p. 404–415. 58

Le, D.-H. 2015, «Network-based ranking methods for prediction of novel disease associated micrornas», *Computational Biology and Chemistry*, vol. 58, p. 139–148. 57

Lee, K., H. Kim, K. An, O.-B. Kwon, S. Park, J. H. Cha, M.-H. Kim, Y. Lee, J.-H. Kim, K. Cho et H.-S. Kim. 2016, «Replenishment of microrna-188-5p restores the synaptic and cognitive deficits in 5xfad mouse model of alzheimer's disease.», *Scientific reports*, vol. 6, n° 2, p. 34 433. 65

LEUNG, Y.-K., Q. K.-Y. CHAN, C.-F. NG, F. M.-T. MA, H.-M. TSE, K.-F. TO, J. MARANCHIE, S.-M. HO et K.-M. LAU. 2014, «Hsa-mirna-765 as a key mediator for inhibiting growth, migration and invasion in fulvestrant-treated prostate cancer.», *PloS one*, vol. 9, n° 5, p. e98 037. 65

LI, M., Y. FAN, J. CHEN, L. GAO, Z. DI et J. WU. 2005, «Weighted networks of scientific communication : The measurement and topological role of weight», *Physica A : Statistical Mechanics and its Applications*, vol. 350, n° 2-4, p. 643–656. 57

LI, Y., C. QIU, J. TU, B. GENG, J. YANG, T. JIANG et Q. CUI. 2014, «Hmdd v2.0 : A database for experimentally supported human microrna and disease associations», *Nucleic Acids Research*, vol. 42, n° D1, p. 1070–1074. 59

LIAKOPOULOS, T., G. PALAIOS, V. PROMPONAS, I. HAMODRAKAS, C. PASQUIER et S. HAMODRAKAS. 2000, «A workbench for computational analysis of protein sequence and structure on the Internet», dans *proceedings of the 22nd Conference of the Hellenic Society for Biological Sciences*, Skiathos island. 4

LIAKOPOULOS, T., C. PASQUIER et S. HAMODRAKAS. 1999, «OrienTM : A novel method to predict transmembrane protein topology», dans *21st conference of the Hellenic Society for Biological Sciences*, Galissas, Syros island. 3

LIAKOPOULOS, T., C. PASQUIER et S. HAMODRAKAS. 2001, «A novel tool for the prediction of transmembrane protein topology based on a statistical analysis of the SwissProt database : the OrienTM algorithm», *Protein Engineering Design and Selection*, vol. 14, n° 6, p. 387–390. 3, 4

LIAO, Y.-C., Y.-S. WANG, E. HSI, M.-H. CHANG, Y.-Z. YOU et S.-H. H. JUO. 2015, «Microrna-765 influences arterial stiffness through modulating apelin expression.», *Molecular and cellular endocrinology*, vol. 411, n° 10, p. 11–9. 65

LIN, D. 1998, «An information-theoretic definition of similarity», dans *15th International Conference of Machine Learning*, ICML 1998, Madison,WI, p. 296–304. 56

LIU, Y., X. ZENG, Z. HE et Q. ZOU. 2016, «Inferring microRNA-disease associations by random walk on a heterogeneous network with multiple data sources», *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 5963, n° October 2014, p. 1–1. 61

LU, M., Q. ZHANG, M. DENG, J. MIAO, Y. GUO, W. GAO et Q. CUI. 2008, «An analysis of human microrna and disease associations», *PLoS ONE*, vol. 3, n° 10, p. 1–5. 57

LV, J., K. XIA, P. XU, E. SUN, J. MA, S. GAO, Q. ZHOU, M. ZHANG, F. WANG, F. CHEN, P. ZHOU, Z. FU et H. XIE. 2014, «mirna expression patterns in chemoresistant breast cancer tissues.», *Biomedicine & pharmacotherapy*, vol. 68, n° 8, p. 935–42. 65

MAGLOTT, D., J. OSTELL, K. D. PRUITT et T. TATUSOVA. 2007, «Entrez gene : gene-centered information at ncbi», *Nucleic Acids Research*, vol. 35, n° suppl_1, p. D26–D31. 16

MANNILA, H. et H. TOIVONEN. 1996, «Multiple uses of frequent sets and condensed representations extended abstract», dans *second International Conference on Knowledge Discovery and Data Mining*, KDD 1996, AAAI Press, p. 189–194. 39

MARTINEZ, R., R. CHRISTEN, C. PASQUIER et N. PASQUIER. 2005, «Exploratory Analysis of Cancer SAGE Data», dans *9th European Conferences on Principles and Practice of Knowledge Discovery in Databases (PKDD'05), Discovery Challenge*, Porto, Portugal. 6

MARTINEZ, R. et M. COLLARD. 2007, «Extracted knowledge : Interpretation in mining biological data, a survey», *International Journal of Computer Science and Applications : Special issue in Research Challenges in Information Science*, vol. 4, n° 2, p. 145–163. 31

MARTINEZ, R., C. PASQUIER et N. PASQUIER. 2007, «GenMiner : Mining Informative Association Rules from Genomic Data», dans *IEEE International Conference on Bioinformatics and Biomedicine (BIBM'07)*, IEEE, p. 15–22. 7, 33

MARTINEZ, R., N. PASQUIER, M. COLLARD, C. PASQUIER et L. LOPEZ-PEREZ. 2006a, «Co-expressed gene groups analysis (CGGA) : An automatic tool for the interpretation of microarray experiments», *Journal of Integrative Bioinformatics*, vol. 3, n° 2, p. 1–12. 6

MARTINEZ, R., N. PASQUIER et C. PASQUIER. 2008a, «GenMiner : mining non-redundant association rules from integrated gene expression data and annotations.», *Bioinformatics (Oxford, England)*, vol. 24, n° 22, p. 2643–4. 7

MARTINEZ, R., N. PASQUIER et C. PASQUIER. 2008b, «Mining Association Rule Bases from Integrated Genomic Data and Annotations», dans *5th international conference on computational intelligence methods for bioinformatics and biostatistics (CIBB'08)*, p. 33–43. 7, 33

MARTINEZ, R., N. PASQUIER, C. PASQUIER, M. COLLARD et L. LOPEZ-PEREZ. 2006b, «Analyse des groupes de gènes co-exprimés (AGGC) : un outil automatique pour l'interprétation des expériences de biopuces», dans *13ème Rencontres de la Société Francophone de Classification (SFC'06)*, p. 267–276. 6

MARTINEZ, R., N. PASQUIER, C. PASQUIER, M. COLLARD et L. LOPEZ-PEREZ. 2008c, «Analyse des groupes de gènes co-exprimés : un outil automatique pour l'interprétation des expériences de biopuces (version étendue)», *Revue des Nouvelles Technologies de l'Information (RNTI-C-2), Classification : points de vue croisés*, vol. 831, p. 263–74. 6

MARTINEZ, R., N. PASQUIER, C. PASQUIER et L. LOPEZ-PEREZ. 2006c, «Interpreting microarray experiments via co-expressed gene groups analysis», dans *9th International Conference of Discovery Science (ICDS'06). Lecture Notes in Computer Science*, vol. 4265, Springer Berlin Heidelberg, p. 316–320. 6

MITCHELL, J. A., A. R. ARONSON, J. G. MORK, L. C. FOLK, S. M. HUMPHREY et J. M. WARD. 2003, «Gene indexing : Characterization and analysis of nlm's generifs», dans *AMIA Annual Symposium Proceedings*, p. 460–464. 16

MIYOSHI, Y., T. OZAKI et T. OHKAWA. 2009, «Frequent pattern discovery from a single graph with quantitative itemsets», dans *ICDM Workshops 2009 - IEEE International Conference on Data Mining*, p. 527–532. 36, 37

MOHAN, P., S. SHEKHAR, J. A. SHINE et J. ROGERS. 2010, «Cascading spatio-temporal pattern discovery : A summary of results.», dans *10th SIAM International Conference on Data Mining*, SDM 2010, p. 327–338. 36

MOOTHA, V., C. LINDGREN, K.-F. ERIKSSON, A. SUBRAMANIAN, S. SIHAG, J. LEHAR, P. PUIGSERVER, E. A. NILSSON, M. RIDDERSTRÅLE, E. LAURILA, N. HOUSTIS, M. DALY, N. PATTERSON, J. MESIROV, T. GOLUB, P. TAMAYO, B. SPIEGELMAN, E. LANDER, J. HIRSCHHORN, D. ALTSHULER et L. GROOP. 2003, «Pgc-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes», *Nature Genetics*, vol. 34, n° 3, p. 267–273. 26

MUIÑOS-GIMENO, M., M. GUIDI, B. KAGERBAUER, R. MARTÍN-SANTOS, R. NAVINÉS, P. ALONSO, J. M. MENCHÓN, M. GRATACÒS, X. ESTIVILL et Y. ESPINOSA-PARRILLA.

2009, «Allele variants in functional microrna target sites of the neurotrophin-3 receptor gene (ntrk3) as susceptibility factors for anxiety disorders.», *Human mutation*, vol. 30, n° 7, p. 1062–71. 65

Nishida-Aoki, N. et T. Ochiya. 2015, «Interactions between cancer cells and normal cells via mirnas in extracellular vesicles», *Cellular and Molecular Life Sciences*, vol. 72, n° 10, p. 1849–1861. 66

Noy, N. F. et M. A. Musen. 2003, «The prompt suite : Interactive tools for ontology merging and mapping», *International Journal of Human-Computer Studies*, p. 983–1024. 17

Pallez, D., J. Gardès et C. Pasquier. 2017, «Prediction of miRNA-disease Associations using an Evolutionary Tuned Latent Semantic Analysis», *Scientific Reports*, vol. 7, n° September. 10, 63, 64

Pan, K., C. Lih et N. Cohen. 2005, «Effects of threshold choice on biological conclusions reached during analysis of gene expression by dna microarrays», *National Academy of Sciences PNAS*, vol. 102, p. 8961–8965. 31

Pandey, S., B. Dwarakanath, S. Singh, A. N. Bhatt, K. Natarajan et V. Anang. 2015, «Pattern Recognition Receptors in Cancer Progression and Metastasis», *Cancer Growth and Metastasis*, p. 25–34. 66

Parigot, D., C. Courbis, P. Degenne, A. Fau, C. Pasquier, J. Fillon, C. Held et I. Attali. 2002, «Aspect and XML-oriented Semantic Framework Generator : Smarttools», *2nd Workshop on Language Descriptions, Tools and Applications LDTA'02, (Satellite Event of ETAPS'02), Electronic Notes in Theoretical Computer Science, Elsevier*, vol. 65, n° 3, p. 97–116. 5

Pasquier, C. 1991, «Projet d'utilisation de la norme ODA à la Caisse des Dépôts et Consignations. Échange et génération de documents», dans *Colloque Gestion des Documents Electroniques*, édité par S. d. c. d. L. P. e. d. F. T. (SEPT), Caen. 2

Pasquier, C. 1992, «BIBLE: a system for Design and Management of Context-Controlled Documents», dans *first international conference on Principles of Document Processing (PODP'92)*, Washington D. C. 2

Pasquier, C. 1994, *Gestion et conception de documents structurés par le contexte*, Phd thesis, University of Nice - Sophia Antipolis, Nice, France. 2

Pasquier, C. 2008, «Biological data integration using Semantic Web technologies», *Biochimie*, vol. 90, n° 4, p. 584–94. 6

Pasquier, C. 2010, «Single Document Keyphrase Extraction Using Sentence Clustering and Latent Dirichlet Allocation», dans *5th International Workshop on Semantic Evaluation*, Association for Computational Linguistics, Uppsala, p. 154–157. URL http://www.aclweb.org/anthology/S10-1032. 74

Pasquier, C. 2011, «Applying Semantic Web technologies to biological data integration and visualization», dans *Data Management in the Semantic Web*, édité par H. Jin et L. Zehua, chap. 6, Nova Science Publishers, Inc., p. 131–151. 6

Pasquier, C., S. Agnel et A. Robichon. 2017a, «The Mapping of Predicted Triplex DNA :RNA in the Drosophila Genome Reveals a Prominent Location in Development- and Morphogenesis-Related Genes», *G3 (Bethesda, Md.)*, vol. 7, n° 7, p. 2295–2304. 10, 72

Pasquier, C., M. Clément, A. Dombrovsky, S. Penaud, M. Da Rocha, C. Rancurel, N. Ledger, M. Capovilla et A. Robichon. 2014a, «Environmentally selected aphid variants in clonality context display differential patterns of methylation in the genome.», *PloS one*, vol. 9, n° 12, doi :10.1371/journal.pone.0115022, p. e115 022. URL http://dx.plos.org/10.1371/journal.pone.0115022http://www.ncbi.nlm.nih.gov/pubmed/25551225. 9

Pasquier, C., F. Flouvat, J. Sanhes et N. Selmaoui-folcher. 2014b, «Extraction de motifs dans des graphes orientés attribués en présence d'automorphisme», dans *14e Conférence Francophone sur l'Extraction et la Gestion des Connaissances (EGC'14), Revue des Nouvelles Technologies de l'Information, volume E-26*, p. 371–382. 8

Pasquier, C., F. Flouvat, J. Sanhes et N. Selmaoui-Folcher. 2017b, «Attributed graph mining in the presence of automorphism», *Knowledge and Information Systems*, vol. 50, n° 2, p. 569–584. 8, 69

Pasquier, C. et J. Gardès. 2016, «Prediction of miRNA-disease associations with a vector space model», *Scientific Reports*, vol. 6, n° June, p. 27 036. 10, 57, 59, 64, 67

Pasquier, C., F. Girardot, K. Jevardat de Fombelle et R. Christen. 2004, «THEA : ontology-driven analysis of microarray data.», *Bioinformatics (Oxford, England)*, vol. 20, n° 16, p. 2636–43. 6, 26

Pasquier, C. et S. Hamodrakas. 1999, «An hierarchical artificial neural network system for the classification of transmembrane proteins», *Protein Engineering Design and Selection*, vol. 12, n° 8, p. 631–634. 4

Pasquier, C., V. Promponas et S. Hamodrakas. 2001, «PRED-CLASS : cascading neural networks for generalized protein classification and genome-wide applications.», *Proteins : Structure, Function, and Bioinformatics*, vol. 44, n° 3, p. 361–9. 4

Pasquier, C., V. Promponas, G. Palaios, I. Hamodrakas et S. Hamodrakas. 1999a, «A novel method for predicting transmembrane segments in proteins based on a statistical analysis of the SwissProt database : the PRED-TMR algorithm», *Protein Engineering Design and Selection*, vol. 12, n° 5, p. 381–385. 3

Pasquier, C., V. Promponas, G. Palaios, I. Hamodrakas et S. Hamodrakas. 1999b, «PRED-TMR2 : An hierarchical neural network to classify proteins as transmembrane and a novel method to predict transmembrane segments», dans *proceedings of the 21st conference of the Hellenic Society for Biological Sciences*, Galissas, Syros island. 4

Pasquier, C., V. Promponas, N. Varvayannis et S. Hamodrakas. 1998a, «A web interface for FT : a tool dedicated to the study of periodicities in sequences», dans *20th conference of the Hellenic Society for Biological Science*, Samos Island. 4

Pasquier, C., V. Promponas, N. Varvayannis et S. Hamodrakas. 1998b, «A web server to locate periodicities in a sequence», *Bioinformatics*, vol. 14, n° 8, p. 749–50. 4

Pasquier, C., J. Sanhes, F. Flouvat et N. Selmaoui-Folcher. 2013a, «Extraction de motifs fréquents dans des arbres attribués», dans *13e Conférence Francophone sur l'Extraction et la Gestion des Connaissances (EGC'13), Revue des Nouvelles Technologies de l'Information, volume E-24*, p. 193–204. 8

Pasquier, C., J. Sanhes, F. Flouvat et N. Selmaoui-Folcher. 2013b, «Frequent Pattern Mining in Attributed trees», dans *J. Pei et al. (Eds.) : PAKDD 2013, Part I, LNAI 7818, pp. 26–37. Springer, Heidelberg (2013)*, p. 26–37. 8

Pasquier, C., J. Sanhes, F. Flouvat et N. Selmaoui-Folcher. 2016, «Frequent pattern mining in attributed trees : algorithms and applications», *Knowledge and Information Systems*, vol. 46, n° 3, p. 491–514. 8, 49

Pasquier, C. et L. Théry. 2000, «A Distributed Editing Environment for XML Documents», dans *first ECOOP Workshop on XML and Object Technology (XOT'00)*, p. 1–10. 5

Pasquier, N., Y. Bastide, R. Taouil et L. Lakhal. 1999c, «Discovering frequent closed itemsets for association rules», dans *7th International Conference on Database Theory*, ICDT 1999, Springer-Verlag, London, UK, UK, p. 398–416. 39, 43

Pasquier, N., C. Pasquier, L. Brisson et M. Collard. 2008, «Mining Gene Expression Data using Domain Knowledge», *International Journal of Software and Informatics (IJSI)*, vol. 2, n° 2, p. 215–231. 6

Pasquier, N., R. Taouil, Y. Bastide, G. Stumme et L. Lakhal. 2005, «Generating a condensed representation for association rules», *Journal of Intelligent Information Systems*, vol. 24, n° 1, p. 29–60. 31

Peck, B. C. E., M. Weiser, S. E. Lee, G. R. Gipson, V. B. Iyer, R. B. Sartor, H. H. Herfarth, M. D. Long, J. J. Hansen, K. L. Isaacs, D. G. Trembath, R. Rahbar, T. S. Sadiq, T. S. Furey, P. Sethupathy et S. Z. Sheikh. 2015, «Micrornas classify different disease behavior phenotypes of crohn's disease and may have prognostic utility.», *Inflammatory bowel diseases*, vol. 21, n° 9, p. 2178–87. 66

Pichler, M., V. Stiegelbauer, P. Vychytilova-Faltejskova, C. Ivan, H. Ling, E. Winter, X. Zhang, M. Goblirsch, A. Wulf-Goldenberg, M. Ohtsuka, J. Haybaeck, M. Svoboda, Y. Okugawa, A. Gerger, G. Hoefler, A. Goel, O. Slaby et G. A. Calin. 2016, «Genome-wide mirna analysis identifies mir-188-3p as a novel prognostic marker and molecular factor involved in colorectal carcinogenesis.», *Clinical cancer research : an official journal of the American Association for Cancer Research*, vol. 6, n° 4, p. 25 997. 65

Pighetti, R., D. Pallez et F. Precioso. 2015a, «Comparative Study of Recent Multimodal Evolutionary Algorithms», dans *Symposium on Computational Intelligence in Multi-criteria Decision-Making(IEEE MCDM)*, SSCI, IEEE, Cape Town, South Africa, p. 1–10. 77

Pighetti, R., D. Pallez et F. Precioso. 2015b, «Improving SVM Training Sample Selection Using Multi-Objective Evolutionary Algorithm and LSH», dans *Symposium on Computational Intelligence and Data Mining(IEEE CIDM)*, SSCI, IEEE, Cape Town, South Africa, p. 1–10. 77

Pizzimenti, S., M. Ferracin, S. Sabbioni, C. Toaldo, P. Pettazzoni, M. U. Dianzani, M. Negrini et G. Barrera. 2009, «MicroRNA expression changes during human leukemic HL-60 cell differentiation induced by 4-hydroxynonenal, a product of lipid peroxidation.», *Free radical biology & medicine*, vol. 46, n° 2, p. 282–8. 64

Promponas, V., G. Palaios, C. Pasquier, I. Hamodrakas et S. Hamodrakas. 1998, «CoPreTHi : a program to combine the results of transmembrane protein segment prediction methods», dans *20th conference of the Hellenic Society for Biological Sciences*, Samos Island. 3

Promponas, V., G. Palaios, C. Pasquier, I. Hamodrakas et S. Hamodrakas. 1999, «CoPreTHi : a Web tool which combines transmembrane protein segment prediction methods», *In silico biology*, vol. 1, n° 3, p. 159–62. 3

Rasmussen, C. E. 2004, *Gaussian Processes in Machine Learning*, chap. 4, Springer Berlin Heidelberg", p. 63–71. 62

Riva, A., A.-S. Carpentier, B. Torrésani et A. Hénaut. 2005, «Comments on selected fundamental aspects of microarray analysis», *Computational biology and chemistry*, vol. 29, p. 319–36. 26

Rousset, R., F. Carballès, N. Parassol, S. Schaub, D. Cérézo et S. Noselli. 2017, «Signalling crosstalk at the leading edge controls tissue closure dynamics in the drosophila embryo», *PLOS Genetics*, vol. 13, n° 2, p. 1–26. 26

Rubin, D. L., N. F. Noy et M. A. Musen. 2007, «Protégé : A tool for managing and using terminology in radiology applications», *Journal of Digital Imaging*, vol. 20, n° 1, p. 34–46. 17

Sanhes, J., F. Flouvat, C. Pasquier, N. Selmaoui-Folcher et J.-F. Boulicaut. 2013a, «Extraction de motifs condensés dans un unique graphe orienté acyclique attribué», dans *13ème Conférence Francophone sur l'Extraction et la Gestion des Connaissances (EGC'13), Revue des Nouvelles Technologies de l'Information, volume E-24*, p. 1–10. 9, 45, 49

Sanhes, J., F. Flouvat, C. Pasquier, N. Selmaoui-Folcher et J.-F. Boulicaut. 2013b, «Weighted Path as a Condensed Pattern in a Single Attributed DAG», dans *23rd International Joint Conference on Artificial Intelligence (IJCAI'13)*, p. 1642–8. 8

Sato, F., E. Hatano, K. Kitamura, A. Myomoto, T. Fujiwara, S. Takizawa, S. Tsuchiya, G. Tsujimoto, S. Uemoto et K. Shimizu. 2011, «MicroRNA profile predicts recurrence after resection in patients with hepatocellular carcinoma within the Milan Criteria.», *PloS one*, vol. 6, n° 1, p. e16 435. 64

van Schooneveld, E., M. C. Wouters, I. Van der Auwera, D. J. Peeters, H. Wildiers, P. A. Van Dam, I. Vergote, P. B. Vermeulen, L. Y. Dirix et S. J. Van Laere. 2012, «Expression profiling of cancerous and normal breast tissues identifies microrna that are differentially expressed in serum from patients with (metastatic) breast cancer and healthy volunteers», *Breast Cancer Research*, vol. 14, n° 1, p. R34. 66

Schrauder, M. G., R. Strick, R. Schulz-Wendtland, P. L. Strissel, L. Kahmann, C. R. Loehberg, M. P. Lux, S. M. Jud, A. Hartmann, A. Hein, C. M. Bayer, M. R. Bani, S. Richter, B. R. Adamietz, E. Wenkel, C. Rauh, M. W. Beckmann et P. A. Fasching. 2012, «Circulating micro-rnas as potential blood-based markers for early stage breast cancer detection», *PLOS ONE*, vol. 7, n° 1, p. 1–9. 66

Shatkay, H., S. Edwards, W. W. et B. M. 2000, «Genes, themes, microarrays : using information retrieval for large-scale gene analysis», dans *Proc. ISMB conf.*, p. 340–347. 30

Siddique, N. et H. Adeli. 2015, «Nature inspired computing : An overview and some future directions», *Cognitive Computation*, vol. 7, n° 6, p. 706–714. 60

Silk, W. K. 1984, «Quantitative descriptions of development», *Annual Review of Plant Physiology*, vol. 35, n° 1, p. 479–518. 36

Sirin, E., B. Parsia, B. C. Grau, A. Kalyanpur et Y. Katz. 2007, «Pellet : A practical owl-dl reasoner», *Web Semantics : Science, Services and Agents on the World Wide Web*, vol. 5, n° 2, p. 51–53. 20

Song, W.-Y., H. Meng, X.-G. Wang, H.-X. Jin, G.-D. Yao, S.-L. Shi, L. Wu, X.-Y. Zhang et Y.-P. Sun. 2017, «Reduced microrna-188-3p expression contributes to apoptosis

of spermatogenic cells in patients with azoospermia.», *Cell proliferation*, vol. 50, n° 1, p. 4953–62. 65

Spellman, P., G. Sherlock, M. Zhang, V. Iyer, K. Anders, M. Eisen, P. Brown, D. Botstein et B. Futcher. 1998, «Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization», *Molecular Biology of the Cell*, vol. 9, n° 12, p. 3273–3297. 32

Storn, R. et K. Price. 1997, «Differential evolution : A simple and efficient heuristic for global optimization over continuous spaces», *Journal of Global Optimization*, vol. 11, p. 341–359. 61

Su, W., M. S. Aloi et G. A. Garden. 2016, «Micrornas mediating cns inflammation : Small regulators with powerful potential», *Brain, Behavior, and Immunity*, vol. 52, p. 1–8. 66

Sun, Y., B. Su, P. Zhang, H. Xie, H. Zheng, Y. Xu, Q. Du, H. Zeng, X. Zhou, C. Chen et W. Gao. 2013, «Expression of miR-150 and miR-3940-5p is reduced in non-small cell lung carcinoma and correlates with clinicopathological features.», *Oncology reports*, vol. 29, n° 2, p. 704–12. 64

Szklarczyk, D., A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth, A. Santos, K. P. Tsafou, M. Kuhn, P. Bork, L. J. Jensen et C. von Mering. 2015, «String v10 : protein–protein interaction networks, integrated over the tree of life», *Nucleic Acids Research*, vol. 43, n° D1, p. D447–D452. 16

Tang, D., Q. Zhang, S. Zhao, J. Wang, K. Lu, Y. Song, L. Zhao, X. Kang, J. Wang, S. Xu et L. Tian. 2013, «The expression and clinical significance of microRNA-1258 and heparanase in human breast cancer.», *Clinical biochemistry*, vol. 46, n° 10-11, p. 926–32. 64

Termier, A., Y. Tamada, K. Numata, S. Imoto, T. Washio et T. Higuchi. 2007, «Digdag, a first algorithm to mine closed frequent embedded sub-dags», dans *Mining and Learning with Graphs*, MLG 2007, p. 1–5. 43

Thum, T., P. Galuppo, C. Wolf, J. Fiedler, S. Kneitz, L. W. van Laake, P. A. Doevendans, C. L. Mummery, J. Borlak, A. Haverich, C. Gross, S. Engelhardt, G. Ertl et J. Bauersachs. 2007, «MicroRNAs in the human heart : a clue to fetal gene reprogramming in heart failure.», *Circulation*, vol. 116, n° 3, p. 258–67. 64

Tsuchiya, M., P. Kumar, S. Bhattacharyya, S. Chattoraj, M. Srivastava, H. B. Pollard et R. Biswas. 2013, «Differential regulation of inflammation by inflammatory mediators in cystic fibrosis lung epithelial cells.», *Journal of interferon & cytokine research : the official journal of the International Society for Interferon and Cytokine Research*, vol. 33, n° 3, p. 121–9. 66

Turney, P. D. et P. Pantel. 2010, «From frequency to meaning : Vector space models of semantics», *Journal of Artificial Intelligence Research*, vol. 37, p. 141–188. 53, 57

Tusher, V. G., R. Tibshirani et G. Chu. 2001, «Significance analysis of microarrays applied to the ionizing radiation response», *Proceedings of the National Academy of Sciences*, vol. 98, n° 9, p. 5116–5121. 26

Tuzhilin, A. et G. Adomavicius. 2002, «Handling very large numbers of association rules in the analysis of microarray data», dans *Proc. ACM SIGKDD conf.*, p. 396–404. 31

UniProt. 2017, «Uniprot : the universal protein knowledgebase», *Nucleic Acids Research*, vol. 45, n° D1, p. D158–D169. 16

WANG, K., C.-Y. LIU, L.-Y. ZHOU, J.-X. WANG, M. WANG, B. ZHAO, W.-K. ZHAO, S.-J. XU, L.-H. FAN, X.-J. ZHANG, C. FENG, C.-Q. WANG, Y.-F. ZHAO et P.-F. LI. 2015, «Apf lncrna regulates autophagy and myocardial infarction by targeting mir-188-3p.», *Nature communications*, vol. 6, n° 2, p. 6779. 65

WANG, L., X. FU, M. I. MENHAS et M. FEI. 2010, «A modified binary differential evolution algorithm», dans *International Conference on Life System Modeling and Simulation and Intelligent Computing*, LSMS/ICSEE'10, p. 49–57. 62

WANG, L. et H. LIU. 2016, «microrna-188 is downregulated in oral squamous cell carcinoma and inhibits proliferation and invasion by targeting six1.», *Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine*, vol. 37, n° 3, p. 4105–13. 65

WILKINSON, K., C. SAYERS, H. KUNO et D. REYNOLDS. 2003, «Efficient rdf storage and retrieval in jena2», dans *Proceedings of the First International Conference on Semantic Web and Databases*, SWDB'03, CEUR-WS.org, p. 120–139. 20

WU, J., Q. LV, J. HE, H. ZHANG, X. MEI, K. CUI, N. HUANG, W. XIE, N. XU et Y. ZHANG. 2014, «Microrna-188 suppresses g1/s transition by targeting multiple cyclin/cdk complexes.», *Cell communication and signaling : CCS*, vol. 12, n° 8, p. 66. 65

XIE, B.-H., X. HE, R.-X. HUA, B. ZHANG, G.-S. TAN, S.-Q. XIONG, L.-S. LIU, W. CHEN, J.-Y. YANG, X.-N. WANG et H.-P. LI. 2016, «Mir-765 promotes cell proliferation by downregulating inpp4b expression in human hepatocellular carcinoma.», *Cancer biomarkers : section A of Disease markers*, vol. 16, n° 3, p. 405–13. 65

XUAN, P., K. HAN, M. GUO, Y. GUO, J. LI, J. DING, Y. LIU, Q. DAI, J. LI, Z. TENG et Y. HUANG. 2013, «Prediction of microRNAs Associated with Human Diseases Based on Weighted k Most Similar Neighbors», *PLoS ONE*, vol. 8, n° 8. 61

XUAN, P., K. HAN, Y. GUO, J. LI, X. LI, Y. ZHONG, Z. ZHANG et J. DING. 2015, «Prediction of potential disease-associated microrna based on random walk.», *Bioinformatics (Oxford, England)*, vol. 31, n° January, p. 1805–1815. 61

XUE, B., M. ZHANG, W. N. BROWNE et X. YAO. 2016, «A survey on evolutionary computation approaches to feature selection», *IEEE Transactions on Evolutionary Computation*, vol. 20, n° 4, p. 606–626. 77

YAMAMOTO, Y., Y. YOSHIOKA, K. MINOURA, R.-u. TAKAHASHI, F. TAKESHITA, T. TAYA, R. HORII, Y. FUKUOKA, T. KATO, N. KOSAKA et T. OCHIYA. 2011, «An integrative genomic analysis revealed the relevance of microRNA and gene expression for drug-resistance in human breast cancer cells.», *Molecular cancer*, vol. 10, p. 135. 64

YAN, X. et J. HAN. 2002, «gspan : Graph-based substructure pattern mining», dans *second IEEE International Conference on Data Mining*, ICDM 2002, p. 721–724. 40, 41

YAN, X. et J. HAN. 2003, «Closegraph : Mining closed frequent graph patterns», dans *9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD 2003, p. 286–295. 43

YAN, X., J. HAN et R. AFSHAR. 2003, «Clospan : Mining : Closed sequential patterns in large datasets», dans *the 2003 SIAM International Conference on Data Mining*, p. 166–177. 43

YANG, S., H. LI, Q. GE, L. GUO et F. CHEN. 2015, «Deregulated microrna species in the plasma and placenta of patients with preeclampsia.», *Molecular medicine reports*, vol. 12, n° 1, p. 527–34. 65

Yao, X. 2003, «Research issues in spatio-temporal data mining», dans *Geographic Information Science (UCGIS) Workshop on Geospatial Visualization and Knowledge Discovery*, p. 1–6. 36

You, Z. H., Z. A. Huang, Z. Zhu, G. Y. Yan, Z. W. Li, Z. Wen et X. Chen. 2017, «PBMDA : A novel and effective path-based computational model for miRNA-disease association prediction», *PLoS Computational Biology*, vol. 13, n° 3. 63

Zhang, H., L. Liu, J. Hu et L. Song. 2015a, «Microrna regulatory network revealing the mechanism of inflammation in atrial fibrillation», *Medical Science Monitor*, vol. 21, p. 3505–3513. 66

Zhang, H., S. Qi, T. Zhang, A. Wang, R. Liu, J. Guo, Y. Wang et Y. Xu. 2015b, «mir-188-5p inhibits tumour growth and metastasis in prostate cancer by repressing laptm4b expression.», *Oncotarget*, vol. 6, n° 8, p. 6092–104. 65

Zhang, J., M. Hu, Z. Teng, Y.-P. Tang et C. Chen. 2014, «Synaptic and cognitive improvements by inhibition of 2-ag metabolism are through upregulation of microrna-188-3p in a mouse model of alzheimer's disease.», *The Journal of neuroscience : the official journal of the Society for Neuroscience*, vol. 34, n° 45, p. 14 919–33. 65

Zhou, C., Q. Yu, L. Chen, J. Wang, S. Zheng et J. Zhang. 2012, «A miR-1231 binding site polymorphism in the 3'UTR of IFNAR1 is associated with hepatocellular carcinoma susceptibility.», *Gene*, vol. 507, n° 1, p. 95–8. 64

Zhou, T., J. Ren, M. Medo et Y. C. Zhang. 2007, «Bipartite network projection and personal recommendation», *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, vol. 76, n° 4, p. 1–7. 57

Annexe A

Liste des acronymes

ARD Association Rule Discovery.
ARNm ARN messager.
ARNnc ARN non codant.
AUC Area Under the ROC Curve.

FPR False Positive Rate.

GFE Groupe Fonctionnellement Enrichi.
GO Gene Ontology.
GOA Gene Ontology Annotation.

KBS Knowledge-Base System.

LSA Latent Semantic Analysis.
LUBM Lehigh University BenchMark.

MeSH Medical Subject Headings.
miARN micro ARN.

OQL Object Query Language.
OWL Ontology Web Language.

RDF Resource Description Framework.
RDFS RDF Schema.
ROC Receiver Operating Characteristic.

SPARQL SPARQL Protocol and RDF Query Language.
SQL Structured Query Language.
SVD Singular Value Decomposition.
SWL Semantic Web Language.

TF-IDF Term Frequency-Inverse Document Frequency.
TFO Triplex-Forming Oligonucleotides.
TPR True Positive Rate.
TTS Triplex-Target Sites.

URI Uniform Resource Identifier.

WWW World Wide Web.