Organisation et exploitation des connaissances sur les réseaux d'interactions biomoléculaires pour l'étude de l'étiologie des maladies génétiques et la caractérisation des effets secondaires de principes actifs

Emmanuel Bresso

UNIVERSITÉ DE LORRAINE

Ecole Doctorale BioSE (Biologie-Santé-Environnement)

Thèse

Présentée et soutenue publiquement pour l'obtention du titre de

DOCTEUR DE l'UNIVERSITE DE LORRAINE

Mention : « Sciences de la Vie et de la Santé »

par **Emmanuel BRESSO**

Organisation et exploitation des connaissances sur les réseaux d'interactions biomoléculaires pour l'étude de l'étiologie des maladies génétiques et la caractérisation des effets secondaires de principes actifs

Date de soutenance : 25 septembre 2013

Membres du jury :

Rapporteurs :

Dr Christine Brun	CR, Laboratoire TAGC, CNRS, Marseille
Dr Julie Thompson	DR, Laboratoire IGBMC, CNRS, Strasbourg

Examinateurs :

Dr Marie-Christine Jaulent	DR, CRC, INSERM U729, Paris
Pr Bruno Lacarelle	PU-PH, INSERM U911, Université Aix-Marseille, Marseille
Dr Malika Smaïl-Tabbone	MC, LORIA, Université de Lorraine, Nancy
Pr Philippe Jonveaux	PU-PH, Laboratoire de Génétique Humaine, Université de Lorraine, directeur de thèse
Dr Marie-Dominique Devignes	CR, LORIA, CNRS, co-directeur de thèse

Invité :

Dr Michel Souchet	PDG, Harmonic Pharma, Villers-lès-Nancy

Laboratoire de génétique-EA4368-IFR 111 Déficiences mentales et anomalies de structure du génome, CHU de Nancy, Rue du Morvan, 54511 Vandœuvre-lès-Nancy

A la mémoire de mamie Gaby qui voulait tellement être là.

Remerciements

Je voudrais commencer ces remerciements en exprimant ma plus profonde reconnaissance à Christine Brun et à Julie Thompson pour avoir accepté d'être les rapporteurs de ma thèse. Je voudrais également remercier Marie-Christine Jaulent ainsi que Bruno Lacarelle pour avoir accepté d'examiner mon travail et de participer à mon jury.

J'exprime toute ma reconnaissance à Monsieur Philippe Jonveaux, sans qui tout cela n'aurait pas été possible. Merci d'avoir accepté d'être mon directeur de thèse, ce qui m'a permis de combiner l'intérêt que je porte à la génétique ainsi qu'à la bioinformatique.

Je remercie particulièrement Marie-Dominique et Malika de m'avoir encadré. Je pense que je ne vous remercierai jamais assez pour tout ce que vous m'avez apporté. Depuis le premier jour où je suis arrivé en stage de M1, c'est un véritable plaisir de travailler avec vous. Merci pour tout : vos conseils, votre patience, votre disponibilité, d'avoir su me redonner le moral quand ça n'allait pas, ...

Je tiens également à remercier toute l'équipe d'Harmonic Pharma, tout d'abord pour avoir accepté de me financer durant mes trois premières années et surtout pour m'avoir fait découvrir le monde des start-up et de l'industrie pharmaceutique.

Je souhaite également remercier Amedeo Napoli ainsi que toute l'équipe Orpailleur pour m'avoir accueilli durant ces quelques années.

Merci également à Céline et Asma du laboratoire de génétique du CHU pour avoir pris le temps de tester toutes les hypothèses de gènes sur le syndrome d'Aicardi que je leur ai envoyé.

Je voudrais particulièrement remercier Renaud, Hugo, Solange et Martin pour les excellents moments que l'on a passés ensemble. J'espère que vous avez bien profité de ces derniers mois pour vous entrainer à CS, j'aimerai bien un peu de chalenge quand on joue ensemble ;-). Merci également Renaud pour toutes les discussions que l'on a eues, que ce soit sur la fouille de données ou sur des sujets moins professionnels, et pour tous les services que tu m'as rendus.

Un grand merci pour Anisah qui a pris le temps de m'expliquer LaTeX et de répondre à toutes mes questions. Merci également d'avoir trouvé le temps pour "papoter" avec moi surtout lors de la rédaction.

Je remercie également tous mes collègues de bureau : Ghania, Lucia, Niruba, Patricia, Yasmine, Birama, Sidahmed et plus particulièrement Florent et Gino avec qui j'ai passé de très agréables moments. Je tiens à remercier Birama pour sa patience et sa disponibilité lors du débogage de MODIM.

Je n'oublie pas tous ceux avec qui j'ai passé du temps au LORIA et surtout à l'extérieur du labo. Je pense en particulier à Aurore, Laura, Victor, Iñaki, César, Juan Pablo, Hernan, Ioanna, François, Thomas, Johanna, Sébastien,...

Enfin, je tiens à remercier tous les membres de ma famille pour leur soutient et ce, même si "le titre de thèse de [m]a thèse n'est pas écrit en français" pour eux.

Préambule

Cette thèse a été financée via une convention CIFRE (Convention Industrielle de Formation par la REcherche) entre le CNRS et l'entreprise Harmonic Pharma (HPh). HPh est une entreprise issue de la recherche académique dans le domaine de l'informatique et des sciences de la vie. Cette entreprise est spécialisée dans le repositionnement de molécules d'intérêt thérapeutique, c'est-à-dire dans la recherche de nouveaux usages thérapeutiques pour des molécules en phase clinique ou déjà mises sur le marché. Pour cela, Harmonic Pharma utilise une représentation des molécules basées sur les harmoniques sphériques et, à partir de cette représentation moléculaire, recherche les similarités existant entre les molécules données en entrée et sa base de molécules de références. L'une des retombées attendues de la thèse consistait à enrichir cette ressource propriétaire en données sur les effets secondaires des molécules et sur les réseaux biologiques concernés par leurs modes d'action. Créée en Juillet 2009 Harmonic Pharma est dirigée par Michel Souchet (PDG) assisté de Arnaud Sinan Karaboga (Directeur scientifique) et de Stéphane Gégout (Directeur Général). La société emploie trois salariés et s'appuie sur les compétences de trois conseillers scientifiques : Marie-Dominique Devignes, Bernard Maigret et Dave Ritchie. En plus du repositionnement moléculaire, Harmonic Pharma propose à ses clients des solutions pour la valorisation de molécule pré-cliniques par prédiction de profil et pour la valorisation de composés naturels dans une perspective d'économie circulaire (www.harmonicpharma.com). La liste des publications dans lesquelles Harmonic Pharma s'est impliqué est donnée ci-dessous.

Par ailleurs, entre 2010 et 2013, Harmonic Pharma a été membre du projet BIOPROLOR (BIOactifs PROduits en LORraine, www.bioprolor.com) qui est destiné à construire en Lorraine une filière industrielle de production de principes actifs thérapeutiques et cosmétiques à partir et par les plantes. Le projet BIOPROLOR regroupe 7 laboratoires académiques et 6 entreprises lorraines travaillant dans les domaines de la conception et/ou la production de substances actives destinées aux marchés pharmaceutiques ou cosmétiques. L'une des tâches de ce projet avait pour objectif d'identifier les voies biologiques et biochimiques qui pourraient décrire l'effet pharmacologique global d'un principe actif. A partir des travaux développés dans cette thèse, cette tâche a pu être remplie et faire l'objet d'un livrable décrivant le modèle de données utilisé pour stocker les informations concernant les molécules, ainsi que quelques statistiques sur la couverture de cette base de données en termes de molécules, catégories, effets secondaires, cibles et pathways. Ce livrable présente également sous forme de tableaux les listes de réseaux biologiques associés à chaque catégorie de médicaments.

Liste des publications d'Harmonic Pharma

Karaboga,A.S., Planesas,J.M., Petronin,F., Teixido,J., Souchet,M. et Pérez-Nueno,V.I. (2013) A highly specific and sensitive pharmacophore model for identifying CXCR4 antagonists. Comparison with docking and shape-matching virtual screening performance. *J Chem Inf Model*, **53**, 1043-1056

Carrieri,A., Pérez-Nueno,V.I., Lentini,G. et Ritchie,D.W. (2013) Recent trends and future prospects in computational GPCR drug discovery : from virtual screening to polypharmacology. *Curr Top Med Chem*, **13**, 1069-1097.

Karaboga,A.S., Petronin,F., Marchetti,G., Souchet,M. et Maigret,B. (2013) Benchmarking of HPCC : A novel 3D molecular representation combining shape and pharmacophoric descriptors for efficient molecular similarity assessments. *J Mol Graph Model*, **41**, 20-30.

Ghemtio,L., Pérez-Nueno,V.I., Leroux,V., Asses,Y., Souchet,M., Mavridis,L. et Ritchie,D.W. (2012) Recent Trends and Applications in 3D Virtual Screening. *Comb Chem High Throughput Screen*, **15**, 749-769.

Pérez-Nueno,V.I., Venkatraman,V., Mavridis,L. et Ritchie,D.W. (2012) Detecting Drug Promiscuity using Gaussian Ensemble Screening. *J Chem Inf Model*, **52**, 1948-1961.

Pérez-Nueno,V.I. et Ritchie,D.W. (2012) Identifying and characterizing promiscuous targets : Implications for virtual screening. *Expert Opin Drug Discov*, **7**, 1-17.

Pérez-Nueno,V.I., Venkatraman,V., Mavridis,L., Clark,T. et Ritchie,D.W. (2011) Using spherical harmonic surface property representations for ligand-based virtual screening. *Mol Inform*,**30**, 151-159.

Pérez-Nueno,V.I. et Ritchie,D.W. (2011) Predicting drug promiscuity using spherical harmonic (SH) shape-based similarity comparisons. *J Chem Inf Model*, **51**, 1233-1248.

Pérez-Nueno,V.I. et Ritchie,D.W. (2011) Using Consensus-Shape Clustering to Identify Promiscuous Ligands and Protein targets and to Choose the Right Query for Shape-Based Virtual Screening. *Expert Opin Drug Discov*, **7**, 1-17.

Pérez-Nueno,V.I., Venkatraman,V., Mavridis,L. et Ritchie,D.W.(2011) Predicting drug polypharmacology using a novel surface property similarity-based approach. *J Cheminform*, **3 Suppl 1**, O19.

Pérez-Nueno, V.I., Venkatraman,V., Mavridis,L. et Ritchie,D.W.(2012) Applying in silico Tools to the Discovery of Novel CXCR4 Inhibitors. *J Chem Inf Model*, **52**, 1948-1961.

Venkatraman,V., Pérez-Nueno,V.I., Mavridis,L. et Ritchie,D.W.(2010) A Comprehensive Comparison of Ligand-Based Virtual Screening Tools Against the DUD Dataset Reveals Limitations of Current 3D Methods. *J Chem Inf Model* **50**, 2079-2093.

Carrieri,A., Pérez-Nueno,V.I., Fano,A., Pistone,C., Ritchie,D.W. et Teixidó,J.(2009) Biological profiling of anti-HIV agents and insights into CCR5 antagonist binding using in silico techniques. *ChemMedChem*, **4**, 1153-1163.

Pérez-Nueno,V.I., Pettersson,S., Ritchie,D.W., Borrell,J.I. et Teixidó,J. (2009) Discovery of Novel HIV Entry Inhibitors for the CXCR4 Receptor by Prospective Virtual Screening. *J. Chem. Inf. Model*, **49**, 810-823.

Pérez-Nueno,V.I., Ritchie,D.W., Borrell,J.I. et Teixidó,J. (2008) Clustering and classifying diverse HIV entry inhibitors using a novel consensus shape based virtual screening approach : Further evidence for multiple binding sites within the CCR5 extracellular pocket. *J Chem Inf Model*, **48**, 2146-2165.

Pérez-Nueno,V.I., Ritchie,D.W., Rabal,O., Pascual,R., Borrell,J.I. et Teixidó,J. (2008), Comparison of Ligand-Based and Receptor-Based Virtual Screening of HIV Entry Inhibitors for the CXCR4 and CCR5 Receptors Using 3D Ligand Shape matching and Ligand-Receptor Docking. *J Chem Inf Model*, **48**, 509-533

Table des matières

Table des figures

Liste des tableaux

Abréviations

2D deux dimensions

3D trois dimensions

ADN Acide désoxyribonucléique

AERS adverse event reporting system

ARNm Acide ribonucléique messager

ARX Homeobox protein ARX

BioPAX Biological pathways exchange

BP Biological process

CC Cellular component

ChIP-seq Chromatin immunoprecipitation sequencing

DAVID Database for Annotation, Visualization and Integrated Discovery

DILX Déficiences intellectuelles liées au chromosome X

EC Enzyme Commission

FDA Food and drug administration

GO Gene ontology

GraphML Graph markup language

GWAS Genome wide association studies

Id Identifiant

KEGG Kyoto encyclopedia of genes and genomes

MBD Methyl-CpG binding domain

MedDRA Medical Dictionary for Regulatory Activities

MeSH Medical Subject Headings

MF Molecular function

MFI Maximal frequent itemset

Nb Nombre

NCBI National Center for Biotechnology Information

NCI National Cancer Institute

OMIM Online Mendelian Inheritance in Man

PDB Protein data bank

PID Pathway interaction database

PLI Programmation logique inductive

PSI-MI Proteomics Standards Initiative - Molecular Interaction

RDF Resource Description Framework

SBML Systems Biology Markup Language

SQL Structured Query Language

STRING Search Tool for the Retrieval of Interacting Genes

TAP-MAS Tandem Affinity Purification - Mass Spectrometry

TC cluster de termes

TXT Texte

XML eXtensible Markup Language

Chapitre 1

Introduction

1.1 Les réseaux biologiques

L'estimation du nombre de gènes chez l'Homme varie entre 20 000 et 25 000 gènes (International Human Genome Sequencing Consortium, 2004). Ces gènes codent des ARNm qui sont traduits en protéines. Les protéines sont un des principaux composants moléculaires de la matière vivante. Afin de pouvoir réaliser leurs fonctions, les protéines ne sont pas isolées au sein des cellules mais sont organisées en un réseau d'interactions appelé interactome. En 2009, Venkatesan *et al.* ont estimé le nombre d'interactions protéine-protéine à 130 000. Ce gigantesque réseau est organisé en modules correspondant à des voies biologiques précises telles que la régulation de l'apoptose ou la réception de signaux synaptiques. De plus en plus de ressources publiques mettent à disposition des données sur les entités biologiques (gènes, protéines, petites molécules), leurs propriétés (annotation fonctionnelles, catégories de molécules, ...) et sur leurs interactions connues ou inférées Par exemple, la base de données IntAct collecte toutes les interactions protéine-protéine présentées dans la littérature (Kerrien *et al.*, 2012). Une autre illustration est la base de données AmiGO qui contient les annotations Gene Ontology (composants cellulaires, fonctions moléculaires et processus biologiques) des gènes (Carbon *et al.*, 2009).

De fait, nous désignerons par réseau biologique tout ensemble de données qu'il est possible de représenter sous forme d'un graphe dont les nœuds sont des entités biologiques et les liens sont des interactions documentées entre ces entités. Cette définition nous permet d'établir une équivalence entre réseaux biologiques et réseaux d'interactions quelqu'ils soient (voies métaboliques, voies de signalisation, voies de régulations, interactions protéine-protéine, ...). Cependant, l'hétérogénéité et la multiplicité croissantes de ces données rendent encore aujourd'hui difficile l'intégration des réseaux biologiques dans les raisonnements des utilisateurs que sont par exemple les cliniciens, confrontés à des phénotypes atypiques, ou les industriels de la pharmacie, désireux de comprendre les effets secondaires de certains médicaments. De nombreux outils sophistiqués de visualisation graphique ont été mis au point afin de les assister dans cette tâche. Il demeure pourtant que l'exploitation des données sur les réseaux biologiques reste difficile, notamment dans la compréhension des phénotypes complexes associés aux maladies génétiques ou des effets secondaires des médicaments.

1.2 Compréhension de l'étiologie des maladies génétiques

Dans le cas de maladies monogéniques, l'identification du gène responsable de la maladie est une étape cruciale dans la compréhension des causes de la maladie. De nos jours, cette étape est facilitée par les techniques de séquençage haut débit qui permettent d'obtenir rapidement les séquences des patients. Cependant, ces techniques fournissent souvent une liste importante de variations toutes susceptibles d'être à l'origine de la maladie. Il est donc nécessaire de filtrer ces variations pour ne retenir que celles provoquant la maladie. En fonction du mode de transmission de la maladie (dominant/récessif, lié ou non au chromosome X), cela peut notamment passer par la comparaison des séquences du malade avec celles de sa famille (Vissers *et al.*, 2010).

Malheureusement, aussi importante soit-elle, l'étape d'identification du gène n'est pas toujours suffisante pour comprendre l'étiologie de la maladie. En effet, identifier la position et le type de mutation affectant un gène ne permet pas systématiquement de comprendre l'ensemble du phénotype de la maladie. Même si les fonctions de la protéine affectée par la mutation du gène sont bien décrites, ce qui est loin d'être toujours le cas, elle n'agit pas de manière isolée dans la cellule. Ainsi, elle peut interagir avec de multiples partenaires appartenant à des processus cellulaires différents. Ces processus étant le plus souvent reliés entre eux, la mutation d'un gène va entrainer des perturbations dans un grand nombre de voies cellulaires et provoquer les différents symptômes de la maladie (Barabasi *et al.*, 2011). La compréhension des phénotypes complexes provoqués par des mutations géniques passe donc par une meilleure prise en compte des différents réseaux biologiques présents dans les cellules.

Le nombre croissant de ressources mettant à disposition ces données d'interactions devrait donc faciliter la compréhension de ces mécanismes. Cependant, l'accès à ces informations ainsi que leur exploitation peut poser problèmes aux non spécialistes. En effet, même si un grand nombre d'outils de visualisation sous forme de réseaux de ces données sont développés, ils ne résolvent pas le problème de l'intégration des données et ne permettent pas une analyse poussée des réseaux biologiques.

1.3 Médicaments et effets secondaires

Le développement de nouveaux médicaments passe par la recherche de molécules inhibant ou activant une protéine ciblée. Au sein de la cellule, les protéines sont en interaction les unes avec les autres et sont impliquées dans divers processus comme la transmission de signaux ou la division cellulaire. La modification de l'activité d'une protéine est donc susceptible d'altérer le fonctionnement d'un ensemble de réseaux. A cela s'ajoute le fait que contrairement à ce que l'on a longtemps pensé, une molécule ne cible pas qu'une seule protéine mais plusieurs (Bolognesi, 2013). Il en résulte qu'un médicament va entrainer des modifications dans le fonctionnement de multiples réseaux. Grâce à cette propriété, un médicament peut être utilisé pour d'autres pathologies que celle pour laquelle il a été développé, on parle alors de repositionnement moléculaire (Wu *et al.*, 2013).

Cependant, si ces multiples cibles peuvent permettre de réutiliser un médicament existant pour une autre indication, elles sont également sources de problèmes. En effet, l'altération du fonction-

nement de multiples réseaux peut provoquer des réponses indésirables à la suite de la prise de médicaments. Lors des essais cliniques nécessaires à la mise au point de nouveaux traitements, l'apparition d'effets secondaires est l'une des principales causes d'arrêt de développement. Ces multiples effets secondaires forment des phénotypes complexes et une meilleure compréhension des réseaux biologiques sous-jacents permettrait d'en limiter l'apparition lors du développement de nouveaux médicaments.

Les effets indésirables de médicaments sont donc de plus en plus étudiés dans le cadre de la pharmacovigilance (Liu *et al.*, 2012*a*). Cependant s'il existe de nombreuses études cherchant à prédire les effets secondaires, seulement un faible nombre étudie les mécanismes provoquant des effets indésirables. A cela s'ajoute le fait que la plupart d'entre elles n'utilisent qu'une fraction des données disponibles. Ainsi, certaines études établissent un lien entre molécules et effets secondaires en utilisant leurs cibles et d'autres n'utilisent que la seule structure des molécules.

1.4 Problématique

La compréhension des maladies génétiques humaines tout comme celle des effets secondaires de médicaments, concerne l'interprétation de phénotypes complexes. Cela passe aujourd'hui par la prise en compte de multiples réseaux d'interactions entre biomolécules (ou réseaux biologiques). Les recherches récentes produisent de plus en plus de données sur ces réseaux gouvernant les différents processus cellulaires. Ainsi, des ressources publiques se développent afin de stocker et donner un accès aux informations sur ces réseaux. A titre d'exemple, les bases de données KEGG PATHWAY et Pathway Interaction Database (PID) fournissent toutes deux des informations sur les réseaux biologiques. Cependant, la variété de ces sources pose le problème de l'intégration des données. En effet, les différentes bases de données existantes ne fournissent pas les mêmes données ni aux mêmes formats. Pour reprendre les deux exemples précédents, KEGG PATHWAY est une base généraliste alors que Pathway Interaction Database se focalise sur les voies de signalisation et de régulation. De fait, la complexité de ces données et des concepts sous-jacents à la notion de réseau biologique rend encore aujourd'hui difficile l'exploitation et l'interprétation de ces réseaux par les cliniciens confrontés à des phénotypes atypiques, ou par les industriels de la pharmacie, désireux de mieux comprendre les effets secondaires liés à certains médicaments. De plus, les grandes quantités de données empêchent toute exploitation manuelle. Il est donc particulièrement utile de développer des solutions suffisamment génériques pour modéliser et stocker des informations hétérogènes afin de faciliter l'accès aux données mais aussi déployer des programmes d'analyse et d'apprentissage afin d'abstraire les données et les rendre compréhensibles par les utilisateurs. Dans cette thèse je propose une approche intégrative qui permet d'exploiter à la fois les entités biologiques, leurs interactions, et les propriétés de ces entités et des interactions, d'abord en les réunissant sous un modèle de données intégré, puis en les visualisant sous forme de graphes étendus dont les noeuds peuvent représenter soit des entités, soit des propriétés de ces entités, enfin en y recherchant des régularités et des règles explicatives.

1.5 Plan du mémoire

Dans le deuxième chapitre, je présente les réseaux biologiques qu'ils soient des réseaux d'interactions protéine-protéine, des voies métaboliques ou encore des voies de signalisation. J'introduis ensuite les différents formats et outils utilisés pour visualiser ces réseaux. Enfin, je résume les travaux ayant porté sur l'utilisation des réseaux pour l'étude des maladies génétiques ainsi que des médicaments.

Le chapitre 3 présente l'entrepôt de données, NetworkDB, que j'ai développé afin de pouvoir caractériser à la fois les gènes de maladies génétiques entrainant une déficience intellectuelle et les effets secondaires de médicaments. Au sein de ce chapitre, je détaille les caractéristiques de l'entrepôt de données ainsi que la méthodologie utilisée pour le peupler. J'y présente également deux manières possibles de visualiser le contenu de NetworkDB.

La caractérisation des gènes responsables de déficiences intellectuelles est présentée dans le chapitre 4. A partir de deux exemples, je commence par montrer comment on peut associer un gène à une maladie en s'appuyant sur les réseaux biologiques. Afin d'explorer de manière plus systématique les mécanismes associés aux déficiences intellectuelles liées à l'X (DILX), j'ai utilisé NetworkDB pour mettre en évidence les mécanismes communs des gènes responsables d'une DILX.

Le cinquième chapitre se concentre sur la compréhension des effets secondaires de principes actifs. J'ai utilisé NetworkDB pour regrouper un certain nombre de données biologiques concernant les médicaments ainsi que leurs cibles. Parmi les informations propres aux médicaments, on trouve les effets secondaires. Une des étapes clef avant la caractérisation de ces effets secondaires a été de regrouper les effets sémantiquement similaires. Ensuite, j'ai défini une méthodologie permettant d'extraire des groupes d'effets secondaires (ou profils d'effets secondaires) partagés par un grand nombre de molécules. Afin de comprendre les mécanismes associés à ces profils, ceux-ci ont été caractérisés en utilisant une méthode de fouille de données relationnelles produisant un ensemble de règles explicites. Ces règles décrivant les caractéristiques des molécules pourront être utilisées afin d'éviter le développement de molécules présentant des effets secondaires graves. Deux articles publiés lors de la conférence KDIR 2011 et dans le journal *BMC Bioinformatics* sont inclus dans ce chapitre.

Enfin, le chapitre 6 résume les contributions de cette thèse et présente des développements futurs.

Chapitre 2

Représentation et exploitation des réseaux biologiques dans l'étude des systèmes biologiques

Sommaire

2.1 Des protéines aux réseaux biologiques

2.1.1 Interactions protéine-protéine (IPP)

Les protéines sont l'un des principaux composants de la matière vivante. En effet, elles constituent la majeure partie de la masse sèche des cellules (Alberts, 1998) et sont impliquées dans de très nombreux processus allant de la protection de l'organisme à la réplication de l'information génétique, en passant par la transduction de signaux cellulaires.

Les protéines ne travaillent pas seules. En effet, la majorité des processus biologiques font intervenir plus d'une dizaine d'entre elles, chaque protéine interagissant avec une ou plusieurs autres protéines et formant ainsi des complexes protéiques transitoires ou permanents. Ainsi, on estime l'interactome humain (l'ensemble des interactions protéine-protéine) à environ 130 000 interactions (Venkatesan *et al.*, 2009). A une échelle moindre, la base de données SynSysNet [1] spécialisée dans les protéines de la synapse recense 4638 interactions connues au sein des synapses (von Eichborn *et al.*, 2013). Au sein de ce réseau d'interactions, toutes les protéines ne sont pas également connectées. En effets certaines n'interagissent qu'avec une protéine, alors que d'autre interagissent avec plusieurs centaines de protéines. Par analogie avec les réseaux de télécommunications, ces protéines centrales sont dénommées "hub" et sont particulièrement importantes pour le fonctionnement des cellules de par leur rôle central dans la formation de complexes (Jeong *et al.*, 2001, Pang *et al.*, 2010).

2.1.2 Mise en évidence d'interactions physiques entre protéines

Il existe de nombreuses techniques expérimentales pour mettre en évidence les interactions physiques entre protéines. L'une des premières méthodes haut débit développée est la méthode du "double hybride". Les paires des protéines dont on veut tester l'interaction sont exprimées sous forme de protéines chimériques. Sur l'une des deux protéines on ajoute un domaine de fixation à l'ADN et sur la seconde protéine un domaine activateur de la transcription. Si les deux protéines interagissent, la présence de ces deux domaines entrainera la transcription d'un gène rapporteur et donc la détection de la transcription (Ito *et al.*, 2001). Cette technique est puissante car elle se déroule *in vivo* et permet de détecter des interactions même transitoires.

Contrairement à la méthode du double hybride qui n'identifie que des couples de protéines interagissant, la méthode TAP-MAS (Tandem Affinity Purification - Mass Spectrometry) permet de mettre en évidence les complexes multiprotéiques (Puig *et al.*, 2001). Cette technique s'appuie sur la création d'une protéine chimère formée d'une séquence tag et de la protéine d'intérêt. Cette séquence tag permettra de retenir la protéine d'intérêt dans une colonne d'affinité. Ainsi, lors du passage des protéines à tester dans la colonne, les protéines formant un complexe avec la protéine chimère seront retenues. La purification des complexes permettra ensuite d'identifier leurs composants par spectrométrie de masse.

1. http://bioinformatics.charite.de/synsysnet

2.1.3 Mise en évidence d'interactions fonctionnelles

Il existe également des méthodes dites indirectes pour détecter des interactions entre protéines. On parle alors plutôt d'interactions fonctionnelles au lieu d'interactions physiques.

Une méthode indirecte utilisée pour les organismes procaryotes est la notion de voisinage génomique. Cette méthodologie est possible grâce à l'organisation en opérons des génomes procaryotes. Les opérons sont des ensembles de gènes voisins qui sont régulés par le même facteur de transcription et impliqués dans les mêmes voies biologiques. Ainsi, en observant que deux gènes sont très fréquemment voisins dans le génome de plusieurs organismes, il est probable que les protéines issues de ces deux gènes aient une interaction fonctionnelle (Overbeek *et al.*, 1999).

Pour les organismes dont les gènes ne sont pas organisés en opérons, il est possible d'étudier les co-expressions de gènes. En effet, une conservation de la co-expression de 2 gènes dans de multiples organismes indique un avantage sélectif lors de l'évolution et donc que les protéines codées par ces gènes interagissent (Stuart *et al.*, 2003).

Une autre manière de détecter des interactions fonctionnelles est d'étudier les évènements de fusion de gènes. En effet, deux protéines d'un organisme peuvent être en interaction si elles sont également présentes dans un autre organisme sous la forme de deux domaines d'une seule protéine (Yanai *et al.*, 2001).

Les interactions protéine-protéine peuvent être conservées entre les organismes proches (Walhout *et al.*, 2000), on parle alors d'interologues (issue de la combinaison d'interaction et d'orthologue). En utilisant cette notion, on peut alors prédire des interactions en recherchant les interactions existantes dans des organismes proches.

2.1.4 Bases de données d'interactions protéine-protéine

Les différentes méthodes mentionnées ci-dessus permettent de mettre en évidence de plus en plus d'interactions protéine-protéine. Devant ce nombre croissant de données, de nombreuses sources de données se développent afin de mettre les informations sur les interactions protéine-protéine à la disposition des biologistes. L'une des plus importantes est certainement la base de données IntAct[2] qui recense les interactions protéine-protéine décrites dans la littérature (Kerrien *et al.*, 2012). Une autre source de données est la base STRING[3] qui reporte les interactions protéine-protéine élucidées expérimentalement ou prédites (Figure 2.1).

2.1.5 Autres réseaux biologiques

L'interactome peut être divisé en modules. Ainsi, Alberts (1998) compare ces sous-réseaux à des machines, dans lesquelles les protéines sont organisées en modules de manière à réaliser des fonctions précises. Au sein d'un module, les protéines sont fortement connectées entre elles tandis que les interactions avec des membres extérieurs au module sont plus rares. Par exemple, on peut

2. www.ebi.ac.uk/intact
3. http://string-db.org

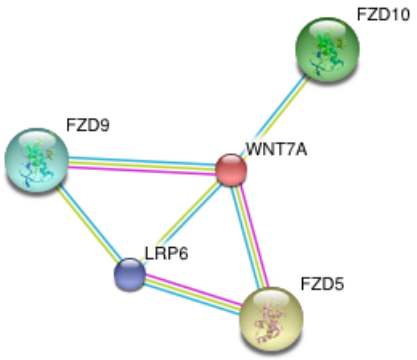


FIGURE 2.1 – Réseau d'interaction de la protéine WNT7A, extrait de la base de données String. Les relations en rose ont été élucidées expérimentalement, les relations bleues sont extraites d'autres bases de données et les relations vertes ont été prédites par fouille de texte.

remarquer sur la figure 2.2 qu'au sein du module correspondant au métabolisme du galactose il existe seulement 5 connections avec d'autres modules. Ainsi, cette organisation peut être utilisée pour définir des réseaux biologiques. Ces réseaux peuvent être définis comme un ensemble d'interactions entre des composants physiques ou génétiques de la cellule, décrivant un processus de cause à effet ou dépendant du temps, et expliquant des phénomènes biologiques observables (Demir *et al.*, 2010). Différents types de processus sont décrits par ces réseaux : la régulation de gènes, le transport de molécules, la transformation de petites molécules ou une interaction entre protéines entrainant la modification de l'une d'entre elle (Schaefer *et al.*, 2009). Ainsi, trois groupes de réseaux sont couramment définis : voies métaboliques, voies de signalisation et voies de régulation. Les voies métaboliques décrivent le métabolisme de la cellule, comme par exemple le métabolisme du galactose (Figure 2.2), et font intervenir des petites molécules qui sont modifiées par les réactions chimiques réalisées par des protéines. Les voies de signalisation correspondent à la perception de signaux cellulaires ou extracellulaires, puis à la transduction du signal dans la cellule. Le troisième groupe de voies est associé à la régulation des gènes. Le plus souvent ces 3 voies sont connectées : la cellule perçoit un signal qui sera transmis jusqu'au noyau et entrainera une modification de la régulation de gènes, ce qui pourra provoquer des changements du métabolisme de la cellule.

2.1.6 Les ressources sur les réseaux biologiques

Il existe plusieurs ressources contenant des réseaux biologiques connus. Les informations qui suivent correspondent à l'état de ces ressources en avril 2013. L'une des plus connues est certainement KEGG PATHWAY [4] contenant 496 réseaux construits manuellement et repartis en 6 catégories : métabolisme, traitement de l'information génétique, traitement des informations environnementales (signalisation), processus cellulaires, systèmes physiologiques et maladies humaines (Ogata *et al.*, 1998). Reactome [5] est une autre source disponible pour décrire les réseaux biologiques (Joshi-Tope *et al.*, 2005). Elle contient 1402 réseaux regroupés en 22 grands groupes et est peuplée manuellement. Une troisième source, BioCarta [6], répertorie 354 réseaux vérifiés manuellement et repartis

4. www.genome.jp/kegg/pathway.html
5. www.reactome.org
6. www.biocarta.com/genes/index.asp

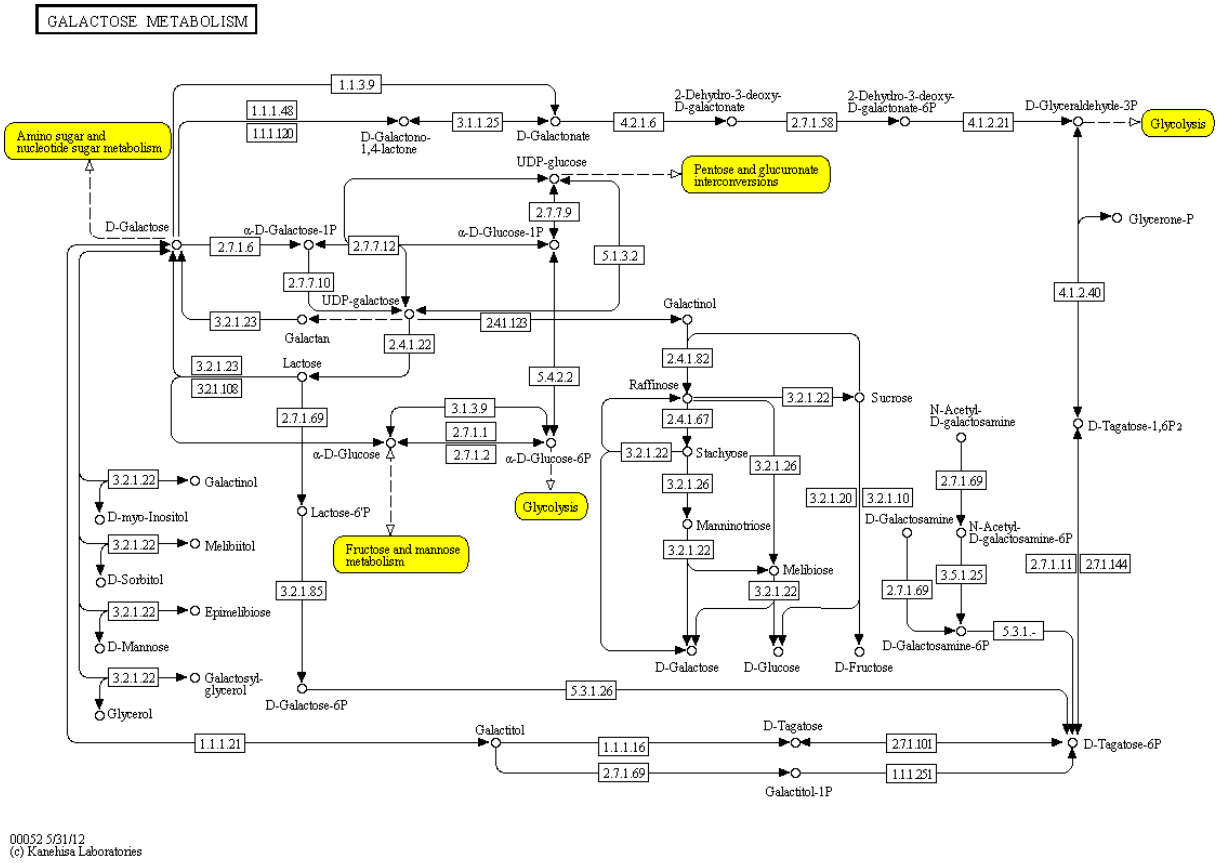


FIGURE 2.2 – Voie métabolique KEGG de référence correspondant au métabolisme du galactose. Les protéines sont représentées sous la forme de rectangles contenant le numéro EC ("Enzyme Commission") de la protéine et les petites molécules sont représentées par des cercles. Les éléments en jaune sont les voies métaboliques d'autres sucres.

en 22 fonctions. La Pathway Interaction Database (PID) est une base de données intégrée. En effet, elle contient 1367 réseaux vérifiés par le groupe NCI-Nature ainsi que 322 réseaux provenant de Reactome et BioCarta. Contrairement aux autres bases qui sont plus générales, PID se concentre sur les voies de signalisation et de régulation (Schaefer *et al.*, 2009) et n'intègre donc que ces deux types de réseaux. Il existe également la ressource BioModels [7] qui contient des modèles informatiques de processus biologiques (Li *et al.*, 2010). C'est à dire des représentations formalisées (par exemple dans le langage SBML) et quantifiées (en vue d'une simulation dynamique de processus) tels que le cycle cellulaire ou les cascades MAP kinases.

2.1.7 Conclusion

Les protéines notamment via leurs interactions ont un rôle majeur dans le fonctionnement des cellules. Les interactions protéine-protéine sont donc particulièrement étudiées et les méthodes de mise en évidence expérimentales et informatiques de ces interactions produisent de plus en plus de données. L'organisation de ces interactions permet de définir des réseaux biologiques qui expliquent

7. www.ebi.ac.uk/biomodels-main

des phénomènes cellulaires tels que la transduction de signaux au sein de la cellule ou la régulation des gènes. Il est important d'aider l'utilisateur biologiste à exploiter la richesse et la profusion des données disponibles. De nombreux formats permettent de décrire ces réseaux et des outils de visualisation des réseaux sont développés pour faciliter l'analyse de ces réseaux d'interactions.

2.2 Représentation et visualisation des réseaux biologiques

2.2.1 Formats de représentations des réseaux

Il existe plusieurs formats permettant la représentation de réseaux biologiques. Certains ont été développés spécifiquement pour les réseaux biologiques (BioPAX, SBML, ...) alors que d'autres sont plus généraux pour représenter des graphes quels qu'ils soient (OXL, GraphML, ...). Notons que la plupart de ces formats ne se limite pas à la représentation des réseaux en tant que tel mais permet la représentation des données sur les réseaux soit sous forme d'attributs des nœuds, soit sous forme de nœuds décrivant les propriétés. Dans ce dernier cas, nous parlerons alors de graphe étendu.

2.2.1.1 Formats de réseaux biologiques

2.2.1.1.1 BioPAX : BioPAX est un modèle XML spécifiquement développé pour permettre l'intégration, l'échange, la visualisation et l'analyse des données de réseaux biologiques (Demir *et al.*, 2010). Il est divisé en trois niveaux : le premier décrit seulement les voies métaboliques, le deuxième prend en compte les voies de signalisation et les interactions moléculaires en plus des voies métaboliques et le troisième niveau décrit les réseaux de régulation génique et les interactions génétiques. BioPAX est basé sur une ontologie de concepts avec des attributs, ce qui permet d'avoir des relations entre concepts plus explicites que pour les autres formats (Pavlopoulos *et al.*, 2008). Ainsi, le type d'interactions des entités physiques (en ajoutant les gènes) est particulièrement bien renseigné. Par exemple, il est possible de faire la différence entre une interaction de type dégradation et de type transport (Figure 2.3). De plus, BioPAX a été développé de manière à être compatible avec les formats existants comme SBML dans leurs domaines d'applications communs (Demir *et al.*, 2010).

2.2.1.1.2 SBML : Le Systems Biology Markup Language (SBML) est un modèle XML décrivant de manière qualitative et quantitative les modèles de réseaux biochimiques (Finney et Hucka, 2003). En effet, il est orienté vers la description des systèmes dans lesquels des entités biologiques sont impliquées et modifiées par des processus au fil du temps. Il contient notamment des éléments permettant de décrire la fonction mathématique associée au modèle ainsi que les réactions et leurs paramètres entre les espèces réagissant. Ainsi, SBML est particulièrement bien adapté pour modéliser les voies de signalisation cellulaire, des voies métaboliques et les régulations géniques.

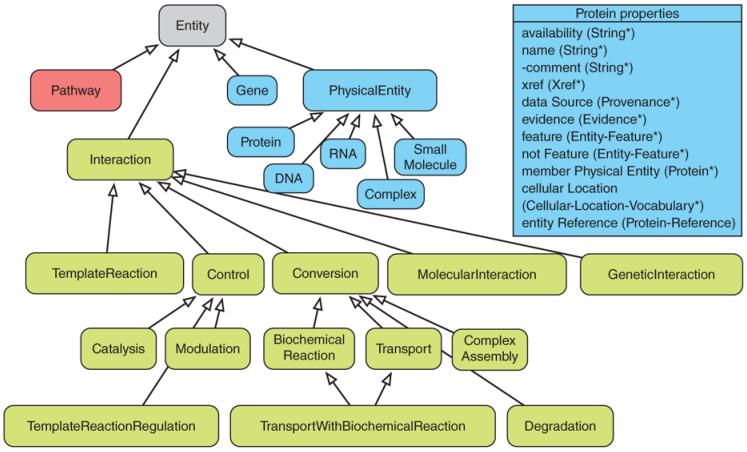


FIGURE 2.3 – Entités composant BioPAX. Les quatre types de classes composant BioPAX sont les réseaux biologiques (en rouge), les interactions (en vert) et les entités physiques avec les gènes (en bleu). Les flèches représentent les relations entre les entités BioPAX. La figure est extraite de Demir *et al.* (2010).

2.2.1.2 Formats de graphes

2.2.1.2.1 GraphML : GraphML est un modèle XML développé spécifiquement pour les graphes (Brandes *et al.*, 2001). Ainsi, ce modèle a été pensé de façon à décrire tous les types de graphes (orientés, non orientés, mixtes, hiérarchiques et hypergraphes). De manière générique, GraphML est composé d'entités nœuds reliées entre elles par des arcs, un arc étant caractérisé par un nœud "source" et un nœud "cible".

2.2.1.2.2 OXL : Le format OXL est un format spécifique à Ondex. Bien que le programme Ondex soit plutôt dédié à l'analyse de données biologiques (Kohler *et al.*, 2006), le format OXL a été développé de manière à couvrir un grand nombre d'applications. Il est ainsi suffisamment flexible et extensible pour combiner différents types de données (Taubert *et al.*, 2007).

OXL est défini comme un schéma XML composé de deux grands types d'éléments : *ondexdataseq* et *ondexmetadata*. *Ondexdataseq* décrit les éléments composant le graphe, tandis qu'*ondexmetada* contient la liste de tous les types de métadonnées utilisées dans le graphe.

Ondexdataseq est composé de 2 groupes d'éléments : les concepts et les relations, représentant respectivement les nœuds et les arcs du graphe (Figure 2.4). Les concepts sont caractérisés par un identifiant unique (*id*), un identifiant textuel alternatif (*pid*), des annotations et une description. La base de données source du concept est représentée par *elementOf* et son type par *ofType*. La méthode de mise en évidence du concept est également stockée. Le(s) nom(s) du concept sont collectés dans *concept_name* et la liste de leurs différents identifiants dans des bases de données dans *concept_accession*. Il est également possible de définir des attributs comme la date de collecte ou l'organisme source. Cette liste des attributs "personnalisés" correspond à *concept_gds*. Les relations entre concepts sont décrites comme allant d'un concept source *fromConcept* vers un concept cible *toConcept*. Ces relations sont précisées par leur type *ofType* (interaction, traduction, ...). Comme les concepts, il est possible de définir des attributs supplémentaires (*relation_gds*) pour

mieux caractériser ces relations.

```
<ondexdataseq>
 <concepts> <!-- Noeuds -->
  <concept> <!-- Premier concept -->
   <id>1</id>
   <pid>
    Aldo-keto reductase family 1 member C1
   </pid>
   <annotation/>
   <description/>
   <elementOf> <!-- BD source -->
    <idRef>Uniprot</idRef>
   </elementOf>
   <ofType> <!-- Type de donnée -->
    <idRef>Protein</idRef>
   </ofType>
   <evidences> <!-- code d'évidence -->
    <evidence>
  <idRef>IMPD</idRef>
    </evidence>
   </evidences>
   <conames> <!-- Noms du concept -->
    <concept_name>
     <name>
      Aldo-keto reductase family 1 member C1
     </name>
    </concept_name>
   </conames>
   <coaccessions> <!-- Identifiants dans BDs -->
    <concept_accession>
     <accession>Q04828</accession>
     <elementOf>
      <idRef>Uniprot</idRef>
     </elementOf>
    </concept_accession>
   </coaccessions>
   <cogds> <!-- Attributs personalisables-->
    <concept_gds> <!-- Organisme -->
     <attrname>
      <idRef>Organism</idRef>
     </attrname>
      <literal>Homo sapiens</literal>
     </value>
    </concept_gds>
   </cogds>
   <contexts/>
  </concept>
  <concept> <!-- Second concept -->
   <id>2</id>
   <!-- Autre concept similaire au concept précédent (par exemple une autre protéine) -->
  </concept>
 </concepts>
 <relations> <!-- Relations -->
  <relation>
   <fromConcept>1</fromConcept>
   <toConcept>2</toConcept>
   <ofType>
    <idRef>int_with</idRef>
   </ofType>
   <evidences>
    <evidence>
     <idRef>IMPD</idRef>
    </evidence>
   </evidences>
   <relgds>
    <relation_gds> <!-- Méthode de mise en évidence -->
     <attrname>
      <idRef>methode</idRef>
     </attrname>
     <value java_class="java.lang.String">
      <literal>two hybrid</literal>
     </value>
    </relation_gds>
   </relgds>
   <contexts/>
  </relation>
 </relations>
</ondexdataseq>
```

FIGURE 2.4 – Extrait d'un fichier OXL représentant les données *Ondexdataseq* et décrivant une interaction entre deux protéines

Les métadonnées (*ondexmetadata*) permettent de décrire les types de données utilisées dans le graphe. Elles sont caractérisées par un identifiant unique *id*, un nom *fullname* et une description *description* en texte libre et sont entièrement personnalisables par l'utilisateur (Figure 2.5). Il existe 6 types de métadonnées : les éléments de type *cvs* ("controlled vocabularies") représentent le vocabulaire décrivant les sources de données utilisées pour décrire les concepts et les relations. Les éléments *units* correspondent aux unités des propriétés des concepts et des relations. Les façons dont les concepts et les relations ont été mis en évidence (mise en évidence expérimentale, ...) sont représentées par les éléments de type *evidences*. Les éléments de types *attrnames* ("attribut names") qui sont des attributs définissables par l'auteur les *concept classes* et les *relation types* correspondent respectivement aux différents types de concepts et aux types de relations qui sont utilisées.

```
<!-- Métadonnées -->
<ondexmetadata>
 <cvs> <!-- bases de données -->
  <cv>
   <id>int</id>
   <fullname>IntAct</fullname>
   <description>
    Interactions protéine-protéine
   </description>
  </cv>
  <cv>
   <id>uni</id>
   <fullname>UniProt</fullname>
   <description/>
  </cv>
 </cvs>
 <attrnames>
  <attrname> <!-- Attributs personalisables -->
   <id>Organism</id>
   <fullname>Organism</fullname>
   <description>Organism</description>
  </attrname>
```

```
</attrnames>
<evidences> <!-- Codes de mise en évidence -->
 <evidence>
  <id>IMPD</id>
  <fullname>Imported from database</fullname>
  <description/>
 </evidence>
</evidences>
<conceptclasses> <!-- Type de concepts -->
 <cc>
  <id>prot</id>
  <fullname>Protein</fullname>
  <description/>
 </cc>
</conceptclasses>
<relation_types> <!-- Type de relations -->
 <relation_type>
  <id>int_with</id>
  <fullname>interact with</fullname>
  <description/>
 </relation_type>
</relation_types>
</ondexmetadata>
```

FIGURE 2.5 – Extrait des métadonnées d'un fichier OXL. Cet extrait correspond aux métadonnées utilisées pour décrire des IPP

2.2.1.2.3 RDF : Le Resource Description Framework (RDF) est un modèle d'échange de données sur le Web développé par le W3C [8]. Un graphe RDF est composé de triplets (sujet, prédicat, objet) ou le sujet est la ressource à décrire, le prédicat est la propriété associée au sujet et l'objet correspond à la valeur de la propriété. Ainsi, le triplet correspondant à une interaction entre la protéine A et la protéine B est (A, interagit avec, B) ou (B, interagit avec, A).

2.2.2 Outils de visualisation

L'ensemble des interactions d'un organisme forme un réseau d'interaction protéine-protéine. Ce réseau est un outil important pour la compréhension des mécanismes cellulaires (Agapito *et al.*, 2013). Étant donné que la visualisation des réseaux peut permettre de mettre en évidence des sous-structures intéressantes, comme des complexes protéiques, la visualisation de réseaux sous forme de graphes est particulièrement répandue (Ciofani *et al.*, 2012, Cheng *et al.*, 2012, Agapito *et al.*, 2013, Campillos *et al.*, 2008). Ainsi, de nombreux outils sont développés afin de filtrer et d'analyser les réseaux biologiques. Ces outils varient notamment sur les formats de données qu'ils utilisent, le mode de visualisation, la possibilité d'interroger des sources de données distantes, la possibilité d'annoter un réseau existant et la capacité de l'utilisateur à développer de nouvelles fonctionnalités au programme via des extensions.

2.2.2.1 Arena3D

Arena3D est un programme de visualisation de réseaux biologiques en trois dimensions (3D) (Secrier *et al.*, 2012). Afin de faciliter la visualisation en 3D, chaque type d'élément (protéine, gène, maladie, ...) est affiché à une hauteur différente sur le graphe (Figure 2.6). Ce programme permet d'importer des fichiers SBML, PSI-MI et TXT. Et il est également capable d'interroger des bases de données distantes, telles que STRING pour les interactions protéine-protéine, OMIM pour les informations concernant les maladies génétiques, PDB pour les informations sur les structures et Gene Ontology pour les annotations fonctionnelles. Il est cependant impossible d'ajouter manuellement des annotations aux nœuds ou aux arcs. De plus, Arena3D ne permet pas le développement d'extensions afin de d'éviter des problèmes d'incompatibilité de technologies pouvant affecter les performances du programme (Agapito *et al.*, 2013).

8. www.w3.org/RDF

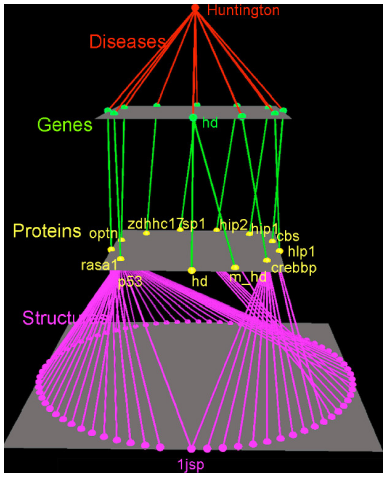


FIGURE 2.6 – Exemple de visualisation 3D avec Arena3D. Le réseau représente les gènes, protéines et structures protéiques associés à la maladie de Huntington et chaque type d'élément est représenté à une hauteur différente sur le graphe. La figure est extraite de Pavlopoulos *et al.* (2008).

2.2.2.2 BioLayout express3D

BioLayout express3D est spécialement conçu pour la visualisation en 3D et l'analyse de grands réseaux de données biologiques (Theocharidis *et al.*, 2009). Ce programme est compatible avec un grand nombre de format de fichiers tels que : le format OWL de la base de données Reactome, les fichiers EXPRESSION, MATRIX, GraphML, mEPN, OXL, LAYOUT, SIF et TXT. Il ne permet pas de formuler des requêtes vers des bases de données externes. BioLayout express3D permet à l'utilisateur d'annoter les arcs notamment pour permettre de relier des nœuds grâce à des arcs ayant différentes annotations Pour les même raisons qu'Arena3D, BioLayout express3D ne permet pas le développement d'extensions.

2.2.2.3 Cytoscape

Cytoscape est l'un des outils de visualisation 2D de réseaux les plus populaires (Smoot *et al.*, 2011, Cheng *et al.*, 2012, Agapito *et al.*, 2013). Ce programme permet de visualiser des réseaux allant jusqu'à des centaines de milliers de nœuds et de liens (Figure 2.7). Le principal objectif de Cytoscape est la visualisation d'interactions moléculaires et leur intégration avec des profils d'expression génique ou d'autres données.

Cytoscape permet d'importer des réseaux existant sous différents formats : XML, RDF, OWL, GraphML, XGMML et SBML. Cytoscape permet également d'importer des réseaux à partir de bases de données distantes notamment via IntAct et les bases du NCBI. Cytoscape est capable d'importer des données locales ou distantes (notamment GO) pour annoter les réseaux affichés. Il existe de nombreuses extensions développées par les utilisateurs, certaines permettant notamment de calculer l'enrichissement en termes GO du graphe étudié ou d'un sous graphe (Maere *et al.*, 2005). Il existe également d'autres extensions permettant d'interroger des bases de données distantes afin d'enrichir le réseau affiché ou pour simplement créer un nouveau réseau, comme par exemple l'extension *IntActWSClient* qui permet d'interroger la base de données IntAct. De ce point de vu,

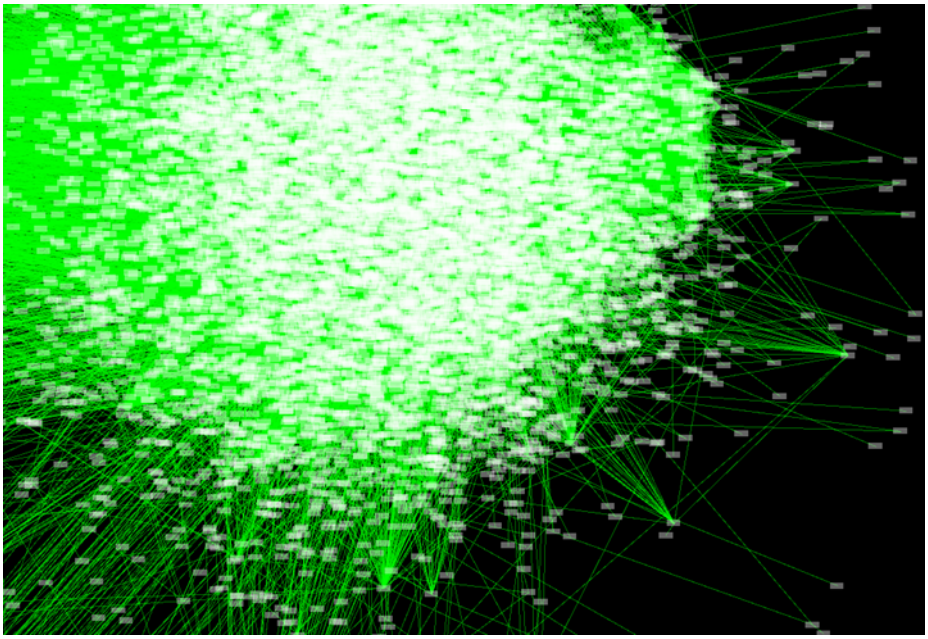


FIGURE 2.7 – Extrait de l'interactome humain (construit par Rual *et al.* 2005 et disponible à http://wiki.cytoscape.org/Data_Sets) sous Cytoscape. Au total, 10 203 protéines présentant 61 262 interactions sont affichées.

l'une des extensions les plus utiles est certainement *BisoGenet* (Martin *et al.*, 2010) puisqu'a partir d'une liste de gènes ou de protéines, cette extension va interroger la base de données distante SysBiomics [9] qui intègre le contenu des bases de données d'interactions DIP, HPRD, IntAct, BioGRID, MINT et BIND. En utilisant *BisoGenet* à partir d'une liste de protéines, on peut obtenir l'ensemble de leurs interactants connus de rang 1, 2 ou 3. De plus, pour chaque élément affiché, les termes GO ainsi que les réseaux biologiques KEGG associés sont également présentés. Il est ensuite possible de rechercher le plus court chemin entre deux protéines du réseau.

2.2.2.4 Ondex

Ondex est un outil d'intégration et de visualisation de graphe 2D dédié aux données biologiques (Kohler *et al.*, 2006). Il est capable d'importer les fichiers au format OXL (qui lui est spécifique), SBML et BioPAX. Ondex est développé comme un outil d'intégration de données biologiques. Il permet donc d'importer des données provenant de nombreuses sources telles que : Aracyc2, ChEBI, ChEMBL, EXPASY ENZYME, Gene Ontology, KEGG, MedLine et UniProt. L'utilisateur peut annoter manuellement ou à partir de sources externes chaque nœud et chaque arc du graphe affiché. Il est possible d'ajouter des extensions à Ondex principalement pour de nouvelles fonctionnalités d'intégration de données.

9. http://biomine.cigb.edu.cu/sysbiomics

2.2.3 Discussion

De nombreux formats existent pour décrire les réseaux biologiques. Parmi les formats existants, certains sont développés spécifiquement pour les données biologiques. Par exemple, BioPAX est particulièrement adapté pour décrire les réseaux biologiques (Demir *et al.*, 2010) et SBML est idéal pour modéliser des réseaux biochimiques (Finney et Hucka, 2003). Ainsi, ces formats sont bien adaptés et complémentaires, car développés pour décrire différentes facettes des mécanismes cellulaires. Cependant, ils ne permettent pas l'ajout de nouveaux concepts. Par exemple, il est impossible d'ajouter un concept maladie à ces formats si on veut décrire les effets d'une maladie génétique. D'un autre côté, les formats "classiques" de description de graphes, tel que GraphML, sont beaucoup trop génériques et ne prennent pas en compte les propriétés biologiques de chaque élément. Ainsi, un nœud GraphML est seulement un nœud et pas une protéine ou un gène annoté. Une solution à ce problème est d'utilisation du format OXL qui est entièrement personnalisable selon les besoins de l'utilisateur en permettant de définir différents types de concepts et de relations. De façon plus générale, une autre solution à ce problème est l'utilisation de triplets RDF qui permettent à la fois de décrire des phénomènes biologiques et d'ajouter de nouveaux types de données en cas de besoin.

Les programmes de visualisation des réseaux biologiques sont assez nombreux. S'il est possible de les différencier par les formats d'entrée qu'ils utilisent ou par le type de visualisation 2D/3D, il est plus intéressant de les étudier selon leurs possibilités d'interroger des bases de données externes et leurs possibilités d'annoter les réseaux existants afin d'en faciliter l'exploitation (Table 2.1). En effet, pour les utilisateurs occasionnels, le fait d'intégrer des bases de données externes est un grand plus car cela permet d'éviter les étapes d'intégration des données puis de leur formatage qui dépend du programme de visualisation utilisé. Selon ces critères, Cytoscape est certainement le programme le plus utile, notamment par le grand nombre de bases externes qu'il peut interroger mais aussi par les nombreuses extensions qui existent permettant notamment de faciliter l'analyse de réseaux. Ondex est également intéressant puisqu'il possède des fonctionnalités d'intégration de données. De plus, son format de fichier OXL est particulièrement utile car générique et particulièrement bien adapté pour décrire les données biologiques.

Programme	Visualisation	Intégration de données	Annotation des données	Ajout d'extensions
Arena3D	2D/3D	oui	non	non
Biolayout Express3D	2D/3D	non	oui	non
Cytoscape	2D	oui	oui	oui
Ondex	2D	oui	oui	oui

TABLE 2.1 – Comparaison des outils de visualisation des réseaux biologiques selon le type de visualisation, les possibilités d'intégrer des sources de données distantes, la capacité de l'utilisateur à annoter les éléments du graphe et la possibilité de développer/utiliser des extensions au programme.

2.3 Utilisation des réseaux biologiques

Les réseaux d'interaction biologiques forment la base des processus cellulaires. De plus en plus d'études les prennent donc en compte afin de mieux appréhender des phénomènes complexes.

Ainsi, ils sont de plus en plus utilisés pour étudier les maladies génétiques et plus particulièrement pour rechercher des gènes candidats pour expliquer ces maladies. Ils sont également particulièrement utiles pour la conception de nouveaux médicaments. En effet, ces réseaux permettent d'identifier et de rechercher des cibles potentielles lors du développement de nouveaux médicaments ou pour repositionner un médicament déjà sur le marché dans une autre indication. Grâce à eux on peut également espérer obtenir une meilleure compréhension des effets indésirables des médicaments.

2.3.1 Étude des maladies génétiques

2.3.1.1 Introduction

Les maladies dites génétiques sont causées par une ou plusieurs anomalies dans le génome. L'origine de ces maladies est assez variée : il peut s'agir d'une "simple" modification de la séquence d'ADN (substitution, insertion, délétion) d'un ou plusieurs gènes ou encore d'un réarrangement chromosomique (duplication de séquence génomique, anomalies du nombre de chromosomes ou de structure). Il est possible de séparer ces maladies en plusieurs groupes selon les mécanismes impliqués.

Si la maladie à pour origine un gène unique, on parle de maladie monogénique. Les maladies monogéniques peuvent être reparties en 6 sous-groupes selon leur mode de transmission. Le premier d'entre eux concerne les gènes à transmission autosomique dominante. Dans ce cas, le gène est situé sur un chromosome non sexuel et un seul allèle muté est suffisant pour provoquer la maladie. Par exemple, le syndrome de Marfan (MIM :154700) est provoqué par la présence d'un seul allèle muté du gène *FBN1*. Le second mode de transmission est autosomique récessif, la maladie n'apparaissant alors que si les 2 allèles d'un gène situé sur un chromosome non sexuel sont mutés. Ainsi, deux versions mutées du gène *CFTR* sont nécessaire pour provoquer la mucoviscidose(MIM :219700). Le troisième groupe correspond aux gènes à transmission dominante liée au chromosome X. Les maladies associées à ce mode transmission sont provoquées par la présence d'un seul allèle portant la mutation. Les garçons comme les filles peuvent être affectés par ce type de maladie. Cependant, dans certain cas comme le syndrome de Rett (MIM :312750), la maladie est létale chez la plupart des garçons ce qui provoque une prédominance des observations chez les filles. Le quatrième groupe concerne les gènes récessifs situés sur le chromosome X. Du fait de la présence d'un seul chromosome X chez les garçons, ceux ci sont plus fréquemment affectés. L'une des affections récessives liées au chromosome X les plus connues est le daltonisme (MIM :303800) qui implique le gène *OPN1MW*. Les gènes du chromosome Y sont associés au quatrième groupe. Du fait du faible nombre de gènes sur ce chromosome (50 gènes codant une protéine [10]), seuls 4 phénotypes lui sont associés dans OMIM. Le dernier mode de transmission concerne les gènes mitochondriaux (transmission mitochondriale). Ce mode de transmission est également appelé transmission maternelle puisque les mitochondries sont uniquement transmises par la mère. A ce jour, 28 phénotypes ayant une transmission mitochondriale sont recensés dans OMIM. Il est important de noter qu'une maladie monogénique peut être provoquée par des mutations dans plusieurs gènes différents à condition qu'une seule mutation soit suffisante pour provoquer la maladie. Ainsi, l'ami-

10. http://vega.sanger.ac.uk/Homo_sapiens/Location/Chromosome?r=Y

loïdose VIII (MIM :105200) peut être provoquée par une mutation soit dans le gène *FGA*, soit dans le gène *APOA1*, soit dans le gène *LYZ*.

Un maladie génétique peut également être provoquée par plusieurs gènes ainsi que par l'environnement. On parle alors de maladies multifactorielles. Certaines maladies sont provoquées par la présence de mutation "simultanée" dans deux gènes. Ainsi, Schaffer (2013) recense une cinquantaine de maladies dites digéniques. La première à avoir été mise en évidence est une forme de rétinite pigmentaire (MIM :608133) et implique des mutations dans les gènes *PRPH2* et *ROM1* (Kajiwara *et al*., 1994). Dans la plupart des maladies multifactorielles, il existe un grand nombre de gènes pour lesquels certains SNP vont provoquer une plus grande susceptibilité d'apparition de la maladie. Par exemple, Chen *et al.* (2010), Neale *et al.* (2010) et Ricci *et al.* (2013) ont montrés les variations présentes dans 5 gènes sont associés à un plus grand risque de développer une dégénérescence maculaire liée à l'age.

L'un des chalenges de l'ère post-génomique est la compréhension des fonctions biologiques de gènes isolés ou des réseaux de gènes qui conduisent à l'apparition de maladies (Chen *et al*., 2008). L'ensemble des maladies monogéniques peut être représenté sous forme de graphe où les nœuds représentent les maladies (Goh et Choi, 2012). Deux maladies sont alors connectées si elles ont un gène en commun (Figure 2.8). On peut ainsi remarquer que les maladies humaines sont fortement interconnectées montrant ainsi une origine génétique commune à de nombreuses maladies. Cependant, si plusieurs maladies sont reliées par un gène commun, une maladie est rarement la conséquence de la perturbation d'un seul gène (Barabasi *et al*., 2011). Ainsi l'étude des réseaux biologiques peut permettre d'identifier de nouveaux gènes responsables de maladies et de mieux comprendre les mécanismes sous-jacents.

2.3.1.2 Recherche de gènes responsables de maladies

A ce jour, il existe environ 5500 maladies monogéniques répertoriées dans la base de données OMIM [11]. Pour environ 30% de ces maladies, les gènes responsables ne sont pas connus.

La recherche des gènes responsables d'une maladie se fait par l'établissement d'une liste de gènes candidats. Un gène candidat peut être un gène situé dans la région suspectée d'être responsable de la maladie ou possédant une fonction pouvant expliquer le phénotype de la maladie (Freudenberg et Propping, 2002). Freudenberg et Propping (2002) étendent cette définition en considérant que les maladies similaires peuvent être expliquées par des gènes similaires. Yilmaz *et al.* (2009) proposent une définition étendue d'un gène candidat comme un gène ayant une relation directe ou indirecte avec la maladie. On considère qu'il a une relation directe si ce gène est "colocalisé" avec la maladie c'est à dire s'il est localisé dans la région chromosomique associée à la maladie, s'il est observé comme étant dérégulé chez les malades ou s'il a une annotation fonctionnelle similaire à celle de la maladie. Par contre, la relation est considérée comme indirecte si un gène intermédiaire intervient. Celui-ci pouvant soit être un interactant soit un orthologue. Ainsi, un gène candidat ayant une relation indirecte avec la maladie peut être un gène dont l'orthologue a une annotation fonctionnelle similaire à celle de la maladie. Des définitions plus complexes peuvent encore être développées. Par exemple, un gène candidat peut être un gène qui a une annotation

11. http://omim.org

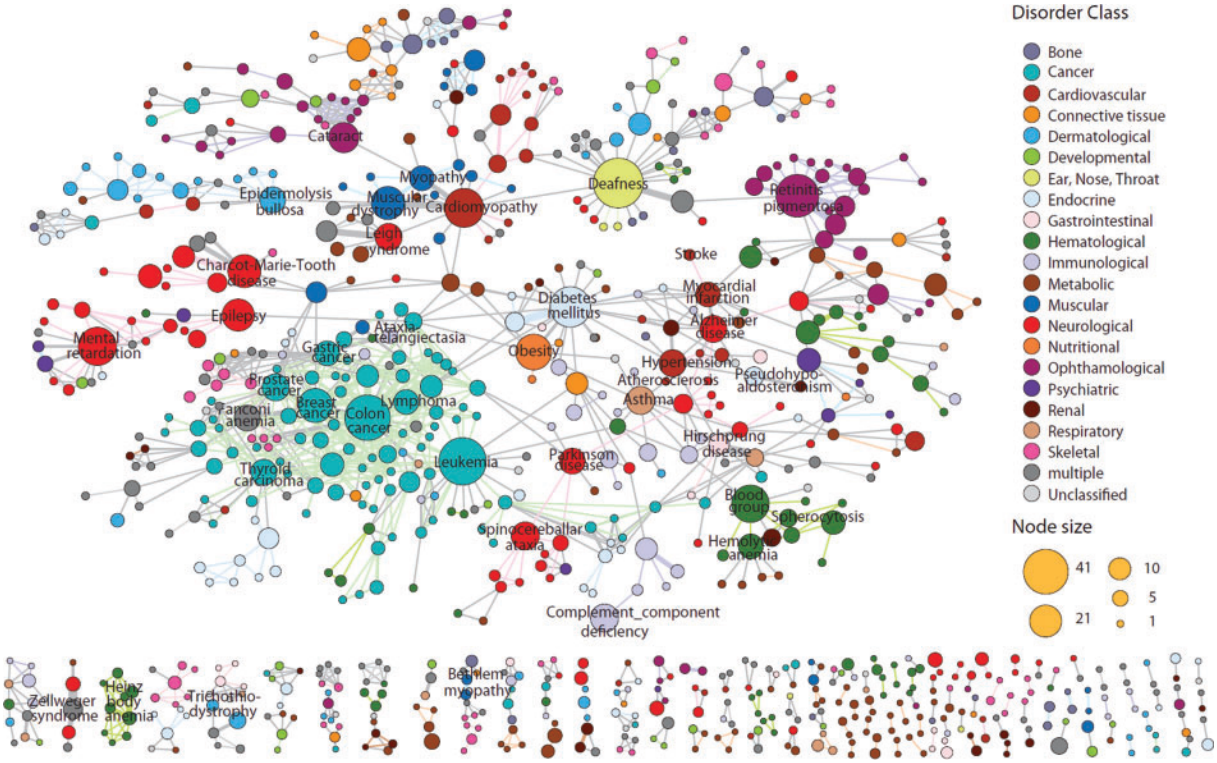


FIGURE 2.8 – Représentation du réseau des maladies humaines. Deux nœuds (maladies) sont reliés s'ils partagent un composant génétique selon la liste maladie-gène définie dans OMIM en 2005. La figure est extraite de Goh et Choi (2012).

fonctionnelle similaire à celle de la maladie et établie par des généticiens et qui interagit avec un gène dérégulé dans la maladie, ou encore un gène colocalisé sur le génome avec le locus de la maladie et qui est un orthologue d'un gène ayant une annotation fonctionnelle similaire à la maladie (Yilmaz *et al.*, 2009).

Des méthodes de priorisation permettant de déterminer des gènes candidats ont été développées. Par exemple, le programme ENDEAVOUR utilise un grand nombre de sources de données différentes afin de proposer des gènes candidats (Aerts *et al.*, 2006, Tranchevent *et al.*, 2008). Ainsi, des données d'annotations fonctionnelles, d'interactions protéine-protéine, d'expression, d'orthologie et de régulation sont intégrées. Des scores sont calculés pour chaque type de données puis une étape de "genomic data fusion" est appliquée afin de fusionner les scores de chaque méthode et ainsi obtenir un résultat global.

Cependant, ces études basées sur les réseaux d'interaction et les annotations fonctionnelles, sont grandement dépendantes de la qualité d'annotation des gènes (Piro et Di Cunto, 2012). Ainsi, on peut passer à côté d'un bon gène candidat si ce gène possède peu d'annotations. Pour éviter ce problème, Wagner *et al.* (2013) proposent d'intégrer différentes sources de données afin détecter les gènes associées aux rétinites pigmentaires. Pour cela, ils ont collecté des données de ChIP-seq, mRNA-seq et de microarray issues respectivement de la rétine de souris, de la rétine humaine et de profils d'expressions provenant de 10 tissus oculaires dont la rétine. Ensuite, à partir d'une liste de gènes comprenant des gènes connus pour être à l'origine de rétinite pigmentaires et d'autres

gènes, Wagner *et al.* (2013) réalisent un apprentissage automatique afin d'apprendre à séparer les deux types de gènes. Ils ont ensuite testé le classifieur ainsi obtenu sur un jeu de données de 13 gènes déjà connus et ont comparé les résultats avec ceux donnés par ENDEAVOUR. Ils montrent ainsi que leur système est plus performant et donc que l'utilisation de données expérimentales est également importante.

Ces deux méthodes de recherche de gènes candidats montrent bien l'intérêt d'utiliser des sources de données variées. Cependant, elles ne fonctionnent pas pour les maladies dont aucun gène responsable n'est connu. En effet, dans ces deux approches il est toujours nécessaire d'utiliser une liste de gènes associés à la maladie étudiée afin de pouvoir générer des candidats.

2.3.1.3 Réseaux biologiques au service de la compréhension des maladies génétiques

Alors que les méthodes de recherche de gènes candidats basées uniquement sur les polymorphismes nucléotidiques ne permettent pas de comprendre les mécanismes provoquant la maladie étudiée, les méthodes basées sur les annotations fonctionnelles ou sur des données d'expression permettent une meilleure compréhension des mécanismes associés à la maladie. En effet, on peut ainsi extraire des processus biologiques ou des gènes dont le fonctionnement est dérégulé lors de la maladie.

Les études d'associations pangénomiques ("genome wide association studies" ou GWAS) sont une stratégie populaire pour essayer d'identifier les sources génétiques des maladies multifactorielles Humaines. Ces études consistent à étudier des centaines de milliers de SNP sur un grand nombre de patients et de personnes saines, afin de pouvoir établir des corrélations entre les SNP et la maladie étudiée. Cependant, dans un très grand nombre d'études, seul un petit nombre des corrélations les plus évidentes gènes-maladie sont expliquées (Bakir-Gungor et Sezerman, 2011). Bakir-Gungor et Sezerman (2011) proposent d'utiliser les connaissances sur les réseaux biologiques dont les réseaux d'interactions afin d'enrichir l'analyse des données GWAS. Leur méthodologie est divisée en 3 étapes. La première consiste à associer les SNP aux gènes puis à calculer un score pour chaque SNP dépendant des propriétés fonctionnelles du transcrit, de l'effet du SNP sur la transcription, l'épissage, la traduction et les modifications post traductionnelles. Un score est calculé pour chaque gène en fonction de ses SNP. La deuxième étape consiste à rechercher des sous-réseaux actifs à partir d'un réseau d'interactions protéine-protéine. Un sous-réseau actif correspond à un ensemble de protéines (probablement associées à la maladie) qui interagissent physiquement et ont une forte chance d'appartenir au même réseau biologique. Cette étape est réalisée en utilisant l'extension *jActiveModule* de Cytoscape qui prend en compte à la fois les scores associés aux gènes et la topologie du réseau pour définir les sous-réseaux actifs (Ideker *et al.*, 2002, Bandyopadhyay *et al.*, 2006). La troisième et dernière étape correspond à l'enrichissement fonctionnel des sous-réseaux. Pour cela, ils recherchent les termes GO ainsi que les réseaux biologiques KEGG et BioCarta qui sont plus fréquemment retrouvés dans le sous-réseau que dans la population générale. Bakir-Gungor et Sezerman (2011) ont appliqué cette méthodologie à la polyarthrite rhumatoïde. Ils ont ainsi pu associer de nouveaux réseaux biologiques à cette maladie, ainsi que des nouveaux gènes appartenant à des réseaux identifiés comme associés à la maladie.

Au lieu d'intégrer interactions et réseaux biologiques, Curtis *et al.* (2012) ont récemment proposé

une approche utilisant des données d'expression pour comprendre les relations qui existent entre les SNP détectés par GWAS et les maladies. Les résultats obtenus sur un ensemble artificiel de données ont montré qu'en plus d'augmenter la capacité de détection des relations gènes-maladies, cette approche permet de mieux comprendre les mécanismes liant les SNP aux maladies en associant les SNP à des dérégulations.

2.3.2 Médicaments, cibles et effets secondaires

2.3.2.1 Étude des cibles de médicaments

Depuis l'avènement de la biologie moléculaire, la recherche de médicaments s'est essentiellement fondée sur l'hypothèse qu'une molécule ne se fixe que sur un seul récepteur. Ainsi les nouveaux médicaments sont le plus souvent des molécules qui se lient spécifiquement sur la cible désirée (Apic *et al.*, 2005). Cependant, cette approche est assez réductrice puisqu'il est maintenant admis qu'une molécule peut se lier à plusieurs cibles (Figure 2.9, Apic *et al.* 2005, Yildirim *et al.* 2007, Scheiber *et al.* 2009*a*, Pujol *et al.* 2010, Vogt et Mestres 2010). Ainsi, la recherche de nouvelles cibles pour des médicaments existants se développe sous le nom de "repositionnement moléculaire". Cette recherche intègre forcement des données sur les réseaux biologiques.

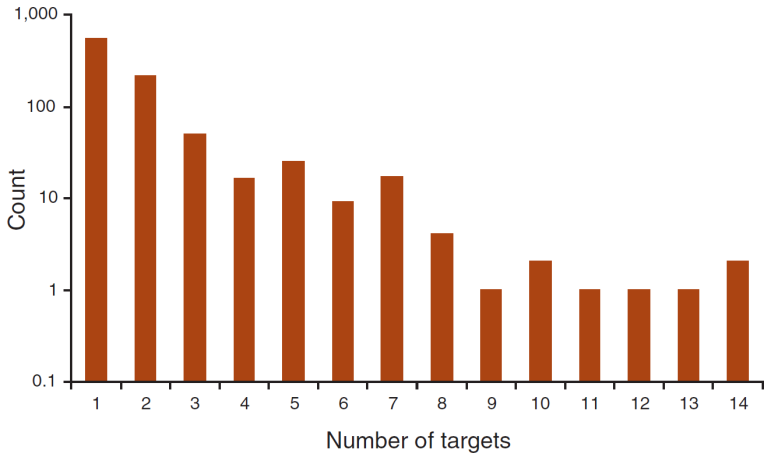


FIGURE 2.9 – Distribution des médicaments en fonction de leur nombre de cibles. Les médicaments et leurs cibles proviennent de DrugBank. La figure est extraite de Yildirim *et al.* (2007).

De nombreuses méthodes de prédiction de cibles de médicaments se basant sur différents types de propriétés ont ainsi été développées. Hopkins et Groom (2002) se basent sur la similarité de séquence entre les cibles pour proposer de nouvelles cibles à des médicaments. En partant du principe que la plupart des médicaments existants se fixent sur le site de liaison d'une molécule endogène, ils montrent que les domaines de liaison des cibles se regroupent en 130 familles et que près de la moitié des cibles représentent seulement 6 familles. Au sein d'une famille de protéines, la séquence et la fonction d'un site de liaison sont généralement bien conservées. Ceci suggère que si un membre d'une famille est capable de se lier à un médicament, alors les autres membres devraient également se lier à ce médicament ou à un autre structurellement similaire.

Au lieu de se baser uniquement sur la séquence des cibles, Li et Lai (2007) prennent en compte

les propriétés de cette séquence protéique. Chaque protéine est décrite par un vecteur à 146 dimensions représentants propriétés suivantes : composition en acide aminés, hydrophobicité, polarité, polarisabilité, charge, accessibilité au solvant, volume de van der Waals normalisé. Ce vecteur est calculé pour chaque protéine de 2 ensembles : un ensemble de protéines cibles et un autre ensemble de protéines non cibles. Ces deux ensembles servent de base d'apprentissage à un SVM (Support Vector Machine) qui permettra de reconnaitre les protéines qui peuvent servir de cibles à des médicaments. A l'issu d'un test de validation croisée (apprentissage sur 9/10 du jeu de donnée et test sur le dixième restant, répété pour les 10 dixièmes), le SVM obtient une précision de l'ordre de 84%. Les tests menés sur des jeux de données n'ayant pas servi à l'apprentissage ont confirmé les résultats obtenus montrant ainsi l'intérêt d'utiliser les propriétés de la séquence des cibles.

Afin de mieux comprendre les caractéristiques des cibles de médicaments, Ma'ayan *et al.* (2007) ont notamment étudié leur enrichissement en termes GO. Ainsi, à partir de 485 cibles issues de la base de données DrugBank, ils ont remarqué que ces cibles étaient enrichies en termes GO associés aux protéines membranaires, aux récepteurs, aux facteurs de transcription et aux composants des voies de signalisation cellulaire. Ainsi, les propriétés fonctionnelles des cibles semblent partagées par un grand nombre d'entre elles pouvant ainsi servir à déterminer de nouvelles cibles à des médicaments existant. Yao et Rzhetsky (2008) se sont intéressés à d'autres propriétés des cibles. Ils se sont notamment posé la question de savoir si les cibles des médicaments correspondent à des gènes plus sujets à des polymorphismes que la moyenne des gènes humains. A partir d'une liste d'environ 16 000 gènes, ils ont montré que les gènes "cibles" possèdent moins de polymorphismes. Ils expliquent cette observation par le fait qu'un grand nombre de polymorphismes dans une protéine cible pourrait diminuer la capacité du médicament à interagir avec cette cible. Yao et Rzhetsky (2008) ont également analysé les relations entre les médicaments et les tissus d'expression de leurs cibles. Ils ont ainsi remarqué que cinq tissus sont fréquemment ciblés (grandes endocrines, système nerveux central, l'appareil urinaire, les glandes excrétrices et les ganglions) alors que d'autres ne le sont que très rarement, notamment les tissus embryonnaires. Ce faible taux de ciblage de certains tissus pouvant s'expliquer par les risques importants d'effets secondaires liés à ces tissus. Par ailleurs, en étudiant le nombre de réseaux biologiques associés aux cibles des médicaments approuvés ou non par la "Food and Drug Administration" (FDA) (Figure 2.10), Sakharkar *et al.* (2008) montrent que le fait de cibler une protéine intervenant dans de multiples processus n'est pas une situation préférentielle pour un médicament.

Afin de prédire de nouvelles associations protéine médicament, Cheng *et al.* (2012) utilisent une méthode basée sur les réseaux d'interactions. Ils utilisent un score basé sur la topologie du réseau médicament-cible. De cette façon, ils obtiennent un score pour chaque couple médicament nouvelle protéine permettant ainsi d'obtenir une liste de cibles candidates à tester. Cheng *et al.* (2012) ont ensuite validé leur approche en testant et en confirmant *in vitro* leurs résultats pour 5 molécules et 2 cibles.

Les approches les plus originales pour la recherche de nouvelles cibles sont certainement celles de Campillos *et al.* (2008) et Takarabe *et al.* (2012). Ces deux approches se basent sur les effets secondaires des médicaments pour prédire de nouvelles cibles, leur hypothèse étant que le partage d'effets secondaires s'explique par des cibles communes. Campillos *et al.* (2008) se sont basés sur les notices des médicaments afin d'extraire les effets secondaires. Une valeur de similarité entre

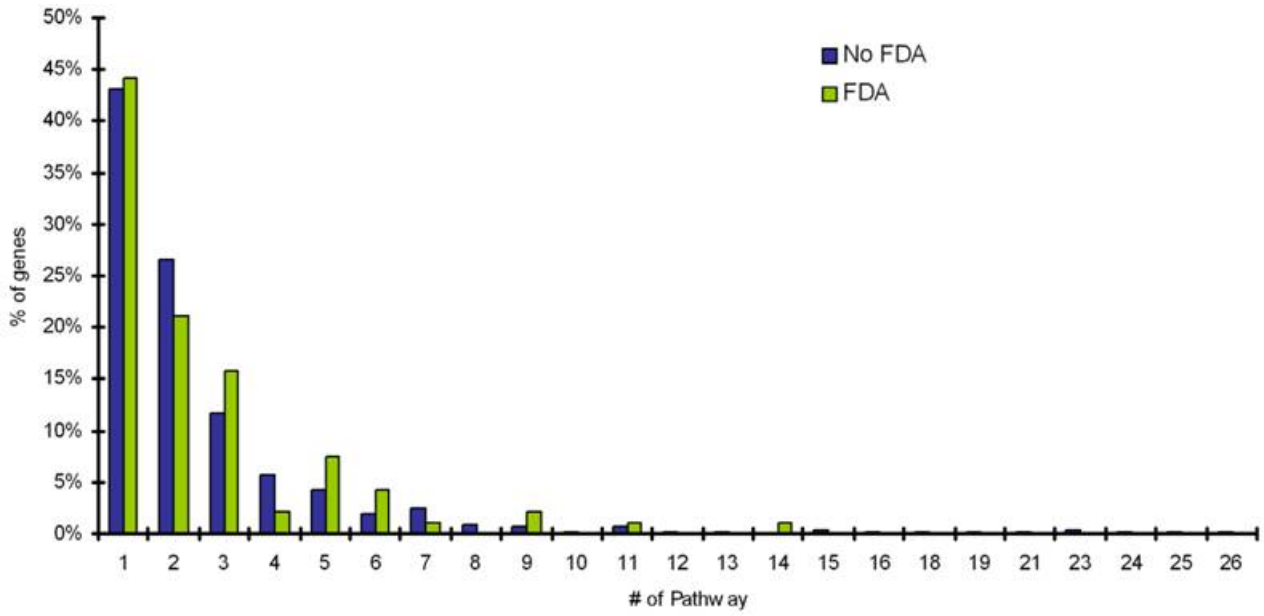


FIGURE 2.10 – Distribution du pourcentage de gènes cibles selon leur nombre de réseaux biologiques extraits à partir de SwissProt. La figure est extraite de Sakharkar *et al.* (2008).

les effets est ensuite calculée en se basant sur la co-ocurence des effets dans les médicaments (deux effets sont considérés comme similaires s'ils sont fréquemment associées aux mêmes molécules). Cette similarité en effets secondaires est ensuite combinée avec une mesure de similarité chimique à 2D afin de détecter de nouvelles cibles. Cette méthode a été testée sur 746 molécules et a permis de former 1018 couples de médicaments similaires fortement susceptibles de partager des cibles. Une vingtaine de ces couples découverts ont été vérifiés *in vitro* et pour 13 d'entre eux, des cibles communes ont été identifiées, confirmant ainsi l'intérêt de la méthode. L'approche utilisée par Takarabe *et al.* (2012) se base sur le système de signalement d'effets secondaire de la FDA (adverse event reporting system, AERS). AERS est un système permettant aux professionnels de la santé de rapporter les effets indésirables des médicaments observés chez des patients et ainsi d'mettre en évidence les effets secondaire de manière plus précoce que sur les notices. En utilisant AERS on est capable d'observer la présence d'un effet indésirable pour une molécule quand au moins un rapport en fait mention. Contrairement à cela, l'industrie pharmaceutique ne fait mention d'un effet secondaire pour une molécule que si cet effet a été observé un grand nombre de fois. Ainsi, Takarabe *et al.* (2012) ont comparé les prédictions de cibles basées sur une similarité d'effets secondaires AERS avec une méthode de prédiction basée sur la structure des molécules. Les résultats de cette comparaison montrent que l'utilisation des effets secondaires donne de bons résultats mais moins bons que ceux basés sur la structure. Cette observation suggère que l'utilisation des effets secondaire pour prédire les cibles peut-être utile quand on ne possède pas de données sur la structure des molécules, notamment comme cela arrive pour les peptides utilisés comme médicaments.

2.3.2.2 Étude des effets secondaires

De même qu'un médicament peut avoir plusieurs cibles, une cible peut intervenir dans plusieurs voies biologiques. L'utilisation des réseaux biologiques permet de replacer les cibles dans leur contexte et donc de mieux comprendre les effets indésirables de ces molécules (Pujol *et al.*, 2010). L'étude des effets secondaires est donc facilitée par l'utilisation de données des réseaux biologiques.

Le lien entre la structure des molécules et leurs mécanismes d'action ayant été établi (Martin *et al.*, 2002), les premières approches de prédiction des effets secondaires se sont basées sur les sous-structures des molécules (Bhavani *et al.*, 2006, Scheiber *et al.*, 2009*b*, Atias et Sharan, 2011, Pauwels *et al.*, 2011). Dans ces approches, les molécules sont décrites à la fois par un vecteur contenant leur composition (sous-structure) et un autre vecteur contenant leurs propriétés (effets secondaires). Ensuite, pour chaque effet secondaire, un apprentissage est effectué pour reconnaitre les molécules ayant cet effet. De plus ces méthodes ont permis d'identifier des sous-structures responsables des effets étudiés (Figure 2.11).

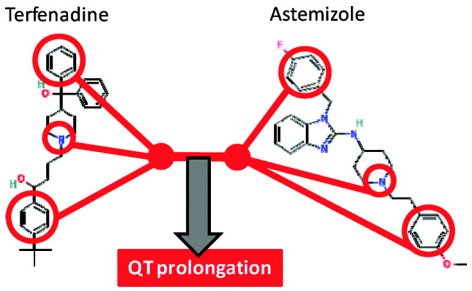


FIGURE 2.11 – Sous-structures moléculaires identifiées par Scheiber *et al.* (2009*b*) comme étant responsables d'une prolongation QT (arythmie cardiaque).

Comme Campillos *et al.* (2008) l'ont souligné, il existe des relations entre les effets secondaires et les cibles des médicaments. Afin de prendre en compte l'aspect molécule et l'aspect cible, Yamanishi *et al.* (2012) ont développé une méthode de prédiction des effets secondaires basée sur l'espace chimique et l'espace biologique. Comme précédemment, l'espace chimique correspond à une description des sous-structures des molécules. Quant à l'espace biologique, il est caractérisé par les cibles des médicaments (Figure 2.12). L'apprentissage nécessaire à la prédiction est réalisé sur 658 médicaments annotés par 969 effets secondaires et décrits par un vecteur chimique de 881 sous-structures et un autre vecteur de 1368 cibles. Les résultats de tests obtenus montrent l'intérêt relatif de combiner à la fois des données biologiques et chimiques pour prédire les effets secondaires (Table 2.2).

Au lieu de se baser sur la structure des médicaments, d'autres études ont préféré prendre en compte les processus biologiques associés à ces derniers (Lee *et al.*, 2011, Huang *et al.*, 2011). Ainsi, Lee *et al.* (2011), ont utilisé les processus biologiques associés aux médicaments par la ressource "Connectivity map" (Lamb *et al.*, 2006). Ces données ont permis de déterminer quels processus étaient dérégulés par les médicaments. Lee *et al.* (2011) proposent ensuite que les médicaments partageant des effets secondaires partagent également des processus biologiques (Figure 2.13). Cette approche a été testée sur l'effet secondaire éruptions cutanées ("rash") et les résultats

Descripteurs	AUPR
Aléatoire	0.04
Chimie	0.18
Biologie	0.19
Chimie + biologie	0.21

TABLE 2.2 – Performance de prédiction basée sur une cross validation à 5 itérations. AUPR (Area Under the Precision-Recall curve) est la surface sous la courbe précision-rappel. La ligne aléatoire correspond aux résultats attendu si la classification est réalisée de manière aléatoire. Les données sont extraites de Yamanishi *et al.* (2012).

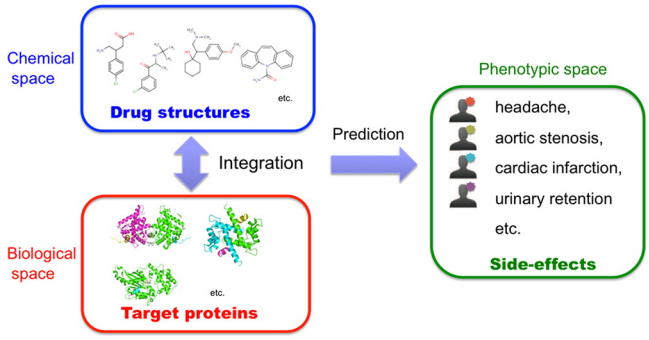


FIGURE 2.12 – Méthodologie proposée par Yamanishi *et al.* (2012) pour prédire les effets secondaires.

ont été comparés à ce qui était déjà connu dans la littérature. Cela a permis monter qu'un grand nombre des processus biologiques découverts par leur méthode étaient déjà reportés comme liés à l'effet secondaire dans la littérature. Au lieu de se baser sur les termes GO associés directement

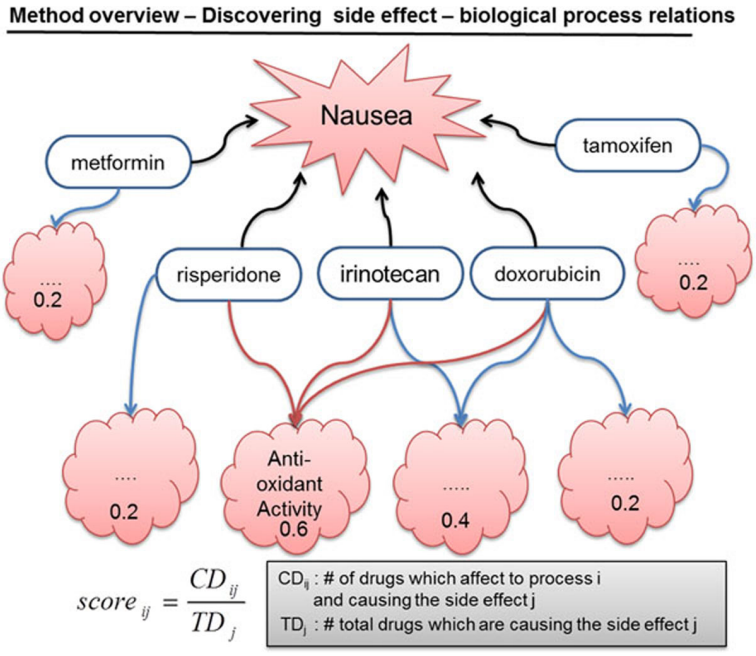


FIGURE 2.13 – Méthodologie proposée par Lee *et al.* (2011) pour associer effets secondaires et processus biologiques.

aux médicaments comme l'ont fait Lee *et al.* (2011), Huang *et al.* (2011) prennent en compte les cibles ainsi que leurs termes GO. Ainsi, de manière similaire à Lee *et al.* (2011), ils montrent que les fonctions associées aux cibles sont fortement corrélés l'apparition d'effets secondaires.

Les travaux de Liu *et al.* (2012*b*) sont particulièrement intéressants car ils combinent les caractéristiques phénotypiques des médicaments (indication et effets secondaires autres que celui étudié), la structure chimique et des propriétés biologiques (cibles et réseaux biologiques). Cette étude a permis de montrer que la combinaison de toutes ces données permet d'augmenter la qualité des prédictions. Il est cependant important de noter que l'utilisation des effets secondaires pour en prédire d'autres augmente fortement les résultats de la prédiction. Ceci pourrait s'expliquer par le fait que les effets secondaires ne sont pas indépendants entre eux (Liu *et al.*, 2012*b*).

2.4 Conclusion

Les réseaux biologiques sont de plus en plus utilisés pour comprendre des phénomènes biologiques complexes, comme les maladies génétiques et la compréhension de l'action des médicaments. De nombreuses ressources publiques mettant à disposition les données concernant les réseaux biologiques sont disponibles. De nombreux formats de description de ces réseaux existent et sont tous plus ou moins spécifiques à un problème donné (Pavlopoulos *et al.*, 2008). A cela s'ajoute différents programmes permettant de visualiser et d'analyser ces réseaux (Agapito *et al.*, 2013). Ainsi, devant cette multitude de ressources et d'outils, l'utilisateur biologiste peut avoir du mal à s'y retrouver. Afin de les aider, il est intéressant d'extraire les informations utiles pour résoudre un problème donné à partir de diverses sources de données(Bell *et al.*, 2011). De plus, il est important de pouvoir stocker et tracer l'origine des données ainsi collectées. Cette étape d'intégration et de structuration des données peut se faire en construisant un entrepôt de données. Les types de questions biologiques auxquels les réseaux sont susceptibles de répondre sont assez différents. De fait, cela nécessite de créer un entrepôt de données dont le modèle est suffisamment générique pour pouvoir s'adapter à ces problèmes variés. L'exploitation de ces données peut ensuite se faire via des outils de visualisation existants tel qu'Ondex. Cependant, des méthodes d'analyse plus poussées comme les méthodes de fouille de données, peuvent être utiles afin d'extraire de nouvelles connaissances.

Chapitre 3

NetworkDB : un entrepôt de données pour l'étude des réseaux biologiques

Sommaire

3.1 Introduction

La compréhension des phénotypes complexes tels que les effets secondaires des médicaments et les maladies génétiques passe par l'étude des réseaux biologiques. En effet, la prise en compte des interactions protéine-protéine, des voies métaboliques et de signalisation, ainsi que des annotations fonctionnelles de chaque protéine peut permettre de comprendre les mécanismes associés à une maladie génétique ou à un effet secondaire. Il existe de nombreuses sources de données publiques qui mettent à disposition ces informations. Même si des formats standards se développent, toutes les sources de données ne les utilisent pas. Ainsi, certaines bases fournissent leurs données au format XML, alors que d'autres se contentent du format TXT, voir HTML. Du fait de l'hétérogénéité des données et de la dispersion des sources, il est intéressant de collecter l'ensemble des données d'intérêt et de les structurer de façon à les stocker dans un entrepôt de données local.

Une approche de type entrepôt de données consiste à construire une base de données réelle, ou entrepôt, contenant les données extraites de différentes sources. Les entrepôts de données sont principalement utilisés en entreprise dans le cadre de l'informatique décisionnelle afin de faciliter la prise de décision. Le schéma d'un entrepôt suit souvent un modèle en étoile (Riazati et Thom, 2011). Dans ce type de modèle, une table centrale contient les faits et des tables satellites décrivent les dimensions des faits (Riazati et Thom, 2011). Par exemple, dans la figure 3.1, les faits sont des ventes qui sont décrites par leur date, le magasin et le produit vendu. Chaque élément de chaque table possède un identifiant unique, ou clef primaire, ainsi que des références vers des clefs primaires d'autres tables, on parle alors de clefs étrangères. Les relations entre deux tables se font grâce à un couple clef primaire-clef étrangère partagé par les deux tables. Ainsi, dans la figure 3.1, la relation entre une vente et un produit se fait grâce à la clef ID_PRODUIT. La définition de contraintes telles que les clés primaires et étrangères n'est pas obligatoire mais permet de maintenir l'intégrité de l'entrepôt. Ainsi, on ne peut pas vendre un produit qui n'est pas répertorié dans la table produit.

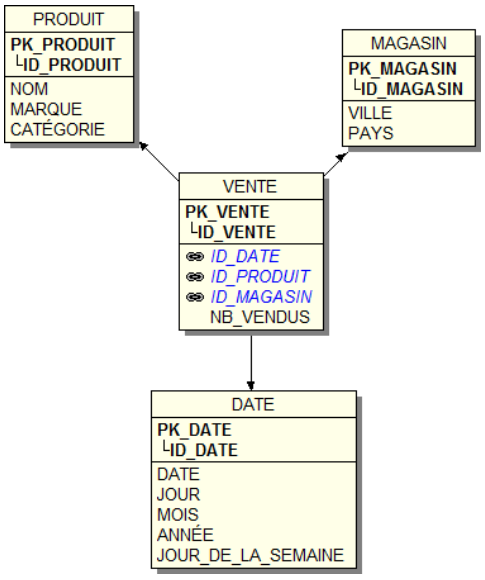


FIGURE 3.1 – Exemple de modèle en étoile pour la vente de produits. La clef primaire de chaque table est en gras et les clefs étrangères en italique.

Le peuplement d'un entrepôt est réalisé par des outils de type "Extract, Tansform, Load" (ETL). Ces outils permettent d'extraire les données de leur source, de les transformer (par exemple, discrétiser des valeurs) puis de les intégrer dans l'entrepôt. L'un des atouts de ces systèmes est de permettre une actualisation des données de l'entrepôt lors des mises à jour des sources.

Afin de construire notre entrepôt, j'ai décidé d'utiliser une base de données relationnelle. En effet, à partir d'un modèle de données défini, les bases de données relationnelles permettent de stocker de grandes quantités de données de façon structurée et en respectant des contraintes d'intégrité. De plus, il est possible de mettre à jour les données de la base en utilisant un langage de requêtes. Ce même langage de requête permet de visualiser le contenu de tout ou partie de la base de données. Les bases de données relationnelles sont également très facilement interfaçables avec des langages de programmation. Le langage PHP est notamment utilisé pour effectuer des requêtes sur une base de données et visualiser les résultats dans une page web. Des interfaces telles que phpMyAdmin [12] ont ainsi été développées pour administrer et interroger les bases de données relationnelles (avec un système de gestion de base de données MySQL).

Ce chapitre décrit la construction d'une base de données générique, NetworkDB, destinée à recueillir les données des réseaux biologiques concernant le problème étudié. J'y décris le modèle utilisé pour stocker les données, les bases de données qui ont été intégrées et les différentes façons d'interroger la base de données.

3.2 Conception et modèle de données

L'implémentation de la base de données est réalisée avec PostgreSQL [13], un système de gestion de base de données relationnelles libre et nécessaire au fonctionnement du programme MODIM (Model-driven data integration for Mining) qui est utilisé pour collecter les données et les intégrer à la base de données (Ndiaye *et al.*, 2011).

J'ai défini un modèle de donnée générique, NetworkDB, permettant de stocker les informations concernant les réseaux biologiques (Figures 3.2 et 3.3 en vert). Pour faire l'analogie avec le modèle en étoile, le point central de mon modèle est la protéine et toutes les informations collectées sont reliées à cette entité. Les protéines sont souvent composées d'une ou plusieurs sous-unités appelées domaines. Un domaine est une région de la protéine qui est généralement formé d'un segment continu d'acides aminés et qui est le plus souvent capable de se replier de manière suffisamment stable pour exister par lui même. Il existe deux grand types de méthodes de définition des domaines protéiques : les définitions basées sur la séquence de la protéine et les définitions basées sur la structure. Étant donné que les données de séquence sont plus nombreuses que les données structurales et que la séquence d'une protéine détermine sa structure, la plupart des définitions de domaines est basée sur l'identification de séquences conservées (Copley *et al.*, 2002). En parallèle, il existe des classifications de domaines basés sur la structure telles que SCOP (Murzin *et al.*, 1995) et CATH (Orengo *et al.*, 1997). Ces deux classifications sont divisées en quatre niveaux : la classe (composition en structure secondaire), le repliement/architecture (arrangement de la struc-

12. www.phpmyadmin.net
13. www.postgresql.org

ture secondaire), la superfamille/topologie (lien entre les structures secondaires) et la famille/homologie (similarité de séquence, structure et fonction). Un domaine étant fréquemment associé à une fonction, l'identification de domaines dans une protéine peut permettre de comprendre sa fonction (Copley *et al.*, 2002). Il est donc important de collecter les domaines composant les protéines. Bien que la composition en domaine des protéines ait un rôle important, les fonctions biologiques des protéines sont aussi réalisées via leurs interactions. En effet, la plupart des processus biologiques sont réalisés par des assemblages d'au moins 10 protéines (Alberts, 1998). Dans la base de données, les interactions protéine-protéine sont donc intégrées. Les complexes protéiques interviennent dans des voies biologiques définies, permettant ainsi de connaitre les rôles associés aux protéines. Il est donc également important de stocker ces réseaux biologiques dans NetworkDB. Le projet Gene Ontology [14] définit et met à disposition un vocabulaire contrôlé afin de décrire la fonction des gènes et leurs produits. Ce vocabulaire contient 38 137 termes (version 1.3499) groupés en trois branches : "Biological Process" (BP) qui décrit des processus cellulaires composés d'un enchainement d'évènements correspondant à des fonctions moléculaires, "Cellular Component" (CC) pour les termes associés à la localisation cellulaire ou intracellulaire du produit d'un gène et la branche "Molecular Function" (MF) décrivant l'activité moléculaire du produit d'un gène (liaison, catalyse, ...). Les termes sont reliés entre eux par des relations de type *is a* et *part of*. La relation "T1 *is a* T2" signifie que le terme T1 est une spécialisation de T2 et la relation "T1 *part of* T2" signifie que T1 est une sous-partie de T2. De plus, vingt codes permettent de caractériser la façon dont un terme GO a été associé à la protéine ou au gène (Table 3.1). Ce vocabulaire est intéressant car il permet de décrire à la fois la localisation, les fonctions et les processus de chaque protéine, tout en permettant de tracer l'origine de l'annotation.

Le modèle développé pour l'entrepôt NetworkDB est générique car il peut être utilisé pour différentes applications avec quelques ajouts nécessaires selon le problème considéré. Par exemple, l'étude de résultats de séquençage de patients atteints de maladies génétiques (chapitre 4) entrainera l'ajout d'une partie dédiée aux patients et à leurs mutations (Figure 3.2 en violet). Le cas d'étude développé dans le chapitre 5 concerne l'étude des effets secondaire de médicaments. Pour cela, il est utile d'ajouter les données concernant les médicaments et leurs effets secondaires (Figure 3.2 en jaune).

14. www.geneontology.org

Code	Signification
Codes pour les analyses expérimentales	
EXP	Inferred from Experiment
IDA	Inferred from Direct Assay
IPI	Inferred from Physical Interaction
IMP	Inferred from Mutant Phenotype
IGI	Inferred from Genetic Interaction
IEP	Inferred from Expression Pattern
Codes pour les analyses informatiques	
ISS	Inferred from Sequence or Structural Similarity
ISO	Inferred from Sequence Orthology
ISA	Inferred from Sequence Alignment
ISM	Inferred from Sequence Model
IGC	Inferred from Genomic Context
IBA	Inferred from Biological aspect of Ancestor
IBD	Inferred from Biological aspect of Descendant
IKR	Inferred from Key Residues
IRD	Inferred from Rapid Divergence
RCA	inferred from Reviewed Computational Analysis
Codes issus de la littérature	
TAS	Traceable Author Statement
NAS	Non-traceable Author Statement
Codes établis par les currateurs	
IC	Inferred by Curator
ND	No biological Data available
Code établis automatiquements	
IEA	Infered from Electronic Annotation

TABLE 3.1 – Liste des codes (“evidence codes”) caractérisant les annotations des gènes par des termes GO.

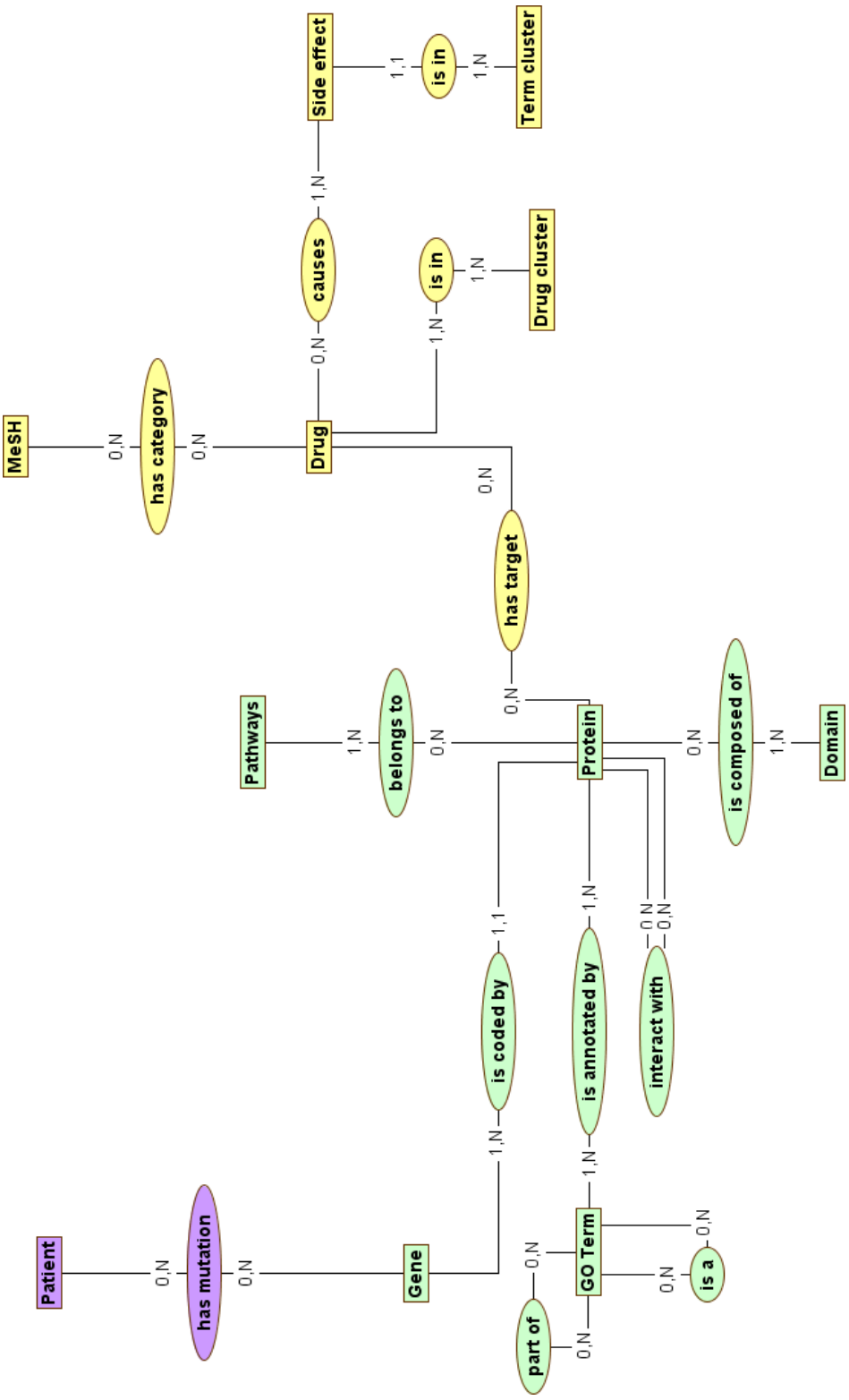


FIGURE 3.2 – Schéma entité-association de NetworkDB. Les entités sont représentées par des rectangles et sont reliées par des associations sous forme d'ellipses. La partie en violet est spécifique à la gestion des données issues de l'étude du Syndrome d'Aicardi et des déficiences intellectuelles liées à l'X DILX. Quant à la partie jaune, elle concerne l'étude des effets secondaires de médicaments.

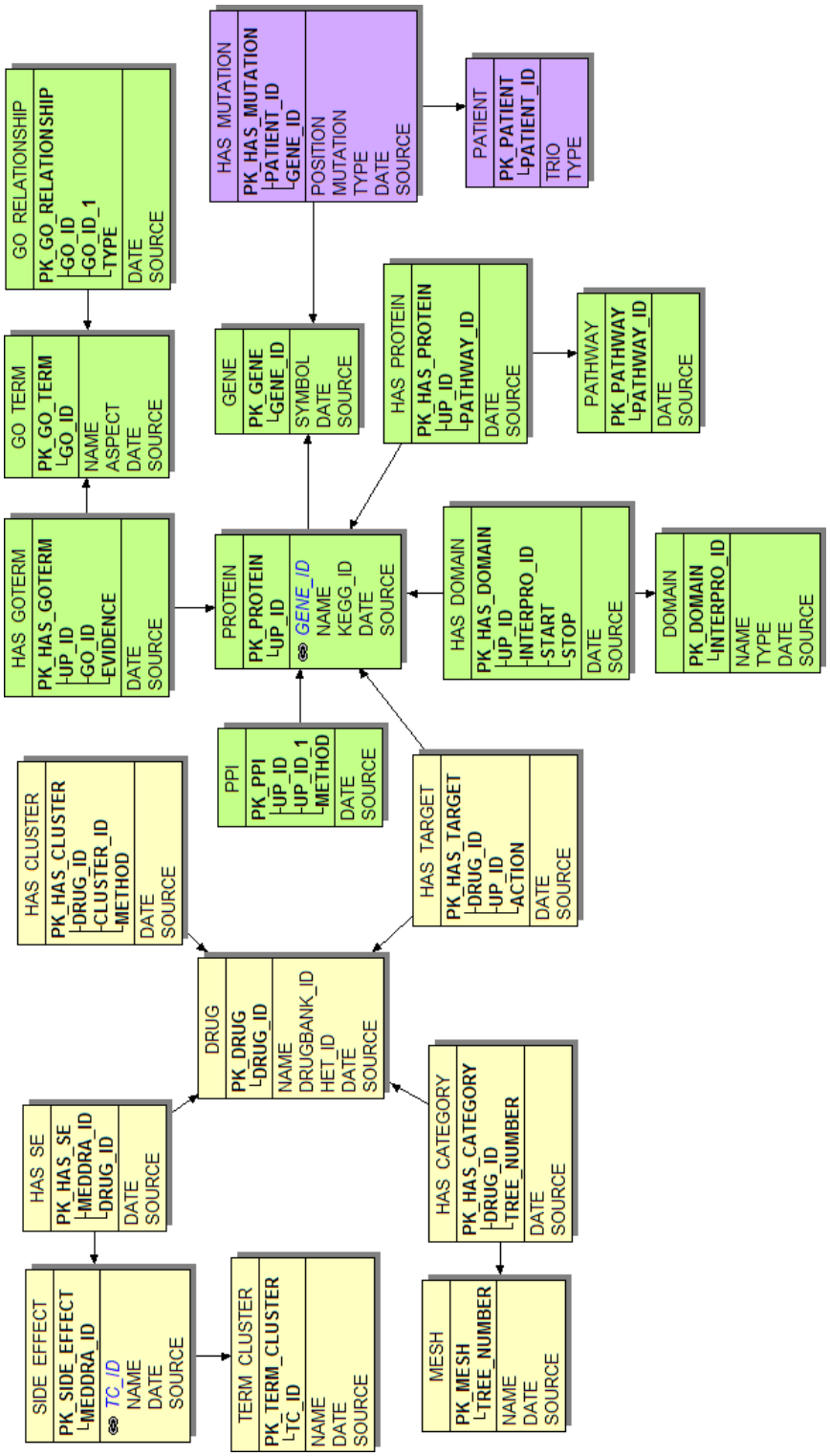


FIGURE 3.3 – Modèle relationnel de NetworkDB. Les clefs primaires de chaque table sont en gras.

3.3 Sources de données utilisées pour le peuplement de l'entrepôt NetworkDB

3.3.1 Bases de données utilisées pour peupler le noyau NetworkDB

3.3.1.1 Protéines

La collecte des principales informations sur les protéines est réalisée à partir de la base de données UniProt [15] (The UniProt Consortium, 2012). UniProt est composée de Swiss-Prot qui contient des séquences vérifiées manuellement et de TrEMBL dont les séquences sont issues de la traduction automatique des bases de séquences nucléotidiques EMBL-Bank/GenBank/DDBJ. La version 2013_02 contient 539 165 séquences provenant de Swiss-Prot et 29 769 971 issues de TrEMBL.

A partir de la fiche UniProt d'une protéine on récupère de nombreuses informations, dont sa séquence, son nom, son numéro EC et ses modifications post-traductionnelles connues. Une entrée UniProt fournit également de nombreux liens vers d'autres bases de données. On peut ainsi collecter les identifiants *gene id* et *kegg id* de la protéine, qui serviront respectivement à interroger les bases de données Entrez Gene et KEGG.

3.3.1.2 Gènes

Les informations sur les gènes proviennent de la base de données Entrez Gene [16] du NCBI. Cette base contient 11 442 877 gènes dont 43 826 chez l'Homme (le 04-03-13). On peut ainsi récupérer le symbole du gène et son organisme. Par exemple, le symbole correspondant à l'identifiant *gene id* 10013 est *HDAC6* et est présent chez Homo sapiens.

3.3.1.3 Domaines

La base de données InterPro [17] contient les domaines prédits et les sites importants de chaque protéine. Pour cela ils intègrent les signatures de onze bases de données (CATH, PANTHER, PIRSF, Pfam, PRINTS, ProDOM, PROSITE, HAMAP, SMART, SUPERFAMILY et TIGRFAMs). InterPro (version 41.0) contient 24 356 entrées annotant l'ensemble des protéines d'UniProt.

3.3.1.4 Termes Gene Ontology (GO)

Deux étapes sont nécessaires pour collecter les termes GO. Dans un premier temps, il faut récupérer les identifiants des termes associés aux protéines d'intérêt. Puis dans un second temps, il est nécessaire d'obtenir le terme associé à chaque identifiant ainsi que les relations sémantiques entre termes.

15. www.uniprot.org
16. www.ncbi.nlm.nih.gov/gene
17. www.ebi.ac.uk/interpro

Les associations protéine-terme GO sont collectées à partir de la base de données QuickGO [18] contenant 141 074 143 annotations GO pour 21 318 366 protéines distinctes (version du 24/02/13). A partir de QuickGO, il est possible de collecter les annotations GO de chaque protéine ainsi que les codes d'évidence associés.

AmiGO [19] est la base de données associée au projet GO (Carbon *et al.*, 2009). On y trouve les labels correspondant aux identifiants des termes collectés ainsi que la branche (BP, CC ou MF) à laquelle ils appartiennent. De plus, chaque terme est relié à ses ancêtres et ses descendants dans structure du vocabulaire.

3.3.1.5 Interactions protéines-protéines

Les interactions protéine-protéine sont collectées à partir d'IntAct [20] qui recense les interactions moléculaires décrites dans la littérature (Kerrien *et al.*, 2012). La version du 06/03/13 de la base de données contient 307 975 interactions et comporte les informations concernant la méthode utilisée pour détecter l'interaction, le type d'interaction ainsi que la référence bibliographique où l'interaction est décrite.

3.3.1.6 Réseaux biologiques

Afin de prendre en compte le contexte biologique lié aux protéines stockées dans la base de données, les réseaux biologiques dans lesquels elles sont impliquées sont collectés à partir des bases de données KEGG PATHWAY [21] et Pathway Interaction Database [22] (PID). KEGG PATHWAY (version 06/03/13) contient 436 réseaux biologiques de référence construits manuellement. PID est une base intégrant 137 réseaux biologiques vérifiés par le groupe NCI-Nature ainsi que 322 réseaux biologiques issus de l'intégration des bases de données BioCarta [23] et Reactome [24]. L'un des intérêts de PID est d'intégrer les données de BioCarta et de Reactome. Malheureusement, PID n'est plus mise à jour depuis septembre 2012, donc en cas de nouvelle collecte de données il faudra récupérer les informations provenant de ces deux bases à partir des deux sources primaires.

3.3.2 Bases de données utilisées pour la caractérisation des effets secondaires des médicaments

3.3.2.1 Molécules et leurs cibles

Afin de pouvoir caractériser les effets secondaires associés aux médicaments, il est nécessaire de récupérer des informations concernant ces molécules.

18. www.ebi.ac.uk/QuickGO
19. http://amigo.geneontology.org
20. www.ebi.ac.uk/intact
21. www.genome.jp/kegg/pathway.html
22. http://pid.nci.nih.gov
23. www.biocarta.com
24. www.reactome.org

Les molécules proviennent de la base de données DrugBank [25] contenant 6712 molécules (version 3.0, Knox *et al.* 2011). Cette base contient notamment des informations sur la structure des molécules mais également la liste des catégories qui leur sont associées. De plus, chaque molécule est associée à la liste de ses cibles. DrugBank fournit également des références croisée notamment sous forme d'identifiants *HET id* référençant les molécules dans la Protein Data Bank [26] (PDB, Berman *et al.* 2000).

Une partie des cibles associée aux molécules proviennent donc de DrugBank. Afin de compléter ces informations, j'ai également collecté les informations sur les associations molécules-cibles en utilisant la base de données de structures PDB. En effet, bien que PDB soit une base de données de structures protéiques, elle recense 14 819 ligands (version du 06/03/13) différents associés à une ou plusieurs structures de protéines. Ces complexes étant obtenus expérimentalement, on peut en déduire l'existence d'une interaction ou liaison physique entre une molécule et une protéine.

3.3.2.2 Effets secondaires

Les effets secondaires associés aux médicaments proviennent de la base de données SIDER (Kuhn *et al.*, 2010). SIDER (version 2) contient environ 100 000 paires médicament-effets secondaires concernant 996 molécules et 4192 effets différents. Les effets décrits par SIDER sont issus des notices provenant de sources publiques telles que la FDA, des sites web publics et les fabricants. Après extraction des termes, Kuhn *et al.* (2010) ont associé ceux-ci à leur équivalent dans le vocabulaire contrôlé "Medical Dictionary for Regulatory Activities" (MedDRA). La terminologie MedDRA est utilisée pour classer les effets indésirables associés à l'emploi de médicaments et vaccins. MedDRA est souvent présenté comme une hiérarchie à 5 niveaux : classes de système d'organes, groupes de termes de haut niveau, termes de haut niveau, termes préférés et termes de bas niveau (Merrill, 2008).

3.3.3 Données utilisées pour l'étude de l'étiologie de maladies génétiques

Aucune nouvelle base de données n'est à proprement parler intégrée dans le cas de l'étude de l'étiologie des maladies génétiques. En effet, comme on peut le voir en violet sur la figure 3.2, la seule entité spécifique à cette étude est le patient et la seule relation entre le patient et le gène (faisant partie de NetworkDB) est la mutation. Les données concernant les patients sous fournies directement par les médecins du CHU et les données de mutation proviennent de l'analyse du séquençage.

3.3.4 Conclusion

De nombreuses sources de données sont utilisées pour peupler l'entrepôt NetworkDB. Certaines sont communes à toutes les applications de NetworkDB et permettent de peupler le cœur de

25. www.drugbank.ca
26. www.rcsb.org

l'entrepôt et d'autres sont spécifiques à une application donnée. Chaque nouvelle application utilisant une nouvelle version de l'entrepôt , il est intéressant d'employer un système qui soit facilement paramétrable et réutilisable tel le système MODIM, décrit dans la section suivante, pour peupler NetworkDB à partir des sources de données développées ici.

3.4 Le système MODIM pour l'intégration de données dirigée par un modèle

3.4.1 Présentation de MODIM

Le système "Model-driven Data Integration for Mining" (MODIM) met en œuvre une approche générique pour collecter et intégrer des données en fonction d'un modèle relationnel défini dans le but de réaliser des expérimentations de fouille de données. Dans cette approche, les besoins et l'expertise de l'utilisateur ont un rôle central et les tâches répétitives sont automatisées.

MODIM se décompose en trois modules : base de données (i), configuration des tâches (ii) et exécution des tâches (iii) (Figure 3.4). Une fois que le modèle relationnel a été défini par l'utilisateur, la base de données correspondante peut être construite en utilisant le premier module. Le deuxième module permet de définir et de configurer les tâches qui serviront à peupler la base de données. Chaque tâche est spécifique d'une entrée (par exemple, un identifiant de protéine) et est composée de sous-tâches, chacune dédiée à une source de données utilisée (le plus souvent sous forme d'une URL). La configuration d'une sous-tâche consiste à décrire comment reconnaitre, dans la page de résultats, les données à collecter et où les stocker dans la base de données. La reconnaissance des données à collecter est réalisée soit par des expressions XPATH dans le cas de fichier XML, soit avec des expressions régulières pour les fichiers texte ou HTML. Un fois terminées, les tâches peuvent être sauvegardées sur le serveur ou exportées au format XML. Dans les deux cas, elles sont ensuite éditables et modifiables en cas de besoins. La collecte et l'intégration des données sont réalisées par le module d'exécution des tâches. Ce module prend en entrée un fichier ou une requête SQL contenant les données nécessaires à la réalisation de la tâche.

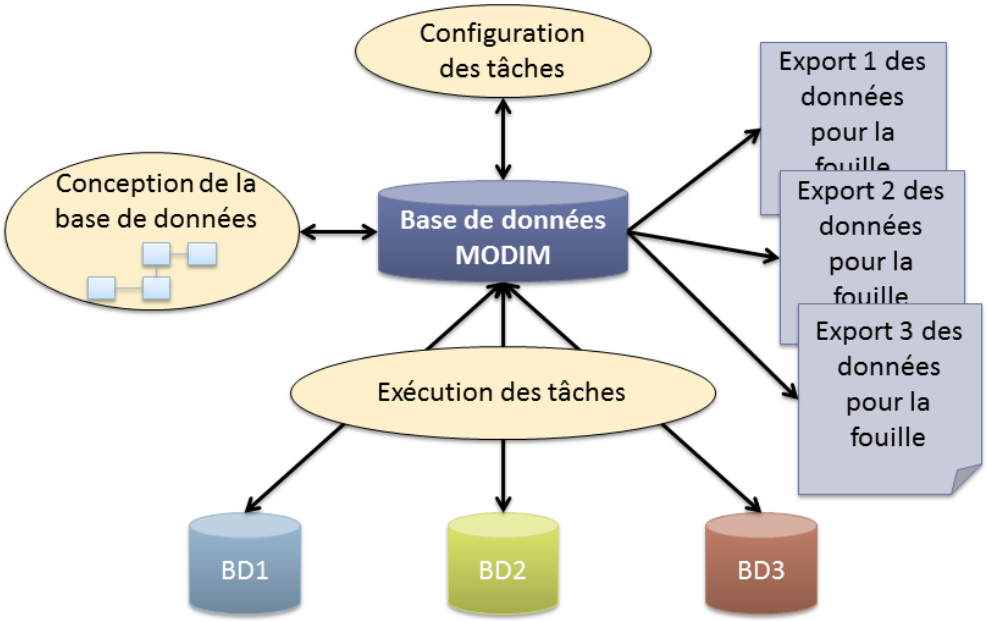


FIGURE 3.4 – Représentation schématique de l'approche MODIM.

3.4.2 Application au peuplement de NetworkDB

Le peuplement des bases de données utilisant le modèle défini dans la figure 3.3, se fait grâce au programme MODIM. Quatre tâches sont nécessaires (Figure 3.5). A partir de la liste des identifiants UniProt donnée en entrée, la première tâche consiste à interroger IntAct et de récupérer toutes leurs interactions protéine-protéine.

La seconde tâche utilise l'ensemble des identifiants UniProt collectés par la première, ce qui inclut les protéines données en entrée et leurs interactants. Cette tâche est divisée en 3 sous-tâches. La première va collecter le nom, le *gene id* et le *kegg id* de chaque protéine à partir d'UniProt (Figure 3.6), la deuxième récupère les termes GO ainsi que les codes d'évidence et la troisième va interroger PID afin d'obtenir les identifiants ainsi que les noms des réseaux biologiques.

La troisième tâche se concentre sur la collecte des réseaux biologiques KEGG en prenant en entrée les identifiants KEGG des protéines et en interrogeant la fiche KEGG de la protéine.

Enfin, la dernière tâche permet d'obtenir les symboles des gènes disponibles sur la fiche Entrez Gene à partir du *gene_id*.

La liste des URL utilisées avec MODIM et les informations collectées correspondantes sont présentées dans la Table 3.2.

Base de données	URL	Données
IntAct	www.ebi.ac.uk/intact/export ?query=q%3D$up_id %26sort %3Drigid%2Basc%26rows%3D30%26 start%3D0%26facet%3Dtrue%26facet.missing %3Dtrue%26facet.field%3Dexpansion%26 facet.field%3DinteractorType_id&format=mitab _intact&conversationContext=	*Uniprot id*, methode, type d'interaction
UniProt	www.uniprot.org/uniprot/$up_id.xml	Nom, *gene id* et *KEGG id*
QuickGO	www.ebi.ac.uk/QuickGO/GAnnotation ?col= proteinID+goID+evidence+splice&count=1000 &protein=$up_id&select=normal&termUse =ancestor &format=tsv&slimTypes=IP=	Termes GO et codes d'évidence
PID	pid.nci.nih.gov/search/intermediate_landing. shtml ?molecule=$up_id&Submit=Go	*Pathway id* et son nom
KEGG	www.genome.jp/dbget-bin/www_bget ?$kegg_id	*Pathway id* et son nom
NCBI GENE	www.ncbi.nlm.nih.gov/sites/entrez ?db=gene& term=$gene_id&dopt=xml	*Gene symbol*

TABLE 3.2 – Liste des URL utilisées avec MODIM et type de données qu'elles permettent de collecter. Les éléments en bleu correspondent aux identifiants utilisés en entrée du 3ème module de MODIM.

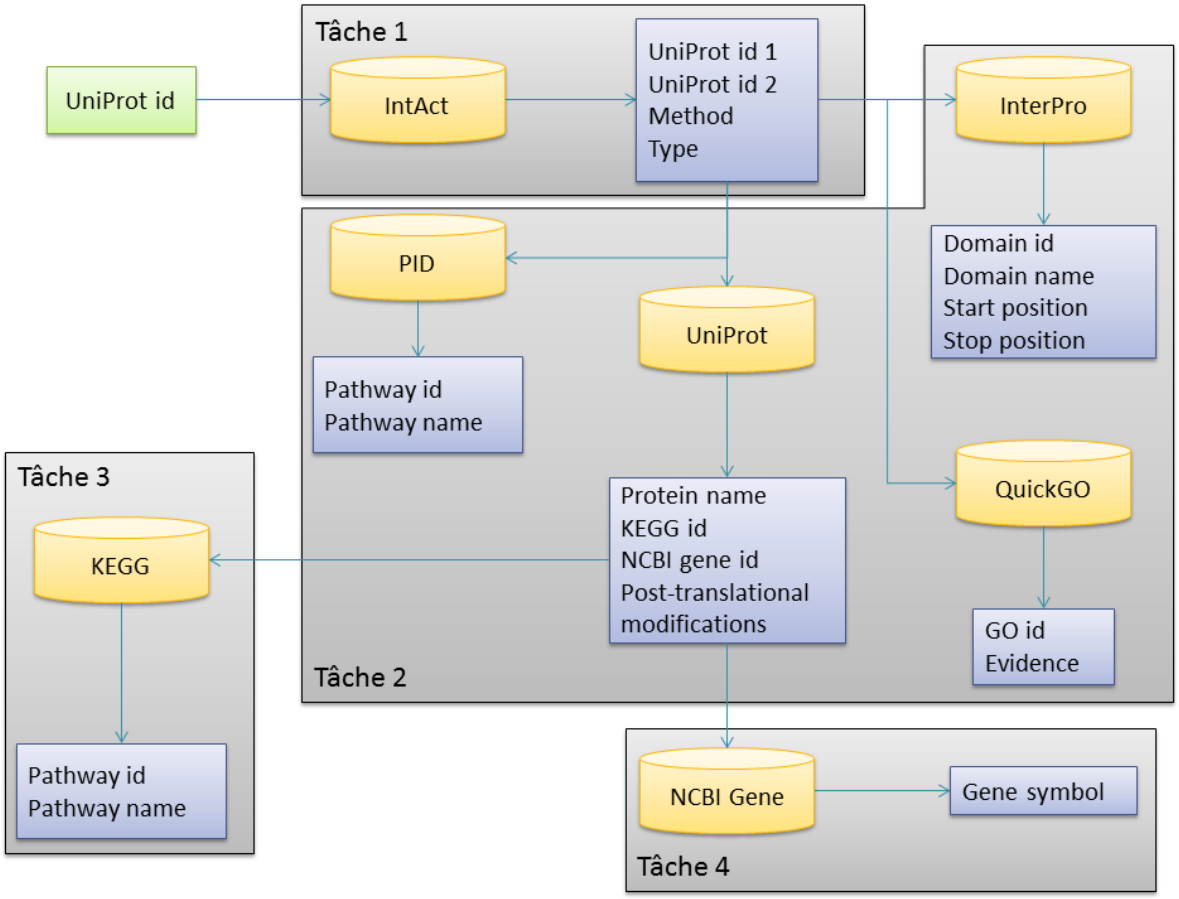


FIGURE 3.5 – Organigramme de collecte des données utilisé pour peupler NetworkDB. Le Point d'initialisation de la collecte est une liste d'*uniprot id*. Les bases de données interrogées sont représentées sous forme de cylindres et les données collectées par des rectangles.

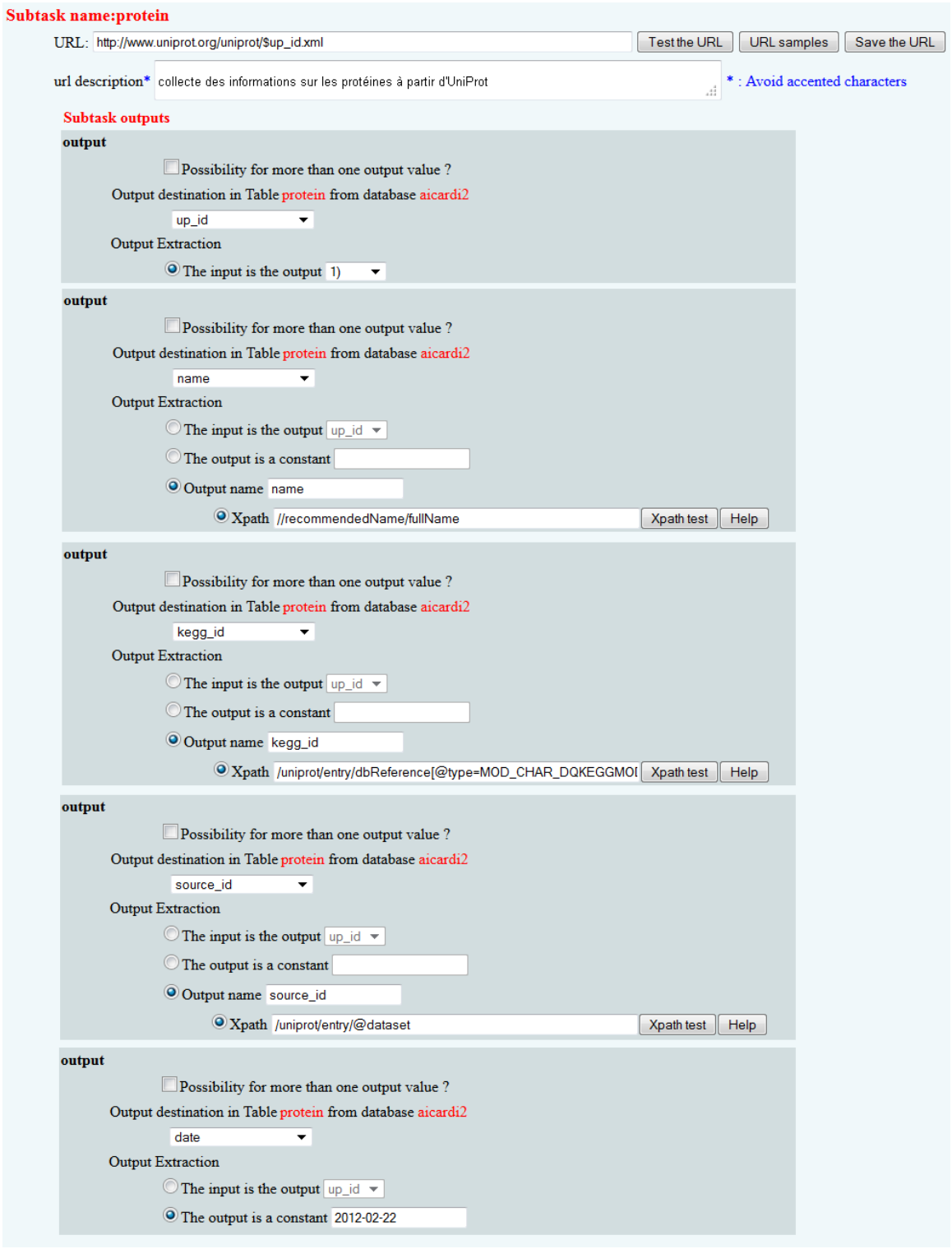


FIGURE 3.6 – Exemple de sous-tâche MODIM utilisée pour collecter des informations à partir de la base de donnée UniProt.

3.4.2.1 Étude de l'étiologie des maladies génétiques

Dans le cadre de l'étude des maladies génétiques, deux scénarios de collecte des données sont possibles : soit on dispose d'informations issues du séquençage de patients, soit on veut étudier une liste de gènes donnés. Dans les deux cas, le scénario de collecte est assez similaire, seul le point d'entrée diffère. En effet, si on étudie des résultats de séquençage, il est important de stocker ces informations.

Par contre dans le cas de l'étude d'un ensemble de gènes prédéfini, on ne dispose pas de données concernant des mutations mais on dispose de la liste des identifiants de gènes. On peut donc interroger directement la base NCBI GENE. A partir des *gene id*, on interroge la base du NCBI et on collecte les identifiants UniProt correspondants grâce à une tâche MODIM. Puis, le scénario de collecte suit le schéma défini dans la section 3.4.2.

3.4.2.2 Caractérisation des effets secondaires

Pour l'étude des effets secondaires, le point d'entrée est le médicament. Les médicaments sont issus de la base de données DrugBank. Cette base de données va permettre de collecter entre autres, les catégories associées aux médicaments (champ "Categories" de DrugBank). Les termes collectés ont été associés aux termes correspondant dans le MeSH (Medical Subject Headings). Le MeSH est un vocabulaire contrôlé principalement utilisé pour indexer les articles scientifiques. Or, j'ai remarqué que les termes utilisés dans le champ "Categories" de DrugBank étaient issus de 3 branches du MeSH : la branche "Molecular mechanism of pharmacological action", la branche "Physiological Effetcs of drugs" et "Therapeutic Uses" décrivant respectivement les mécanismes d'action, les effets physiologiques et les usages thérapeutiques des médicaments. Les annotations issues de ces 3 branches ont été collectées en tant que catégories. Pour chaque molécule, il est également possible de rechercher si elle est référencée dans SIDER afin de pouvoir connaitre ses effets secondaires. Les informations sur les cibles des médicaments sont importantes pour mieux étudier leurs effets secondaires. Ainsi, les données de DrugBank sur les cibles sont très utiles. On retrouve ainsi le type d'action du médicament sur sa cible. Afin de limiter le nombre de types d'actions possible dans la suite des études je les ai regroupés en 3 classes : Inhibiteurs, activateurs et autres. Le détail des 3 classes est donné en Table 3.3. De plus, afin de compléter ces données, j'ai interrogé PDB en utilisant les HET id, donnés par DrugBank, pour obtenir une liste plus complète des cibles associées à chaque médicament. Malheureusement, le type d'action de la molécule sur la protéine n'est pas toujours documenté dans PDB. Ainsi, toutes les interactions cible-médicaments extraites de PDB sont classées en "autres".

Une fois les identifiants UniProt des cibles extraits soit à partir de PDB soit à partir de DrugBank, on peut collecter les données concernant les protéines, leurs interactions, leurs réseaux biologiques, leurs termes GO et leurs domaines selon la façon décrite par la section 3.4.2.

Inhibiteurs	activateurs	autres
antagonist	activator	adduct
blocker	agonist	antibody
inhibitor	partial agonist	binder
inverse agonist	potentiator	cleavage
partial antagonist	stimulator	cofactor
suppressor		intercalation
		multitarget
		modulator
		negative modulator
		other
		other unknown
		unknown

TABLE 3.3 – Classification utilisée pour caractériser les effets de médicaments sur leur(s) cible(s)

3.5 Résultats de la collecte et interrogation de l'entrepôt NetworkDB

Lorsque l'ensemble des données ont été collectées, il est possible d'interroger la base de données en formulant des requêtes SQL (Structured Query Language). Une requête SQL est de la forme : Select From Where. La liste des attributs que l'on veut visualiser est donnée après Select, From correspond à la table contenant les données à visualiser et Where contient les conditions à remplir pour être visualiser. A titre d'exemple, la requête "`Select up_id From protein Where kegg_id is not NULL`" permet d'obtenir les identifiants UniProt des protéines dont on connait le *Kegg id*. Il est également possible d'interroger plusieurs tables en réalisant des jointures, c'est à dire en faisant des jonctions entre les tables en utilisant des attributs partagés. Cependant, même si la formulation de requêtes est un outil puissant pour interroger la base de données, cela reste relativement compliqué pour les personnes ne maitrisant pas le langage SQL. J'ai donc créé une interface web permettant d'interroger la base de données à partir de modèles de requêtes prédéfinies puis de visionner les résultats sous forme de graphes.

3.5.1 Exemple d'interrogation par SQL : relations entre catégories de molécules et réseaux biologiques.

D'après le modèle de données utilisé pour construire NetworkDB (Figure 3.3), il est possible d'associer les catégories des molécules ayant des effets secondaires et les réseaux biologiques. Pour chaque catégorie, on recueille la liste des molécules annotées, on se limite ensuite aux molécules ayant des effets secondaires, puis on collecte l'ensemble de leurs cibles et enfin, on recueille la liste des réseaux biologiques contenant ces cibles. La requête SQL correspondante (Figure 3.7) permet de collecter 3750 associations différentes entre catégories et réseaux biologiques. Ces associations concernent 67 catégories pour 502 réseaux biologiques.

Il est possible d'analyser ces associations réseaux biologiques-catégories en fonction du nombre de réseaux biologiques associés à chaque catégorie et inversement (Table 3.4). On peut ainsi remarquer que la catégorie impliquant le plus grand nombre de réseaux biologiques est celle des agents antinéoplasiques, montrant ainsi qu'un très grand nombre de réseaux biologiques sont "ci-

blés" afin de traiter des cancers. En parallèle, de nombreuses catégories de médicaments vont cibler les voies de signalisation du Calcium, ou les interactions des ligands neuroactifs avec leur récepteur.

Catégorie	Nb réseaux biologiques	Réseau biologique(Id)	Nb Catégories
Antineoplastic Agents	327	Neuroactive ligand-receptor interaction (KEGG : hsa04080)	105
Protein Kinase Inhibitors	240	Calcium signaling pathway (KEGG : hsa04020)	90
Sympathomimetics	221	Metabolic pathways (KEGG : hsa01100)	86
Cardiotonic Agents	216	Gap junction (KEGG : hsa04540)	63
Bronchodilator Agents	215	Salivary secretion (KEGG : hsa04970)	58
Adrenergic beta-Agonists	197	Vascular smooth muscle contraction (KEGG : hsa04270)	57
Immunosuppressive Agents	194	N-cadherin signaling events (PID : 200175)	50
Antihypertensive Agents	191	Pathways in cancer (KEGG : hsa05200)	49
Vasodilator Agents	188	Role of Calcineurin-dependent NFAT signaling in lymphocytes (PID : 200073)	48
Anesthetics	180	EPHA2 forward signaling (PID : 200180)	48

TABLE 3.4 – Aperçu des 10 catégories associées au plus grand nombre de réseaux biologiques (à gauche) et des 10 réseaux biologiques impliqués dans le plus grand nombre de catégories.

3.5.2 Interface d'interrogation

Les requêtes SQL permettent d'interroger avec précision le contenu de la base de données mais un utilisateur ne maitrisant pas ce langage sera dans l'incapacité d'accéder aux données. Ainsi, j'ai créé une interface web permettant de répondre à des questions bien précises définies avec les utilisateurs de la base NetworkDB.

La première de ces questions est : "quelles sont les informations disponibles sur une molécule et ses cibles ?" (Figure 3.8). Par exemple si on s'intéresse à la molécule d'aspirine (DB00945), on obtient les informations synthétisées dans la table 3.5. Il est ensuite possible de sauvegarder les résultats au format XML afin de les visualiser sous Ondex. Ondex est un outil de visualisation de réseaux sous forme de graphes. Comme expliqué précédemment (section 2.2.2.4), le principal intérêt de ce programme est d'utiliser un format XML d'entrée entièrement personnalisable. J'ai pu créer les concepts correspondant aux données étudiées et ainsi permettre de visualiser les réseaux extraits de NetworkDB sous forme de graphe puis de naviguer dans ces réseaux. Pour en revenir à

FIGURE 3.7 – Requête SQL utilisée pour associer les catégories des médicaments (ayant au moins un effet secondaire) aux pathways des cibles de ces mêmes médicaments

```
select
        distinct m.name as category, p.name as pathway, p.pathway_id
from
        mesh m,
        has_category c,
        has_se s,
        has_target t,
        has_protein pp,
        pathway p
where
        m.tree_number=c.tree_number
        c.drug_id=s.drug_id and
        s.drug_id=t.drug_id and
        t.up_id=pp.up_id and
        pp.pathway_id=p.pathway_id
```

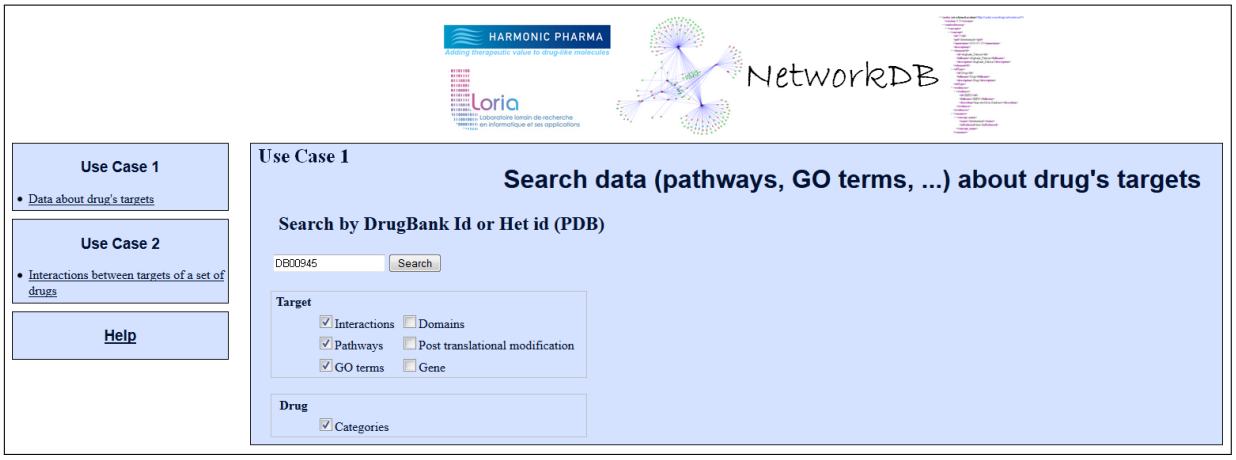


FIGURE 3.8 – Interface d'interrogation de NetworkDB pour la recherche d'informations sur une molécule et ses cibles.

Catégories	Nb Éléments
Catégories	6
Cibles	7
IPP des cibles de l'aspirine	2
Processus biologiques des cibles et de leurs interactants	96
Termes GO des cibles et de leurs interactants	157
Domaines des cibles et de leurs interactants	24

TABLE 3.5 – Informations contenues dans NetworkDB concernant la molécule d'aspirine.

l'exemple de l'aspirine, le réseau correspondant est représenté par la figure 3.9. On peut notamment

voir que seules deux cibles (cercles rouges) partagent un réseau biologique (carré violet, sur la droite). Par contre, 13 termes GO sont partagés par au moins deux des 5 cibles. Il en est de même pour 5 domaines partagés par deux cibles.

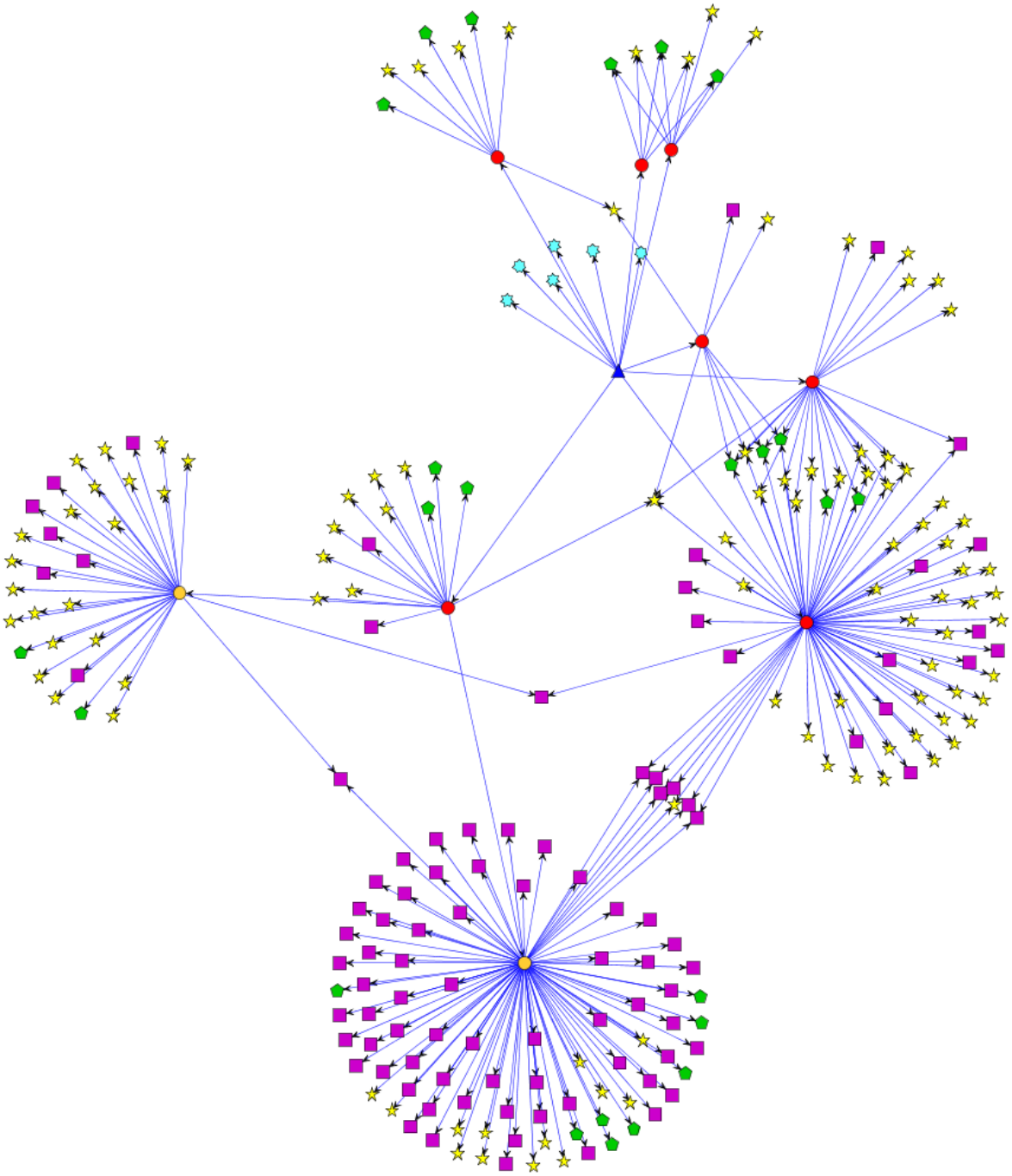


FIGURE 3.9 – Graphe étendu représentant les informations connues sur la molécule d'aspirine (triangle bleu) et ses cibles (rond rouges). Les interactants des cibles sont représentés en orange, les réseaux biologiques en carrés violets, les termes GO BP en étoiles jaunes les domaines protéiques en pentagones verts. Les catégories associées à l'aspirine sont en étoiles bleues claires.

La seconde question fréquemment posée par les utilisateurs de NetworkDB concerne la présence de cibles communes et de propriétés communes des cibles d'un ensemble de molécules. Par exemple, pour les molécules d'aspirine (acide acétylsalicylique), d'ibuprofène et de paracétamol (acétaminophène), les informations présentées dans la table 3.6 sont contenues dans NetworkDB. Parmi les 10 cibles de ces 3 molécules, 2 sont partagées par l'ibuprofène et l'aspirine et 2 sont

Catégories	Nb Éléments
Cibles	10
IPP des cibles	155
Processus biologiques des cibles et de leurs interactants	256
Termes GO des cibles et de leurs interactants	1052

TABLE 3.6 – Informations contenues dans NetworkDB concernant les molécules d'aspirine, d'ibuprofène et de paracétamol.

communes à trois molécules. De plus, parmi les 6 cibles restantes, aucune ne partage d'interactions (Figure 3.10). Cependant, deux de ces cibles "isolées" possèdent des interactants connus. La première de ces protéines est l'albumine sérique (140 interactant) qui est la principale protéine du plasma sanguin et qui se lie facilement à de nombreuses molécules. La seconde protéine est l'Aldo-keto reductase family 1 member C1 qui inactive la progestérone et qui possède un rôle dans le transport de la bile. Si on étudie les réseaux biologiques des cibles, on peut voir que les deux prostaglandines synthases partagent le réseau biologique "Arachidonic acid metabolism" et que la prostaglandine synthase G/H appartient au "metabolic pathway" tout comme l'Acyl-coenzyme A synthétase mitochondriale ACSMA (Figure 3.11).

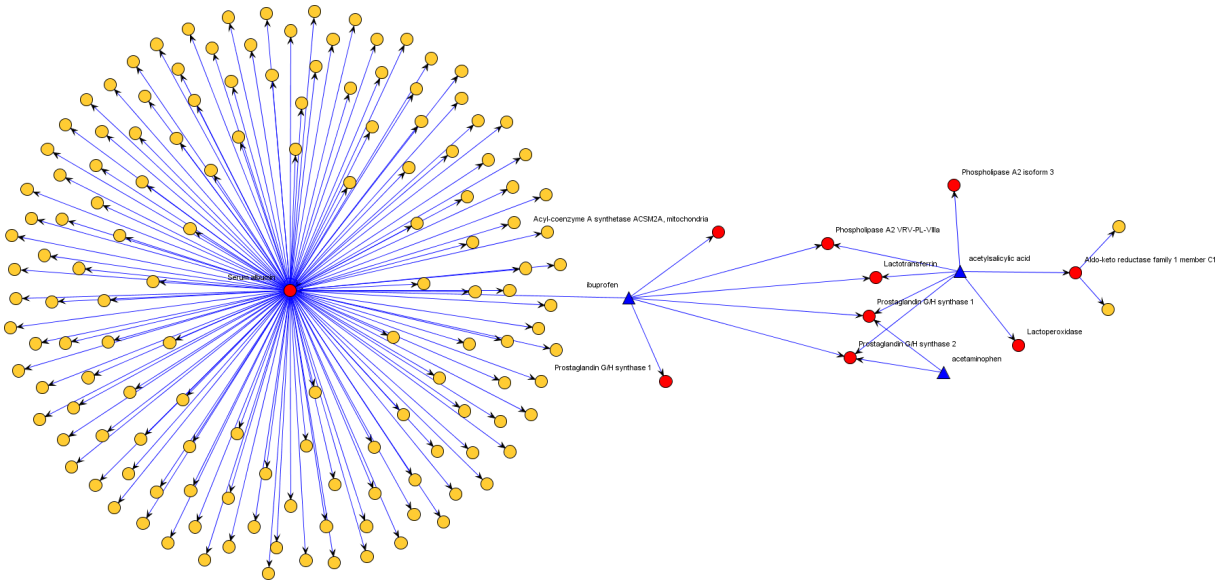


FIGURE 3.10 – Réseau des interactions protéine-protéine des cibles de l'ibuprofène, l'aspirine et du paracétamol (acetaminophen). Les cibles de ces molécules sont en rouge et leurs interactants en orange.

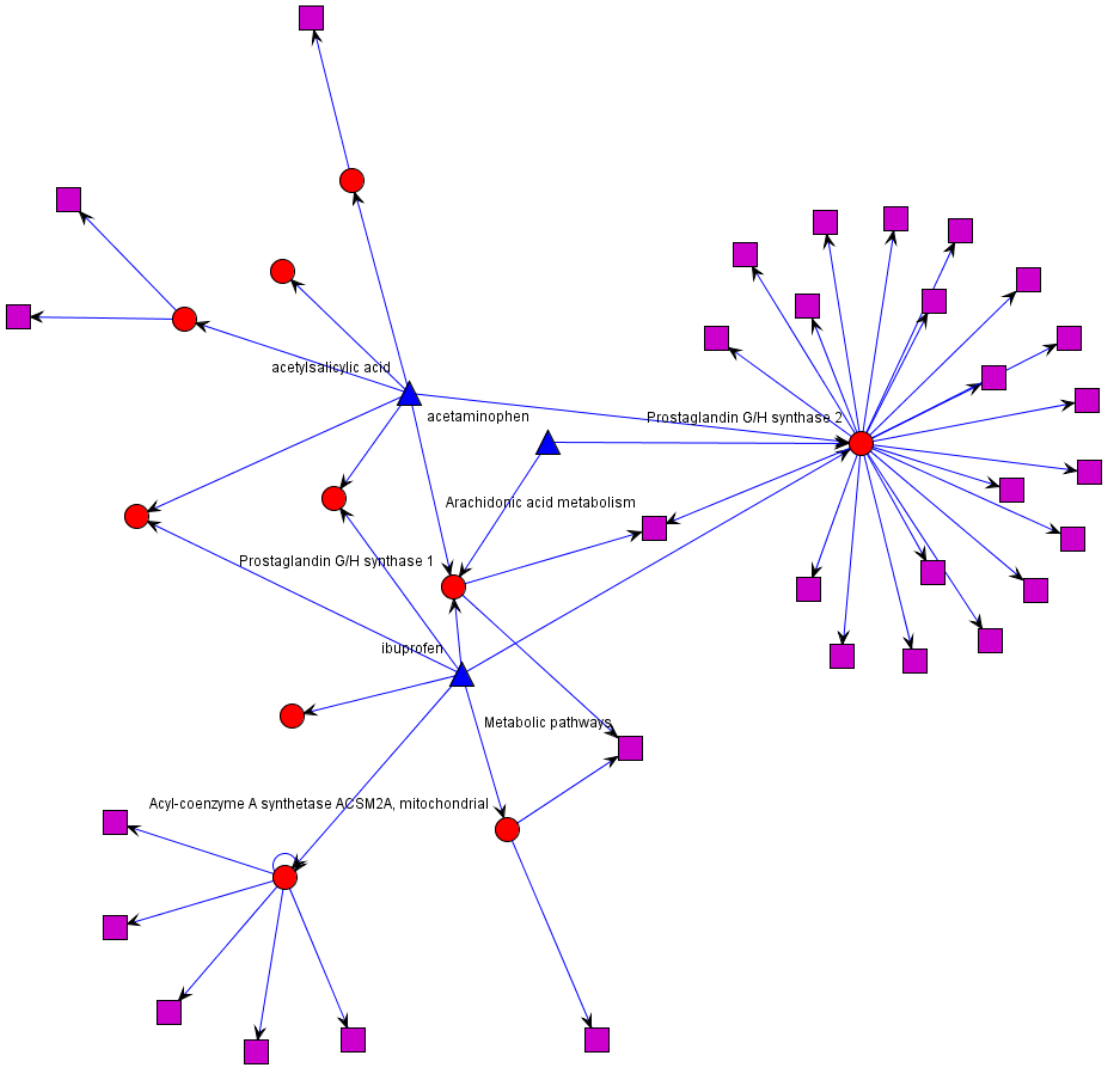


FIGURE 3.11 – Graphe des cibles de l'ibuprofène, l'aspirine et le paracétamol (acetaminophen). Les cibles de ces molécules sont en rouge et leurs réseaux biologiques en violet.

3.6 Conclusion

Dans ce chapitre, j'ai présenté le modèle de NetworkDB, un modèle générique de données permettant de structurer les informations concernant les réseaux biologiques. L'intérêt de ce modèle de données est qu'il peut être étendu pour traiter des problèmes variés, tels que l'étude des effets secondaires de médicaments ou la compréhension de l'étiologie des maladies génétiques. Pour cela, il suffit d'ajouter au modèle de NetworkDB les tables contenant les informations spécifiques au problème considéré. Le peuplement de la base de données se fait grâce à une série de tâches MODIM réutilisables qui permettent d'intégrer les données provenant de sources variées telles que UniProt, KEGG PATHWAY ou encore IntAct. La visualisation du contenu de la base peut ensuite se faire en utilisant des requêtes SQL pour les utilisateurs pouvant formuler des requêtes complexes. Pour les demandes plus fréquentes, une interface web couplée à une visualisation avec Ondex ont été mises en place.

Je détaillerai deux applications exploitant ces données dans les deux chapitres suivants. La première application concerne l'étude de l'étiologie des maladies génétiques et plus particulièrement des retards mentaux (chapitre 4). Dans la seconde application, je chercherai à comprendre les mécanismes associés aux effets secondaires de médicaments (Chapitre 5).

Chapitre 4

Réseaux biologiques pour caractériser les gènes responsables de déficiences intellectuelles

Sommaire

4.1 Introduction

La recherche de gènes candidats est une étape majeure dans la compréhension des maladies génétiques. De nos jours, grâce au développement des techniques de séquençage à haut débit, il est possible de séquencer l'exome des patients atteints afin de pouvoir localiser la mutation responsable. Cependant, des difficultés subsistent pour identifier cette mutation parmi tous les polymorphismes. De plus, dans un certain nombre de cas, l'identification du gène portant la mutation ne permet pas de comprendre le mécanisme associé à cette maladie. En effet, le gène muté n'est pas forcément bien décrit dans la littérature ou dans les bases de données. Une approche basée sur l'intégration de données provenant de différentes sources et prenant en compte les connaissances sur les réseaux d'interactions peut permettre d'aider à mieux comprendre l'étiologie des maladies génétiques.

Dans ce chapitre, je présenterai d'abord deux cas particuliers étudiés au laboratoire de génétique Humaine : les gènes *KIAA1468* et *MBD5*, qui illustrent bien notre approche. Ensuite je montrerai comment l'utilisation de la base de données intégrée NetworkDB et l'outil de visualisation ONDEX peuvent aider à structurer et préciser les connaissances sur les déficiences intellectuelles liées à l'X (DILX).

4.2 Méthode empirique d'exploration d'un gène isolé

4.2.1 Le gène *KIAA1468*

Une délétion *de novo* intragénique du gène *KIAA1468* a été identifiée chez un enfant présentant un syndrome de West (épilepsie infantile) et des anomalies cérébrales sous forme d'hétérotopies nodulaires et d'une agénésie du corps calleux. Le gène *KIAA1468* code une protéine de 1216 acides aminés. Cette protéine n'est annotée par aucun terme GO. Ainsi, si nous nous limitons aux données d'annotation des banques de données (QuickGO), nous n'avons aucune indication quant à sa fonction.

Selon OMIM (entrée MIM 308350), le seul gène connu pour être associé au syndrome de West est le gène *ARX*. J'ai donc essayé de rechercher les points communs entre le produit du gène *KIAA1468* et la protéine ARX afin de pouvoir comprendre la fonction de la protéine KIAA1468. La première étape a consisté à rechercher des interactions communes aux 2 protéines en utilisant l'extension Bisogenet de Cytoscape (Martin *et al.*, 2010). A partir des noms de gène ou de protéine, cette extension permet d'interroger en une seule fois les bases de données suivantes : DIP, BIND, HPRD, BioGRID, MINT et IntAct puis de visionner sous forme de graphe les interactions entre protéines. Comme on peut le voir dans la Figure 4.1, KIAA1468 ne possède qu'une seule interaction avec UBC c'est-à-dire la Polyubiquitin-C, la protéine ARX ne possède aucune interaction connue. On ne peut donc pas rapprocher ces 2 protéines en utilisant leurs interactions. J'ai ensuite comparé la composition en domaine des deux protéines.

KIAA1468 est composée de 3 domaines : un motif de dimérisation LisH et 2 domaines "Armadillo-like helical" (Figure 4.2). La protéine ARX est composée d'un domaine "Homeodomain" et d'un do-

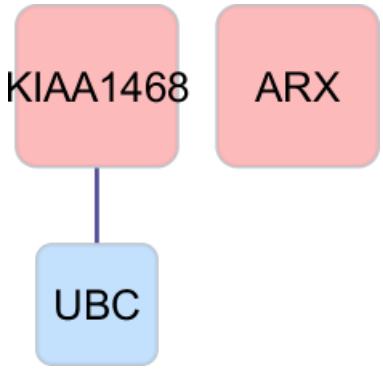


FIGURE 4.1 – Interaction protéine-protéine de KIAA1468 et ARX.

maine "OAR" (figure 4.2). Les 2 protéines n'ont pas de domaine commun. On ne peut donc pas déduire la fonction de KIAA1468 à partir d'ARX. J'ai donc exploré les fonctions connues des domaines de KIAA1468.

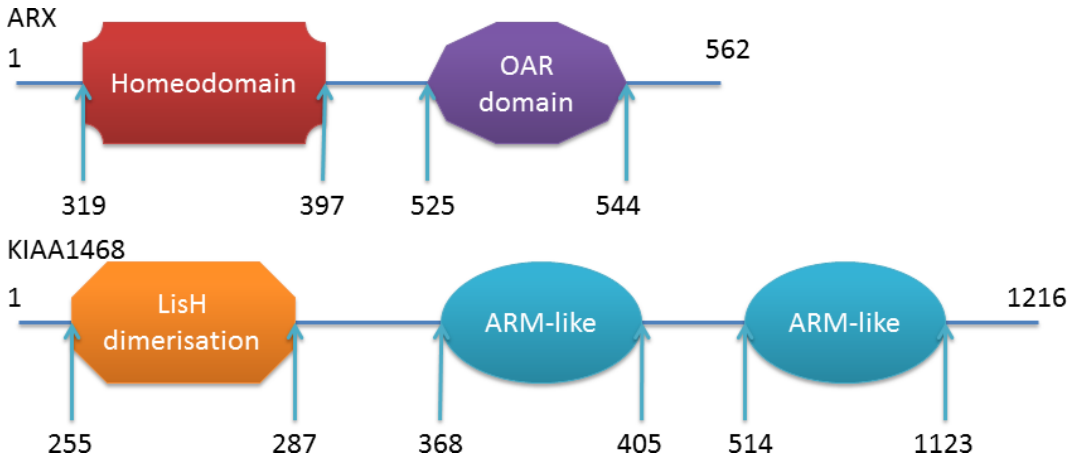


FIGURE 4.2 – Composition en domaines de KIAA1468 et ARX. Les domaines sont extraits par le programme InterProScan (Quevillon *et al.*, 2005).

Les domaines "Armadillo-like helical" sont des domaines principalement composés d'hélices α (Figure 4.3) présentant une large surface accessible au solvant permettant la liaison de protéines (Andrade *et al.*, 2001). Le domaine LisH est un domaine permettant la dimérisation de la protéine (Mateja *et al.*, 2006) et est certainement impliqué dans le complexe dyneine/dynactine lié aux microtubules (Emes et Ponting, 2001). Les mutations de ce domaine observées dans différentes protéines, ont entrainé un défaut de migration cellulaire dont par exemple un défaut de migration de neurones quand la protéine LIS1 ("Lissencephaly-1") est mutée (Emes et Ponting, 2001). Les études de Mateja *et al.* (2006) ont montré que le domaine LisH est indispensable à la dimérisation de la protéine LIS1. Ils ont également remarqué que la présence d'hélices à la suite du motif LisH est nécessaire à la dimérisation de la protéine. De plus, il est suggéré que les protéines possédant un domaine LisH suivi d'hélices possèdent un rôle similaire à celui de LIS1 qui est notamment impliquée dans la lissencéphalie de Dieker (MIM 607432). Cette pathologie inclut une hypoplasie du corps calleux, comme le syndrome de West.

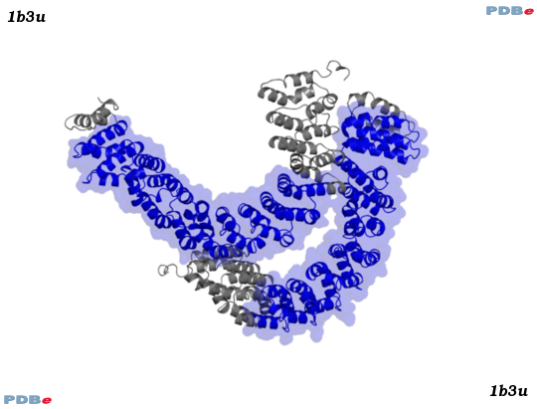


FIGURE 4.3 – Structure, en bleu, du domaine Armadillo-like helical de la protéine phosphatase PP2A (P30153).

La protéine KIAA1468 possédant une organisation de type LisH-hélices, on peut supposer que son mécanisme est proche de celui de LIS1. Ainsi, comme LIS1, KIAA1468 pourrait être impliquée dans un processus de dimérisation et de formation de complexes avec les microtubules, ce qui pourrait expliquer les symptômes impliquant une mauvaise migration neuronale lors du développement du cerveau.

4.2.2 Le gène *MBD5*

Des délétions de la région chromosomique 2q23.1 ont été décrites comme étant associées à des présentations cliniques proches du syndrome du Rett (MIM 312750, Table 4.1). Dans l'intervalle génomique délété est inclus le gène *MBD5*, gène qui partage un domaine fonctionnel avec le gène *MECP2* responsable du syndrome de Rett. Le séquençage du gène *MBD5* a été réalisé dans une population d'individus avec déficience intellectuelle sévère. Chez un garçon âgé de 10 ans une mutation non-sens a été identifiée dans le gène *MBD5*. J'ai donc essayé de rechercher les points communs entre le gène *MBD5* et le gène *MECP2*.

Symptômes	Mutation MBD5	Syndrome de Rett
Retard psycho moteur sévère	oui	oui
Hypotonie	oui	oui
Absence de langage	oui	oui
Pas de marche acquise	oui	oui
Epilepsie	oui	oui
Rires immotivés	oui	oui

TABLE 4.1 – Comparaison des symptômes du patient présentant une mutation dans *MBD5* et le syndrome de Rett.

Les protéines issues de ces 2 gènes appartiennent toutes deux à la famille des protéines à "Methyl-CpG binding domain" MBD et contiennent donc un domaine de liaison aux Methyl-CpG qui confie à la protéine la capacité de se fixer sur l'ADN methylé. (Figure 4.4). Les protéines de cette famille permettent de réguler la transcription de l'ADN en compactant la structure de la chromatine. Contrairement aux autres membres de cette famille (dont MECP2), MBD5 ne se fixe pas spécifique-

ment sur l'ADN méthylé (Laget *et al.*, 2010).

FIGURE 4.4 – Composition en domaine des protéines MBD5 et MECP2. Les domaines sont extraits par InterProScan.

Alors que MECP2 n'est composée que du domaine MBD, MBD5 contient également un domaine PWWP. C'est un domaine qui est caractérisé par la séquence suivante : Proline-Tryptophane-Tryptophane-Proline et qui a été décrit comme pouvant se fixer à l'ADN double brin de manière non spécifique (Laguri *et al.*, 2008) et avec la lysine 20 monométhylée de l'histone H4 (Wang *et al.*, 2009).

Comme précédemment, la recherche d'interaction entre les 2 protéines s'est faite en utilisant l'extension Bisogenet de Cytoscape. Les protéines MBD5 et MECP2 n'ayant pas d'interactant direct commun (Figure 4.5), j'ai augmenté le "rayon" de recherche afin de récupérer les interactants des interactants des deux protéines. Étant donné le nombre important de protéines (Figure 4.6), il est difficile de relier de façon optimale MECP2 et MBD5. Ainsi, en utilisant le chemin le plus court entre MECP2 et MBD5 donnée par Bisogenet, j'ai filtré le réseau pour ne garder que les protéines permettant de relier le plus rapidement possible MECP2 et MBD5 (Figure 4.7).

Bien que MBD5 et MECP2 ne partagent pas d'interaction, elles sont reliées via 2 protéines. Cette proximité d'interaction entre les 2 protéines pourrait expliquer la proximité entre les symptômes observés (Table 4.1). A cette proximité d'interaction peut s'ajouter l'effet de la mutation non-sens (figure 4.8) qui entraine la disparition de près de 90% de la protéine MBD5. Ainsi, seul le domaine MBD est présent. Ce dernier n'étant pas connu pour interagir avec des protéines, il y a probablement perte de l'interaction avec la protéine APC ("Adenomatous polyposis coli protein"). La perte de cette interaction est d'autant plus importante pour la compréhension de la maladie, que la protéine APC est impliquée dans la stabilisation des microtubules des cellules du cortex notamment dans les processus de motilité cellulaires (Zaoui *et al.*, 2010).

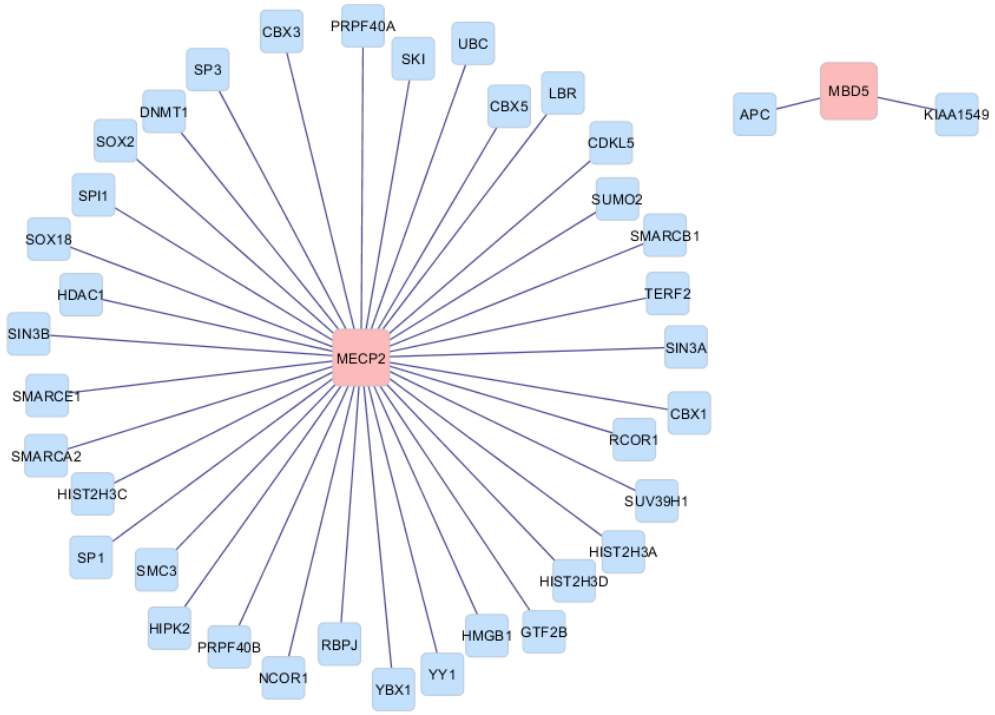


FIGURE 4.5 – Interactions des protéines MECP2 et MBD5. La protéine MBD5 possède 2 interactants et la protéine MECP2 interagit avec 36 protéines.

4.2.3 Généralisation de la méthode utilisée pour les gènes *KIAA1468* et *MBD5* à d'autres cas

La méthodologie appliquée au deux exemples développés précédemment est généralisable pour l'étude d'autres gènes. Elle est divisible en 4 étapes :

1. Rechercher dans OMIM les gènes responsables de la maladie ayant un phénotype similaire à celui observé chez le patient présentant la mutation.
2. Collecter les termes GO associés au gène muté et aux gènes des maladies similaires. S'ils partagent un/des processus communs, alors il est probable que la similarité observée s'explique par ce(s) processus.
3. S'il n'y a pas de termes GO communs, alors, on peut s'intéresser à la recherche d'interactants communs. La recherche d'interactant peut se faire grâce à l'extension BisoGenet de Cytoscape. En effet, cette extension va chercher tous les interactants de niveau 1, 2 ou 3 de chaque protéine donnée en entrée et permettre la visualisation du réseau obtenu sous forme de graphe. Si on observe une interaction directe, alors le phénotype observé peut être dû à la perte de l'interaction entre les 2 protéines.
4. Si les protéines ne peuvent pas être reliées entre elles ou si le lien n'est pas évident, il est intéressant d'étudier la composition en domaine de chaque protéine. Pour cela, il est nécessaire de les soumettre au programme InterProScan. Si les protéines partagent un domaine dont la fonction est connue et qui pourrait expliquer la maladie, alors la similarité peut être établie. Dans le cas contraire, les protéines partageant le(s) même(s) domaine(s) que la pro-

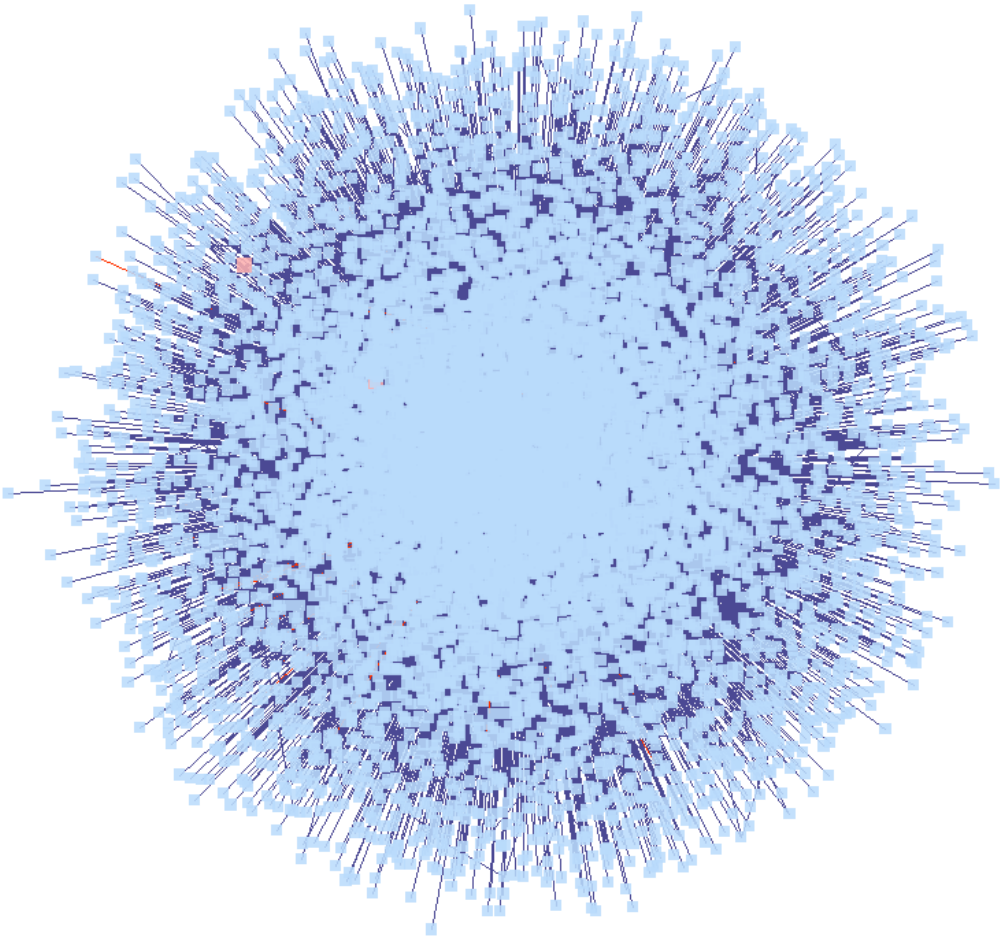


FIGURE 4.6 – Réseau d'interactions entre MBD5 et MECP2. Le réseau contient les interactants des interactions directes des deux protéines. Au total 8922 protéines sont affichées.

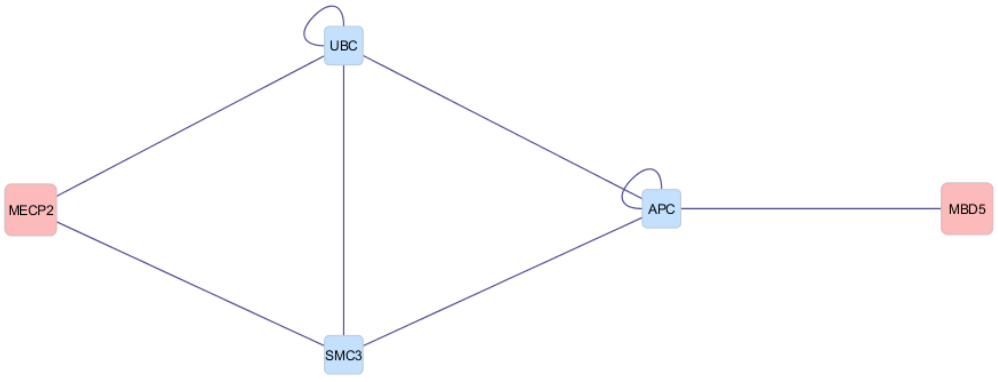


FIGURE 4.7 – Chemin le plus court entre MBD5 et MECP2.

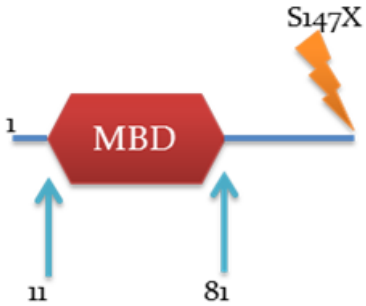


FIGURE 4.8 – Effet de la mutation non-sens S147X sur la protéine MBD5.

téine mutée peuvent permettre de comprendre par analogie le rôle de la protéine mutée. En parallèle, l'étude de l'effet de la mutation sur la protéine peut permettre d'appuyer sur une théorie (par exemple, la mutation entraîne la perte d'un domaine susceptible d'interagir avec une autre protéine).

4.3 Analyse d'un ensemble de gènes : réseaux biologiques impliqués dans les déficiences intellectuelles liées à l'X

4.3.1 Fonctions associées aux gènes des DILX

Des études récentes ont montré que les gènes du chromosome X pourraient être à l'origine de 10% à 15% des déficiences intellectuelles (Ropers, 2010). Ainsi, 105 gènes parmi les 826 gènes du chromosome X codant une protéine[27], soit environ 13% des gènes du chromosome X, sont reconnus pour être impliqués dans une Déficience Intellectuelle Liée à l'X (DILX). La liste des 105 gènes est disponible en annexe D.

Les gènes responsables de DILX participent à des processus cellulaires variés. Parmi ces processus certains sont très généraux tels que la régulation de l'expression génique et d'autres sont plus spécifiques (régulation du cytosquelette et morphogenèse neuronale). Le fonctionnement et la structure de la synapse semblent jouer un rôle majeur tout comme la dérégulation de la transcription et le remodelage de la chromatine (van Bokhoven, 2011).

4.3.2 Peuplement de NetworkDB

Étant donné qu'il est peu envisageable d'étudier individuellement les 105 gènes responsables de DILX, l'analyse de leurs fonctions connues et des réseaux biologiques dans lesquels ils interviennent peut permettre de confirmer les grandes fonctions associées aux gènes DILX ainsi que de mettre en évidence de nouvelles fonctions ayant un rôle dans les DILX. Pour cela, j'ai peuplé NetworkDB à partir de ces 105 gènes et en suivant la procédure décrite dans le chapitre 3.3 (Table 4.2).

Table	Nombre d'éléments
Gene	105
Protein	159
Protein_protein_interaction	67
Pathway	341
Pathway_has_protein	583
GO_term	1426
Protein_has_goterm	3757

TABLE 4.2 – Statistiques de la base de donnée NetworkDB.

Parmi les 159 protéines collectées, 105 proviennent des gènes des DILX et 54 sont des interactants d'au moins une protéine parmi les 105. Seules 38 protéines parmi les 105 ne possèdent pas d'interactants dans IntACT. On dénombre également 101 protéines (DILX et interactants) associées à au moins un réseau biologique et un réseau annote de 1 à 19 protéines différentes. Les termes GO sont groupés dans trois branches : "Biological Process", "Molecular Function" et "Cellular Component" (BP, MF et CC). Ainsi, au sein des 1426 termes GO, 872 sont des BP, 326 appartiennent à la branche MF et 228 correspondent à des CC. Au maximum, un terme GO annote respectivement 31, 72 et 67 protéines.

27. http://vega.sanger.ac.uk/Homo_sapiens/Location/Chromosome?r=X, version 50, décembre 2012

4.3.3 Enrichissement fonctionnel en termes GO

L'enrichissement fonctionnel en terme GO permet d'obtenir la liste des termes GO les plus spécifiquement associés à une liste de gènes. Les termes "enrichis" sont ceux dont la distribution au sein des gènes étudiés est la plus différente de leur distribution au sein de la population.

Afin de déterminer si ces gènes ont une fonction similaire, un outil en ligne tel que DAVID (Database for Annotation, Visualization and Integrated Discovery), permettant de calculer l'enrichissement en annotation fonctionnelles est intéressant car il va permettre de déterminer quelles annotations sont spécifiques à un groupe de gènes (Huang *et al.*, 2009). En soumettant la liste des 105 gènes à DAVID, on obtient un enrichissement significatif (P-Value$\leq 10^{-3}$) pour 10 termes GO BP impliqués notamment dans la morphogenèse des neurones et le développement du système nerveux (Table 4.3). Ceci confirme que les gènes impliqués dans des DILX ont un rôle dans le développement du système nerveux, mais cela ne donne pas plus d'information sur des fonctions plus spécifiques. Ceci peut être expliqué par le grand nombre de gènes utilisés pour calculer l'enrichissement. Ainsi, leurs fonctions plus précises pourraient être perdues lors du calcul de l'enrichissement car ils ne partagent pas tous la même fonction.

Terme GO (BP)	P-Value
central nervous system neuron development	$1,3.10^{-3}$
cell projection morphogenesis	$1,3.10^{-3}$
cell part morphogenesis	$1,7.10^{-3}$
central nervous system neuron differentiation	$2,4.10^{-3}$
cell morphogenesis	$2,710^{-3}$
neuron projection morphogenesis	$3,1.10^{-3}$
cell projection organization	$3,3.10^{-3}$
cellular component morphogenesis	$5,2.10^{-3}$
neuron projection development	$7,6.10^{-3}$
dopamine metabolic process	$7,9.10^{-3}$

TABLE 4.3 – Enrichissement en termes GO BP par rapport à l'ensemble de l'annotation des gène de l'Homme calculés par DAVID pour les 105 gènes DILX. Seuls les termes GO ayant une P-Value inférieure ou égale à 10^{-3} sont affichés.

Une solution à ce problème est de regrouper les gènes ayant une fonction similaire, avant de rechercher un enrichissement fonctionnel. J'ai ainsi calculé la similarité fonctionnelle des 105 gènes en utilisant IntelliGO, une mesure de similarité sémantique entre termes GO (Benabderrahmane *et al.*, 2010). IntelliGO est une mesure de similarité qui prend en compte les relations sémantiques entre les termes GO, le contenu en information de ces termes (inversement proportionnel à leur fréquence dans l'ensemble des annotation de gènes), ainsi que la provenance de ces annotations (décrite dans la littérature, inférées automatiquement, ...). La similarité entre deux gènes une adaptation du cosinus généralisée proposée par Ganesan *et al.* (2003) pour les vocabulaires hiérarchiques aux graphes dirigés acycliques (DAG). IntelliGO a été définie pour quantifier la similarité entre les termes GO. La similarité entre deux termes t_i et t_j est fonction de la profondeur maximale de leur ancêtre commun le plus spécifique (LCA, Least Common Ancestor) et du plus court chemin

(SPL, Shortest Path Length) entre les deux termes dans le DAG :

$$Sim_{IntelliGO}(t_i, t_j) = \frac{2 \cdot Profondeur(LCA)}{SPL(t_i, t_j) + 2 \cdot Profondeur(LCA)} \tag{4.1}$$

La similarité entre deux gènes g et h, représentés par leur vecteur de termes GO $\vec{g}$ et $\vec{h}$,respectivement est alors définie par :

$$Sim_{IntelliGO}(g, h) = \frac{\vec{g} * \vec{h}}{\sqrt{\vec{g} * \vec{g}}\sqrt{\vec{h} * \vec{h}}} \tag{4.2}$$

Où :

- ▷ $\vec{g} = \sum_i \alpha_i * \vec{e_i}$: la représentation vectorielle du gène g dans l'espace vectoriel de la mesure IntelliGO.
- ▷ $\vec{h} = \sum_j \beta_j * \vec{e_j}$: la représentation vectorielle du gène h dans l'espace vectoriel de la mesure IntelliGO.
- ▷ α_i et β_j sont les coefficients des termes t_i et t_j pour les gènes g et h respectivement selon la mesure IntelliGO.
- ▷ $\vec{g} * \vec{h} = \sum_{i,j} \alpha_i * \beta_j * \vec{e_i} * \vec{e_j}$: représente le produit scalaire entre les deux vecteurs des gènes g et h.
- ▷ $\vec{e_i} * \vec{e_j} = Sim_{IntelliGO}(t_i, t_j)$ telle que définie ci dessus.

La visualisation sous forme de heatmap de la matrice de similarité, obtenue à partir de termes BP et grâce à IntelliGO, permet de déterminer visuellement quels groupes de gènes sont similaires (Figure 4.9). Ainsi, en considérant le cluster noté 2 sur la figure 4.9 comme étant suffisamment homogène et que l'on coupe toutes les branches du dendrogramme au niveau ayant permis d'obtenir ce cluster, on obtient 40 "groupes de gènes" dont 23 sont des singletons (Figure 4.10).

La mesure de similarité IntelliGO nous a permis de grouper des gènes appartenant à un processus biologique proche mais ne nous renseigne pas sur ce processus. En soumettant à l'outil DAVID les 17 clusters ayant au moins 2 gènes, on obtient un enrichissement significatif en termes GO BP pour seulement 3 clusters (2, 15 et 24 sur la figure 4.9). Ainsi, les gènes du cluster 2 ont plutôt une action sur la transcription de l'ADN et sa régulation, les gènes du cluster 15 sont impliqués dans la transduction de signaux et ceux du cluster 24 ont un rôle dans le métabolisme de petites molécules parmi lesquelles on trouve des neurotransmetteurs (Table 4.4). Pour ces 3 clusters, on retrouve donc au moins deux des principales fonctions associées aux déficiences intellectuelles liées à l'X. Pour les 14 autres clusters testés (fourni en annexe C.1), DAVID n'obtient pas d'enrichissement significatif bien qu'IntelliGO nous dise que les gènes ont une fonction similaire. Cela peut être expliqué par le fait que DAVID n'utilise pas les relations sémantiques entre les termes GO pour calculer l'enrichissement. Ainsi, DAVID ne trouvera pas enrichissement pour un terme particulier car les gènes sont annotés par des termes différents mais ayant un sens proche.

4.3.4 Étude des réseaux biologiques

En parallèle à l'étude de l'enrichissement fonctionnel, j'ai tenté de regrouper les gènes DILX grâce à leurs interactions et aux réseaux biologiques dans lesquelles ils interviennent. Pour cela,

Terme GO (BP)	P-Value	Cluster
GO :0006350 transcription	$2,94.10^{-4}$	2
GO :0045449 regulation of transcription	$9,84.10^{-4}$	2
GO :0046578 regulation of Ras protein signal transduction	$7,13.10^{-5}$	15
GO :0051056 regulation of small GTPase mediated signal transduction	$1,23.10^{-4}$	15
GO :0035023 regulation of Rho protein signal transduction	$7,8.10^{-4}$	15
GO :0008624 induction of apoptosis by extracellular signals	$9,97.10^{-4}$	15
GO :0006576 biogenic amine metabolic process	$3,54.10^{-6}$	24
GO :0006575 cellular amino acid derivative metabolic process	$1,78.10^{-5}$	24
GO :0042417 dopamine metabolic process	$2,07.10^{-5}$	24
GO :0006584 catecholamine metabolic process	$6,10.10^{-5}$	24
GO :0034311 diol metabolic process	$6,10.10^{-5}$	24
GO :0009712 catechol metabolic process	$6,10.10^{-5}$	24
GO :0018958 phenol metabolic process	$6,47.10^{-5}$	24
GO :0044271 nitrogen compound biosynthetic process	$1,33.10^{-4}$	24

TABLE 4.4 – Enrichissement des 17 clusters en termes GO BP par rapport à l'ensemble de l'annotation des gène de l'Homme calculés par DAVID. Seuls les termes GO ayant une P-Value inférieure ou égale à 10^{-3} sont affichés.

j'ai généré le fichier XML permettant de visualiser le contenu de la base de données sous forme de graphe avec Ondex. Chaque élément de NetworkDB (protéine, gène, réseau biologique, ...) est représenté sous forme d' un nœud, les relations entre les éléments sont représentées par des arcs reliant les nœuds (Figure 4.11).

En essayant de grouper les protéines DILX par leur interactions (Figure 4.12), on remarque que l'on peut former seulement quatre réseaux ne rassemblant chacun que 2 protéines DILX. Ainsi, le regroupement des protéines DILX en utilisant seulement leurs interactants est très limité.

Une autre approche pour mieux comprendre les fonctions des gènes des déficiences mentales liées à l'X consiste à utiliser les réseaux biologiques. Ainsi, en associant les protéines à leurs réseaux biologiques (Table 4.5) et en explorant le réseau ainsi formé, on peut former trois grands groupes (A, B, C) et un quatrième plus petit (D) sur la base des similarités entre réseaux biologiques (Figure 4.13).

Les gènes du groupe A sont plutôt impliqués dans la signalisation cellulaire. En effet, la plupart des gènes de ce groupe sont impliqués dans des voies de signalisation ou appartiennent au type de voie KEGG "traitement de l'information environnementale". Les gènes du groupe B sont associés à la régulation de l'expression génétique. Au sein de ce groupe, on peut remarquer que les réseaux biologiques permettant de relier les protéines entre elles sont principalement les voies de traitement de l'information génétique et les voies de type "expression génique". Ce groupe contient également des gènes liés à la dégradation des protéines. En effet, bien que cette fonction ne soit pas directement liée à l'expression des gènes, elle permet néanmoins de réguler les processus biologiques en agissant sur la quantité de protéines disponible pour réaliser ces processus. Les gènes du groupe C sont plutôt impliqués dans le fonctionnement des synapses. En effet, les gènes appartenant au groupe C sont impliqués dans le métabolisme de petites molécules parmi lesquelles on trouve des neurotransmetteurs. Le groupe D permet de regrouper des gènes liés au métabolisme des stéroïdes et du cholestérol.

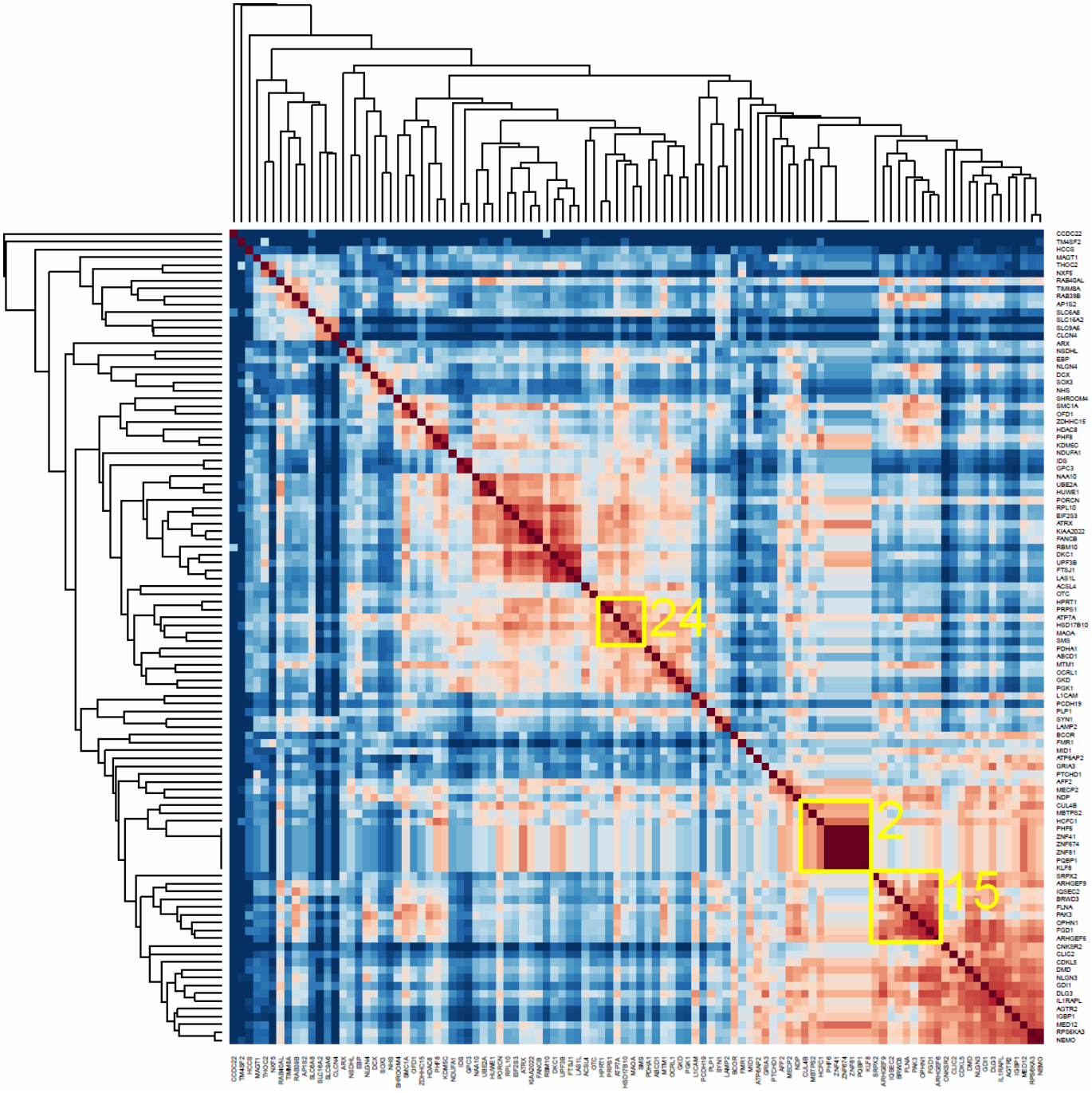


FIGURE 4.9 – Clustering hiérarchique et visualisation par heatmap de la similarité IntelliGO entre les 105 gènes DILX. Les trois groupes de gènes en jaune présentent un enrichissement en termes GO selon DAVID.

Les quatre groupes A, B, C et D permettent de retrouver les grandes "fonctions" associées aux déficiences intellectuelles liées à l'X. On y retrouve par exemple la régulation de l'expression génique (groupe B) et le fonctionnement des synapses (groupe C). Cependant, d'autres fonctions semblent avoir un rôle important. Ainsi les gènes du groupe A, impliqués dans les voies de signalisations, représentent 14 des 43 gènes représentés sur la Figure 4.13. Au sein des groupes A et B il est

Réseau biologique	Type PID ou KEGG	Source
Adipocytokine signaling pathway	Organismal Systems	KEGG
Arginine and proline metabolism	Metabolism	KEGG
Axon guidance	Organismal Systems	KEGG
Cap-dependent Translation Initiation	Gene expression	PID
Cell adhesion molecules (CAMs)	Environmental Information Processing	KEGG
Cholesterol biosynthesis	metabolism of lipids and lipoproteins	PID
Dopaminergic synapse	Organismal Systems	KEGG
Focal adhesion	Cellular Processes	KEGG
Glycolysis / Gluconeogenesis	Metabolism	KEGG
HIF-1 signaling pathway	Environmental Information Processing	KEGG
Inositol phosphate metabolism	Metabolism	KEGG
Lysosome	Cellular Processes	KEGG
MAPK signaling pathway	Environmental Information Processing	KEGG
mRNA Splicing - Major Pathway	Gene expression	PID
mRNA surveillance pathway	Genetic Information Processing	KEGG
Neuroactive ligand-receptor interaction	Environmental Information Processing	KEGG
Oocyte meiosis	Cellular Processes	KEGG
Peroxisome	Cellular Processes	KEGG
Phosphatidylinositol signaling system	Environmental Information Processing	KEGG
PPAR signaling pathway	Organismal Systems	KEGG
Purine metabolism	Metabolism	KEGG
Regulation of actin cytoskeleton	Cellular Processes	KEGG
Ribosome biogenesis in eukaryotes	Genetic Information Processing	KEGG
RNA transport	Genetic Information Processing	KEGG
Spliceosome	Genetic Information Processing	KEGG
Steroid biosynthesis	Metabolism	KEGG
T cell receptor signaling pathway	Organismal Systems	KEGG
TCR signaling in naïve CD4+ T cells	Regulatory pathway	PID
Ubiquitin mediated proteolysis	Genetic Information Processing	KEGG

TABLE 4.5 – Liste des réseaux biologiques n'étant pas du type KEGG "Human Diseases" et permettant de relier au moins 2 protéines DILX.

également possible de définir des sous-groupes ayant des fonctions plus précises. Ainsi, les gènes *FLNA, NLGN4, NLGN3, PAK3, L1CAM, FGD1, ARHGEF6* dans le groupe A sont plutôt impliqués dans le contrôle du cytosquelette qui est une des fonctions traditionnellement associées aux DILX. De même, le groupe B peut être divisé en deux. Le premier sous-groupe est directement impliqué dans la régulation de l'expression des gènes (*RPL10, EIF2S3, FMR1, THOC2, PQBP1, DKC1, NXF5, UPF3B, RP56KA3, SMC1A*) et le second sous-groupe est associé à la dégradation des protéines (*HUWE1, UBE2A, MID1, CUL4B, IDS, LAMP2, AP1S2*). Il est intéressant de noter que même si ce deuxième sous-groupe est homogène au niveau des fonctions cellulaires, il forme deux réseaux distincts qui ne sont pas reliés entre eux.

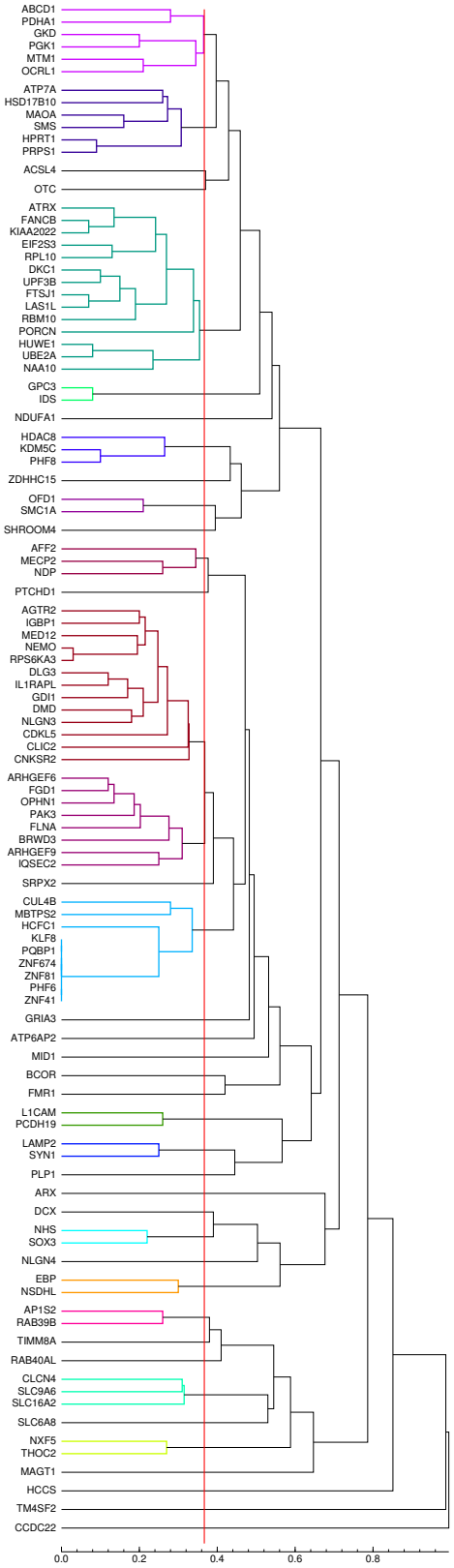


FIGURE 4.10 – Clustering hiérarchique (selon la méthode UPGMA) des 105 gènes DILX en utilisant la similarité IntelliGO. Le dendrogramme a été coupé afin d'obtenir 40 groupes de gènes. La ligne rouge indique le seuil utilisé pour définir chaque cluster.

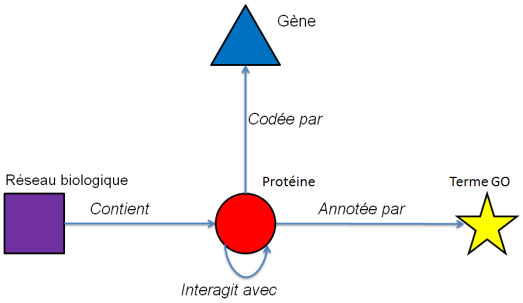


FIGURE 4.11 – Métagraphe Ondex représentant les relations entre les éléments de la base de données.

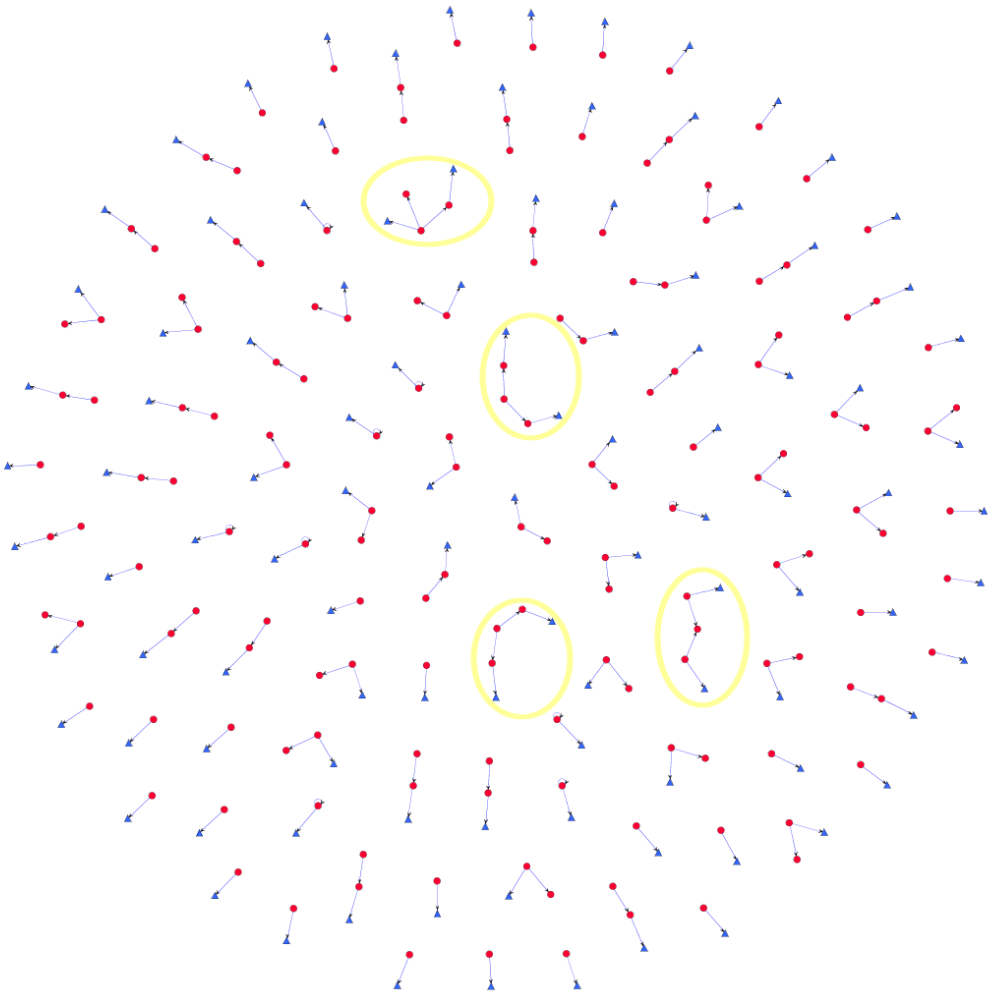


FIGURE 4.12 – Interactions protéine-protéine des 105 protéines codées par les gènes DILX. Les protéines DILX sont celles reliées à des gènes (triangles bleus). Les groupes en jaunes sont les réseaux faisant intervenir au moins 2 gènes DILX. Seules les protéines DILX ayant au moins une interaction protéine-protéine sont affichées.

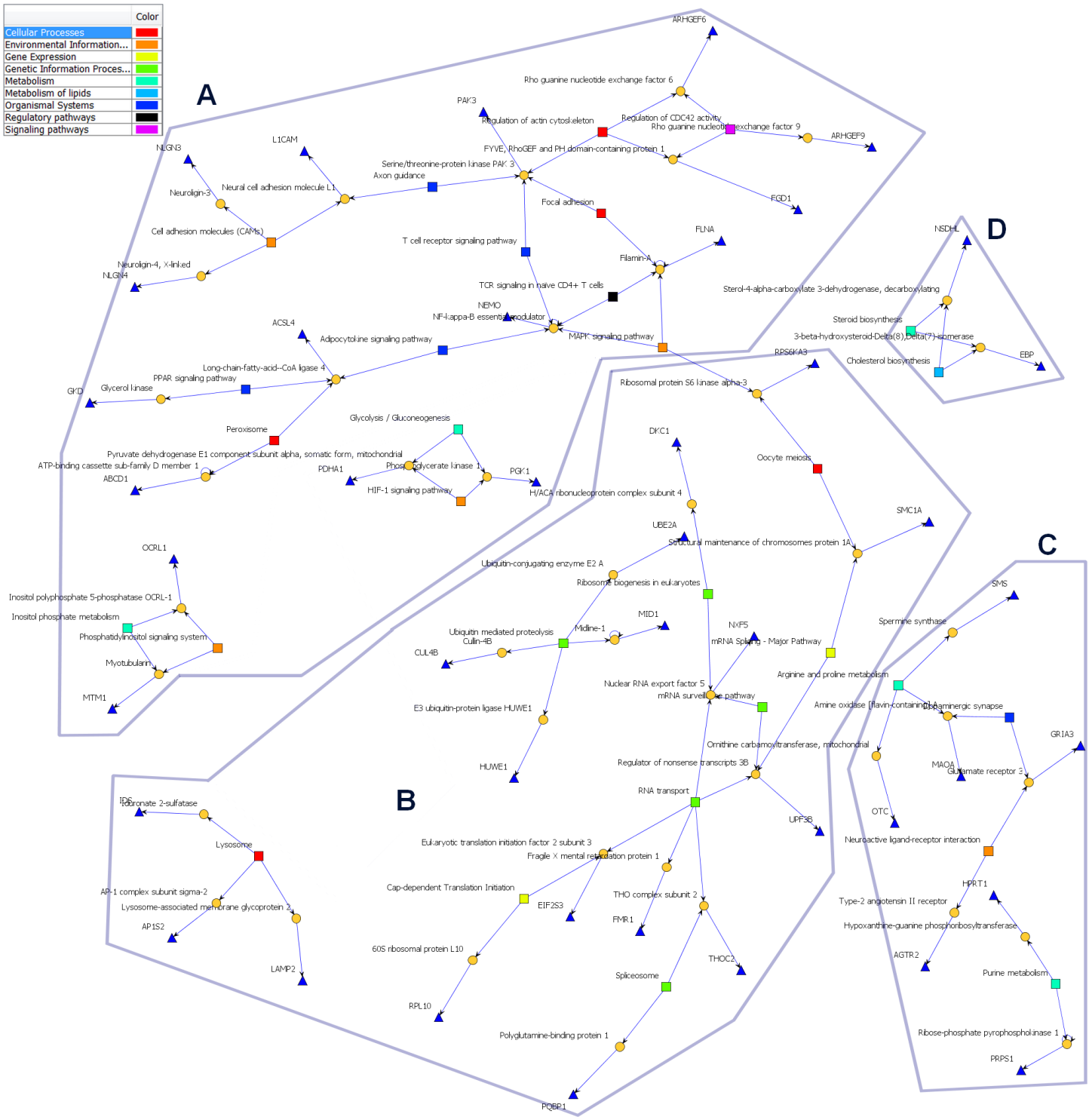


FIGURE 4.13 – Gènes DILX (sous forme de triangles bleus) regroupés par les processus biologiques (représentés par des carrés) associés à leur protéine (ronds orange). La couleur des processus biologiques dépend de leur type référencé dans les bases de données KEGG PATHWAY et PID. Seuls les 29 réseaux biologiques associés à au moins deux protéines sont affichés et de même, seules les 42 protéines associées à au moins un réseau biologique restant sont affichées. Les groupes en bleu sont des groupes de gènes associés à des fonctions similaires.

Groupe	Réseaux biologiques	Gènes	Processus proposé
A	Focal adhesion, axon guidance, cell adhesion molecules (CAMs), Regulation of actin cytoskeleton ,PPAR signaling pathway, peroxysome, Glycolysis/Gluconeogenesis, HIF-1 signaling pathway, Phosphatidylinositol signaling system, Inositol phosphate metabolism	*FLNA, NLGN4, NLGN3, PAK3, L1CAM, FGD1, ARHGEF6, ACSL4, ABCD1, GKD, PGK1, PDHA1, OCRL1, MTM1*	Signalisation cellulaire
B	Cap-dependent translation Initiation, RNA transport, Spliceosome, Ribosome biogenesis in eukaryotes, mRNA surveillance pathways, mRNA splicing - major pathway, Oocyte meiosis, MAPK signaling pathway, Ubiquitin mediated proteolysis, Lysosome	*RPL10, EIF2S3, FMR1, THOC2, PQBP1, DKC1, NXF5, UPF3B, RP56KA3, SMC1A, HUWE1, UBE2A, MID1, CUL4B, IDS, LAMP2, AP1S2,*	Contrôle de l'expression
C	Neuroactive ligand receptor interaction, Dopaminergic synapse, arginine and proline metabolism, Purine metabolism,	*AGTR2, GRIA3, MAOA, OTC, SMS, PRPS1, HPRT1*	Métabolisme des petites molécules
D	Steroid biosynthesis, Cholesterol biosynthesis	*NSDHL, EBP*	Synthèse des stéroïdes

TABLE 4.6 – Composition des groupes de gènes issus la représentation en réseaux des gènes DILX et de leurs réseaux biologiques.

4.3.5 Conclusion

L'enrichissement fonctionnel comme l'étude des réseaux biologiques impliquant 53 des 105 gènes connus pour être responsables de déficience intellectuelle liée au chromosome X, ont permis de mettre en évidence les grands mécanismes associés à ces maladies. Ainsi, on retrouve les mécanismes déjà connus tel que la régulation de l'expression génique, le contrôle du cytosquelette et le fonctionnement des synapses. Des analyses révèlent également d'autres mécanismes tels que la dégradation protéique, qui semblent avoir un rôle important dans ces maladies. De plus, le fait que seuls 11 gènes sont retrouvés dans les deux approches montre la complémentarité qui existe entre ces deux méthodes. Cependant, à peine plus de la moitié des gènes étudiés ont pu être caractérisés. La prise en compte des interactants, des réseaux biologiques dans lesquelles ils interviennent et/ou de leurs termes GO pourrait permettre d'augmenter le nombre gènes caractérisés. Malheureusement, du fait du nombre important d'éléments que cela ajoute au graphe, il serait difficilement possible d'analyser ce graphe de manière visuelle. Afin d'analyser ces données une solution serait d'utiliser des méthodes de fouille de données. Certaines de ces méthodes, comme la programmation logique inductive (PLI), sont particulièrement bien adaptées aux données représentées sous forme relationnelle. L'utilisation de la PLI permettrait ainsi de caractériser les gènes provoquant des DILX en prenant en compte l'ensemble de leurs propriétés et des propriétés de leurs interactants.

Chapitre 5

Caractérisation de profils d'effets secondaires de médicaments

Sommaire

5.1 Introduction

Les effets secondaires sont des réponses indésirables obtenues à la suite de prise de médicaments. La plupart d'entre eux sont détectés durant les essais cliniques et sont responsables du fort taux d'attrition observé durant ces phases. Ainsi, en 2008 le ministère de l'industrie estimait que seulement un médicament sur 1250 est approuvé par la FDA pour une mise sur le marché [28]. De plus, même si ce taux d'attrition est assez important, tous les effets secondaires ne sont pas détectés lors des essais cliniques. Par exemple, les effets cardiotoxiques du benfluorex (Mediator) ont été mis en évidence récemment alors que la molécule est utilisée depuis les années 70 (Frachon *et al.*, 2010, Derumeaux *et al.*, 2012). Au-delà de la toxicité, certains effets secondaires sont particulièrement indésirables pour certaines prescriptions. Ainsi, il est peu envisageable de prescrire un médicament entrainant des nausées ou des maux de tête pour un traitement de longue durée. Il est donc important de pouvoir connaitre les mécanismes associés aux effets secondaires afin de limiter leur apparition lors du développement de nouveaux médicaments.

Les principales sources référençant les effets secondaires de médicaments sont AERS et SIDER (Cami *et al.*, 2011, Harpaz *et al.*, 2010, Huang *et al.*, 2011, Lee *et al.*, 2011, Liu *et al.*, 2012*b*, Tatonetti *et al.*, 2012, Yamanishi *et al.*, 2012). La base de données SIDER est issue du jeu de données utilisé par Campillos *et al.* (2008). Les effets secondaires de SIDER sont extraits des notices associées aux molécules lesquelles sont issues des systèmes d'alerte ou de surveillance sanitaire des États-Unis (gérés par la FDA). Les données issues de AERS sont totalement brutes car les rapports sont directement rédigés par les médecins (nom de médicaments et de molécules mélangés, lien avec les effets secondaires pas toujours évident). Contrairement à cela, les données de SIDER ont été préparées pour être facilement analysables et la nomenclature des effets secondaires a été mise en correspondance avec la terminologie MedDRA par Kuhn *et al.* (2010).

Afin de comprendre les mécanismes associés aux effets secondaires des médicaments, il est nécessaire d'utiliser un schéma étendu de NetworkDB (Figure 5.1). Pour cela, on ajoute au modèle décrit en section 3.2 (Figure 5.1, en vert), les informations sur les médicaments (cibles, catégories et appartenance à un cluster de molécules) et sur leurs effets secondaires (Figure 5.1 en jaune). Mis à part pour le clustering des effets secondaires et des molécules, la collecte des données suit la procédure décrite dans le chapitre 3.3.2.1, et est initiée par une liste de molécules provenant de SIDER. Afin de faire le lien entre les médicaments de SIDER et leurs cibles, j'ai utilisé les identifiants PubChem fourni par DrugBank et SIDER. Ainsi, en croisant les deux sources de données, on obtient une liste de 554 molécules partagées dont on connait les cibles grâce à DrugBank. En exécutant la tache MODIM collectant les interactions protéine-protéine (chapitre 3.3.2.1) à partir des 768 cibles des 554 médicaments, dont on connait les effets secondaires, on obtient un ensemble de 5156 interactions protéine-protéine. Ces interactions ajoutent 2868 nouvelles protéines. Pour l'ensemble des protéines collectées, on récupère 1403 réseaux biologiques, 4650 domaines et 6494 termes GO. Quatre mesures de similarité ont été utilisées pour réaliser 4 regroupements différents des 554 molécules. On obtient 60 clusters avec une mesure dite de Tanimoto prenant en compte les groupements fonctionnels et sous-structures présents dans les molécules, 53 clusters

28. www.dgcis.redressement-productif.gouv.fr/files/files/archive/www.industrie.gouv.fr/biblioth/docu/dossiers/sect/etude_pharma.pdf

avec la mesure HPCCgeo qui compare les formes 3D des molécules, représentées par leurs coefficients harmoniques sphériques, 21 clusters avec la mesure HPCCchem qui compare les propriétés physico-chimiques des molécules et 34 clusters avec la mesure HPCCcombo qui combine les deux précédentes (Karaboga *et al.*, 2013).

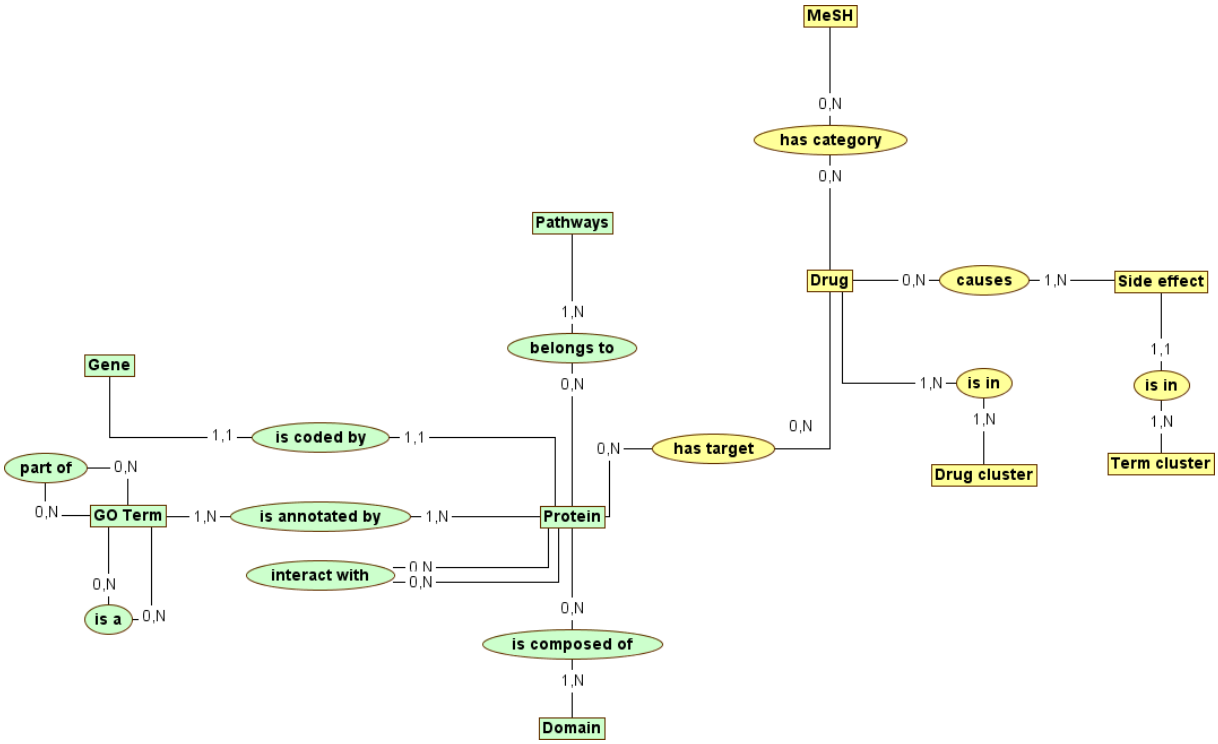


FIGURE 5.1 – Schéma entité-association de NetworkDB pour l'étude des effets secondaires de médicaments. Le cœur du modèle est représenté en vert et les ajouts sont en jaune.

Au vu de ce très grand nombre d'éléments, il apparait extrêmement difficile d'extraire les caractéristiques de médicaments (et de leurs cibles) partageant des effets secondaires, d'autant plus qu'il existe près de 50 000 associations entre médicaments et effets secondaires dans NetworkDB. La visualisation du réseau molécules-effets secondaires permet de se rendre compte de cette complexité (Figure 5.2). Ainsi, si on ajoute les différentes propriétés de cibles et des médicaments à ce réseau, il sera beaucoup trop complexe pour être analysé. Il en est de même pour une interrogation de NetworkDB via des requêtes SQL. En effet, le nombre très élevé de résultats aux requêtes empêchera toute interprétation manuelle des résultats. Par contre, il existe des méthodes de fouille de données adaptées pour analyser ce type de données, nous permettant de mieux comprendre les mécanismes d'apparition des effets secondaires.

Le problème de la caractérisation des molécules provoquant un effet secondaire peut être assimilé à un problème de classification supervisée. En effet, à partir d'un ensemble de molécules dont on connait la classe, il s'agit de déterminer quelles sont les propriétés spécifiques des molécules associées à cette classe. La classe d'une molécule peut être un effet secondaire isolé, ou alors un ensemble d'effets secondaires, les médicaments provoquant rarement un seul effet indésirable. En fonction du type de données utilisées en entrée, il existe deux types de méthodes permettant d'extraire de nouvelles connaissances à partir d'un jeu de données. Les premières méthodes utilisent

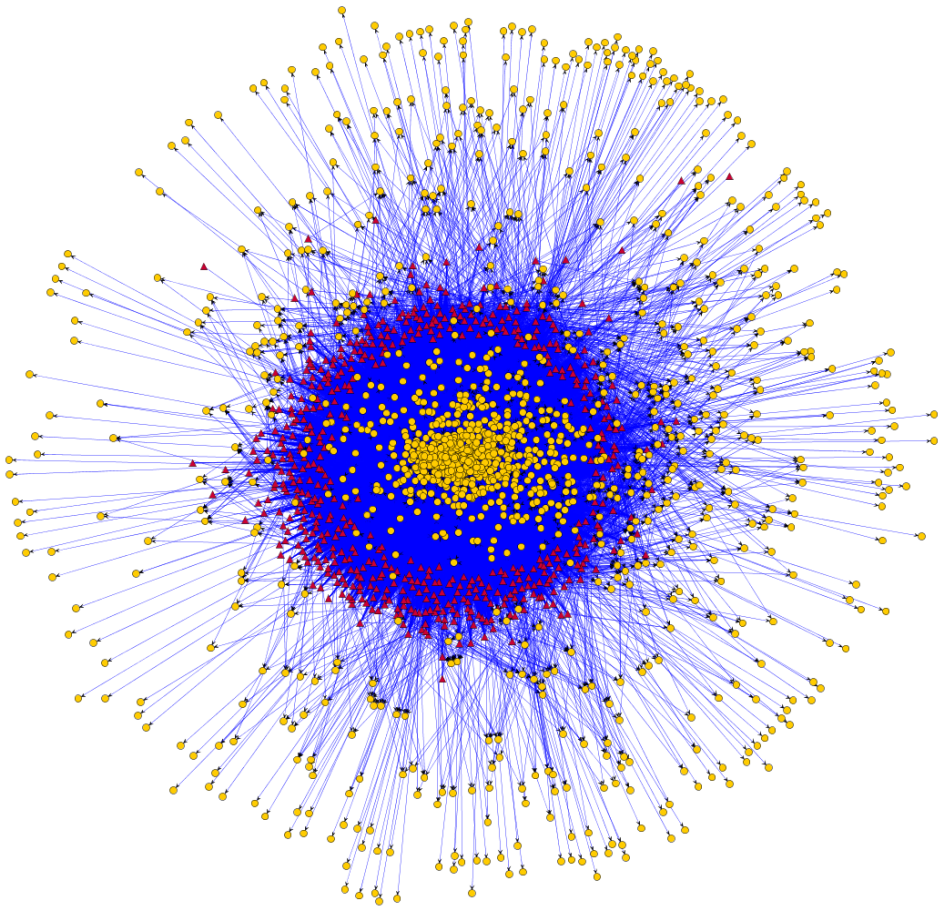


FIGURE 5.2 – Graphe des associations entre les 554 médicaments et les 1288 effets secondaires. Les médicaments sont représentés par des triangles rouges et les effets secondaires par des ronds orange. Un total de 49 471 associations est représenté. Les effets secondaires au centre de la figure sont ceux qui annotent un grand nombre de molécules. Au contraire, les effets secondaires à la périphérie sont associés à peu de molécules.

des données sous forme d'une matrice objets×attributs où chaque objet est décrit par un ensemble de propriétés. Ainsi, dans le cas de l'étude des effets secondaires, une molécule peut être décrite par ses cibles, par ses catégories, et par son appartenance à un cluster de molécules (Table 5.1). Ce

Molécule	Cible 1	Cible 2	Cible 3	Catégorie 1	Catégorie 2	Catégorie 3	...
M1	X			X		X	
M2		X	X			X	
M3	X	X		X	X		

TABLE 5.1 – Exemple de matrice objets×attributs représentant des molécules décrites par leurs effets secondaires et leurs cibles. Un X représente une association entre la molécule et la propriété correspondante.

type de représentation est notamment utilisé pour construire des arbres de décisions. Cependant, ce mode de représentation est limité. En effet, si un objet est décrit par un autre objet (par exemple, un médicament décrit par ses cibles), il n'est pas possible de prendre en compte les propriétés de ces cibles (par exemple, les termes GO des cibles). Pour cela, il existe d'autres méthodes de fouille

basées sur une représentation relationnelle des données. Ainsi, dans le cas de l'étude des effets secondaires, il devient possible de prendre en compte les cibles et leurs propriétés (composition en domaines, interactions protéine-protéine, ...). L'une des méthodes appliquée avec succès à de nombreux domaines dont la bioinformatique est la programmation logique inductive (PLI, Muggleton *et al.* 1998, Page et Craven 2003, Santos *et al.* 2012). Pour la PLI, les données relationnelles sont représentées par des prédicats décrivant à la fois les objets et leurs propriétés. Les prédicats utilisés pour caractériser les molécules sont présentés dans la table 5.2.

Prédicat	Définition
Propriétés des médicaments	
`category(A, C)`	Le médicament A appartient à la catégorie C
`drug_cluster(A,C,M)`	Le médicament A est un membre du cluster de molécules C déterminé par la méthode M
`drug_has_target(A,T,I)`	Le médicament A à une action I sur la protéine T
Propriétés des cibles	
`pathway(T,P)`	La protéine T appartient au réseau biologique P
`goterm(T,G)`	La protéine T est annoté par le terme GO G
`domain(T,D)`	La protéine T est composé du domaine D
`interact(T,N)`	La protéine T interagit avec la protéine N
Connaissances du domaine	
`go_relation(`G_1`,R,`G_2`)`	La relation entre les termes GO G_1 et G_2 est de type R

TABLE 5.2 – Définition des prédicats utilisés pour caractériser les molécules partageant des effets secondaires.

Dans ce chapitre, j'exposerai la méthodologie utilisée pour diminuer la complexité du problème en réduisant le nombre de termes décrivant les effets secondaires. Ensuite, j'introduirai la notion de fingerprints à base d'effets secondaires. Puis, j'expliquerai comment utiliser ces fingerprints pour extraire des profils d'effets secondaires partagés par un nombre important de molécules. Enfin, je montrerai comment une méthode de fouille de données relationnelle peut être utilisée conjointement avec NetworkDB pour comprendre les mécanismes d'apparition des profils d'effets secondaires.

5.2 Définition d'empreinte moléculaire à base d'effets secondaires

Parmi les 1288 termes décrivant les effets secondaires, un grand nombre semble assez redondant. Ainsi, certains médicaments sont annotés par "nausées", d'autres par "vomissement" et encore par "symptômes de nausée et vomissements". Il est également possible qu'un médicament soit annoté par plusieurs de ces termes comme, par exemple, la molécule "abacavir" (un antiviral utilisé contre le VIH) qui provoque ces trois effets. Par ailleurs, il est connu que l'analyse de données peut être perturbée lorsque les dimensions (ou variables) du jeu de données ne sont pas indépendantes les unes des autres. De plus, certains algorithmes de fouille de données symboliques sont peu adaptés aux jeux de données ayant un grand nombre de dimensions (voir article présenté en section 5.2.1). Cette situation complexe m'a obligé à trouver une solution pour réduire le nombre de dimensions (ici le nombre de termes MedDRA) du jeu de données.

5.2.1 Clustering des termes décrivant les effets secondaires

Deux approches étaient possibles pour regrouper les effets secondaires similaires. La première est une similarité de distribution. Elle consiste à considérer comme semblables des effets secondaires qui annotent fréquemment les mêmes molécules. Cette approche a notamment été utilisée par Campillos *et al.* (2008) afin de rechercher de nouvelles cibles à des médicaments partageant des effets secondaires similaires. La seconde approche, développée ici, se base sur les connaissances du domaine. Les connaissances du domaine sont les relations sémantiques qui existent entre les différents termes composant MedDRA (Figure 5.3). Ainsi, en utilisant la mesure de similarité sémantique IntelliGO (définie sur Gene Ontology) adaptée à MedDRA, il est possible de regrouper les termes ayant un sens similaire (Figure 5.4, Benabderrahmane *et al.* 2010).

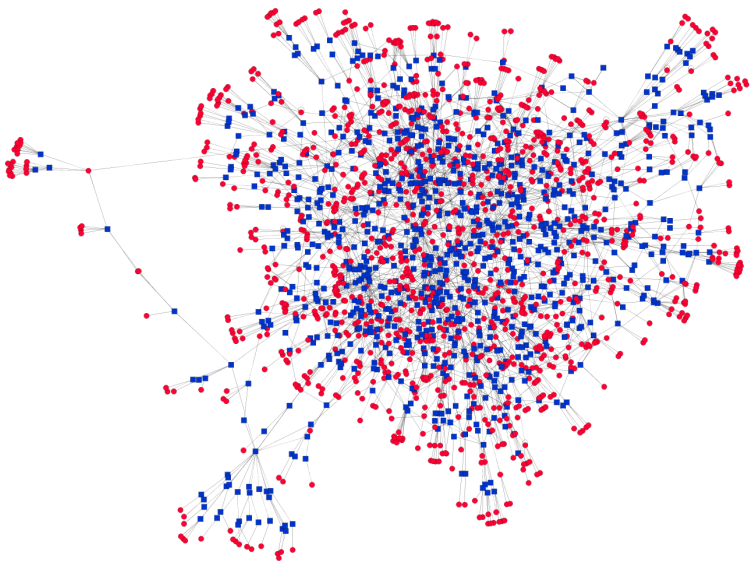


FIGURE 5.3 – Relations entre les 1288 termes de MedDRA correspondant à des effets secondaires. Les termes associés aux effets secondaires sont illustrés par des ronds rouges et les autres termes par des carrés bleus.

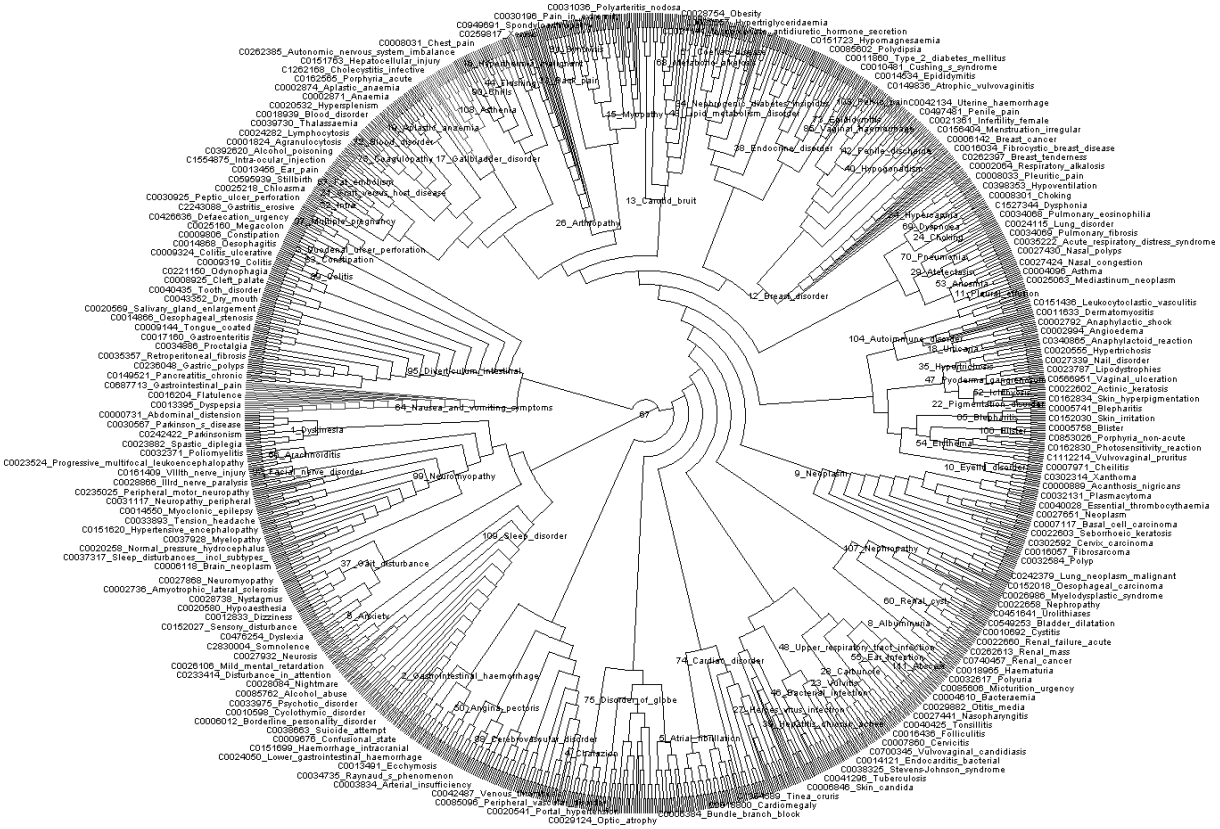


FIGURE 5.4 – Clustering hiérarchique des 1288 termes MedDRA correspondant aux effets secondaires de SIDER. La similarité calculée entre les éléments est basée sur la méthode IntelliGO

La méthodologie utilisée pour regrouper en cluster les effets secondaires à étudier a fait l'objet d'une communication présentée à la conférence *Knowledge Disovery and Information Retrival (KDIR)* en 2011 à Paris. Dans ce papier, nous avons évalué l'apport de ce clustering sur les résultats de deux tâches de fouille de données (recherche de motifs fréquent et de sous-groupes). L'article correspondant est inclus ci-après.

USE OF DOMAIN KNOWLEDGE FOR DIMENSION REDUCTION
Application to Mining of Drug Side Effects

Emmanuel Bresso[1,2], Sidahmed Benabderrahmane[1], Malika Smail-Tabbone[1], Gino Marchetti[1], Arnaud Sinan Karaboga[1], Michel Souchet[2], Amedeo Napoli[1] and Marie-Dominique Devignes[1]
[1]*LORIA UMR 7503, CNRS, Nancy Université, INRIA NGE, 54503 Vandoeuvre-les-Nancy, France*
[2]*Harmonic Pharma, Espace Transfert INRIA NGE, 615 Rue du Jardin Botanique, 54600 Villers-les-Nancy, France*

Keywords: Dimension reduction, Clustering, Semantic similarity, Drug side effects.

Abstract: High dimensionality of datasets can impair the execution of most data mining programs and lead to the production of numerous and complex patterns, inappropriate for interpretation by the experts. Thus, dimension reduction of datasets constitutes an important research orientation in which the role of domain knowledge is essential. We present here a new approach for reducing dimensions in a dataset by exploiting semantic relationships between terms of an ontology structured as a rooted directed acyclic graph. Term clustering is performed thanks to the recently described *IntelliGO* similarity measure and the term clusters are then used as descriptors for data representation. The strategy reported here is applied to a set of drugs associated with their side effects collected from the SIDER database. Terms describing side effects belong to the MedDRA terminology. The hierarchical clustering of about 1,200 MedDRA terms into an optimal collection of 112 term clusters leads to a reduced data representation. Two data mining experiments are then conducted to illustrate the advantage of using this reduced representation.

1 INTRODUCTION AND MOTIVATION

In some domains such as biology, data complexity is ubiquitous and constitutes a major challenge for Knowledge Discovery from Databases (KDD) approaches. One can anticipate that the more complex the data, the more relevant and interesting the extracted knowledge is. However, human expertise is then heavily required to supervise the KDD process especially for the data preparation and the interpretation steps. Thus, an interesting research orientation consists in exploring how domain knowledge can be used to help the expert and contribute to the KDD process in an automated way.

Complex data are usually voluminous and high-dimensional. In a classical *ObjectsXAttributes* representation, the number of objects reflects the volume of data to handle and the number of attributes provides the number of dimensions in the dataset. Most data mining methods become rather inefficient when the dimensionality is too large. Moreover, the patterns extracted from the data may also reveal inappropriate for interpretation by the expert due to plethoric, redundant, or non informative attributes. Data reduction is therefore a crucial issue for data preparation in complex datasets. Several methods exist and are reviewed in section 2. An interesting situation is the case where attributes are terms of a controlled and structured vocabulary. In biology and biomedicine such vocabularies (*e.g.* the Gene Ontology) have been created to annotate biological objects in databases (genes, proteins, etc.) in order to facilitate the retrieval of objects sharing similar annotations. Actually, these vocabularies represent basic domain knowledge in the form of semantic relationships between terms which enhance subsequent data analyses such as classification or characterization.

The present case-study deals with drugs annotated with respect to their side effects. The objects are drugs retrieved from the Side Effects Repository (SIDER) database, which compiles all side effects described in the drug package inserts (Kuhn et al., 2010). The annotation terms used in SIDER pertain from the Medical Dictionary for Regulatory Activities (MedDRA) terminology. Not less than 1,200 MedDRA terms are used in SIDER to annotate the drugs.

Dealing with annotation terms taken from a hierarchical vocabulary allows to use existing strategies for attribute reduction such as generalization (Han and

Bresso E., Benabderrahmane S., Smail-Tabbone M., Marchetti G., Sinan Karaboga A., Souchet M., Napoli A. and Devignes M. (2011).
USE OF DOMAIN KNOWLEDGE FOR DIMENSION REDUCTION - Application to Mining of Drug Side Effects.
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval*, pages 271-276
DOI: 10.5220/0003662602710276

Kamber, 2001). However, available domain vocabularies ore often organized as a rooted Directed Acyclic Graph (DAG) rather than a tree structure in which generalization is intractable. In the present study, we propose alternatively to cluster annotation terms based on their similarity in the rooted DAG using the *IntelliGO* similarity measure that was initially defined on the Gene Ontology (Benabderrahmane et al., 2010).

The application of the *IntelliGO* measure to the MedDRA vocabulary resulted in the clustering of the 1,200 MedDRA terms into an optimal collection of 112 term clusters. These term clusters led to a reduced data representation in which drugs are annotated with term clusters instead of MedDRA terms (Section 3). Several data mining experiments are then conducted to show the advantage of using the reduced representation of the data (Section 4).

2 DATA REDUCTION FOR KDD: A BRIEF STATE OF THE ART

Methods for data reduction divide into two types: feature selection and dimension reduction. Feature selection (or attribute selection) is a means of data reduction without altering the original data representation (Guyon and Elisseeff, 2003). Feature selection methods fall into two categories: filters and wrappers. Filter methods perform feature selection as a pre-processing step based on a search in the feature space. Feature subset evaluation relies on measures that use the training data properties and is carried out independently of any classifier (John et al., 1994). A typical example is the correlation-based feature selection method which eliminates redundant and irrelevant attributes by selecting those that individually correlate with the class but have little inter-correlation (Witten et al., 2011). Concerning the wrapper methods, they evaluate candidate feature subsets on the basis of their predictive accuracy (classification performance) using a learning algorithm (Kohavi and John, 1997). Most feature selection methods work with supervised classification and use the class information of the training examples to select the relevant features. However, Wrapper methods were recently proposed for feature selection upstream unsupervised learning, namely clustering (Kim et al., 2000; Dy and Brodley, 2004). Such studies focus on how to evaluate the results of clustering for each candidate feature subset. Also,a recent study suggests a filter method independent of both the learning algorithm and any predefined classes by guiding the attribute selection with formalized domain knowledge (Coulet et al., 2008).

Alternative data reduction methods alter the data representation by encoding the data into a smaller representation space. Such dimension reduction is also called feature compression. The principal component analysis is a popular example of such methods which deals with numeric and continuous data. Clustering was proposed for grouping the attributes in order to improve the classification or the clustering of textual documents (Kyriakopoulou, 2008). Here the attributes are binary and correspond to words annotating textual documents. Word clustering then relies on comparing their joint distributions in the documents over the classes (Koller and Sahami, 1996; Slonim and Tishby, 2000). Thus, the similarity measures used for clustering the word attributes are corpus-based.

In this paper, we propose that when the attributes belong to a domain-specific structured vocabulary, a better clustering of these attributes could be achieved by using a suitable semantic similarity measure.

3 SEMANTIC CLUSTERING OF ATTRIBUTES

3.1 The MedDRA Terminology

The MedDRA medical terminology is used to classify adverse event information associated with the use of drugs and vaccines. MedDRA is a part of the Unified Medical Languages System (UMLS) and is often presented as a hierarchy consisting of five levels: System Organ Class, High Level Group Term, High Level Term, Preferred Term, Lowest Level Term (MedDRA, 2007). Lowest level terms correspond to different terms for the same preferred term. In the MedDRA terminology, each term has an identifier and all the paths to the root can be download. For example, for the term C0000733, we have two paths: C1140263.C0017178.C0947761.C0947846 and C1140263.C0947733.C0021502.C0851837. A careful review of the parent-child relationships shows that the MedDRA is actually not a hierarchy: about 37% of the MedDRA terms have more than one direct parent. This together with the natural oriented of term-term relationship and the absence of cycle, confers to the MedDRA terminology the status of a rooted DAG.

3.2 Term-term Semantic Similarity Measure

Two approaches exist for term-term semantic similarity measures: structure-based approaches which

exploit the structure of the vocabulary (depth, path length) and corpus-based approaches which exploit the term distribution in a corpus (annotation frequency, information content). The *IntelliGO* measure is a structure-based approach in which the generalized cosine similarity measure proposed by Ganesan for hierarchical vocabularies has been adapted to rooted DAGs (Benabderrahmane et al., 2010). The *IntelliGO* measure was initially defined for quantifying similarity between Gene Ontology (GO) terms. For two terms t_i, t_j, it takes into account the maximal depth of their common ancestors (CA) and the minimal path-length (SPL) between them:

$$Sim_{IntelliGO}(t_i,t_j) = \frac{2MaxDepth(CA)}{MinSPL(t_i,t_j)+2MaxDepth(CA)} \quad (1)$$

To calculate the semantic similarity between two MedDRA terms, the algorithm starts by retrieving for each term all their paths to the root node. Then, the set of common ancestors is defined as the intersection between the two sets containing the ancestors of the two terms. In the next step, the algorithm identifies the common ancestors having the maximal depth from the root node (MaxDepth(CA)). Note that the Depth of a MedDRA term can be calculated as the maximal length of a path from this term to the root node. After that, the algorithm calculates the shortest path length (MinSPL) separating the two terms. Finally, the semantic similarity between the two terms is computed using the equation (1).

As the values of $Sim_{IntelliGO}$ range from 0 to 1, we also define the distance $Dist_{IntelliGO}$ as its complement to 1.

3.3 Clustering MedDRA Terms

A total of 1,288 terms from the 20,037 MedDRA terms are used in the SIDER database for annotating drug side effects. Pairwise distances were calculated for these 1,288 terms. We then used the Ward's hierarchical agglomeration algorithm (Ward, 1963) with an optimization step necessary to select the best level where to cut the dendrogram in order to obtain a set of clusters (Kelley et al., 1996). This method defines a penalty value which is function of the cluster number and the intra-cluster distance. When this value is minimal, the resulting clusters are as highly populated as possible while simultaneously maintaining the smallest average intra-cluster distance. In our case the minimal penalty value was obtained with 112 clusters which were subsequently inspected and validated by two experts. In the rest of this paper, these clusters will be defined as the term clusters (TC) which will be used as attributes for data mining.

In order to label each TC with its most representative term, we introduce a function $AvgDist_{TC}$ associating to each TC term its average distance to all the TC terms:

$$AvgDist_{TC}(t_i) = \frac{1}{|TC|}\sum_{j=1}^{N_{TC}} dist(t_i,t_j) \quad (2)$$

Then the representative element R of a TC is the term which minimizes $AvgDist_{TC}$. For example in Table 1, *Erythema* is the representative element of the TC. The label of a given TC is the concatenation of the TC number and the representative term (*e.g.*, 54_Erythema for the TC described in Table 1). Once TC are built, they can be used for dimension reduction.

Table 1: Example of TC with the average distance function calculated for each term t.

Term Cluster Element t	$AvgDist_{T_{54}}(t)$
Decubitus ulcer	0.35
Rash	0.35
Lichen planus	0.32
Parapsoriasis	0.32
Pruritus	0.35
Psoriasis	0.37
Sunburn	0.35
Erythema	**0.31**
Pityriasis alba	0.32
Photosensitivity reaction	0.37
Rash papular	0.32
Dandruff	0.37
Lupus miliaris disseminatus faciei	0.35
Vulvovaginal pruritus	0.35

4 EVALUATION OF THE IMPACT OF FEATURE CLUSTERING ON DATA MINING

4.1 Experimental Design

In order to evaluate the impact of our dimension reduction strategy, two data mining experiments were conducted. The first experiment aims at retrieving frequent associations of side effects shared by drugs in a given drug category. The second experiment aims at discriminating drugs belonging to two categories in terms of side effects. Datasets consist of binary (*Objects X Attributes*) relations between drugs (*Objects*) and their side effects (*Attributes*). The side effects are represented either as individual MedDRA terms or as TC leading to two data representations.

In the first experiment we search for Frequent Closed Itemsets (FCIs) in order to compare the two data representations with respect to computation time, number and relevance of the extracted FCIs. In our context, a FCI of length *n* and support *s* corresponds to an association of *n* terms, respectively term clusters, shared by the maximal group of drugs corresponding to a percentage value *s* of the whole dataset. The Zart program was used for FCI extraction on the Coron platform (Szathmary et al., 2007). The experiment was ran on a 2.6GHz processor with 1GB memory.

As for the second experiment we use the CN2-SD subgroup discovery algorithm (Lavrac et al., 2004) with the two data representations in order to check the impact of term clustering on the computation time and the produced subgroups. Given a population of objects and a property of those objects that we are interested in, subgroup discovery allows to find subgroups of objects that are statistically most interesting, *i.e.*, as large as possible and having the most unusual distributional characteristics with respect to the property of interest. In our case, two categories of drugs are investigated with this method for identifying subgroups of drugs sharing discriminative side-effects in one category versus the other. The CN2-SD algorithm implementation used is the one of the Keel software (Alcala-Fdez et al., 2009) and was executed on a 8-core 1.86GHz processor with 8GB memory.

4.2 Datasets Description

The category of a drug refers to its therapeutic uses. The categories of the drugs present in the SIDER DB are available in the DrugBank DB (Knox et al., 2011). We chose to perform the data mining experiments on the drugs corresponding to the two largest categories: the Cardiovascular Agents (CA) and the Anti-Infective Agents (AIA) containing respectively 94 and 76 drugs.

For each category, we built two datasets : the *All* dataset has for attributes the 1,288 MedDRA terms annotating the drugs in SIDER and the *TC* dataset has for attributes the 112 TC (as described in section 3) note that a TC is assigned to a drug if at least one member of the TC is reported as annotating the drug in SIDER. This gives four datasets CA_{All}, AIA_{All}, CA_{TC}, and AIA_{TC}.

4.3 Frequent closed Itemset Extraction with and without Term Clustering

The Zart program was executed on each dataset with a minimal support varying from 50 to 100%. Table 2 summarizes the number of FCIs produced in each case.

Table 2: Number of FCIs for each dataset when varying the minimal support value.

Minimal support	50%	60%	70%	80%	90%	100%
AIA_{all}	178	41	9	2	0	0
AIA_{TC}	654	154	30	3	0	0
CA_{all}	386	94	41	11	1	0
CA_{TC}	5,564	1,379	256	62	6	0

The first observation is the increase in the number of FCIs for a given minimal support when term clusters are used. For the AIA_{All} and AIA_{TC} datasets, this increase varies from more than 3-fold for minimal supports between 50 and 70% to about 1.5-fold for higher minimal supports. For CA_{All} and CA_{TC} datasets this increase goes from about 14-fold at minimal support 50 and 60% to 6-fold for higher minimal supports. This increase clearly reflects the expected role of clustering in feature reduction, namely increasing the density of the binary (*Objects X Attributes*) relation by aggregating object properties. Accordingly, the computation time of the program was two-fold longer with the *TC* than with the *All* representation.

Content analysis of the FCIs was done after ranking FCIs according to their support. The five top-ranked FCIs (plus *ex-aequo*) are listed for each dataset in Figure 1. With the *All* representation, FCIs can be very redundant. For example in the AIA_{All} dataset (left panel), three from the five displayed FCIs contain very similar term: *nausea*, *vomiting*, *nausea and vomiting symptoms*. On the contrary with the TC representation (right panel), a unique FCI contains the attribute *64_nausea and vomiting symptoms* which represents the cluster of terms containing all three attributes cited before. Thus, data reduction by term clustering allows less redundant FCI extraction and therefore leads to the presentation of more potentially interesting itemsets to the expert.

Figure 1 also shows that the FCI supports are generally higher with the *TC* than with the *All* data representation. A correspondence can be established between individual terms in the *All* representation and the term clusters in the *TC* representation as illustrated in Figure 1 (remember that a term cluster is labeled with its number and its most representative term). For example, the $\{54\_Erythema\}$ FCI found in the AIA_{TC} dataset with a support of 88% contains the term cluster labeled *54_Erythema* that includes the *Pruritus* individual term and has therefore been matched with the $\{Pruritus\}$ FCI found in the AIA_{All} dataset with a support of 79%.

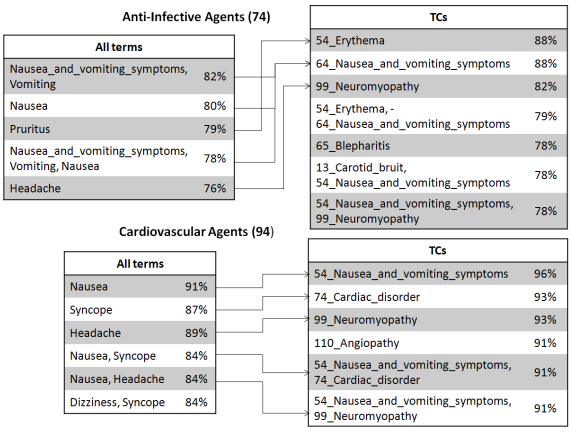


Figure 1: The five most frequent FCIs are extracted from the total list of FCIs obtained for each dataset. In case of ex-aequo, all itemsets with the same support are listed.

Furthermore, combined with the lack of redundancy, this increase of FCI support leads to the discovery in the *TC* datasets of frequent term clusters for which all individual terms are less frequent and therefore not considered in the *All* datasets.

4.4 Subgroup Discovery with and without Term Clustering

The second experiment aims at discriminating the drugs belonging to two categories in terms of presence or absence of side effects. Indeed, the absence of side effects may also be important for drug characterization. This is possible with the CN2-SD algorithm as the input data is in attribute-value format. We ran the CN2-SD program on the following unions of datasets in which an additional attribute was added for the category information: CA_{All} versus AIA_{All}, and CA_{TC} versus AIA_{TC}.

The first observation concerns the computation time. When term clusters are used the execution time is less than ten minutes whereas it does not resume within six days when all side effects are used. Thus, data reduction is necessary for successful execution of the CN2-SD algorithm.

In a second stage, the rules extracted from CA_{TC} versus AIA_{TC} dataset were analyzed. The left part of a rule is verified for a number of drugs (support) among which a certain fraction (coverage) are of the category indicated in the right part of the rule. The resulting subgroup therefore identifies a subset of drugs from this category sharing a specific profile of side effects with regard to the other category. The best rules in terms of coverage are shown in Table 3.

To sum up, these results show that the data reduction used allows subgroup discovery which was impossible with the extended data representation. Further investigation by domain experts is ongoing.

Table 3: Best rules (with coverage / support) extracted by the CN2-SD program for CA_{TC} versus AIA_{TC}.

50_Angina_pectoris = T AND 93_Bacteraemia = F AND 52_Ichthyosis = F AND 54_Erythema = T AND 49_Folate_deficiency = F => CA (0.96/56)
31_Splenic_infarction = T AND 41_Neutropenia = T AND 42_Penile_discharge = F AND 77_Facial_pain = T AND 79_Cachexia = T => AIA (0.88/26)

5 DISCUSSION AND PERSPECTIVES

In this paper we have reported a method for dimension reduction guided by domain knowledge. The method is based on attribute clustering using a semantic similarity measure. We took advantage of our recently defined *IntelliGO* similarity measure which applies to the rooted DAG structure encountered in many vocabularies.We believe that our strategy can be applied in various other biomedical context (Leva et al., 2005; Pakhomov et al., 2007). We tested our method on a dataset of 170 drugs annotated with 1,288 terms taken from the MedDRA terminology and representing the drugs possible side effects. Using IntelliGO-based term-term distances and hierarchical clustering, we reduced data representation from 1,288 individual terms down to 112 term clusters. In this work we adopted a binary representation for the reduced data representation, *i.e.*, a TC is assigned to a drug if at least one of its elements has been associated with the drug in the SIDER database. This representation ignores the impact of multiple associations between a drug and TC elements. A many-valued relation could be produced to take into account such situations. Recently described extension of formal concept analysis may help us handling such data representation (Messai et al., 2008; Kaytoue-Uberall et al., 2009).

The dimension reduction method we have developed was tested with two data mining algorithms: FCI extraction and subgroup discovery. The results show that FCIs extracted from the *TC* data representation are less redundant and display higher supports than from the *All* representation. Another consequence of the data reduction is that the expert's task is facilitated because more relevant and explicit side effects are found among FCIs displaying high support. As for subgroup discovery, dimension reduction revealed to play a crucial role. Indeed, the program was unable to resume with the *All* data representation whereas it

provided, with the reduced *TC* representation, quite interesting rules characterizing subgroups of one drug category versus another one. Complementary experiments can now be carried out to identify rules specific of a given category versus all other categories.

REFERENCES

Alcala-Fdez, J., Snchez, L., Garca, S., del Jesus, M., Ventura, S., Garrell, J., Otero, J., Romero, C., Bacardit, J., Rivas, V., Fernndez, J., and Herrera, F. (2009). KEEL: a software tool to assess evolutionary algorithms for data mining problems. *Soft Computing - A Fusion of Foundations, Methodologies and Applications*, 13(3):307–318–318.

Benabderrahmane, S., Smail-Tabbone, M., Poch, O., Napoli, A., and Devignes, M. (2010). IntelliGO: a new vector-based semantic similarity measure including annotation origin. *BMC Bioinformatics*, 11(1):588.

Coulet, A., Smail-Tabbone, M., Benlian, P., Napoli, A., and Devignes, M. (2008). Ontology-guided data preparation for discovering genotype-phenotype relationships. *BMC Bioinformatics*, 9(Suppl 4):S3.

Dy, J. G. and Brodley, C. E. (2004). Feature selection for unsupervised learning. *J. Mach. Learn. Res.*, 5:845–889.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182.

Han, J. and Kamber, M. (2001). *Data Mining: Concepts and Techniques*. Morgan Kaufmann, San Francisco, 1 edition.

John, G. H., Kohavi, R., and Pfleger, K. (1994). Irrelevant Features and the Subset Selection Problem. In *International Conference on Machine Learning*, pages 121–129.

Kaytoue-Uberall, M., Duplessis, S., Kuznetsov, S. O., and Napoli, A. (2009). Two fca-based methods for mining gene expression data. In *ICFCA*, pages 251–266.

Kelley, L. A., Gardner, S. P., and Sutcliffe, M. J. (1996). An automated approach for clustering an ensemble of NMR-derived protein structures into conformationally related subfamilies. *Protein Engineering*, 9(11):1063 –1065.

Kim, Y., Street, W. N., and Menczer, F. (2000). Feature selection in unsupervised learning via evolutionary search. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 365–369, Boston, Massachusetts, United States. ACM.

Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., Pon, A., Banco, K., Mak, C., Neveu, V., Djoumbou, Y., Eisner, R., Guo, A. C., and Wishart, D. S. (2011). DrugBank 3.0: a comprehensive resource for Omics research on drugs. *Nucleic Acids Research*, 39(suppl 1):D1035 –D1041.

Kohavi, R. and John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324.

Koller, D. and Sahami, M. (1996). Toward optimal feature selection. In Saitta, L., editor, *Proceedings of the Thirteenth International Conference on Machine Learning (ICML)*, pages 284–292. Morgan Kaufmann Publishers.

Kuhn, M., Campillos, M., Letunic, I., Jensen, L. J., and Bork, P. (2010). A side effect resource to capture phenotypic effects of drugs. *Mol Syst Biol*, 6.

Kyriakopoulou, A. (2008). Text classification aided by clustering: a literature review. In *Tools in Artificial Intelligence*, chapter 14. Paula Fritzsche, intech edition.

Lavrac, N., Kavsek, B., Flach, P., and Todorovski, L. (2004). Subgroup discovery with CN2-SD. *J. Mach. Learn. Res.*, 5:153–188.

Leva, A. D., Berchi, R., Pescarmona, G., and Sonnessa, M. (2005). Analysis and prototyping of biological systems: the abstract biological process model. *International Journal of Information and Technology*, 3(4):216–224.

MedDRA (2007). Meddra maintenance and support services organization. introductory guide, meddra version 10.1.

Messai, N., Devignes, M.-D., Napoli, A., and Smaïl-Tabbone, M. (2008). Many-valued concept lattices for conceptual clustering and information retrieval. In *ECAI*, pages 127–131.

Pakhomov, S. S., Hemingway, H., Weston, S. A., Jacobsen, S. J., Rodeheffer, R., and Roger, V. L. (2007). Epidemiology of angina pectoris: Role of natural language processing of the medical record. *American Heart Journal*, 153(4):666–673.

Slonim, N. and Tishby, N. (2000). Document clustering using word clusters via the information bottleneck method. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 208–215, Athens, Greece. ACM.

Szathmary, L., Napoli, A., and Kuznetsov, S. O. (2007). Zart: A multifunctional itemset mining algorithm. In *CLA*, pages 26–37.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244. ArticleType: research-article / Full publication date: Mar., 1963 / Copyright 1963 American Statistical Association.

Witten, I. H., Frank, E., and Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, Burlington, MA, 3 edition.

5.2.2 Exploration des empreintes à base d'effets secondaires

Dans la suite de ce travail, nous considérons toujours les effets secondaires regroupés en clusters de termes ou TC. Les médicaments peuvent donc être caractérisés par les 112 TC définis précédemment, sous forme d'un vecteur à 112 dimensions, chaque dimension représentant un TC. Ce vecteur définit ce que j'ai appelé une empreinte à base d'effets secondaires ou "fingerprint". La valeur associée à chaque dimension de ce vecteur correspond au pourcentage d'effets du TC annotant la molécule dans la base de données SIDER.

Afin d'examiner les empreintes associées à chaque molécule de la base de données NetworkDB, j'ai développé une interface web, nommée HPfingerprint, que j'ai développée. A partir de cette interface, il est possible d'interroger NetworkDB avec le code SMILES d'une molécule ou son identifiant DrugBank. Si cette molécule est présente dans NetworkDB, l'utilisateur a accès à son fingerprint. L'interface donne accès aux TC qui sont spécifiques à la molécule. En effet, nous considérons qu'un TC annote de manière spécifique une molécule s'il annote la molécule et peu d'autres molécules. De façon duale, les absences spécifiques d'associations sont relevées pour les TC n'annotant pas la molécule mais qui annotent fréquemment le reste des autres médicaments de NetworkDB.

Par exemple, la molécule de Paliperidone est associée au TC "36_Connective_tissue_disorder" alors que seulement 12% des molécules de la base le sont. Cette même molécule n'est pas annotée par "54_Lupus_miliaris_disseminatus_faciei" alors que 87% des médicaments de NetworkDB sont caractérisés par ce TC (Figure 5.5). Pour chaque TC annotant la molécule, il est également possible d'explorer les effets secondaires qui sont à l'origine de cette annotation. Ainsi, la molécule de paliperidone est annotée par un seul terme, "Connective tissue disorder", du TC 36.

Clusters d'effets secondaires associés à la molécule

Clusters d'effets secondaires	Pourcentage des molécules de référence associées à ce cluster
36_Connective_tissue_disorder	11.6%
102_Mediastinal_disorder	12.8%
105_Dysmenorrhoea	19.1%

Clusters d'effets secondaires non associés à la molécule

Clusters d'effets secondaires	Pourcentage des molécules de référence associées à ce cluster
54_Lupus_miliaris_disseminatus_faciei	86.6%
65_Dermatitis	76.7%
83_Diarrhoea	76.4%

FIGURE 5.5 – TC les plus spécifiques de la molécule de paliperidone. Les molécules de référence sont les 554 médicaments de NetworkDB dont on connait les effets secondaires.

Une autre fonctionnalité de HPfingerprint est de pouvoir explorer les TC. Ainsi, il est possible d'afficher le contenu en effets secondaires de chaque TC puis pour chaque effet secondaire, il est possible de connaitre les molécules présentant l'effet et leurs catégories.

Cette interface permet également de rechercher le médicament, dont les effets secondaires sont connus, le plus similaire à une molécule donnée en entrée au format SMILES. Scheiber *et al.* (2009*b*) ont montré qu'il existait un lien entre les sous-structures d'une molécule et ses effets secondaires. Ainsi, en utilisant une mesure de similarité basée sur les sous-structures, il est possible d'obtenir une indication sur les effets secondaires d'une nouvelle molécule. Pour cela, l'interface web utilise une similarité de type Tanimoto qui calcule le nombre de sous-structures communes aux deux molécules divisé par le nombre total de sous-structures réunies dans les deux molécules (sans les doublons).

Ces empreintes à base d'effets secondaires permettent de mettre en évidence le rôle de l'action des médicaments sur leur cible. En effet, bien que Scheiber *et al.* (2009*b*) aient démontré le rôle des sous-structures dans l'apparition des effets secondaires, il existe des cas où deux molécules parfaitement similaires possèdent des fingerprints d'effets secondaire totalement différents. Ainsi, quand on étudie les molécules de risperidone et de paliperidone, on remarque qu'elle ne différent que d'un groupement hydroxyle (Figure 5.6). Or ces deux molécules ont deux empreintes en effet secondaires fortement divergente. L'explication à cette observation ne peut pas venir de l'existence de cibles différentes pour les deux molécules puisque ces dernières agissent sur exactement les mêmes protéines (Table 5.3). Cependant, quand on étudie le type d'action sur leur cible on peut remarquer que les deux médicaments ont un effet opposé pour le récepteur H1 à l'histamine. En effet, la paliperidone va activer cette protéine alors que la risperidone en est un inhibiteur. Cette différence étant la seule observée entre les deux molécules, il est fort probable qu'elle soit à l'origine de la différence d'empreinte. Suite à cette observation, j'ai ajouté une fonctionnalité à HPfingerprint permettant de visualiser pour chaque cible et type d'action, les empreintes de chaque molécule. Par exemple, on peut voir sur la figure 5.7 que les molécules activant la sous unité ρ-3 du récepteur à l'acide γ-aminobutyrique (GABA), peuvent être organisées en trois grands groupes. Le groupe central correspond à des molécules ayant peu d'effets secondaires, le groupe inférieur correspond à des molécules fortement annotées par des TC et le groupe supérieur correspond à des molécules intermédiaires. Comme on peut le voir ici, l'information sur une cible ne suffit pas pour réaliser des prédictions de TC puisqu'on observe à la fois des molécules associées à peu de TC et des molécules associées à de nombreux TC.

Cible	Paliperidone	Risperidone
5-hydroxytryptamine receptor 1A	-	-
5-hydroxytryptamine receptor 1D	-	-
5-hydroxytryptamine receptor 2A	-	-
5-hydroxytryptamine receptor 2C	-	-
Alpha-1A adrenergic receptor	-	-
Alpha-1B adrenergic receptor	-	-
Alpha-2A adrenergic receptor	-	-
Alpha-2B adrenergic receptor	-	-
Alpha-2C adrenergic receptor	+	+
D(1A) dopamine receptor	-	-
D(2) dopamine receptor	-	-
D(3) dopamine receptor	-	-
D(4) dopamine receptor	-	-
Histamine H1 receptor	**+**	**-**

TABLE 5.3 – Action des molécules de Paliperidone et de Risperidone sur leurs cibles. Un + indique une activation de la cible et un - une inhibition.

5.2.3 Exploration des fingerprints pour une étude des médicaments retirés du marché

Des molécules sont régulièrement retirées du marché suite à la découverte d'effets secondaires graves. Ainsi, le thalidomide, qui était utilisé comme anti-nauséeux chez les femmes enceintes dans

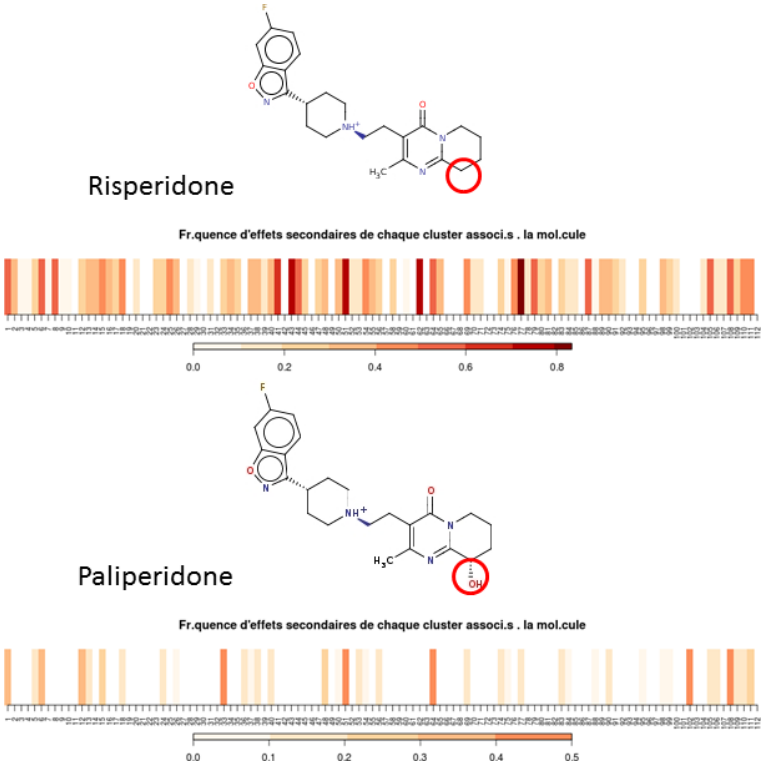


FIGURE 5.6 – Empreintes d'effets secondaires des molécules de Risperidone et de Paliperidone. Les deux molécules ne diffèrent que par un groupement hydroxyle (entouré en rouge). Les couleurs de l'empreinte correspondent à la fréquence d'effets secondaires associés à la molécule pour chaque TC. Plus la couleur est sombre, plus la molécule possède d'effets secondaires du TC.

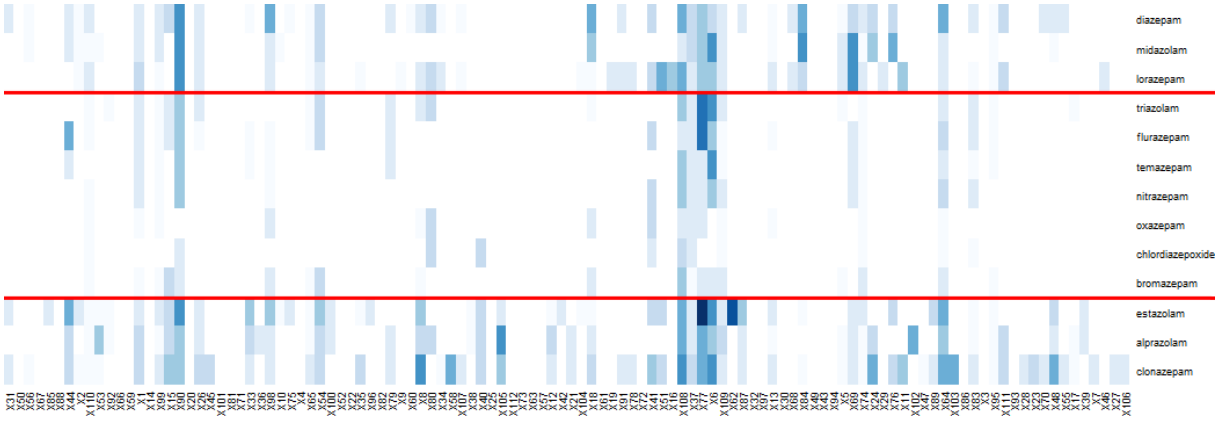


FIGURE 5.7 – Empreintes d'effets secondaires des molécules activant la sous unité ρ-3 du récepteur au GABA. Les lignes rouges correspondent à la séparation entre les 3 groupes de molécules selon leur nombre de TC associés.

les années 50, a été retiré du marché suite aux graves malformations congénitales qu'il provoquait. Plus récemment, la molécule de benfluorex (Mediator) a été retirée suite aux problèmes cardiaques qu'elle pouvait provoquer.

Nous nous sommes demandés s'il existait des effets communs aux molécules qui ont été reti-

rées. Pour cela, j'ai extrait de la liste des 554 médicaments de mon jeu de données, les 14 molécules annotées comme "withdrawn" dans DrugBank (Table 5.4) avec le motif de leur retrait. Afin de ne pas perdre d'information lors de l'association entre TC et médicaments, j'ai choisi la solution qui consiste à établir l'association si au moins un terme du TC est associé à la molécule, permettant ainsi d'obtenir les 14 empreintes binaires présentées dans la figure 5.8.

Étant donné que plusieurs médicaments ont été retirés pour les mêmes raisons, il est intéressant de regarder si en regroupant les molécules ayant une empreinte similaire, on retrouve les mêmes groupes. En utilisant une similarité de type Tanimoto entre les fingerprints puis en appliquant une méthode de clustering hiérarchique sur les scores calculés, on obtient le dendrogramme présenté en figure 5.9. Par comparaison avec la table 5.4, on observe que seules les molécules de pemoline et de troglitazone d'une part, de nefazodone et de tolcapone d'autre part, sont trouvées similaires en fingerprint alors qu'elles ont été retirées pour la même raison. Par contre, le thalidomide et le vadecoxib sont trouvés très similaires alors qu'elles ont été retirées pour deux raisons différentes. Ainsi, utiliser l'ensemble du fingerprint pour essayer de comprendre pourquoi des molécules ont été retirées ne semble pas une méthode satisfaisante.

Cependant, il est possible qu'un sous-ensemble de TC permette d'expliquer le retrait des molécules. Pour cela, il est intéressant de comparer les TC associés aux molécules retirées avec ceux des molécules non retirées du marché. L'une des façons de réaliser cette comparaison est de construire des arbres de décision. En effet, cette méthode de fouille de données permet de construire un arbre de classification des données. A partir d'une matrice molécule$\times$TC pour laquelle la classe (retirée ou non) des molécules est connue, l'algorithme de construction de l'arbre va déterminer le TC qui sépare le mieux les deux classes et former ainsi deux branches. Pour chaque nouvelle branche, il va ensuite à nouveau déterminer le meilleur TC pour séparer les données et répéter cette étape tant qu'il augmente le pourcentage d'éléments bien classés. En utilisant les 14 molécules retirées (exemples positifs) contre les 540 autres (exemples négatifs), l'algorithme estime qu'il ne faut pas séparer les données. En effet, si toutes les molécules sont classées comme non retirées, alors le taux d'erreur de classement sera de 2,5%. Il n'est donc pas possible d'utiliser l'ensemble des molécules des molécules restées sur le marché pour construire les arbres. Une solution à ce problème consiste à utiliser un nombre restreint d'exemples négatifs, assez proche du nombre d'exemples positifs, le risque étant que l'arbre construit soit dépendant de ces molécules. Il a donc été nécessaire de tester plusieurs jeux de données avec de exemples négatifs différents afin de déterminer si les arbres ne varient pas trop en fonction des exemples négatifs utilisés pour l'apprentissage. Ainsi, en répétant 5 constructions d'arbres utilisant chaque fois 28 molécules non retirées choisies aléatoirement, on obtient une précision de validation croisée (apprentissage sur 9/10 du jeu de donnée et test sur le 1/10 restant, répété dix fois) moyenne de 74%. Cependant, comme on peut le voir dans la figure 5.10 sur deux exemples, les arbres peuvent être très différents les uns des autres. Seul le TC "17_Hepatitis" est conservé dans 4 arbres sur 5, ce qui semble indiquer que les molécules ayant un effet sur le foie sont le plus souvent retirées. Cette observation est compatible avec les raisons réelles qui ont entrainé le retrait des molécules (Table 5.4) d'autant plus que seules 5% des molécules du jeu de données sont annotées par ce TC. Ainsi, le faible nombre de molécules retirées du marché et répertoriées dans SIDER limite l'exploration automatique des raisons qui ont poussé l'industrie pharmaceutique à cesser de commercialiser ces médicaments. Il reste que l'analyse manuelle du tableau 5.4 permet de remarquer que les principales raisons de ces

retraits sont la toxicité hépatique et l'apparition d'effets cardiotoxiques et d'accidents vasculaires.

Molécule	Raison du retrait
Alosetron	effets gastro-intestinaux sérieux
Cisapride	arythmies cardiaques
Encainide	arythmies cardiaques
Metoprolol	défaillances cardiaques
Nefazodone	hépatotoxicité
Pemoline	hépatotoxicité
Rofecoxib	infarctus du myocarde
Tegaserod	cardiotoxicité
Thalidomide	malformations congénitales
Tolcapone	hépatotoxicité
Triazolam	effets secondaires psychiatriques
Troglitazone	hépatotoxicité
Trovafloxacin	hépatotoxicité
Valdecoxib	infarctus du myocarde, accident vasculaire cérébral

TABLE 5.4 – Molécules annotées comme "withdrawn" dans DrugBank et présentes dans SIDER.

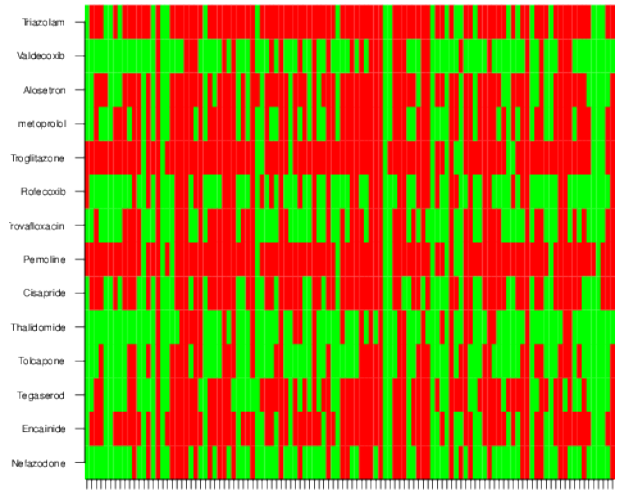


FIGURE 5.8 – Fingerprints binaires des 14 molécules "withdrawn". Un TC associé à la molécule est représenté en vert, sinon il est en rouge.

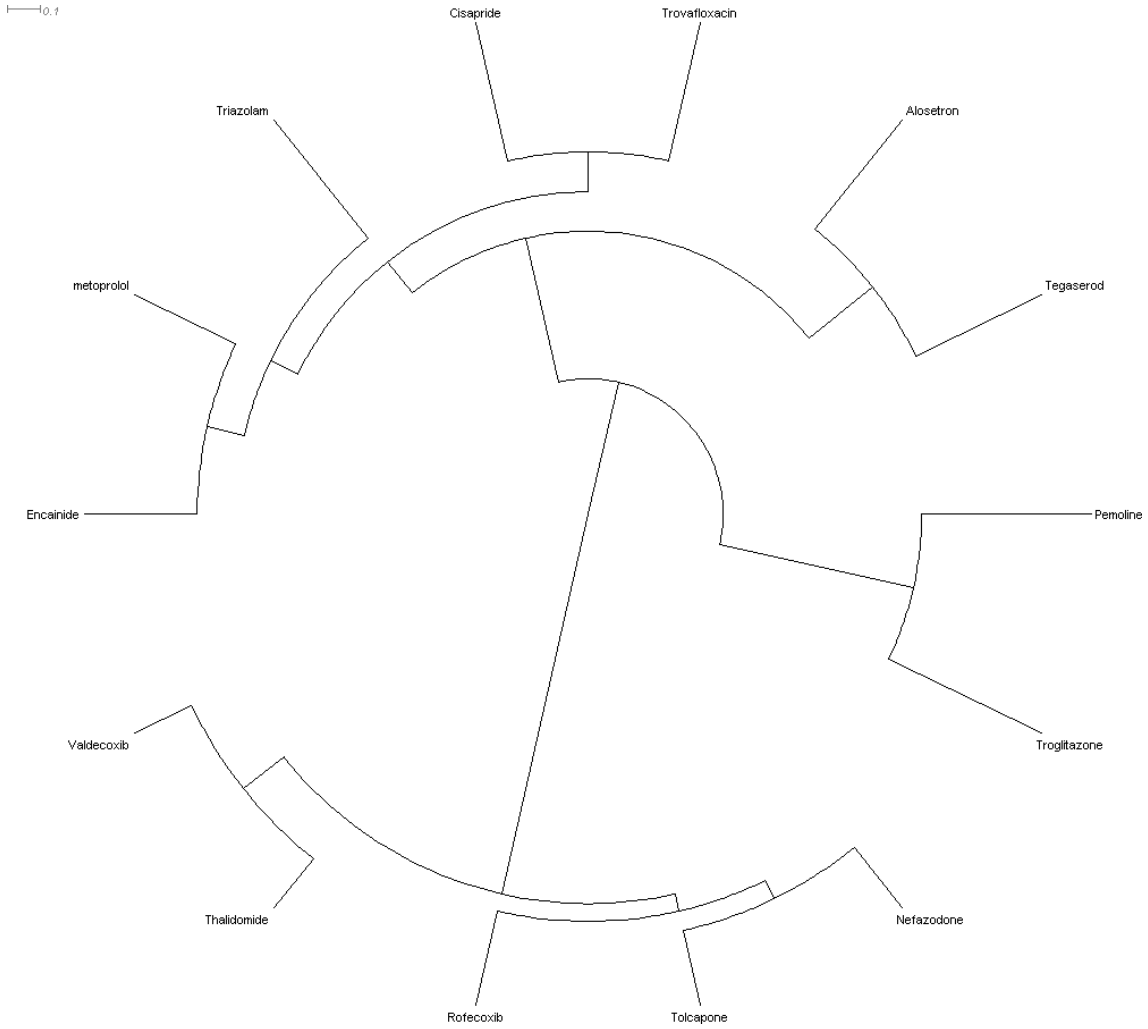


FIGURE 5.9 – Clustering hiérarchique (méthode de Ward) des 14 molécules retirées, basé sur les fingerprints de TC et calculé avec une similarité Tanimoto

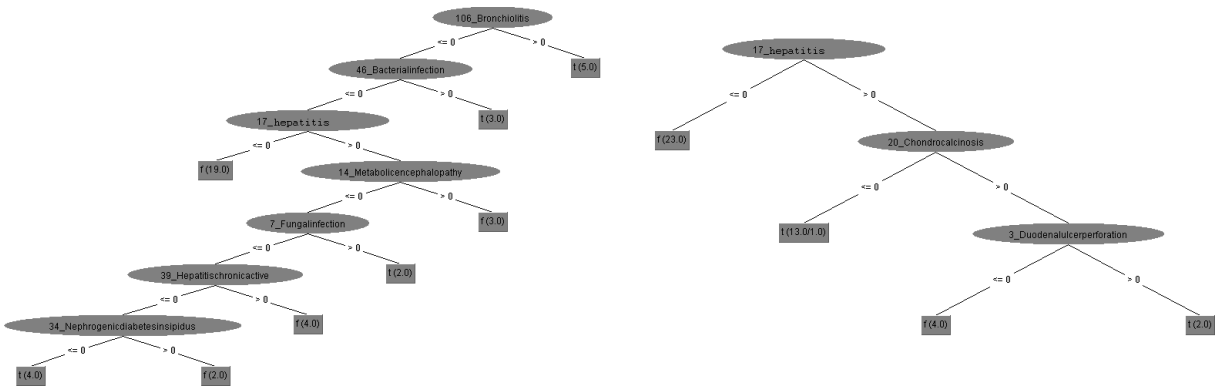


FIGURE 5.10 – Arbres de décisions (Algorithme J48 de Weka) obtenus à partir des 14 molécules retirés du marché et de deux jeux de 28 molécules non retirées du marché tirées aléatoirement. Les choix se font en fonction de la présence (>0) ou l'absence (<=0) d'effets secondaires dans les TC représentés dans les nœuds. Un t ("true") représente une prédiction positive : retirée du marché et un f ("false") une prédiction négative : non retirée du marché

Cet échec relatif de l'utilisation des arbres de décisions pour comprendre les raisons qui ont conduit au retrait de médicaments du marché, a permis de mettre en place une réflexion sur la compréhension des effets secondaires. En effet, si des effets secondaires sont à l'origine de l'arrêt de la commercialisation de certains médicaments, la compréhension des mécanismes provoquant leur apparition permettrait d'éviter des retraits pour les médicaments en cours de développement. Cette compréhension peut se faire à la lumière des connaissances biologiques, permettant ainsi d'avoir plus de descripteurs pour les molécules et leurs cibles et nécessitant une méthode de fouille relationnelle afin de prendre en compte ces nouveaux descripteurs.

5.3 Définition et extraction de profils d'effets secondaires

5.3.1 Définition

Les médicaments provoquent le plus souvent une série d'effets secondaires plutôt qu'un effet isolé. Ainsi, les molécules présentes dans NetworkDB sont associées au minimum à 1 effet secondaire (pour le malathion et le sulfanilamide) et 483 au maximum (pour la molécule de ropinirole). Plusieurs effets secondaires sont donc susceptibles d'être partagés par un nombre important de molécules. Cette observation m'a amené à définir des profils d'effets secondaires comme étant les plus grand groupes de TC partagés par un nombre significatifs de molécules.

D'un point de vue informatique, cette définition des profils d'effets secondaires correspond aux motifs fréquent maximaux (MFI). Dans une matrice binaire de type objets$\times$attributs, les MFI sont les plus longs ensembles d'attributs partagés par un nombre significatif (ou support minimum) d'objets. Si le support minimum est faible, on obtient des MFI longs (un grand nombre de TC est partagé par peu de molécules) et inversement, un support minimum élevé entrainera l'apparition de MFI courts, voire ne contenant qu'un TC. De plus, afin que les programmes de fouille de données fonctionnent correctement, il est nécessaire d'avoir un nombre suffisant d'exemples pour l'apprentissage. Nous avons choisi d'utiliser un seuil de 100, représentant un peu moins de 20% du jeu de données total, afin de disposer d'un nombre correct d'exemples et de motifs pas trop courts.

Afin de pouvoir extraire les profils d'effets secondaires, il est nécessaire d'utiliser une matrice molécules$\times$effets secondaires binaire. L'association des TC avec les médicaments est donc une étape importante pour permettre la caractérisation des effets secondaires partagés par des médicaments.

5.3.2 Binarisation de la matrice molécules$\times$effets secondaires pour la fouille de données

La méthode d'association utilisée en première intention est de considérer qu'une molécule est associée à un TC si elle est annotée par au moins un effet secondaire composant ce TC. Cette façon de faire est particulièrement sensible et permet d'éviter des pertes d'information. Néanmoins, cette méthode pose un problème majeur : du fait du grand nombre d'associations médicaments-TC, les algorithmes de fouille de données ne terminent pas. La fixation d'un seuil unique plus grand que

1 pour tous les TC n'est cependant pas une solution envisageable. La taille des TC variant de 2 à 59 (Figure 5.11), cela entrainerai la disparition des TC dont la taille est inférieure au seuil.

Ainsi, j'ai proposé d'utiliser un seuil variable en fonction de la taille des TC : plus un TC contient d'effets secondaires, plus le nombre de ses effets qui annotent une molécule devra être important pour associer le TC à la molécule. Cette procédure d'attribution peut être définie de la manière suivante : soit k_i le nombre de termes composant le TC$_i$ et n_i le nombre minimal d'effets secondaires nécessaire pour associer le TC$_i$ à un médicament. Si $n_i = 1$ pour tous les TC$_i$, alors l'association médicament-TC est très faible (un seul effet est suffisant) et la matrice correspondante est trop dense pour être exploitée. Alors que si $n_i = k_i$ pour tous les TC$_i$, cela entraine une association très forte entre les TC et les médicaments mais risquant de provoquer la disparition d'effets secondaires importants. Un compromis entre ces deux solutions est donc nécessaire.

Pour cela, j'ai regroupé les valeurs k_i en intervalles de taille t. Pour chaque valeur de t testée, n_i varie de 1 au nombre d'intervalles formés. J'ai ensuite testé l'effet de t sur la densité de la matrice médicaments$\times$TC et sur le nombre de motifs fréquents maximaux (MFI) extraits. Dans une matrice binaire de type objets$\times$attributs, la densité se calcule comme le rapport entre le nombre de couples (objets-attributs) associés et le nombre total de couples (objets-attributs) dans la matrice. Dans notre cas, la densité est donc le rapport entre le nombre de couples (médicaments-TC) associés entre eux, définis par la méthode expliquées précédemment, et le nombre total de couples (médicaments-TC) soit 544 médicaments$\times$112 TC.

Taille t de l'intervalle	# Intervalles	Tous les TC		TC annotant moins de 50% des molécules		
		Densité	# MFI	Densité	# MFI	# TC conservés
1	59	0.04%	0	0.04%	0	112
2	11	9%	57	8%	42	111
3	8	14%	820	12%	148	107
4	6	18%	3603	14%	169	103
5	5	21%	14 617	15%	117	97
6	5	23%	23 529	17%	295	97
7	4	25%	33 884	18%	312	95
59	1	30%	NC	21%	2024	75

TABLE 5.5 – Évaluation de l'effet de la taille des intervalles utilisés pour associer médicaments et TC. NC : non calculé 15 jours après avoir lancé l'extraction des MFI.

On peut ainsi voir dans la partie gauche de la table 5.5 que pour $t = 59$ (équivalent à l'association de type au moins un effet), le programme d'extraction des MFI ne se termine pas avec une densité de 30%. A l'opposé, avec $t = 1$ (une molécule doit être annotée par tous les effets secondaires d'un TC pour être associée à ce TC), il n'y a plus que 27 associations molécules-TC sur les 62 048 (554×112) possibles, la densité est donc proche de 0 et il n'y a pas de MFI à étudier. Ainsi, en diminuant la taille des intervalles, on s'aperçoit que la densité et le nombre de MFI à étudier par la suite diminuent. Cependant, le nombre de MFI reste très élevé pour t supérieur à 2 et la densité de 9% pour $t = 2$ indique une forte perte d'informations. Pour résoudre ces deux problèmes, il est possible de limiter le nombre de TC à prendre en compte. Ainsi, en ne prenant pas en compte les TC annotant plus de 50% de molécules, on obtient les statistiques décrites dans la partie droite de la table 5.5. De cette façon, les MFI sont toujours extractibles quelle que soit la valeur de t utilisée. La taille d'intervalle la plus intéressante est 5. En effet, la densité observée correspond à la moitié

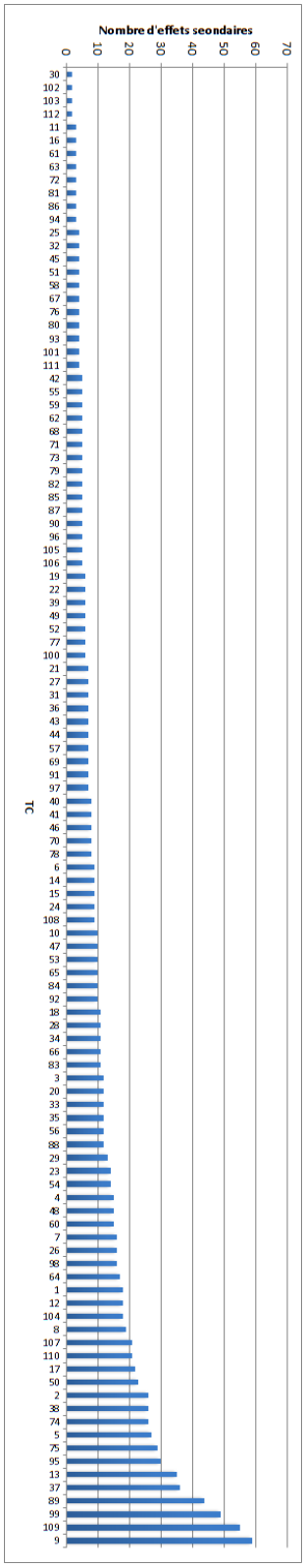


FIGURE 5.11 – Nombre d'effets secondaires par TC.

de celle déterminée par la méthode au moins un effet secondaire et le nombre de MFI à étudier est optimal. De plus, on élimine seulement 15 TC couvrant plus de la moitié des molécules. Cette valeur de t sera ensuite utilisée pour construire la matrice molécules$\times$TC qui servira à extraire des profils d'effets secondaires.

5.4 Extraction de règles explicites pour la compréhension de profils d'effets secondaires

L'article reproduit dans cette section décrit en détail la méthodologie utilisée pour obtenir des profils d'effets secondaires et les résultats obtenus par PLI pour caractériser les molécules présentant ces profils. Il a été publié dans journal *BMC Bioinformatics* le 26 juin 2013.

5.4.1 Résumé de l'article

Dans cet article, les annotations des médicaments sont extraites des bases de données SIDER et DrugBank. Les termes décrivant les effets secondaires sont regroupés en clusters de termes (TC) grâce à la mesure de similarité sémantique IntelliGO. A partir de la matrice médicaments$\times$TC nous avons extrait des motifs fréquents maximaux qui nous ont permis d'identifier des profils d'effets secondaires. Un profil d'effet secondaire est défini comme la plus longue combinaison de TC partagée par un nombre significatif de médicaments. Les profils les plus fréquents ont été explorés par deux méthodes d'apprentissage artificiel sur la base de descripteurs des médicaments et de leurs cibles. Les méthodes utilisées sont les arbres de décision et la programmation logique inductive. Bien que les deux méthodes produisent des modèles explicites, seule la PLI est capable d'exploiter les données relationnelles et ainsi d'utiliser non seulement les propriétés de médicaments mais aussi les connaissances du domaine. L'efficacité de l'apprentissage est évaluée par validation croisée et testée sur des nouvelles molécules. La comparaison des deux méthodes d'apprentissage montre que la PLI possède une plus grande sensibilité que les arbres de décision et exploite les connaissances telles que les annotations fonctionnelles et les réseaux des cibles des médicaments, produisant ainsi des règles riches et expressives. Les tests des règles sur de nouvelles molécules ont permis de prédire de façon satisfaisante des profils d'effets secondaires associés à ces molécules.

Integrative relational machine-learning approach for understanding drug side-effect profiles

Emmanuel Bresso[1,2,3*]
*Corresponding author
Email: emmanuel.bresso@loria.fr

Renaud Grisoni[2]
Email: renaud.grisoni@gmail.com

Gino Marchetti[2,4]
Email: gino.marchetti@loria.fr

Arnaud Sinan Karaboga[3]
Email: karaboga@harmonicpharma.com

Michel Souchet[3]
Email: souchet@harmonicpharma.com

Marie-Dominique Devignes[2,4]
Email: marie-dominique.devignes@loria.fr

Malika Smaïl-Tabbone[1,2*]
*Corresponding author
Email: malika.smail@loria.fr

[1]Université de Lorraine, LORIA, UMR 7503, Vandoeuvre-lès-Nancy, 54506, France

[2]INRIA, Villers-lès-Nancy, 54600, France

[3]Harmonic Pharma, Espace Transfert INRIA NGE, Villers-lès-Nancy, 54600, France

[4]CNRS, LORIA, UMR 7503, Vandoeuvre-lès-Nancy, 54506, France

Abstract

Background

Drug side effects represent a common reason for stopping drug development during clinical trials. Improving our ability to understand drug side effects is necessary to reduce attrition rates during drug development as well as the risk of discovering novel side effects in available drugs. Today, most investigations deal with isolated side effects and overlook possible redundancy and their frequent co-occurrence.

Results

In this work, drug annotations are collected from SIDER and DrugBank databases. Terms describing individual side effects reported in SIDER are clustered with a semantic similarity measure into term clusters (TCs). Maximal frequent itemsets are extracted from the resulting drug x TC binary table, leading to the identification of what we call side-effect profiles (SEPs). A SEP is defined as

the longest combination of TCs which are shared by a significant number of drugs. Frequent SEPs are explored on the basis of integrated drug and target descriptors using two machine learning methods: decision-trees and inductive-logic programming. Although both methods yield explicit models, inductive-logic programming method performs relational learning and is able to exploit not only drug properties but also background knowledge. Learning efficiency is evaluated by cross-validation and direct testing with new molecules. Comparison of the two machine-learning methods shows that the inductive-logic-programming method displays a greater sensitivity than decision trees and successfully exploit background knowledge such as functional annotations and pathways of drug targets, thereby producing rich and expressive rules. All models and theories are available on a dedicated web site.

Conclusions

Side effect profiles covering significant number of drugs have been extracted from a drug$\times$side effect association table. Integration of background knowledge concerning both chemical and biological spaces has been combined with a relational learning method for discovering rules which explicitly characterize drug-SEP associations. These rules are successfully used for predicting SEPs associated with new drugs.

Keywords

Relational machine learning, Data integration, Drug discovery, Data mining, Drug side-effects

Background

Side effects are unwanted responses to drug treatment. Some side effects are adverse, while others are more tolerable. Many side effects are detected during clinical trials, and adverse side effects are often responsible for the high attrition rate of drug candidates. For example in 2008, the French Department of Industry estimated that only 1 drug out of 250 was approved by the FDA [1]. Beside toxicity, it is not desirable to prescribe for a long period drugs having side effects like nausea or headache. Moreover, not all side effects are detected during clinical trials. For example, the cardiotoxicity of benfluorex was only recently highlighted [2] even though benfluorex was approved in the 1970's. Thus, early recognition of side effects is an important issue for drug development and safety.

To support side effect exploration, two main resources reporting their association with drugs have been developed. The FDA Adverse Event Reporting System (FAERS) stores the observed side effects reported directly by health care professionals and consumers. The SIDER database stores side-effect information mentioned on drug package inserts [3].

Two groups of studies have been conducted on side effects. On the one hand, side-effect information has been exploited for drug repositioning. For example, [4] used a corpus-based side-effect similarity approach to show that pairs of drugs sharing similar side effects can have common targets. Thus, they use side-effect similarity to predict new targets for a drug. In a similar spirit, [5] used FAERS to define pharmacological drug-drug similarity and to predict unknown drug-target interactions from the integration of the pharmacological similarity and genomic sequence similarity of target proteins. At the disease level, [6] proposed an approach based on the hypothesis that drugs sharing side effects could be indicated for the same disease. Drug side-effect associations and drug-disease relationships were used to develop a systematic drug repositioning method and to suggest, for instance, an antidiabetic effect for drugs causing porphyria.

On the other hand, other studies focus on understanding how side effects occur. As described above, relationships may exist between side effects and drug targets. Moreover, the link between chemical structure and side effects was shown by [7]. From a more mechanistic point of view, [8] showed that side effects can be correlated with the biological processes in which the drug targets are involved. For instance, they showed that nausea is correlated to an up-regulation of the deaminase activity. A very recent paper aims at predicting the side-effect profiles of molecules based on their chemical structures (defining the chemical space) and the information of their target proteins (defining the biological space) [9]. The so-called side-effect profile of a molecule is simply defined as its binary fingerprint with respect to the side-effect terms. However, such earlier studies have several limitations. For example, (i) they consider only individual side effects, and ignore the fact that often more than one side effect is associated with a drug, (ii) the biological space is over-simplified, and (iii) the resulting prediction models are "black boxes" which do not provide any explicit and reusable knowledge.

Here, we study in a systematic way drug side-effect associations, and we propose a method for identifying and characterizing side-effect profiles (SEPs) shared by several drugs.

Our approach is composed of five main steps, as illustrated in Figure 1. The first step (Figure 1A) consists of grouping the terms used for side effects in SIDER using a semantic similarity measure in order to build Term Clusters (TC) corresponding to groups of semantically related SEs [10]. In parallel, drugs from SIDER are mapped to DrugBank in order to retrieve information about drugs themselves and their targets (Figure 1B). Then, TCs and drugs are associated in order to represent each drug by a side-effect fingerprint (Figure 1C). SEPs are extracted as maximal frequent itemsets from side effect fingerprints (Figure 1D). The aim is then to characterize each SEP in terms of drug and target properties. This can be addressed as a supervised classification task. Two machine-learning methods are chosen for this task: Decision Trees (DTs) and Inductive Logic Programming (ILP) (Figure 1E). These two methods provide easily readable results which can then be exploited for understanding SEPs. Decision trees use a single table as input in which each row corresponds to a drug and each column to a drug descriptor. Inductive Logic Programming uses relational descriptors to learn a first-order-logic concept definition from observations. Relational descriptors encoding characteristics of both drugs and their targets are retrieved from our "NetworkDB" integrated database, which is built from several data sources including DrugBank, UniProt, KEGG, and GO. The models obtained for a set of selected SEPs with these two machine-learning methods are then evaluated by cross-validation and tested directly with new drugs. Finally, some elements are provided for model interpretation.

Figure 1 Overview of our approach for characterizing drug-SEP associations. Terms used for describing side effects in SIDER DB are grouped using a semantic similarity measure in order to build Term Clusters or TCs **(A)**. Drugs are mapped to DrugBank in order to retrieve information about drugs themselves and their targets **(B)**. TCs are associated to drugs to represent each drug by a side-effect fingerprint **(C)**. SEPs are extracted as maximal frequent itemsets from side effect fingerprints **(D)**. Two machine-learning methods are used to characterize each SEP in terms of drug and target properties **(E)**.

Methods

The NetworkDB resource

NetworkDB is a relational database which integrates data about molecules and their targets. These data are collected from various public data sources mentioned in the following sections. Figure 2 shows the conceptual model of the database.

Figure 2 NetworkDB conceptual model. In this entity-relationship schema, entities are in boxes and relationships in ellipses.

Chemical space: drugs and their properties

The SIDER database contains drug side-effect relationships [3]. DrugBank is used to collect data such as categories and targets [11] . The join between SIDER and DrugBank is based on the PubChem Compound identifier given by SIDER and DrugBank. A total of 554 drugs from SIDER are referenced in DrugBank v3.0.

Each drug is described by its category and a set of clusters it belongs to. In fact, various structural representations and associated similarity measures were used to cluster drugs. The first similarity measure is based on SMILES representation. The SMILES codes are converted thanks to Open Babel program into fingerprints which allows linear and ring substructures to be identified [12]. Then, the structural similarity between two molecular fingerprints is calculated using the Tanimoto measure. In addition, we calculated three other similarity scores using spherical harmonics representation of molecules. This parametric representation of macromolecular surface was originally proposed and applied by [13] and [14]. The proprietary program HPCC (Harmonic Pharma) supports three variants of the spherical harmonic representation. HPCCgeo uses spherical harmonic coefficients (shape information) to calculate similarity between drugs, HPCCchem is based on chemical properties mapped on the spherical harmonic representation, and HPCCcombo combines shape and chemical information. Ward's method is used to perform four hierarchical clusterings of drugs [15]. The optimal numbers of clusters is determined by the method of [16]. Thus, 60 clusters are obtained with Tanimoto, 53 with HPCCgeo, 21 with HPCCchem and 34 with HPCCcombo measures.

Drug categories are retrieved from DrugBank. These categories are mapped on the descendants of three MeSH concepts, namely "Molecular Mechanisms of Pharmacological Action" (*D27.505.519*), "Physiological Effects of Drugs" (*D27.505.696*) and "Therapeutic Uses" (*D27.505.954*).

Biological space: proteins and their properties

Drug targets are extracted from both DrugBank and PDB [17]. The outer join between PDB and DrugBank (retaining all DrugBank targets) is based on SMILES code identity. Drug targets are associated with their UniProt accession numbers. Thus, 768 targets are collected, representing an average of four targets per drug. Then, target annotations are retrieved from different databases. Protein-protein interactions are retrieved from the IntAct database [18] and 5959 interactions were collected which correspond to 2827 new proteins. For all the proteins (drug targets and their interactants), 1403 pathway names are extracted from the KEGG database and the Pathway Interaction Database which integrates data from NCI-Nature, BioCarta and Reactome [19,20]. For the same proteins, GO terms are also collected from QuickGO database [21]. Thus, 6494 GO terms annotating the 3595 proteins are stored in NetworkDB. Moreover, the "is_a" and "part_of" relationships between GO terms are stored in NetworkDB. Finally, 4650 protein domains associated with the targets and their interactants are retrieved from InterPro [22].

Grouping side-effect terms into term clusters

Side effects are extracted from SIDER. As shown previously [23], the use of all terms describing side effects in SIDER (about 1500) impairs the execution of data mining programs and produces numerous and redundant patterns which are inappropriate for expert interpretation. As SIDER side effects terms belong to the Medical Dictionary for Regulatory Activities [24] , a semantic similarity between these terms can be calculated based on the structure of MedDRA [10]. Next, a hierarchical clustering method

is applied to obtain 112 Term Clusters (TCs) which are then validated by experts [23]. For instance, TC named 65_Dermatitis is the 65th TC and has Dermatitis as representative term.

Datasets

Association of drugs with side effects

The association between drugs and TCs is an important step for the characterization of drugs sharing side effects. As the TC size varies from 2 to 59 terms, it seems consistent to use a heuristic procedure depending on the TC size. Let k_i be the number of terms in TC_i and n_i be the minimal number of side effects required for assigning TC_i to a drug. Considering $n_i = 1$ for any TC_i results in a very loose association yielding a very dense binary table hampering further computation, whereas considering $n_i = k_i$ for any TC_i results in a very stringent association which might skip over important drug side effects. In fact a trade-off between these two extreme solutions is required. Grouping the k_i values into 5-range intervals with the last interval from 21 to 59 allows to set up a simple association procedure ranging n_i from 1 to 5. The resulting association between drugs and TCs is shown in Figure 3 where each row represents the side-effect binary fingerprint associated with a drug. This binary table (drug×TC) is then used to discover interesting side-effect profiles defined here as the longest combinations of TCs shared by significant sets of drugs.

Figure 3 Drug side-effect binary table. This table is presented as a heatmap (produced with R) where rows and columns are grouped by distribution similarity. Each row represents the side-effect fingerprint of a drug and each column is a side-effect term cluster.

Single-table datasets

Single table datasets designed for DT learning represent each drug by an attribute-value vector. Four types of descriptors retrieved from NetworkDB are used to generate these attributes: the first is the class information, *i.e.* the studied SEP, the second one includes drug categories, the third one lists all drug targets with for each target, three attributes referring to the type of action of the drug (activation, inhibition and other) and the fourth concerns clusters of similar drugs according to the four similarity measures described above. Because of target and category multiplicity, the total dimension of this dataset varies between 741 and 924 depending on the SEP.

Relational datasets

Relational datasets designed for Inductive Logic Programming (ILP) consist in a set of tables extracted from NetworkDB describing drugs properties and background knowledge. Drugs properties are the same as in the single-table dataset, *i.e.* categories, targets and clusters. Background knowledge includes GO annotations, domain composition, interactants and pathways of each drug target. Relationships between GO terms constitute an additional table.

Data mining

Maximal frequent itemsets

In a binary table (object×attribute), a frequent itemset is a group of attributes shared by a number of objects greater than a threshold support. A frequent itemset is considered as a maximal frequent itemset (MFI) if all its proper supersets are not frequent [25]. It follows that two maximal frequent itemsets (MFIs) cannot be shared by a number of objects greater than the threshold support. In our case, MFIs

are the largest combinations of TCs shared by a number of drugs greater than 100. This threshold was chosen as a trade-off between high values yielding short MFIs limited to one or two TCs and low values yielding numerous MFIs covering only a few molecules. MFIs are extracted from the binary table (Figure 2) using the Coron program [26] after excluding TCs which cover more than 50% of the molecules.

Decision trees

Decision tree (DT) construction is a machine-learning method which uses (object×attribute) table to classify objects. Results given by this method are easily readable. Decision trees are built here with the J48 implementation of C4.5 tree learner in the Weka toolbox using single table datasets converted into the ARFF format [27]. We use the default parameters except for two of them: we use $minNumObj = 5$ and $binarySplits = true$.

Inductive Logic Programming (ILP)

ILP is a machine-learning method which uses relational data as input and has been successfully applied to various areas including bioinformatics [28-30]. It allows us to learn a concept definition from observations, i.e, a set of positive examples (E+) and a set of negative examples (E-), and background knowledge (B) [31]. The ILP experiments produce theories as sets of first-order logic rules. They where conducted here with the Aleph Program [32] . Many parameters can be tuned for theory construction. The three main parameters are the *min-pos*, the *noise* and the *induce-type*. The *min-pos* parameter is the minimal number of positive examples that a rule must cover. The *noise* corresponds to the maximal number of negative examples that an acceptable rule may cover (in our case, one is never sure that a drug does not have a given side effect). The third parameter is *induce-type* which directs theory construction. When this parameter is set to *induce-cover*, overlapping rules are produced (*i.e.*, a drug can be covered by several rules). Based on previous experience [33], we used the following settings: $min\text{-}pos = 5$, $noise = 1$ and $induce\text{-}type = induce\text{-}cover$.

Model evaluation

Cross-validation

Both ILP theories and decision trees are evaluated with 10 runs of a 10-fold stratified cross-validation. DT cross-validation is performed with the Weka experimenter interface. For ILP, we took advantage of our recent integration of Aleph into the KNIME platform [34] . KNIME cross-validation meta-node is adapted for theory evaluation. An example is predicted as positive if it is covered by at least one rule. Each cross-validation assay yields a confusion matrix counting true and false positives, as well as true and false negatives. Each assay is then evaluated by the calculation of accuracy (ratio of correctly classified instances), specificity (true negative rate) and sensitivity (true positive rate).

Direct test

Theories and decision trees are also evaluated by direct test. Drugs used for testing are those present in SIDER 2 and DrugBank (v3.0) but not present in SIDER. For these drugs all descriptors are retrieved and stored in the NetworkDB. Furthermore, the reports of FAERS from 2004 to 2011 were imported as a database and used as an external information source for checking the false positives predicted by our models. We consider that a molecule is associated with a SEP in FAERS if for each TC of the SEP there is at least one report that states that the molecule is the primary suspect of an observed side effect belonging to the TC. Our checking procedure is just an anticipation as it relies on the fact that updating

the package insert of a drug (stored in SIDER) requires that sufficient amount of adverse effect incidents occur (especially for new drugs).

Results and discussion

Overall distribution of side effects

A drug is associated with a TC (group of semantically related side effects) if it is annotated by a minimum number of side effects of this TC (see Methods). The resulting binary table is shown in Figure 3, where each row represents the side effect fingerprint of one of the 554 drugs considered here, and each column represents one of the 112 TC. In this representation, drugs and TCs have been grouped by distribution similarity. On the right part of the figure, we can see TCs associated with a limited number of drugs, whereas highly represented TCs are on the left. In the same way, drug fingerprints involving few TCs are on the top of Figure 3 and drugs with high number of TCs are on the lower part. Zooming on adjacent columns reveals that some TCs seem to be frequently associated with the same drugs as for example the pair TC 39_Stevens-Johnson_syndrome and TC 100_Erythema_multiforme.

However, apart from providing a general idea about the complexity of TC association with drugs, this visualization cannot be exploited easily. More precise information can be retrieved by querying NetworkDB. For example, the maximal number of TCs associated with a drug is 89 for the ropinirole (an anti-Parkinson agent). Conversely, 18 drugs are associated with only one TC. For instance, bretilium (an anti-hypertensive agent) is only associated with TC 110_Shock. From the TC point of view, the number of drugs associated with a TC ranges from 1 to 410. The 13 TCs covering more than 50% of the molecules are excluded in the rest of the study.

Side-effect profiles

The overall intuition provided by Figure 3 is that groups of TCs shared by drugs exist and should be extracted. In fact, extracting patterns from such binary table is the purpose of itemset search algorithms [35]. We thus perform MFI extraction and we define side-effect profiles (SEPs) as maximal groups of TCs covering at least 20% of the drug set (110 drugs). The resulting 26 SEPs are listed in Table 1. Regarding length, 3 SEPs have only one TC, 13 combine 2 TCs, 9 combine 3 TCs, and only one combines 4 TCs. These 26 SEPs concern 372 molecules (67% of the drug set) and involve 18 distinct TCs of which the most frequent are 99_Headache and 90_Feeling_abnormal which appear 8 times each, whereas 7 TCs appear in only one SEP. These 26 most frequent SEPs are considered in the rest of the study. By construction, although two SEPs can have common TCs, they cannot cover more than 100 molecules in common.

Table 1 Maximal frequent itemsets covering 20% of drugs (support) extracted from the drug×TC table

SEP	Profile composition	Support	Avg overlap
SEP_1	41_Leukopenia, 90_Feeling_abnormal, 99_Headache	123	69
SEP_2	90_Feeling_abnormal, 99_Headache, 110_Shock	123	73
SEP_3	58_Gout	120	60
SEP_4	70_Pneumonia, 99_Headache	117	71
SEP_5	110_Shock, 111_Infection	117	68
SEP_6	76_Asthma, 90_Feeling_abnormal, 99_Headache	117	68
SEP_7	65_Dermatitis	116	53
SEP_8	2_Haemorrhage, 76_Asthma	115	65
SEP_9	41_Leukopenia, 76_Asthma	115	62
SEP_10	48_Rhinitis, 99_Headache, 111_Infection	115	69
SEP_11	41_Leukopenia, 110_Shock	114	66
SEP_12	39_Stevens-Johnson_syndrome, 41_Leukopenia, 100_Erythema_multiforme	114	52
SEP_13	41_Leukopenia, 48_Rhinitis	113	67
SEP_14	99_Headache, 100_Erythema_multiforme	113	56
SEP_15	31_Lymphadenopathy	112	59
SEP_16	70_Pneumonia, 90_Feeling_abnormal	112	71
SEP_17	41_Leukopenia, 70_Pneumonia	112	64
SEP_18	76_Asthma, 111_Infection	112	64
SEP_19	80_Jaundice, 100_Erythema_multiforme	112	45
SEP_20	41_Leukopenia, 111_Infection	111	63
SEP_21	8_Haematuria, 90_Feeling_abnormal, 99_Headache	111	68
SEP_22	13_Pyrexia, 33_Musculoskeletal_discomfort, 48_Rhinitis, 99_Headache	111	69
SEP_23	13_Pyrexia, 70_Pneumonia	110	69
SEP_24	48_Rhinitis, 90_Feeling_abnormal, 110_Shock	110	70
SEP_25	13_Pyrexia, 90_Feeling_abnormal, 110_Shock	110	70
SEP_26	48_Rhinitis, 90_Feeling_abnormal, 111_Infection	110	69

Avg overlap: average of overlap size between the SEP and other SEPs.

Characterization of frequent SEPs

Our hypothesis is that a SEP shared by a large number of drugs can be explained in terms of drug properties and background knowledge. Thus, two machine-learning methods, decision trees and ILP, are applied on the drugs associated with each SEP. For both methods, the positive examples are taken to be all the drugs associated with a SEP, and those drugs that are not associated with any of the TCs composing the SEP are taken as negative examples. Negative examples represent 60% of the learning set.

For each profile, classification efficiency is evaluated using a 10×10 cross-validation by accuracy (Acc), specificity (Spec) and sensitivity (Sens). The results presented in Table 2 show that for both methods, generated models are good classifiers with an average accuracy of 67% for DTs and 65% for ILP. For 23/26 SEPs, accuracy is better for DTs than with ILP mostly reflecting the higher specificity values obtained with DTs. On the contrary, sensitivity values are always higher with ILP than with DTs with only one exception for SEP_17 where ILP sensitivity value is 0.1 lower than DTs sensitivity. Thus, ILP provides more sensitive theories whereas DTs provide more specific models. In fact, sensitivity is probably more important than specificity for drug development as it is for medical diagnostic. Indeed, low sensitivity means that some SEPs can be skipped over, although they are truly associated with the tested drug. Thus, ILP theories display attractive qualities for SEP prediction. Five SEPs (1, 3, 12, 15, and 19) are particularly well characterized with ILP since sensitivity values are greater than 60%. The amount and quality of available data may explain the observed differences of results between SEPs. It should be noted that comparison with other reported methods is uneasy due to the fact that we aim to characterize and predict SEPs rather than isolated side effects. In fact the closest study is the one of [9]

whose objective is to predict isolated side effects using multi-class statistical methods. Therefore these authors do not produce comparable accuracy values.

Table 2 Evaluation of learning results by 10×10 stratified cross-validation of DT and ILP programs

SEP	DT			ILP		
	Acc	Spec	Sens	Acc	Spec	Sens
SEP_1	0.65	0.86	0.39	0.61	0.63	0.6
SEP_2	0.69	0.88	0.4	0.63	0.69	0.54
SEP_3	0.71	0.88	0.47	0.71	0.77	0.63
SEP_4	0.66	0.89	0.32	0.62	0.7	0.51
SEP_5	0.68	0.88	0.38	0.64	0.7	0.54
SEP_6	0.68	0.87	0.39	0.61	0.69	0.49
SEP_7	0.65	0.86	0.32	0.6	0.67	0.49
SEP_8	0.7	0.87	0.44	0.67	0.73	0.57
SEP_9	0.69	0.84	0.46	0.69	0.75	0.59
SEP_10	0.7	0.89	0.4	0.65	0.76	0.47
SEP_11	0.7	0.88	0.44	0.7	0.82	0.45
SEP_12	0.71	0.88	0.45	0.7	0.76	0.61
SEP_13	0.67	0.88	0.35	0.66	0.74	0.54
SEP_14	0.69	0.89	0.39	0.63	0.71	0.51
SEP_15	0.71	0.9	0.43	0.69	0.76	0.6
SEP_16	0.69	0.89	0.39	0.66	0.72	0.57
SEP_17	0.74	0.89	0.52	0.65	0.74	0.51
SEP_18	0.65	0.87	0.34	0.61	0.69	0.5
SEP_19	0.74	0.91	0.47	0.72	0.77	0.64 0
SEP_20	0.71	0.89	0.44	0.64	0.73	0.51
SEP_21	0.72	0.9	0.46	0.64	0.72	0.54
SEP_22	0.65	0.88	0.32	0.61	0.69	0.48
SEP_23	0.71	0.89	0.43	0.63	0.7	0.51
SEP_24	0.68	0.87	0.4	0.62	0.71	0.5
SEP_25	0.71	0.9	0.43	0.65	0.72	0.56
SEP_26	0.69	0.88	0.4	0.62	0.69	0.52
Average	0.67	0.83	0.43	0.65	0.72	0.54

Acc: accuracy, Spec: specificity, Sens: sensitivity.

Table 3 shows the results obtained with the set of test molecules. Among the novel drugs present in SIDER 2, only 20 are associated with at least one of the 26 studied SEPs. These drugs have been tested with decision trees and ILP theories obtained for each SEP. The total number of drugs in the test set that are associated which each SEP is indicated (column Positives) and compared to the true positive values (TP columns) obtained with test set using either DT model or ILP theory relative to this SEP. Clearly the prediction results are better with ILP theories than with DTs. Indeed 22 true positives (covering 16 SEPs) were detected with ILP theories whereas only 9 true positives (covering 8 SEPs) were detected with DTs. The number of false positives are also reported for each SEP and each model (FP columns). The checking procedure was applied on false positives and the number of confirmed molecules according to FAERS is reported (FAERS columns). Thus, 33 molecules were extracted for ILP theories versus 37 for DTs raising the total number of probable true positives to 55 for ILP and 46 for DTs. Nevertheless, as the variability in cross-validation results suggest, many positive molecules still escape prediction especially for three SEPs: SEP_2, SEP_7, and SEP_21 with both DTs and ILP theories.

Table 3 Direct testing results with 20 new molecules

SEP	Positives	DT			ILP		
		TP	FP	FAERS	TP	FP	FAERS
SEP_1	4	0	5	1	2	3	1
SEP_2	11	1	2	1	0	1	1
SEP_3	2	0	3	1	0	5	1
SEP_4	3	0	3	1	1	2	1
SEP_5	5	0	2	1	1	2	1
SEP_6	5	1	5	3	2	3	1
SEP_7	15	2	2	1	2	1	1
SEP_8	4	1	1	1	1	3	1
SEP_9	3	0	3	2	0	3	1
SEP_10	5	0	1	0	0	3	1
SEP_11	4	1	3	2	1	3	1
SEP_12	0	0	5	1	0	4	1
SEP_13	4	0	6	2	1	4	1
SEP_14	4	1	0	0	1	5	2
SEP_15	1	0	2	1	0	5	2
SEP_16	3	0	6	2	2	7	3
SEP_17	1	0	5	3	0	2	1
SEP_18	2	0	3	1	0	2	1
SEP_19	1	0	4	2	0	5	2
SEP_20	3	0	3	1	1	4	1
SEP_21	8	1	2	1	1	2	1
SEP_22	5	0	3	1	1	5	1
SEP_23	3	0	3	1	1	4	1
SEP_24	5	1	4	2	2	5	2
SEP_25	8	0	3	2	2	5	2
SEP_26	4	0	6	3	0	3	1

Positives: number of positive examples in the test set according to SIDER, TP/FP: number of predicted true/false positives, FAERS: number of fished out molecules based on FAERS data.

Interpretation of decision trees and theories

Quantitative characteristics of DT models and ILP theories for the 26 selected SEPs are presented in Table 4 (the decision trees and ILP theories are available at http://plateforme-mbi.loria.fr/side-effect-profiles). The first observation concerns model coverage. We can see that in average 83% of the drugs are covered by at least one rule in an ILP theory whereas DT models cover in average only 58% of the drugs composing the learning set. The second observation is the use of almost all descriptor types in each DT model or ILP theory. The most represented descriptors are drug categories and clusters for DTs, respectively drug targets and GO terms for ILP theories. This illustrates the importance of using background knowledge about drug targets and GO semantic relationships for the characterization of SEPs.

Table 4 Quantitative characteristics of DT models and ILP theories

	DT (# nodes per model)		ILP (# rules per theory)	
	Avg (min-max)	**% total**	**Avg (min-max)**	**% total**
Model coverage (%)	58 (32–67)	-	83 (77–88)	-
Model size	11 (6–15)	-	33 (16–40)	-
Drug descriptors				
Categories	4 (1–7)	34	6 (2–13)	19
Targets	3 (0–5)	26	30 (23–39)	90
Clusters	4 (1–9)	40	9 (4–14)	27
Target descriptors				
GO terms	NA	NA	24 (16–31)	73
Domains	NA	NA	1 (0–2)	1
Interactions	NA	NA	8 (2–16)	24
Pathways	NA	NA	4 (1–8)	12
GO relationships	NA	NA	6 (3–9)	19

Model coverage is the percentage of positive examples covered, averaged over the 26 DT models and 26 ILP theories. Avg: average. Model size corresponds to the average number of nodes in a DT model or of rules in a ILP theory. Occurrence of each type of descriptor is estimated by counting the number of nodes (rules respectively) involving them (NA: not applicable).

It is worth noting that some rules contained in theories were confirmed using peer-reviewed literature. For example, considering the SEP_7 (65_Dermatitis) theory, rule 11 says that a drug is associated with this SEP if its target interacts with a protein belonging to the KEGG pathway "Focal adhesion" and to the PID pathway "Signaling events mediated by focal adhesion kinase" (Table 5). By searching the list of genes implied in dermatitis [36] and confronting them to the 2 pathways, we extract 7 genes (*THBS1*, *COL1A2*,*COL3A1*, *COL4A1*, *COL5A*, *ITGB4* and *LAMA5*) dysregulated in dermatitis which belong to the KEGG pathway "Focal adhesion". In the same way, two genes (*BDKRB2* and *PTGFR*) are known to be dysregulated in dermatitis and belong to the "Neuroactive ligand-receptor interaction" KEGG pathway mentioned in rule 14. Finally, if we consider rule 16 we could verify that the gene *ERBB3* belonging to the "Endocytosis" KEGG pathway is indeed down regulated in dermatitis.

Table 5 Theory obtained for 65_Dermatitis SEP (SEP_7)

Rule #	Condition part of the rule	P	N
3	drug_has_target(A,B,inhibitor), goterm(B,'cellular response to insulin stimulus')	15	1
18	drug_has_target(A,B,inhibitor), goterm(B,C), go_relation(C,part_of,go:21543)	13	1
1	drug_has_target(A,B,activator), interact(B,C), goterm(C,'central nervous system development')	12	1
30	drug_has_target(A,B,inhibitor), interact(B,C), pathway(C,'BCR signaling pathway',pid), drug_cluster(A,'17_quinine',hpcc)	12	0
24	drug_has_target(A,B,inhibitor), interact(B,C), goterm(C,'translation'), interact(C,D)	10	1
20	drug_has_target(A,B,inhibitor), interact(B,C), pathway(C,'BCR signaling pathway',pid), pathway(C,'EPO signaling pathway',pid)	9	1
25	drug_has_target(A,B,activator), goterm(B,'lipid binding'), goterm(B,'ligand-dependent nuclear receptor activity')	9	1
35	drug_has_target(A,B,activator), interact(B,C), goterm(C,'identical protein binding'), goterm(C,'DNA binding')	9	1
6	drug_has_target(A,B,inhibitor), goterm(B,'protein homodimerization activity'), drug_cluster(A,'16_gliclazide',hpcc)	8	0
8	drug_has_target(A,B,activator), interact(B,C), interact(C,'Serine/threonine-protein phosphatase 2A 55 kDa regulatory subunit B beta isoform')	8	1
15	drug_has_target(A,B,inhibitor), goterm(B,'response to ethanol'), goterm(B,'signal transduction')	8	1
19	drug_has_target(A,B,inhibitor), goterm(B,C), go_relation(C,is_a,go:8227), drug_cluster(A,'16_Flavoxate',hpcombo)	8	0
31	drug_has_target(A,B,inhibitor), interact(B,C), interact(C,'Dedicator of cytokinesis protein 1')	8	0
5	drug_has_target(A,B,activator), goterm(B,'receptor activity'), interact(B,C), goterm(C,'mitosis')	7	1
10	drug_has_target(A,B,inhibitor), goterm(B,C), go_relation(C,is_a,'cation channel activity'), goterm(B,'serotonin receptor activity')	7	1
14	**drug_has_target(A,B,activator), pathway(B,'Neuroactive ligand-receptor interaction',kegg), goterm(B,'transcription, DNA-dependent'), goterm(B,'signal transduction')**	**7**	**0**
16	**drug_has_target(A,B,inhibitor), pathway(B,'Endocytosis',kegg)**	**7**	**0**
21	drug_has_target(A,B,activator), interact(B,C), interact(C,'RNA polymerase-associated protein CTR9 homolog')	7	1
22	drug_has_target(A,B,inhibitor), pathway(B,'Role of Calcineurin-dependent NFAT signaling in lymphocytes',pid), goterm(B,'signal transduction')	7	1
23	drug_has_target(A,B,inhibitor), interact(B,C), domain(C,' Protein synthesis factor, GTP-binding')	7	1
28	drug_cluster(A,'7_marinol',hpcombo)	7	1
7	category(A,'Topoisomerase Inhibitors'), drug_has_target(A,B,inhibitor), goterm(B,'transferase activity')	6	1
12	drug_cluster(A,'29_norfloxacin',hpcf)	6	1
17	category(A,'Cyclooxygenase 2 Inhibitors'), drug_cluster(A,'2_estazolam',hpcc)	6	0
32	drug_has_target(A,B,activator), goterm(B,'inflammatory response'), goterm(B,'protein binding')	6	0
2	category(A,'Serotonin Uptake Inhibitors')	5	0
4	drug_has_target(A,B,inhibitor), goterm(B,'synapse assembly'), drug_cluster(A,'14_fentanyl',hpcombo)	5	1
9	drug_has_target(A,B,activator), goterm(B,'protein heterodimerization activity'), goterm(B,'cell-cell signaling')	5	1
11	**drug_has_target(A,B,other), interact(B,C), pathway(C,'Focal adhesion',kegg), pathway(C,'Signaling events mediated by focal adhesion kinase',pid)**	**5**	**0**
13	category(A,'HIV Protease Inhibitors'), drug_has_target(A,B,inhibitor), goterm(B,C), go_relation(C,is_a,D), go_relation(D,is_a,'catalytic activity')	5	1
26	drug_has_target(A,B,inhibitor), goterm(B,'heart development')	5	1
27	drug_has_target(A,B,inhibitor), goterm(B,C), go_relation(C,is_a,go:65008), drug_cluster(A,'55_thiothixene',tanimoto)	5	0
29	category(A,'HIV Protease Inhibitors'), drug_has_target(A,B,inhibitor), goterm(B,'oxidation reduction')	5	0
33	drug_has_target(A,B,other), goterm(B,C), go_relation(C,is_a,go:51240)	5	1
34	drug_has_target(A,B,inhibitor), goterm(B,C), go_relation(C,is_a,'binding'), drug_cluster(A,'27_quinine',hpcombo)	5	1

The condition parts of the 35 rules contained in SEP_7 theory are given with the number of positive (P) and negative (N) covered examples. The 3 rules confirmed using peer-reviewed literature are in bold. Rules are ordered by number of positive covered examples.

Finally, from a more global point of view the drugs can be represented according to the rules they satisfy resulting in a drug×rule binary table. This table constitutes a kind of abstraction of the initial drug×TC binary table (Figure 3) based on extracted knowledge. Interestingly this new representation leads to improved clustering results for the drug set (not shown) and could be further exploited for prediction studies of particular SEPs.

Conclusions

Our study proposes an integrative machine-learning approach for predicting side-effect profiles (SEPs) and understanding their mechanisms. We integrate drug characteristics and background knowledge such as functional annotation, interactions and pathways in a relational database. An extensive learning set is built by associating drugs with clusters of side effects (TCs) according to SIDER information. Our first contribution consists of extracting SEPs from this complex table of fingerprints as the longest groups of TC shared by more than one hundred drugs. We also set up two machine-learning methods, namely decision trees and inductive logic programming in order to learn which combination of properties of drugs and their targets leads to a given SEP. After evaluating the learning models, our general observation is that ILP models have a higher sensitivity than DT models. Because higher sensitivity means predicting fewer false negatives, this means that ILP predicts SEPs more often than decision trees. This was confirmed on a small test set including a checking procedure using FAERS as external and complementary information source. Indeed, more sophisticated prediction procedures can be designed integrating FAERS and based on selected rules. This should improve the prediction accuracy at least for specific SEPs displaying good quality data. The results obtained with ILP also show that background knowledge is well exploited during rule induction. Thus, in addition to targets, chemical structure and biological process annotation already studied by other groups [4,7,8], we show that information about pathways, protein-protein interaction and to a lower extent protein domains also plays an important role in side effect characterization. Further experiments may include other types of background knowledge such as clinical data and/or polymorphisms.

In our approach we characterize SEPs instead of individual TCs. Indeed as drugs are frequently associated with more than one TC, studying separately each TC implicitly assumes that side effects occur independently one from the other. This likely corresponds to a simplified view of side-effect occurrence and the existence of SEPs shared by more than 20% of the drug set strongly suggests that side effects are correlated. Moreover our approach can be applied to any user-defined SEP or TC of interest.

We believe that our approach represents a valuable methodology for understanding and predicting side-effect profiles. Our results suggest that the first-order logic theories can already be used during the drug discovery process in order to early anticipate side-effect apparition and thus decrease the attrition rate.

Availability of supporting data

All decision trees and ILP theories are available at http://plateforme-mbi.loria.fr/side-effect-profiles.

Abbreviations

Acc: Acccuracy; DT: Decision tree; ILP: Inductive logic programming; MFI: Maximal frequent itemset; Sens: Sensibility; SE: Side effect; SEP: Side effect profile; Spec: Specificity; TC: Term cluster.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

EB participated to the conception and design of the study and acquisition of data. He carried out the machine learning experiments. RG designed and developed programs for automatizing machine learning experiments and cross validations. GM carried out the clustering experiments on molecules. ASK and MS participated in the conception of the study and the interpretation and critical analysis of the results. MDD and MST conceived the study and carried out its design and coordination and helped EB to draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements

EB benefited from a CIFRE contract (ANRT) including the Harmonic Pharma Company. RG was funded by Inria. GM was funded by CNRS via the BioProLor project. Thanks to Dave Ritchie and Anisah Ghoorah for their careful reading of the paper.

References

1. **U.S. Food and Drug Administration** [http://www.fda.gov]
2. Derumeaux G, Ernande L, Serusclat A, Servan E, Bruckert E, Rousset H, Senn S, Van Gaal L, Picandet B, Gavini F, Moulin P: **Echocardiographic evidence for valvular toxicity of benfluorex: a double-blind randomised trial in patients with type 2 diabetes mellitus.** *PLoS ONE* 2012, **7**(6):e38273.
3. Kuhn M, Campillos M, Letunic I, Jensen LJ, Bork P: **A side effect resource to capture phenotypic effects of drugs.** *Mol Syst Biol* 2010, **6**. [http://dx.doi.org/10.1038/msb.2009.98]
4. Campillos M, Kuhn M, Gavin AC, Jensen LJ, Bork P: **Drug target identification using side-effect similarity.** *Science* 2008, **321**(5886):263–266.
5. Takarabe M, Kotera M, Nishimura Y, Goto S, Yamanishi Y: **Drug target prediction using adverse event report systems: a pharmacogenomic approach.** *Bioinformatics* 2012, **28**(18):i611.
6. Yang L, Agarwal P: **Systematic drug repositioning based on clinical side-effects.** *PLoS ONE* 2011, **6**(12):e28025.
7. Scheiber J, Jenkins JL, Sukuru SC, Bender A, Mikhailov D, Milik M, Azzaoui K, Whitebread S, Hamon J, Urban L, Glick M, Davies JW: **Mapping adverse drug reactions in chemical space.** *J Med Chem* 2009, **52**(9):3103–3107.
8. Lee S, Lee KH, Song M, Lee D: **Building the process-drug-side effect network to discover the relationship between biological processes and side effects.** *BMC Bioinformatics* 2011, **12**(Suppl 2):S2.
9. Yamanishi Y, Pauwels E, Kotera M: **Drug side-effect prediction based on the integration of chemical and biological spaces.** *J Chem Inf Model* 2012, **52**(12):3284–3292.
10. Benabderrahmane S, Smail-Tabbone M, Poch O, Napoli A, Devignes MD: **IntelliGO: a new vector-based semantic similarity measure including annotation origin.** *BMC Bioinformatics* 2010, **11:**588.

11. Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, Pon A, Banco K, Mak C, Neveu V, Djoumbou Y, Eisner R, Guo AC, Wishart DS: **DrugBank 3.0: a comprehensive resource for 'omics' research on drugs.** *Nucleic Acids Res* 2011, **39**(Database issue):D1035–1041.

12. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR: **Open Babel: An open chemical toolbox.** *J Cheminform* 2011, **3:**33.

13. Ritchie DW, Kemp GJL: **Fast computation, rotation, and comparison of low resolution spherical harmonic molecular surfaces.** *J Comput Chem* 1999, **20**(4):383–395. [http://dx.doi.org/10.1002/(SICI)1096-987X(199903)20:4<383::AID-JCC1>3.0.CO;2-M]

14. Cai W, Xu J, Shao X, Leroux V, Beautrait A, Maigret B: **SHEF: a vHTS geometrical filter using coefficients of spherical harmonic molecular surfaces.** *J Mol Model* 2008, **14**(5):393–401.

15. Ward JH: **Hierarchical grouping to optimize an objective function.** *J Am Stat Assoc* 1963, **58**(301):236–244. [http://amstat.tandfonline.com/doi/abs/10.1080/01621459.1963.10500845]

16. Kelley LA, Gardner SP, Sutcliffe MJ: **An automated approach for clustering an ensemble of NMR-derived protein structures into conformationally related subfamilies.** *Protein Eng* 1996, **9**(11):1063–1065.

17. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The protein data bank.** *Nucleic Acids Res* 2000, **28:**235–242.

18. Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, Duesbury M, Dumousseau M, Feuermann M, Hinz U, Jandrasits C, Jimenez RC, Khadake J, Mahadevan U, Masson P, Pedruzzi I, Pfeiffenberger E, Porras P, Raghunath A, Roechert B, Orchard S, Hermjakob H: **The intAct molecular interaction database in 2012.** *Nucleic Acids Res* 2012, **40**(Database issue):D841–D846.

19. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M: **KEGG for integration and interpretation of large-scale molecular data sets.** *Nucleic Acids Res* 2012, **40**(Database issue):D109–D114.

20. Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, Buetow KH: **PID: the Pathway Interaction Database.** *Nucleic Acids Res* 2009, **37**(Database issue):D674–679.

21. Binns D, Dimmer E, Huntley R, Barrell D, O'Donovan C, Apweiler R: **QuickGO: a web-based tool for Gene Ontology searching.** *Bioinformatics* 2009, **25**(22):3045–3046.

22. Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, Bateman A, Bernard T, Binns D, Bork P, Burge S, de Castro E, Coggill P, Corbett M, Das U, Daugherty L, Duquenne L, Finn RD, Fraser M, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, McMenamin C, Mi H, Mutowo-Muellenet P, Mulder N, Natale D, Orengo C, Pesseat S, Punta M, Quinn AF, Rivoire C, Sangrador-Vegas A, Selengut JD, Sigrist CJ, Scheremetjew M, Tate J, Thimmajanarthanan M, Thomas PD, Wu CH, Yeats C, Yong SY: **InterPro in 2011: new developments in the family and domain prediction database.** *Nucleic Acids Res* 2012, **40**(Database issue):D306–D312.

23. Bresso E, Benabderrahmane S, Smail-Tabbone M, Marchetti G, Karaboga AS, Souchet M, Napoli A, Devignes MD: **Use of domain knowledge for dimension reduction - application to mining of drug side effects.** In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval*. SciTePress Digital Library; 2011:271–276.

24. **Medical Dictionary for Regulatory Activities** [http://www.meddramsso.com]

25. Szathmary L: **Symbolic data mining methods with the Coron platform.** *PhD Thesis in Computer Science*, Univ. Henri Poincaré – Nancy 1, France 2006.

26. **Coron** [http://coron.loria.fr]

27. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH, Witten IH: **The WEKA data mining software: an update.** *SIGKDD Explorations* 2009, **11:**10–18.

28. Muggleton S, Srinivasan A, King RD, Sternberg MJE: **Biochemical knowledge discovery using inductive logic programming.** In *Discovey Science*, *Volume 1532 of* Lecture Notes in Computer Science. Edited by Arikawa S, Motoda H, Springer Berlin Heidelberg 1998:326–341.

29. Page D, Craven M: **Biological applications of multi-relational data mining.** *SIGKDD Explorations* 2003, **5:**69–79.

30. Santos JC, Nassif H, Page D, Muggleton SH, Sternberg MJ: **Automated identification of protein-ligand interaction features using inductive logic programming: A hexose binding case study.** *BMC Bioinformatics* 2012, **13:**162.

31. Muggleton S: **Inductive logic programming.** *New Generat Comput* 1991, **8**(4):295–318.

32. **The Aleph Manual** [http://www.cs.ox.ac.uk/activities/machlearn/Aleph/aleph.html]

33. Bresso E, Grisoni R, Devignes MD, Napoli A, Smail-Tabbone M: **Formal concept analysis for the interpretation of relational learning applied on 3D protein-binding sites.** In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval*. SciTePress Digital Library; 2012:111–120.

34. **KNIME** [http://www.knime.org]

35. Napoli A: **A smooth introduction to symbolic methods for knowledge discovery.** In *Handbook of Categorization in Cognitive Science.* Edited by Cohen H, Lefebvre C. Elsevier, Amsterdam; 2005:913–933.

36. Dolcino M, Cozzani E, Riva S, Parodi A, Tinazzi E, Lunardi C, Puccetti A: **Gene expression profiling in dermatitis herpetiformis skin lesions.** *Clin Dev Immunol* 2012, **2012:**198956.

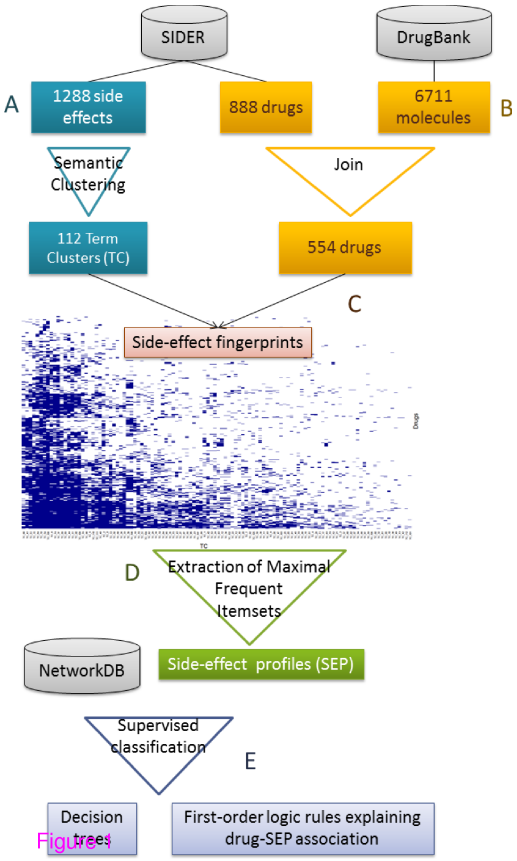

SIDER
DrugBank
A
1288 side effects
888 drugs
6711 molecules
B
Semantic Clustering
Join
112 Term Clusters (TC)
554 drugs
C
Side-effect fingerprints
D
Extraction of Maximal Frequent Itemsets
NetworkDB
Side-effect profiles (SEP)
Supervised classification
E
Decision trees
First-order logic rules explaining drug-SEP association

Figure 1

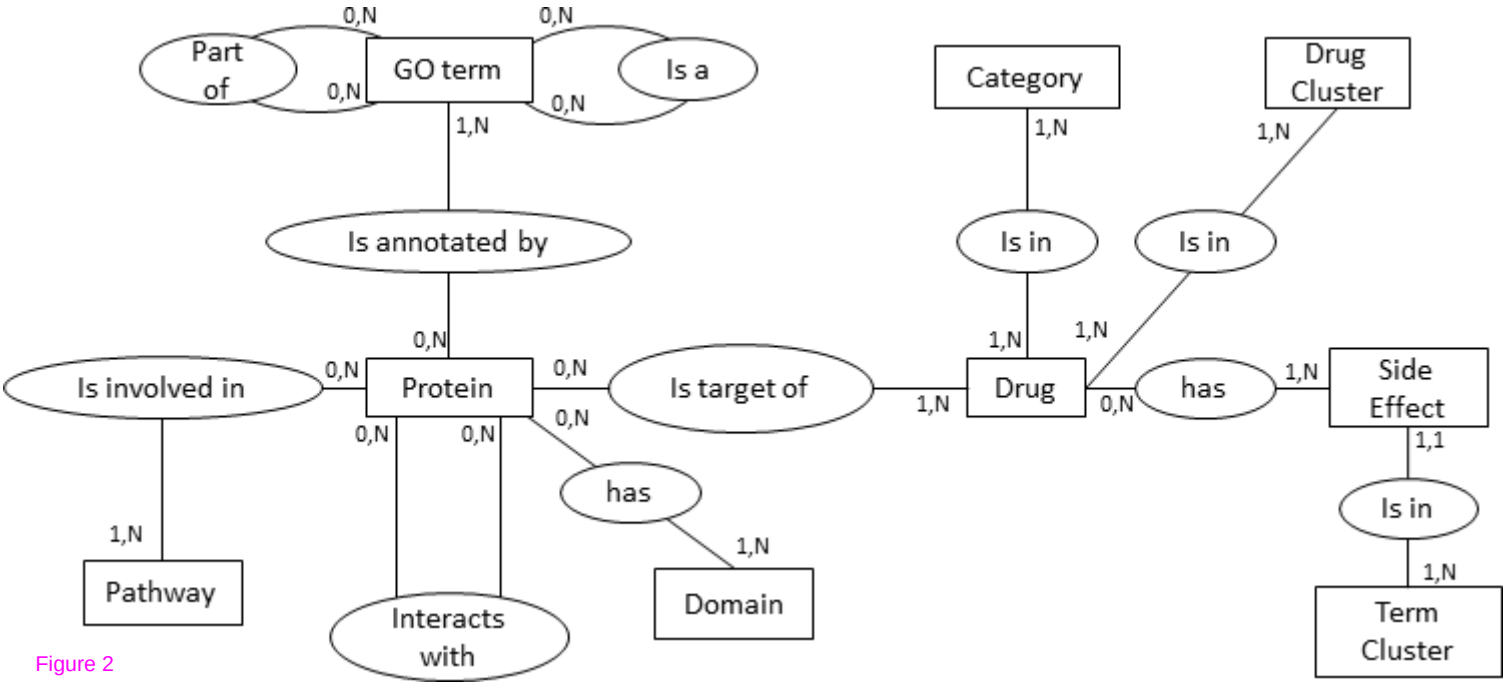
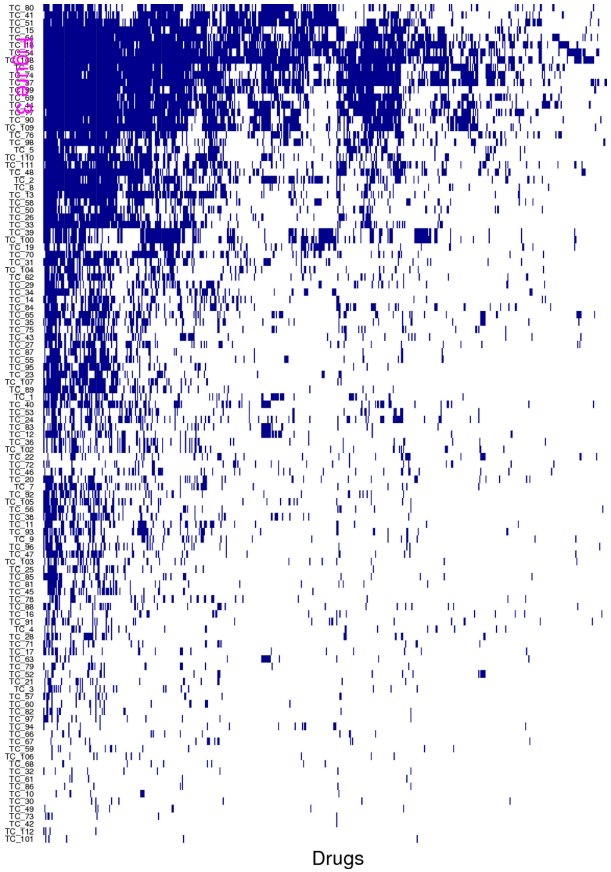

Part of
GO term
Is a
0,N
0,N
0,N
0,N
1,N
Is annotated by
0,N
Is involved in
Protein
Is target of
Drug
has
Side Effect
Category
Drug Cluster
Is in
Is in
1,N
1,N
1,N
1,N
0,N
1,N
1,N
1,N
0,N
0,N
0,N
0,N
1,N
1,N
1,N
1,1
1,N
Pathway
Interacts with
has
Domain
Is in
Term Cluster

Figure 2


Drugs

5.5 Conclusion

J'ai présenté dans ce chapitre l'utilisation de NetworkDB pour la compréhension des effets secondaires de médicaments. Afin de limiter le trop grand nombre d'effets secondaire et la redondance existant entre les termes, les effets secondaires ont été regroupés en clusters de termes en utilisant une mesure de similarité sémantique.

L'association entre les médicaments et ces TC permet de définir des empreintes d'effets secondaires. Ces empreintes peuvent être visualisées et explorées grâce à l'application HPfingerprint. Il est ainsi possible de savoir quels sont les TC qui sont spécifiques à une molécule. Afin d'avoir une idée des TC associés à une nouvelle molécule, il est possible d'utiliser une fonction utilisant les sous-structures de la molécule pour rechercher le médicament de la base de données ayant la structure la plus similaire. Afin de comprendre pourquoi certains médicaments étaient retirés de marché par l'industrie pharmaceutique, ces empreintes ont été utilisées pour explorer un jeu de molécules retirées du marché. J'ai tenté d'extraire par la méthode des arbres de décisions pour extraire les TC partagés et spécifiques d'une liste de 14 molécules. Cette étude a permis de mettre en évidence le rôle important du nombre d'exemples à utiliser pour les algorithme de fouille de données ainsi que celui des descripteurs à utiliser. En effet, la méthode utilisée pour associer médicaments et effets secondaires (au moins un effet secondaire du TC pour faire l'association) est beaucoup trop laxiste et a pour effet de générer du bruit empêchant l'émergence de résultats significatifs.

L'étude des mécanismes sous-jacent aux effets secondaires permet de mieux les comprendre et ainsi limiter le développement de médicaments susceptible de produire des effets secondaires. Les effets secondaires étant rarement isolés (un molécule ne provoque que rarement un seul effet indésirable) j'ai construit des profils d'effets secondaire, définis comme les plus grand groupes de TC partagés par un nombre significatif de médicaments. Cette étape ayant pour base un matrice binaire de type objets$\times$attributs, j'ai mis en place une méthodologie d'association entre TC et médicaments diminuant le bruit observé précédemment et permettant d'utiliser des algorithmes de fouille de données. Les molécules présentant les profils ainsi construits, ont ensuite été caractérisées par programmation logique inductive en utilisant les données contenues dans NetworkDB. Les résultats obtenus par PLI montrent que l'ensemble des informations contenus dans NetworkDB et concernant les molécules ainsi que leurs cibles sont utiles pour mieux comprendre les profils d'effets secondaires. La construction de NetworkDB et son exploitation par des méthodes de fouille de données permettent donc la compréhension de phénomènes complexes tels que les mécanismes liés à l'apparition des effets secondaires de médicaments.

Une perspective intéressante de ce travail est l'étude des effets combinés de plusieurs médicaments pris simultanément. Cela peut se faire en prenant en compte les données d'alertes collectées dans les systèmes de surveillance nationaux.

Chapitre 6

Conclusion - Perspectives

6.1 Conclusion

6.1.1 Résumé des principales contributions

Au cours de cette thèse et afin de répondre au problème de l'exploitation insuffisante des données sur les réseaux biologiques dans la compréhension des phénotypes complexes associés aux maladies génétiques ou des effets secondaires des médicaments, j'ai proposé une approche intégrative, guidée par les connaissances du domaine, et mettant en œuvre des techniques informatiques de gestion de données, de visualisation de graphes et de fouille de données. Cette approche a nécessité de mettre en place un entrepôt de données intégrant des informations sur les protéines et les molécules, puis à exploiter le contenu de cet entrepôt de différentes manières. La première façon d'exploiter les données de l'entrepôt consiste à extraire les informations pertinentes pour l'utilisateur et à les afficher sous la forme de graphes interactifs, permettant ainsi une meilleure compréhension du problème étudié. L'analyse globale des données de NetworkDB m'a conduit à utiliser des méthodes de fouille de données pour mettre en œuvre un processus de découvertes de nouvelles connaissances.

6.1.2 NetworkDB en tant que ressource intégrée

Le cœur de ce travail a consisté à construire un entrepôt de données, NetworkDB. Cet entrepôt repose sur un modèle de données intégrant les protéines, leurs annotations fonctionnelles, leurs interactions les unes avec les autres, leurs domaines ainsi que les réseaux biologiques dans lesquelles elles interviennent. Afin de répondre aux problèmes posés, ce modèle a été étendu avec les entités appropriées. Le peuplement de l'entrepôt se fait grâce à une série de tâches MODIM réutilisables qui permettent d'intégrer les données provenant de sources variées telles que UniProt, KEGG PATHWAY ou encore IntAct. Le contenu de l'entrepôt peut ensuite être visualisé soit à l'aide de requêtes SQL, soit grâce à une interface web qui permet d'exporter les données sous un format compatible avec l'outil Ondex qui permet la visualisation du réseau sous forme de graphe.

6.1.3 Étude de l'étiologie de maladies génétiques

La première approche intégrative a pour but de faciliter l'accès aux réseaux biologiques pour l'étude de l'étiologie des maladies génétiques, notamment celles qui entrainent une déficience intellectuelle. L'identification d'une mutation dans un gène ne permet pas toujours de comprendre les mécanismes associés à la maladie. Ainsi, à partir de deux exemples fournis par les généticiens du CHU de Nancy, j'ai défini une méthodologie permettant d'explorer les mécanismes d'une maladie en exploitant les données connues sur une maladie génétique présentant des symptômes similaires.

L'analyse de l'ensemble des gènes de déficiences intellectuelles liées à l'X nécessite quant-à elle de recourir à l'entrepôt NetworkDB, lequel a été peuplé à partir de la liste de gènes connus pour provoquer ces maladies. J'ai ensuite utilisé la mesure de similarité sémantique IntelliGO pour réaliser une classification fonctionnelle de ces gènes, puis j'ai étudié l'enrichissement en termes GO des groupes de gènes définis précédemment. Ceci m'a permis de montrer que la pré-classification des gènes DILX est essentielle pour obtenir des enrichissements significatifs. Ensuite, en utilisant ONDEX, j'ai identifié des sous-réseaux connectés soit par des interactants communs, soit par des réseaux biologiques. Ces sous-réseaux correspondent bien aux quatre grands types de fonctions connues pour être associées aux gènes DILX. Notre approche fournit des éléments complémentaires sur chaque groupe fonctionnel et donne accès aux réseaux biologiques connectant plusieurs gènes.

6.1.4 Effets secondaires de médicaments

La seconde application de notre approche intégrative a pour but de mieux comprendre les mécanismes sous-jacents aux effets secondaires des médicaments en intégrant dans NetworkDB les médicaments (issus de DrugBank), leurs effets secondaires (SIDER) et différentes propriétés de leurs cibles (termes GO, domaines, ...). Afin de limiter le trop grand nombre d'effets secondaires et la redondance existant entre les termes, les effets secondaires ont été groupés en clusters de termes en utilisant une mesure de similarité sémantique.

L'association entre les médicaments et ces TC permet de définir des empreintes d'effets secondaires. Ces empreintes sont explorables grâce à l'interface HPfingreprint. Il est ainsi possible de savoir quels TC sont spécifiques à une molécule. Afin d'avoir une idée des TC associés à une nouvelle molécule, il est possible d'utiliser une fonction utilisant les sous-structures de la molécule pour rechercher le médicament de la base de données ayant la structure la plus similaire.

Ces empreintes ont été utilisées pour explorer les raisons qui ont poussé l'industrie pharmaceutique à retirer certains médicaments du marché. Pour cela, j'ai utilisé des méthodes de fouilles de données pour extraire les TC partagés et spécifiques d'un échantillon de ces molécules. Malheureusement, le faible nombre de médicaments à étudier est limitant pour les méthodes de fouille de données. Cependant, cela nous a conduit à aller plus loin notamment en modifiant la façon dont les TC et les médicaments sont associés.

Afin d'anticiper le retrait de médicaments, la compréhension des mécanismes sous-jacents aux effets secondaires est une étape importante. Les effets secondaires étant rarement isolés, j'ai défini une méthodologie permettant d'extraire des profils d'effets secondaires partagés par un nombre im-

portant de médicaments. Nous avons défini les profils d'effets secondaires comme les plus grands groupes de TC partagés par au moins 100 médicaments. Cette méthodologie a nécessité de mettre en place une heuristique afin d'associer de manière robuste médicaments et TC. Ces profils ont ensuite été caractérisés par programmation logique inductive en utilisant les données contenues dans NetworkDB. Les résultats obtenus par PLI montrent que l'ensemble des informations contenues dans NetworkDB concernant les molécules et leurs cibles sont utiles pour mieux comprendre différents profils d'effets secondaires. De plus, certaines des règles obtenues ont pu être confirmées par des données issues de la littérature prouvant ainsi la pertinence de ces règles. Ainsi, l'approche mise en œuvre ici permet de mieux comprendre des phénomènes complexes tels que les mécanismes qui entrainent l'apparition des effets secondaires des médicaments.

6.2 Perspectives

6.2.1 Automatisation du rafraichissement de NetworkDB

Une amélioration possible de NetworkDB serait de permettre les mises à jour automatiques de l'entrepôt. En effet, les données biologiques sont produites en continu et les bases de données publiques sont continuellement mises à jour. Ainsi, si l'on veut garder des données à jour, il est important de réaliser des mises à jour régulières. Même si le programme MODIM permet la mise jour de l'entrepôt, il peut être nécessaire d'actualiser les tâches existantes en fonction des modifications survenues dans les sources de données. De plus, il faut ensuite effectuer les tâches dans un ordre précis afin de collecter correctement l'ensemble des données. Cette automatisation nécessiterait l'exécution et l'enchainement périodique des tâches et permettrait aux utilisateurs non experts de MODIM de maintenir l'entrepôt de données à jour.

6.2.2 Facilitation du peuplement de NetworkDB : utilisation des données du projet "Linked Open Data"

L'utilisation des données du projet "Linked Open Data[29]" pourrait permettre de faciliter le peuplement de NetworkDB. Ce projet a pour but de rassembler les données disponibles sur le web et de les convertir au format RDF (Bizer *et al.*, 2009). Parmi les données collectées et formatées, on trouve à côté des données gouvernementales, géographiques et autre, de nombreuses données biologiques (Figure6.1). Les bases de données SIDER, UniProt, Gene Ontology, InterPro ou encore KEGG PATHWAY sont ainsi disponibles au format RDF. Des initiatives plus restreintes, consacrées au Sciences de la Vie sont en train de prendre de l'essor, comme par exemple Bio2RDF[30]. La mise à disposition de nombreuses données biologiques au format RDF est un atout pour peupler facilement des entrepôts de données biologiques. La constitution de tels entrepôts reste nécessaire cependant, notamment pour faire coexister les données RDF et d'autres données encore non disponible dans ce format, comme par exemple des données expérimentales. Les triplets RDF se prêtent assez naturellement à une représentation en graphe et sont compatibles avec l'application développée dans

29. http://linkeddata.org
30. https://github.com/bio2rdf

le chapitre 4 de cette thèse. Concernant la fouille de donnés par arbres de décision ou par programmation logique inductive, l'utilisation directe de données représentées sous forme de triplets RDF représente une perspective intéressante qu'il faudrait expérimenter.

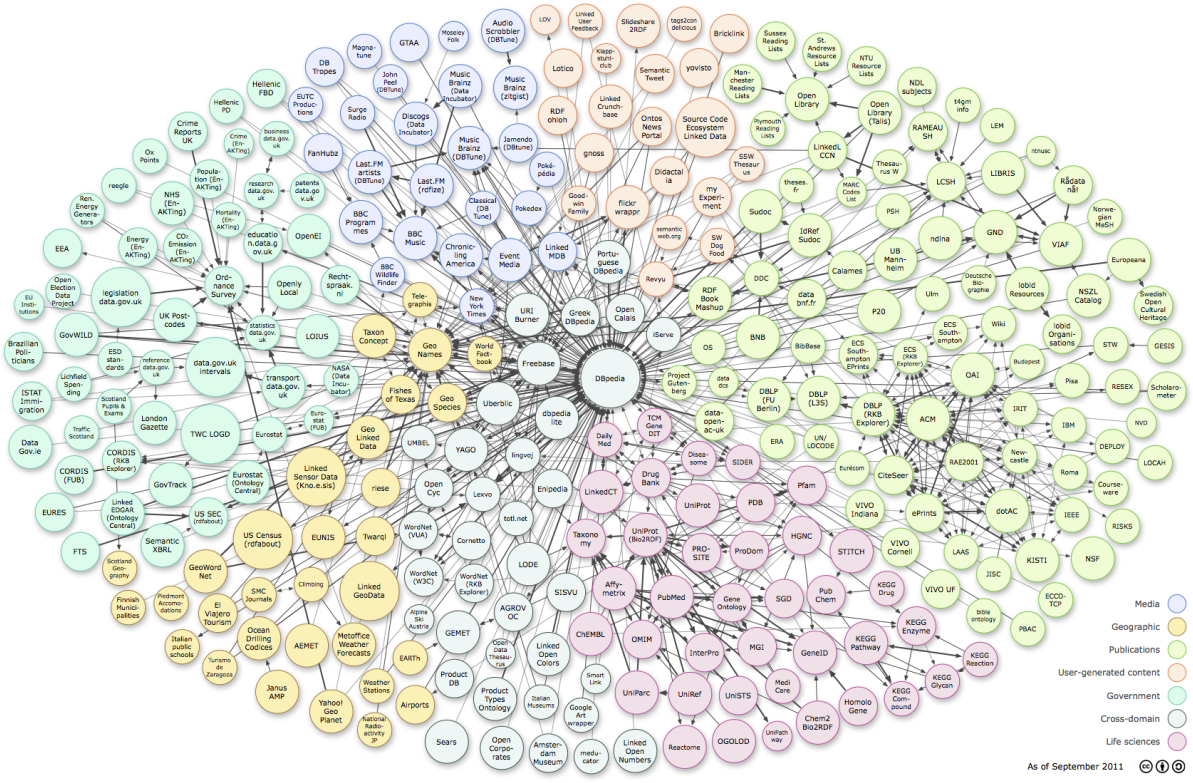


FIGURE 6.1 – Diagramme des ressources intégrées dans le projet "Linked Open Data". Les ressources liées aux sciences de la vie sont en rose. Le diagramme (daté de septembre 2011) est issu de http://lod-cloud.net.

6.2.3 Automatisation de la recherche de propriétés communes aux gènes provoquant des déficiences intellectuelles liées à l'X

L'une des principales limites actuelles de la méthodologie pour étudier les réseaux de gènes des déficiences intellectuelles liées au chromosome X est l'extraction manuelle des sous-réseaux. En effet, même si cette extraction est effectuée par l'utilisateur expert, cela nécessite d'analyser chaque réseau biologique et chaque gène pour essayer de construire des groupes ayant une fonction commune.

Une solution possible serait une approche itérative basée sur l'enrichissement en terme GO. A partir de deux gènes reliés par un réseau biologique, on peut calculer un enrichissement fonctionnel. Si cet enrichissement est significatif, alors on ajoute un nouveau gène relié par un réseau biologique à l'un des deux autres gènes. On peut ensuite calculer un enrichissement fonctionnel pour ce groupe de trois gènes et si on obtient un résultat significatif, on continue d'agrandir le groupe. L'arrêt de l'agrandissement se fera alors lorsqu'il n'y aura plus de gène à ajouter ou alors parce que il n'y a plus d'enrichissement significatif quand on ajoute un gène. Un des intérêts de cette méthodologie est

de ne pas construire une partition des gènes. En effet, cette solution permet à un gène d'appartenir à plusieurs groupes.

Une autre solution possible serait d'utiliser des méthodes de clustering de graphe. L'approche proposée par Becker *et al.* (2012) est particulièrement intéressante car elle permet d'extraire à partir d'un réseau d'interaction protéine-protéine des sous-groupes formant des modules fonctionnels. La particularité de cette approche est d'extraire des sous-groupes chevauchants, permettant ainsi de prendre en compte la capacité des protéines à avoir plusieurs fonctions. Cependant, cette méthode n'est pas adaptée aux graphe étendus puisqu'elle ne prend pas en compte les différents statuts des nœuds. Ainsi, entre des méthodes guidées par la sémantique des nœuds et des arrêtes, et des méthodes basées uniquement sur la structure des graphes, une place existe sans doute pour des méthodes hybrides.

De la même façon que les profils d'effets secondaires de médicaments ont été caractérisés, il est possible d'utiliser la programmation logique inductive afin de prendre en compte à la fois les propriétés des gènes DILX et les propriétés de leurs interactants. Le jeu de données des gènes DILX, qui pourrait être étendu à l'ensemble des gènes des DI, constitue certainement un bon ensemble pour explorer ce genre de méthodes.

6.2.4 Sélection des profils d'effets secondaires

Une extension à la caractérisation des profils d'effets secondaires serait de modifier la façon dont les profils sont sélectionnés. En effet, actuellement ce sont les profils les plus fréquents qui ont été caractérisés. Ainsi, certaines molécules ont été prises en compte plusieurs fois alors que d'autres n'appartiennent à aucun profil d'effets secondaires sélectionné. La solution que je propose serait de sélectionner les profils qui permettent de couvrir l'ensemble des molécules disponibles pour l'apprentissage, cela serait réalisable notamment en diminuant le seuil utilisé pour extraire les motifs fréquents maximaux.

6.2.5 Prédiction des profils d'effets secondaires pour une nouvelle molécule

Actuellement, on prédit comme positive une molécule qui satisfait à au moins une règle issue de l'apprentissage. Cette étape pourrait être remplacée par une étape de classification supervisée utilisant l'ensemble des règles de l'ensemble des molécules de référence. Il faudrait partir de la matrice binaire complète molécules$\times$règles pour "apprendre" quels sont les ensembles de règles qu'une molécule doit vérifier pour être annotée par un profil d'effet secondaires donné. Cette méthodologie de recherche de modèles globaux à partir de motifs locaux correspond au cadre conceptuel LeGo définie Knobbe *et al.* (2008).

6.3 Conclusion générale

Les travaux développés durant cette thèse illustrent différentes façons d'aller des données vers les connaissances. Les méthodes expérimentales biologiques produisent de plus en plus de don-

nées et il devient incontournable de traiter ces données de telle sorte qu'elles puissent être analysées en vue d'en extraire des connaissances utiles et nouvelles. L'intégration des données, soit dans des entrepôts soit dans le "cloud", comme pour le projet "Linked Open Data" ou Bio2RDF, est la première étape. La découverte de connaissances peut ensuite se faire à partir des données ainsi structurées, en utilisant notamment des méthodes de fouille de données. Une fois validées, ces connaissances peuvent à leur tour être stockées dans un entrepôt et servir de base pour d'autres explorations des données. Dans tout ce processus, l'outil informatique doit se faire l'allié de l'expert biologiste ou clinicien en réalisant pour lui les tâches fastidieuses devenues impossibles à accomplir manuellement. Comme illustré dans cette thèse, des dispositifs interactifs de visualisation et d'apprentissage automatique doivent aussi permettre à l'expert de se consacrer à son rôle essentiel et irremplaçable d'interprétation et de validation des connaissances à partir des données.

Publications

Articles

- Bonnet, C., Ali Khan, A., Bresso, E., Vigouroux, C., Béri, M., Lejczak, S., Deemer, B., Andrieux, J., Philippe, C., Moncla, A., Giurgea, I., Devignes, M.D., Leheup, B., Jonveaux, P. (2013). Extended spectrum of MBD5 mutations in neurodevelopmental disorders. *Eur. J. Hum. Genet.*
- Bresso E, Grisoni R, Devignes M-D, Napoli A Smaïl-Tabbone M. ILP Characterization of 3D Protein-Binding Sites and FCA-based Interpretation. *Communications in Computer and Information Science* (Sous Presse).
- Bresso, E., Grisoni, R., Marchetti,G., Karaboga, A.S., Souchet, M., Devignes M.D. et Smaïl-Tabbone M. (2013) Integrative relational Machine-Learning Approach for Understanding Drug Side-Effect Profiles. *BMC Bioinformatics*, **14**, 207.

Conférences internationales avec comité de relecture

- Bresso E, Benabderrahmane S, Smail-Tabbone M, Marchetti G, Karaboga A S, Souchet M, Napoli A et Devignes M.D. Use of domain knowledge for dimension reduction : application to mining of drug side effects. *Knowledge Discovery and Information Retrieval (KDIR 2011)*, Paris, 2011.
- Bresso E., Grisoni R., Devignes M.D., Napoli A. et Smaïl-Tabbone M. Formal Concept Analysis for the Interpretation of Relational Learning applied on 3D Protein-Binding Sites. *Knowledge Discovery and Information Retrieval (KDIR 2012)*, Barcelone, 2012.

Posters

- Bresso E, Smaïl-Tabbone, M. et Devignes M.D. Définition de patchs 3D et fouille relationnelle pour la caractérisation et la prédiction des sites d'interactions protéine-protéine. *Journées Ouvertes de Biologie, Informatique et Mathématiques (JOBIM 2010)*, Montpellier, 2010.
- Ndiaye, B., Bresso, E., Smaïl-Tabbone, M., Souchet, M. et Devignes, M.D. Modim : model driven data integration for mining. *Journées Ouvertes de Biologie, Informatique et Mathématiques (JOBIM 2011)*, Paris, 2011.

- Bresso, E., Benabderrahmane, S., Smaïl-Tabbone, M., Marchetti, G., Karaboga, A.S., Souchet, M., Napoli, A. et Devignes M.D. Use of domain knowledge for dimension reduction : application to drug side-effect. *BeNeLux Bioinformatics Conference (BBC 2011)*, Luxembourg, 2011.
- R. Grisoni, E. Bresso, M.D. Devignes et M. Smail-Tabbone Méthodologie et outils pour l'extraction de connaissances par Programmation Logique Inductive (PLI) *Conférence Francophone sur l'Extraction et la Gestion des Connaissances (EGC'2013)*, Toulouse, 2013.

Annexe A

Mise en évidence du rôle de *MBD5* dans des déficiences intellectuelles

European Journal of Human Genetics (2013), 1–5

www.nature.com/ejhg

SHORT REPORT

Extended spectrum of *MBD5* mutations in neurodevelopmental disorders

Céline Bonnet[1,8], Asma Ali Khan[1,8], Emmanuel Bresso[2], Charlène Vigouroux[1], Mylène Béri[1], Sarah Lejczak[1], Bénédicte Deemer[3], Joris Andrieux[4], Christophe Philippe[1], Anne Moncla[5], Irina Giurgea[6], Marie-Dominique Devignes[2], Bruno Leheup[1,7] and Philippe Jonveaux*[,1]

Intellectual disability (ID) is a clinical sign reflecting diverse neurodevelopmental disorders that are genetically and phenotypically heterogeneous. Just recently, partial or complete deletion of methyl-CpG-binding domain 5 (*MBD5)* gene has been implicated as causative in the phenotype associated with 2q23.1 microdeletion syndrome. In the course of systematic whole-genome screening of individuals with unexplained ID by array-based comparative genomic hybridization, we identified *de novo* intragenic deletions of *MBD5* in three patients leading, as previously documented, to haploinsufficiency of *MBD5*. In addition, we described a patient with an unreported de novo *MBD5* intragenic duplication. Reverse transcriptase-PCR and sequencing analyses showed the presence of numerous aberrant transcripts leading to premature termination codon. To further elucidate the involvement of *MBD5* in ID, we sequenced ten coding, five non-coding exons and an evolutionary conserved region in intron 2, in a selected cohort of 78 subjects with a phenotype reminiscent of 2q23.1 microdeletion syndrome. Besides variants most often inherited from an healthy parent, we identified for the first time a *de novo* nonsense mutation associated with a much more damaging phenotype. Taken together, these results extend the mutation spectrum in *MBD5* gene and contribute to refine the associated phenotype of neurodevelopmental disorder.
European Journal of Human Genetics advance online publication, 20 February 2013; doi:10.1038/ejhg.2013.22

Keywords: *MBD5*; nonsense mutation; intragenic duplication; intellectual disability

INTRODUCTION

Methyl-CpG-binding domain 5 (MBD5) protein (OMIM *611472) is a member of the MBD protein family in which MECP2 (OMIM *300005) is involved in Rett syndrome, a prototypical neurodevelopmental disorder. *MBD5* contains five non-coding exons at its 5′-end, followed by 10 coding exons. Two isoforms have been described,[1] the longer one contains 1494 amino acids and is encoded by exons 6–15, the second one contains 851 amino acids and is encoded by exons 6–9. Functional studies suggested that MBD5 is likely to contribute to the formation or function of heterochromatin.[1] Isoform 1 of *MBD5* is highly expressed in brain and testis and isoform 2 is highly expressed in oocytes, which suggest a possible role in cerebral functions and in epigenetic reprogramming after fertilization. Recently, deletions encompassing *MBD5*, as well as intragenic *MBD5* deletions have been identified in individuals with a phenotype of intellectual disability (ID), seizures, significant speech impairment, and behavioral problems.[2–8] In this study, we used pangenomic array-comparative genomic hybridization (array-CGH) and capillary sequencing of *MBD5* gene to investigate DNAs from patients with unexplained ID. We further extend the mutational spectrum of *MBD5* with damaging intragenic duplication and nonsense mutation associated with a clinical spectrum of neurodevelopmental disorder.

SUBJECTS AND METHODS

Ascertainment of the patients

Patients with an unexplained developmental delay/ID as isolated symptom or in association with behavioral problems took part in a clinical diagnostic testing for genomic imbalance using array-CGH, following initial testing for karyotype (results normal), thanks to the national array-CGH network funded by the French Ministry of Health. To further elucidate the involvement of *MBD5* point mutations, we collected a clinically defined cohort of 78 individuals with moderate to severe ID without a known genetic cause (genomic copy number variants larger than 200 kb were previously excluded) and with significant clinical overlap with 2q23.1 deletion syndrome, reminiscent of Angelman-like phenotype or Smith–Magenis-like syndrome. More specifically, we included patients with ID, severe speech impairment, seizures, behavioral problems and in particular with autistic-like features. Informed consents were available for all tested patients.

Array-CGH analysis

Microarray-CGH analysis was carried out using 44K or 105K-oligonucleotide array (Agilent Technologies, Santa Clara, CA, USA) as previously described.[9] The array was analyzed with the Agilent scanner and the Feature Extraction software (v9.5.3.1; Agilent Technologies). A graphical overview was obtained using the CGH analytics software (v3.5.14; Agilent Technologies).

[1]Laboratoire de Génétique, EA 4368, Université de Lorraine, Centre Hospitalier Universitaire de Nancy, Vandoeuvre les Nancy, France; [2]LORIA UMR7503, CNRS, INRIA, Nancy-Université, BP239, 54506 Vandoeuvre les Nancy, France; [3]Service de Pédiatrie et Génétique, Hôpital Nord, Amiens, France; [4]Laboratoire de Génétique Médicale, Hôpital Jeanne de Flandre, CHU de Lille, Lille, France; [5]Département de Génétique Médicale, Hôpital d'Enfants de la Timone, Marseille, France; [6]Service de Biochimie-Génétique, APHP, Groupe hospitalier Henri Mondor, Créteil, France; [7]Service de Médecine Infantile III et génétique clinique, Hôpital d'enfants, Centre Hospitalier Universitaire de Nancy, Vandoeuvre les Nancy, France
[8]These authors contributed equally to the work.
*Correspondence: Professor P Jonveaux, Laboratoire de Génétique, EA 4368, Université de Lorraine, CHU de Nancy, rue de Morvan, Vandoeuvre les Nancy 54511, France. Tel: +33 3 83 15 37 71; Fax: +33 3 83 15 37 72; E-mail: p.jonveaux@chu-nancy.fr
Received 16 July 2012; revised 27 December 2012; accepted 24 January 2013

Genomic quantitative PCR

Quantitative PCR (qPCR) was performed on genomic DNA, using an ABI PRISM 7500 Sequence Detection System (Applied Biosystems, Foster City, CA, USA). We designed primer sets in *MBD5* gene (all primer sequences used in this study are available on request). qPCR was carried out as previously described.[9] The *RPPH1* gene was selected as the control amplicon. Validation experiments demonstrated that amplification efficiency of the control and all target amplicons were approximately equal. All samples were run in triplicate. The dosage of each amplicon relative to *RPPH1* and normalized to control male DNA was determined using the $2^{-\Delta\Delta Ct}$ method.

Genomic sequencing

MBD5 ten coding, five non-coding exons (NM_018328) and one evolutionary conserved region in intron 2 were PCR amplified using standard procedures (available on request). PCR products were then purified and subjected to sequencing using BigDye Terminator kit (Applied Biosystems).

mRNA isolation, reverse transcriptase-qPCR

Total RNAs were isolated from PaxGen blood RNA tubes using RNeasy mini kit (Qiagen, Hilden, Germany). Family samples were collected for patients A and B (mother, father and brother). Male and female controls were collected for patients C, D and E. RNA was reverse transcribed through the use of random primers (Superscript, Invitrogen, Life technologies, Paisley, UK). Reverse transcriptase quantitative real-time PCR (RT-qPCR) was performed on an ABI PRISM 7500 Sequence Detection System (Applied Biosystems). We designed primer sets within *MBD5* (available on request). RT-qPCR was carried out in a total volume of 20 μl containing 10 μl of SYBR Green Master Mix (Applied Biosystems), 0.4 mM of each primer and 5 μl of complementary DNA (cDNA). Thermal cycling conditions were 95 °C for 20 s, followed by 40 cycles with 95 °C for 3 s and 60 °C for 30 s. The *ESD* and *ABL1* genes were selected as control amplicons. Validation experiments demonstrated that amplification efficiency of control and all target amplicons were approximately equal. All samples were run in triplicate. The dosage of each amplicon relative to *ESD* and *ABL1* and normalized to control male cDNA was determined using the $2^{-\Delta\Delta Ct}$ method.

cDNA sequencing

Primers were selected in *MBD5* exons. RT-PCR products were electrophoresed on agarose gels, purified with NucleoSpin Extract II kit (Macherey–Nagel, SARL, Düren, Germany) and sequenced using BigDye Terminator kit (Applied Biosystems).

RESULTS

Clinical reports

The clinical characteristics of individuals with MBD5-specific disruption are summarized in Table 1.

Patients A and B are monozygotic twin sisters. The father and the two first siblings are healthy. The mother was treated for epilepsy but treatment was interrupted during pregnancies. They were born prematurely without fetal distress. *Z*-scores of birth weight and length were at −1, and head circumference was in the normal range. They were noted to have global developmental delay. Patient A sat independently at 16 months and walked at 2 years 6 months, patient B sat independently at 17 months and walked at 3 years. Both spoke only single words and presented with stereotypies and autistic features. A brain MRI was normal. At the age of 3 years 6 months, both heights were at −3 SD, whereas weights and head circumferences were in the normal range. There is a isolated nostril anteversion on facial examination (Figure 1a).

Patient C,a male proband, was born following an uncomplicated, full-term pregnancy. Parents were non-consanguineous and healthy. Family history is otherwise unremarkable. Neonatal adaptation was normal. Birth weight (3300 g), birth length (49 cm) and head circumference (37 cm) were within the normal range. He presented with hypospadias and developed multiple bronchiolitis. He had an inguinal hernia repaired. He was noted to have global developmental delay. He walked at the age of 22 months, and language milestones were delayed. At the age of 4 years, height was 99 cm (median), weight 15 kg (median) and head circumference 52 cm (+1 SD). He presented with stereotypies. No specific dysmorphic facial features were observed. He had fifth finger clinodactyly. Although his intelligence had not been formally evaluated, his ID was estimated to be mild to moderate.

Patient D is the only child of healthy non-consanguineous parents. Pregnancy was uneventful to the exception of hemorrhage related to partial placental detachment at 3 months of gestation. She was born at term with normal growth parameters. Initial developmental milestones were reported normal. She walked at 19 months. The first words were pronounced at 13 months. Between 24 and 30 months of age, regression of language skills occurred with concomitant regression of response to social overture. She gradually developed problematic behavior, with stereotyped movements of the arms, and periods of hyperactivity and attention deficit. She was seen at the neuropediatric department at 2 years and 4 months and at 3 years and 6 months. There was no motor deficit. Slight symptoms of cerebellar syndrome were noted with oral dyspraxia. She was also seen at the outpatient genetics clinic at 3 years and 11 months. Growth parameters were within the normal range and clinical examination showed a round face, nostril anteversion and down-turned corners of the mouth (Figure 1a).

Patient E is a 10-year-old boy first seen at the age of 14 months because of developmental delay. He is the second of three children of healthy unconsanguineous parents. Pregnancy was reported as normal. He was born at 40 weeks by cesarean section because of placenta praevia. Neonatal adaptation was normal. Birth weight (2560 g), birth length (48 cm) and head circumference (34.5 cm) were in the low normal range. Since the first days of life, parents reported feeding difficulties. The boy developed opisthotonos during the first months of life. Unmotivated laughter was also reported. When first seen at 14 months, sitting was unstable, hand movements were poor, and language was absent. Eye contact was reported as easy. Some jerky movements were described. Length, weight, and head circumference were at the 50th percentile. Craniofacial examination was not specific with slightly broad forehead. EEG and cerebral MRI were normal. A screen for metabolic abnormalities and methylation analysis for Angelman syndrome were normal. *UBE3A* gene analysis (Dr Moncla, Marseille) was normal. At the age of 2 years 7 months, there was still no verbal language. Hypotonia was severe without walking. He presented with generalized tonico–clonic seizures at the age of 4 years. Treatment with valproate was initiated. At the age of 8 years, clonus of both legs were reported, which was associated with tongue and mouth clonus. EEG reported focal spikes and spike-wave complexes in the frontal and temporal left area, leading to the diagnosis of partial epilepsy. He was last seen at the age of 10 years 3 months. Height was at 147.5 cm (+2 SD), weight and head circumference were in the normal range. He stood independently for a short period of time but did not walk. There was no verbal language. His parents reported him as happy with very frequent smiles. Craniofacial examination showed unspecific hypotonic characteristics with long face, open mouth and slightly everted lower lip. Ear lobules were large (Figure 1a).

Molecular investigations

Array-CGH analysis demonstrated, according to UCSC build 36/hg18 (Figures 1b and c): (i) In patients A and B, an interstitial deletion at 2q23.1: arr 2q23.1(148 447 496–148 515 776) × 1, with a minimal size

Table 1 Clinical characteristics of the present patients in comparison with previously reported patients with *MBD5*-specific disruption (*Talkowski *et al.*[10])

	Previous cases % (frequency)*	*Present cases*				
Patients		*Patient A* *Female*	*Patient B* *Female*	*Patient C* *Male*	*Patient D* *Female*	*Patient E* *Male*
Current age (years)		$3^{6/12}$	$3^{6/12}$	4	4	$10^{3/12}$
Type and size of genomic anomaly	*MBD5*-specific disruption	Partial deletion 68 kb	Partial deletion 68 kb	Partial duplication 34 kb	Partial deletion 19 kb	Nucleotide substitution
Psychomotor retardation/ Intellectual disability	100% (14/14)	+	+	+	+	+
Language impairment	71.4% (5/7)	+	+	+	+	+
Autistic-like symptoms	100% (14/14)	+	+	−	+	+
Stereotypic repetitive behavior	60% (3/5)	+	+	+	+	−
Sleep disturbances	50% (3/6)	−	−	−	−	−
Short stature	40% (2/5)	+	+	−	−	−
Microcephaly	0% (0/5)	−	−	−	−	−
Seizures	85.7%(6/7)	−	−	−	−	+
Ataxia	0% (0/3)	−	−	−	+	−
Craniofacial features	66.6% (4/6)	+	+	−	+	+
Hand/foot anomalies	66.6% (4/6)	−	−	+	−	−
Urogenital abnormalities	0% (0/4)	−	−	+	−	−

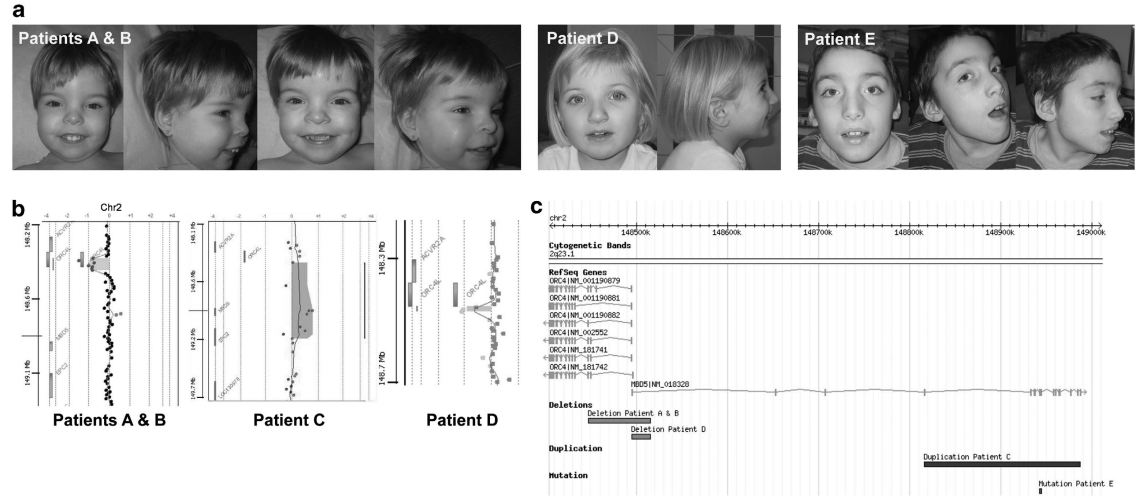


Figure 1 (**a**) Frontal and lateral view of patients A and B (3 years 6 months), patient D (3 years 11 months) and patient E (10 years), demonstrating a broad nasal bridge and hypoplastic nares. (**b**) 105K array-based CGH results showing the extent of *MBD5* intragenic deletion in patients A and B (7 probes) and for patient D (3 probes) and the extent of intragenic duplication for patient C (3 probes using 44K array-CGH) (**c**) Map of genomic alterations: deletions, duplication and nonsense mutation (snapshot of Database of Genomic Variants (http://projects.tcag.ca/variation/).

of 68 280 bp. The region includes the end of *ORC4* and the two first non-coding exons of *MBD5*. (ii) In patient C, an interstitial duplication at 2q23.1: arr 2q23.1(148 944 718–148 979 574) × 3 with a minimal size of 34 856 bp. This duplication affects only *MBD5*, the minimal duplicated region including four exons (5–8) and the maximal region including 11 exons (nc5–10). (iii) In patient D, an interstitial deletion at 2q23.1: arr 2q23.1(148 496 551–148 515 776) × 1 with a minimal size of 19 225 bp, including the end of *ORC4* and the two first non-coding exons of *MBD5*. This region had never been described as a copy number polymorphism in the database of genomic variants (http://projects.tcag.ca/variation/?source=hg18). Except for polymorphic regions, no copy number alterations were observed in other chromosomes. Using qPCR analysis on genomic DNA from patients A, B, D and their respective parents, we confirmed the biological relationships and revealed that genomic imbalances arose *de novo*. For patient C, parental DNAs were not available. However, qPCR on his genomic DNA allowed determining more precisely the extent of the duplication from non-coding exon 5 to coding exon 10. We used Sanger sequencing to screen *MBD5* for point mutations in the selected cohort of 78 individuals with ID. We identified a nonsense mutation (c.440C>G (p.Ser147*); NM_018328.3) within coding exon 4 in patient E (Figure 2a). Analysis of parental DNA confirmed the biological relationships and *de novo* occurrence of the mutation. In this series of patients, we also detected nine variants in protein-coding exons, not annotated in dbSNP (build 137), three intronic variants, three synonymous variants and three missense variants. In evolutionary conserved region and non-coding exons, five different variations were found. Detailed

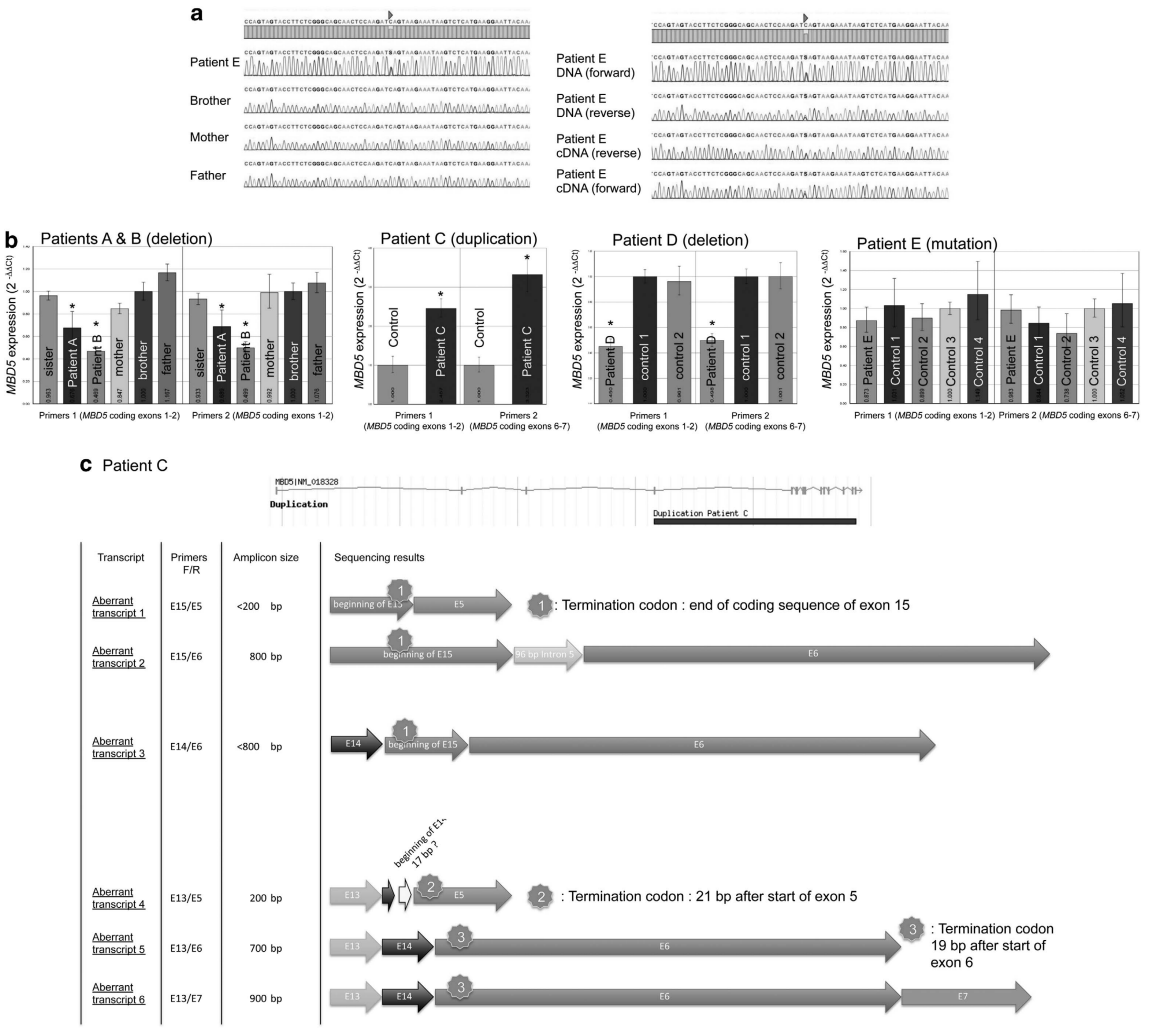


Figure 2 (**a**) Genomic sequencing results for patient E, his brother and parents showing (arrowhead) a *de novo* nonsense mutation in coding exon 4 (NM_018328.3:c.440C>G). Right panel: comparison of DNA and cDNA sequencing results in patient E showing that both normal and mutated alleles are expressed (arrowhead). (**b**) RT-qPCR results with primer set 1 (*MBD5*-coding exons 1–2) and primer set 2 (*MBD5*-coding exons 6–7): left and third panel, a decreased level of expression of MBD5 in patients A and B (compared with that healthy sister and parents), and in patient D (compared with that two controls), respectively, second panel, an increased level of expression for duplicated MBD5 exons in patient C (compared with that control), right panel, a normal level of expression of MBD5 in patient E (compared with that four controls). *indicates a significant difference. (**c**) Aberrant transcripts characterized by RT-PCR and sequencing analysis for patient C.

sequencing results are displayed in Supplementary Tables I and II. When parental material was available, we were able to show transmission from a healthy parent in all cases. RT-qPCR analysis showed (Figure 2b): (i) a notable reduction of *MBD5* expression for both sisters A and B and for patient D, (ii) a significantly increased level for duplicated MBD5 exons in patient C, and (iii) a normal level of expression for patient E. RT-PCR analysis in patient C, with forward primers in coding exons 8, 9 and 10 and reverse primer in exon nc5, coding exons 1 and 2 of *MBD5*, amplified different aberrant transcripts. Sequencing analysis of these fragments (Figure 2c) showed that all aberrant transcripts led to premature termination codon. For patient E, RT-PCR and sequencing analysis of exon 4 showed that both normal and mutated alleles were expressed (Figure 2a).

DISCUSSION

Recently, Talkowski *et al*[10] suggested a mixed model of deleterious, fully penetrant *MBD5* deletions causing a neurodevelopmental disorder associated with features of 2q23.1 microdeletion syndrome, and reduced penetrance missense variants that significantly increase risk for autism spectrum disorder. In our work, we identified five patients with *de novo MBD5*-specific disruption (for patient C, we are aware that parental DNA was not available to confirm *de novo* occurrence of the intragenic duplication) with clinical characteristics similar to previously reported patients[10] (Table 1), mainly with psychomotor retardation/ID, language impairment, and autistic-like symptoms. For patients A, B, C and D the phenotype is overlapping, less specific than patients with 2q23.1 microdeletion, which includes

more frequently microcephaly, small hands and feet, short stature, and broad-based ataxic gait. Developmental delay/ID is isolated in patient C, and associated with behavioral problems in patients A, B and D. Seizures were not observed in these four patients, at this time in development. At the opposite, the phenotype of patient E is much more damaging without walking and verbal speech at the age of 10 years.

Three patients (A, B and D) had a deletion including the last exons of *ORC4* and only the two first untranslated exons of the brain-expressed isoform 1 of *MBD5*. A similar deletion has been reported.[5,10] Expression level of *MBD5* mRNA in patients A and B was significantly reduced in comparison to their non-deleted parents, sister and brother. This result proves that heterozygous deletion of the two first non-coding exons of *MBD5* isoform 1 specifically leads to extinction of its expression on deleted allele. Interestingly, two novel *MBD5* genetic alteration types were identified, an intragenic duplication and a nonsense mutation. Patient C intragenic duplication affects non-coding exon 5 to coding exon 10 of *MBD5*. Transcriptional studies showed the presence of six aberrant transcripts, and sequencing analysis showed in each of these transcripts a premature termination codon (at the end of the coding sequence of exon 15, 21 bp after the start of exon 5, 19 bp after the start of exon 6) in favor of a modified MBD5 protein with putative altered function. Patient E *de novo* nonsense mutation leads to premature termination codon in *MBD5* gene, and is predicted to result in a truncated protein that lacks the Proline-rich domain in addition to the putative nuclear localization signal. This mutation was not reported in the 1000 Genomes project (http://browser.1000genomes.org/) or in dbSNP (builds 137). RT-PCR analysis showed a normal level of expression of *MBD5*, suggesting that RNA decay did not occur. Notably, *MBD5* transcripts sequencing showed *in vivo* expression of both normal and mutated transcripts. Translation of this mutated transcript might lead to a truncated protein with a dominant negative effect or this aberrant protein with the lack of the putative nuclear localization signal might impair the protein function. Complementary functional studies will help to appreciate the pathogenicity of these mutations. This fully penetrant mutation represents 1.2% (1/78) of our selected cohort. Interestingly, a *MBD5* frameshift mutation (c.150del (p.Thr52His*fs**31); NM_018328.4), resulting in a premature stop codon has been reported in a patient with Kleefstra syndrome phenotypic spectrum.[11] This frameshift mutation, predicted to be deleterious, is in favor of the implication of MBD5 mutations in an extended spectrum of neurodevelopmental disorders. Finally, regarding *MBD5* point mutations, missense variants have been reported,[2,10] mainly inherited from a healthy parent. We also identified in our selected cohort of patients (Supplementary Table I) previously reported missense variants. More specifically, for two patients (33 and 64) we detected the variants p.1048Thr>Ile and p.Ile752Val, respectively, each inherited from a healthy parent. As suggested by Talkowski *et al*,[10] these variants might participate as a potential risk factor for autism spectrum disorder. In conclusion, these findings confirm the involvement of *MBD5* mutations in neurodevelopmental disorders and extend the mutational spectrum of *MBD5*. Additional observations will be needed to establish fine genotype–phenotype correlations.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We thank the patients and their family for their kind cooperation.
We thank the cytogenetics and molecular genetics staff at the Nancy Univesity Hospitals for their expert technical assistance. This study was supported by grants from the French Ministry of Health (DGOS) and the 'Fondation Jérôme Lejeune'.

1 Laget S, Joulie M, Le Masson F *et al*: The human proteins MBD5 and MBD6 associate with heterochromatin but they do not bind methylated DNA. *PLoS ONE* 2010; **5**: e11982.
2 Wagenstaller J, Spranger S, Lorenz-Depiereux B *et al*: Copy-number variations measured by single-nucleotide-polymorphism oligonucleotide arrays in patients with mental retardation. *Am J Hum Genet* 2007; **81**: 768–779.
3 Jaillard S, Dubourg C, Gerard-Blanluet M *et al*: 2q23.1 microdeletion identified by array comparative genomic hybridisation: an emerging phenotype with Angelman-like features? *J Med Genet* 2009; **46**: 847–855.
4 Williams SR, Mullegama SV, Rosenfeld JA *et al*: Haploinsufficiency of MBD5 associated with a syndrome involving microcephaly, intellectual disabilities, severe speech impairment, and seizures. *Eur J Hum Genet* 2010; **18**: 436–441.
5 Van Bon BW, Koolen DA, Brueton L *et al*: The 2q23.1 microdeletion syndrome: clinical and behavioural phenotype. *Eur J Hum Genet* 2010; **18**: 163–170.
6 Chung BH, Stavropoulos J, Marshall CR *et al*: 2q23 de novo microdeletion involving the *MBD5* gene in a patient with developmental delay, postnatal microcephaly and distinct facial features. *Am J Med Genet A* 2011; **155**: 424–429.
7 Noh GJ, Graham Jr JM: 2q23.1 microdeletion of the *MBD5* gene in a female with seizures, developmental delay and distinct dysmorphic features. *Eur J Med Genet* 2012; **55**: 354–357.
8 Chung BH, Mullegama S, Marshall CR *et al*: Severe intellectual disability and autistic features associated with microduplication 2q23.1. *Eur J Hum Genet* 2012; **20**: 398–403.
9 Bonnet C, Masurel-Paulet A, Khan AA *et al*: Exploring the potential role of disease-causing mutation in a gene desert: duplication of noncoding elements 5′ of GRIA3 is associated with GRIA3 silencing and X-linked intellectual disability. *Hum Mutat* 2012; **33**: 355–358.
10 Talkowski ME, Mullegama SV, Rosenfeld JA *et al*: Assessment of 2q23.1 microdeletion syndrome implicates *MBD5* as a single causal locus of intellectual disability, epilepsy, and autism spectrum disorder. *Am J Hum Genet* 2011; **89**: 551–563.
11 Kleefstra T, Kramer JM, Neveling K *et al*: Disruption of an EHMT1-associated chromatin-modification module causes intellectual disability. *Am J Hum Genet* 2012; **91**: 1–10.

Supplementary Information accompanies the paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)

Annexe B

Liste des gènes impliqués dans des déficiences intellectuelles liées à l'X

Cette liste est la liste établie par le Greenwood Genetic Center [31]

31. http://www.ggc.org/research/molecular-studies/xlid.html

Genes Involved in X-Linked Intellectual Disability by Order of Discovery (revised November 2012)

Year	Gene Name	Gene Symbol	XLID Entity	Function	How Found
1983	Hypoxanthine guanine phosphoribosyl transferase	*HPRT*	Lesch-Nyhan	Enzyme	Met-Fu
1983	Phosphoglycerokinase 1	*PGK1*	Phosphoglycerokinase Deficiency	Enzyme	Met-Fu
1985	Proteolipid protein	*PLP1*	PMP, SPG1	Myelination	Mol-Fu
1986	Ornithine transcarbamoylase	*OTC*	Ornithine Transcarbamoylase Deficiency	Enzyme	Met-Fu
1987	Dystrophin	*DMD*	Duchenne Muscular Dystrophy	Structure of skeletal muscle membrane	Chr-rea
1989	Pyruvate dehydrogenase	*PDHA1*	Pyruvate Dehydrogenase Deficiency	Enzyme	Met-Fu
1990	Iduronate sulfatase	*IDS*	Hunter	Lysosomal enzyme	Met-Fu
1991	Fragile X mental retardation 1	*FMR1*	Fragile X	RNA-binding protein, gene regulation	Chr-rea L-can
1992	Cell adhesion molecule, L1	*L1CAM*	Hydrocephaly-MASA, SPG2, XL-ACC	Neuronal migration, cell adhesion	L-can
1992	Norrie	*NDP*	Norrie	Neuroectodermal cell interaction	Chr-rea
1992	Oculorenal	*OCRL1*	Lowe	Enzyme	Chr-rea
1993	Adrenoleukodystrophy protein	*ABCD1 (ALDP)*	Adrenoleukodystrophy	Peroxisomal transport protein	L-can
1993	Copper transporting ATPase 7A	*ATP7A*	Menkes, Occipital Horn	Copper transport	Chr-rea
1993	Monoamine oxidase A	*MAOA*	Monoamine Oxidase A Deficiency	Enzyme	L-can
1995	X-linked nuclear protein, X-linked helicase 2	*ATRX (XNP, XH2)*	Alpha-Thalassemia Intellectual Disability, Carpenter-Waziri, Chudley-Lowry, Holmes-Gang, XLID-Hypotonic Facies, XLID-Spastic Paraplegia, XLID-Arch Fingerprints-Hypotonia	Transcription factor, helicase activities	L-can
1996	Dystonia-deafness peptide	*TIMM8A (DDP)*	Mohr-Tranebjaerg, Jensen	Transcription factor	L-can
1996	Faciogenital dysplasia	*FGD1 (FGDY)*	Aarskog-Scott	Guanine nucleotide exchange factor	Chr-rea
1996	Fragile X mental retardation 2	*AFF2 (FMR2)*	Fragile XE	Unknown	Chr-rea
1996	Glycerol kinase deficiency	*GKD*	Glycerol Kinase Deficiency	Metabolism, glycerol uptake	Met-Fu
1996	Glypican 3	*GPC3*	Simpson-Golabi-Behmel	Cell adhesion, motility	L-can
1996	Myotubularin	*MTM1*	Myotubular Myopathy	Tyrosine phosphatase	L-can
1997	Midline 1	*MID1*	Telecanthus-Hypospadias, Opitz G/BBB	Zinc finger gene	L-can, Chr-rea
1997	Rab GDP-dissociation	*GDI1*	MRX41, 48	Stabilizes GDP bound	L-can

	inhibitor 1			conformations	
1997	Threonine-serine kinase 2	*RPS6KA3 (RSK2)*	Coffin-Lowry, MRX19	Kinase signaling pathway	L-can
1998	Doublecortin	*DCX*	Lissencephaly, X-linked	Neuronal migration	Chr-rea (del)
1998	Dyskerin	*DKC1*	Dyskeratosis Congenita	Cell cycle and nucleolar functions	L-can
1998	Filamin 1	*FLNA (FLN1)*	Periventricular Heterotopias, OPD I, OPD II	Actin-binding protein	L-can
1998	Oligophrenin 1	*OPHN1*	MRX60	GTPase activating protein	L-can
1998	P21-activated kinase	*PAK3*	MRX30, 47	Rac/Cdc 42 effector	Chr-rea
1999	IL-1 receptor accessory protein-like	*IL1RAPL*	MRX21	Unknown	Chr-rea
2000	Lysosomal-associated membrane protein 2	*LAMP2*	Danon Cardiomyopathy	Membrane, lysosome	L-can
2000	NF-$_K$B essential modulator	*NEMO (IKB6KG)*	Incontinentia Pigmenti	Activates the transcription factor NF-$_K$B	L-can
2000	Rho guanine nucleotide exchange factor 6	*ARHGEF6 (α-PIX)*	MRX46	Effector of the rho GTPases	Chr-rea
2000	Transmembrane 4 superfamily member 2	*TM4SF2*	MRX58	Interacts with integrins	Chr-rea
2001	Creatine transporter	*SLC6A8*	XLID with Seizures	Creatine transporter	Met-Fu
2001	Methyl-CpG binding protein 2	*MECP2*	Rett, MRX16, 79	Binds methylated CpGs	L-can
2001	Oral-facial-digital syndrome 1	*OFD1*	Oral-Facial-Digital I	Unknown	L-can
2002	Angiotensin-II receptor type 2	*AGTR2*	Optic Atrophy, X-linked, MRX	Angiotensin II receptor	Chr-rea
2002	Aristaless-related X chromosome gene	*ARX*	Hydranencephaly, Partington, Proud, West, Lissencephaly and Abnormal Genitalia, X-linked, MRX29, 32, 33, 36, 43, 54, 76	Neuronal migration	Chr-rea (del)
2002	Fatty acid acyl CoA synthetase type 4	*ACSL4 (FACL4)*	MRX63, 68	Fatty acid CoA ligase 4	Chr-rea (del)
2002	Kruppel-like factor 8	*KLF8 (ZNF741)*	MRX		Chr-rea
2002	PHD-like zinc finger gene 6	*PHF6*	Börjeson-Forssman-Lehmann	Unknown	L-can
2002	Serine-threonine kinase 9	*CDKL5 (STK9)*	Rett-Like Seizures-Hypotonia	Unknown	Chr-rea
2002	SRY-box 3	*SOX3*	XLID-Growth Hormone Deficiency	Pituitary function, transcription factor	Chr-rea, L-can
2003	Immunoglobulin-binding protein 1	*IGBP1*	Graham Coloboma		L-can
2003	Nance-Horan syndrome gene	*NHS*	Nance-Horan	-	L-can
2003	Neuroligin 3	*NLGN3*	Autism	Cell adhesion	L-can

2003	Neuroligin 4	*NLGN4*	Autism	Cell adhesion	L-can
2003	Polyglutamine tract binding protein 1	*PQBP1*	Renpenning, Sutherland-Haan, Hamel Cerebro-Palato-Cardiac, Golabi-Ito-Hall, Porteous, MRX55	Polyglutamine binding, regulates transcription	L-can
2003	Spermine synthase	*SMS*	Snyder-Robinson	Synthesis of spermine	L-can
2003	Zinc finger 41	*ZNF41*	MRX	Zinger finger	Chr-rea
2003	Zinc finger 81	*ZNF81*	MRX45	Zinc finger	Chr-rea
2004	BCL6 corepressor	*BCOR*	Lenz Microphthalmia (1 type)	Histone/protein deacetylation	L-can
2004	Jumonji, AT-rich interactive domain 1C	*KDM5C (JARID1C, SMX)*	MRX	Regulates transcription, chromatin remodelling	L-can
2004	K1AA1202 protein	*SHROOM4 (KIAA1202)*	Stoccos dos Santos	Roles in cellular architecture, neurulation, and ion channel function	Chr-rea
2004	KIAA2022 protein	*KIAA2022*	Cantagrel Spastic Paraplegia	DNA synthesis, DNA polymerase activity	Chr-rea
2004	Methyl transferase	*FTSJ1*	MRX9	Methylase	L-can
2004	Neuroendocrine DLG	*DLG3*	MRX	NMDA-receptor, mediated signaling, synaptic plasticity	X seq
2004	PFD finger protein 8	*PHF8*	XLID-Cleft Lip-Cleft Palate	Regulates transcription, binds DNA	L-can
2004	Renin receptor	*ATP6AP2 (ATP6A8-9)*	XLID-Infantile Epilepsy	Renin receptor	L-can
2004	Rho guanine nucleotide exchange factor 9	*ARHGEF9*	XLID-Hypotonia-Seizures	Regulation of Rho protein signal transduction	Chr-rea
2004	Synapsin 1	*SYN1*	Epilepsy-Macrocephaly	Synaptic vesicle protein	L-can
2004	T3 transporter	*SLC16A2 (MCT8)*	Allan-Herndon-Dudley	T3 receptor	L-can
2005	Zinc finger DHHC domain-containing protein 15	*ZDHHC15*	MRX91		Chr-rea
2006	Fanconi anemia complementation group B protein	*FANCB*	VACTERL-Hydrocephaly	DNA repair	Mol-Fu
2006	Holocytochrome C synthase	*HCCS*	MIDAS	Energy production, cytochrome homolyase	Chr-rea (del)
2006	Sigma 2 subunit of adaptor protein/complex	*AP1S2*	Turner XLID, Hydrocephaly-Basal Ganglia Calcification	Assembly of endocytic vesicles	X-seq
2006	SMC1 structural maintenance of chromosomes 1-like	*SMC1A/ SMC1L1*	Cornelia de Lange, X-linked	Cell cycle, mitotic spindle organization and biogenesis, chromosome segregation	Mol-Fu
2006	Sushi repeat containing protein, X-linked	*SRPX2*	XLID-Rolandic Seizures	Signal transduction, growth factor 2	Mol-Fu
2006	Ubiquitin-conjugating enzyme E2A	*UBE2A*	XLID-Nail Dystrophy-Seizures	Ubiquitin cycle, ubiquitin-protein ligase	L-can
2006	Zinc finger protein 674*	*ZNF674*	XLID-Retinal Dystrophy-Short Stature and MRX92	Transcription regulation	Chr-rea (del)

2007	Bromodomain and WD repeat domain-containing protein 3	*BRWD3*	XLID-Macrocephaly-Large Ears	Transcription factor	X-seq
2007	Cullin 4B	*CUL4B*	XLID-Hypogonadism-Tremor	Cell cycle, ubiquitin cycle, E3 ubiquitin ligase	X-seq
2007	Drosophila porcupine homolog	*PORCN*	Goltz	Wnt receptor signaling pathway, acyltransferase activity, integral to membrane of endoplasmic reticulum	Chr-rea (del)
2007	Glutamate receptor ionotropic AMPA 3	*GRIA3*	Chiyonobu XLID	Signal transduction, ion transport, glutamate signaling pathway	Chr-rea, Exp-Arr, X seq
2007	Hydroxyacyl-coenzyme A dehydrogenase, type III	*HSD17B10 (HADH2)*	XLID-Choreoathetosis	Lipid metabolism	L-can
2007	Mediator of RNA polymerase II transcription, subunit 12	*MED12 (HOPA)*	Opitz FG, Lujan	Transcription regulation, RNA polymerase II transcription mediator activity, ligand-dependent nuclear receptor transcription coactivator activity, vitamin D receptor and thyroid hormone receptor binding	L-can
2007	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex	*NDUFA1*	Mitochondrial Complex 1 Deficiency	Energy production, oxidoreductase activity	Mol-Fu
2007	Nuclear RNA export factor 5	*NXF5*	XLID-Short Stature-Muscle Wasting	mRNA processing, mRNA export from nucleus	Chr-rea
2007	Phosphoribosyl pyrophosphate synthetase 1	*PRPS1*	Arts, *PRPS1* Superactivity	Ribonucleotide monophosphate biosynthesis	L-can
2007	Ribosomal protein L10	*RPL10*	Autism	Protein synthesis, ribosomal protein	X seq
2007	UPF3 regulator of nonsense transcript homolog B	*UPF3B*	MRX, Lujan/FG Phenotype	mRNA catabolism, nonsense-mediated decay	X-seq
2007	Zinc finger, DHHC-domain containing protein 9	*ZDHHC9*	XLID-Macrocephaly-Marfanoid Habitus	?	X-seq
2008	E3 ubiquitin-protein ligase	*HUWE1*	MRX, XLID-Macrocephaly, Juberg-Marsidi-Brooks	Ubiquitin-protein ligase, mRNA transport	M-CGH
2008	Protocadherin 19	*PCDH19*	Epilepsy-Intellectual Disability Limited to Females		L-can
2008	Sodium-hydrogen exchanger NHE6	*SLC9A6*	Christianson, X-linked Angelman-like	Sodium-hydrogen antiporter activity, lysosome organization and biogenesis, regulation of endosome volume	L-can
2009	Magnesium transporter 1	*MAGT1*			
2009	Intramembrane zinc	*MBTPS2*	Ichthyosis Follicularis, Atrichia,	Protease activity, activates	L-can

	metalloprotease		Photophobia (IFAP)	signaling proteins	
2009	NAD(P)H steroid dehydrogenase-like	*NSDHL*	CK (microcephaly, pachygyria, facial dysmorphism, seizures), also in CHILD syndrome	Sterol metabolism	L-can
2010	Small GTPase gene	*RAB39B*	MRX72 and a syndrome with macrocephaly, seizures, and autism	Formation and maintenance of synapse	L-can
2010	Guanine nucleotide exchange factor	*IQSEC2*	MRX1, MRX18, and other Nonsyndromal XLID	Regulation of vesicular transport and organelle structure	X-seq
2010	Patched domain-containing 1	*PTCHD1*	Autism-XLID	Transmembrane protein related to hedgehog receptors	Array CGH
2011	Ras-associated protein RAB40A-like	*RAB40AL*	Martin-Probst	Ras-like GTPase protein	X-seq
2011	RNA-binding motif protein	*RBM10*	TARP	RNA-binding	X-seq
2011	N-acetyltransferase subunit 10	*NAA10*	N-Alpha-Acetyltransferase Deficiency	N-terminal acetylation	X-seq
2011	Las1-like protein	*LAS1L*	Wilson-Turner	Nucleolar protein, cell proliferation and ribosome biogenesis	X-seq
2011	Eukaryotic translation initiation factor 2	*EIF2S3*	MEHMO	Initiates translation	X-seq
2011	Host cell factor C1	*HCFC1*	MRX3	Cell proliferation	X-seq
2011	THO complex, subunit 2	*THOC2*	MRX12	mRNA transcription or export	X-seq
2011	Chloride channel voltage-gated 4	*CLCN4*	MRX49	Chloride transport	X-seq
2011	Histone deacetylase 8	*HDAC8*	Cornelia de Lange, X-linked	Chromatin cohesion	Mol-Fu, X-seq
2011	Connector enhancer of KSR-2	*CNKSR2*	XLID-Microcephaly-Seizures	Stimulates MAPK signalling	Array
2012	Coiled-coil domain-containing protein 22	*CCDC22*	Cardiofacioskeletal	Unknown	X-seq
2012	Emopamil binding protein	*EBP*	XLID-Aggression	Enzyme in cholesterol metabolism	--
2012	Chlorine intracellular channel 2	*CLIC2*	XLID-Cardiomegaly-Seizures	Regulating ryanodine receptor channel activity	X-seq

*Association of *ZNF674* and XLID considered uncertain
Chr-rea = chromosome rearrangement
Exp-Arr = expression array
L-can = linkage and candidate gene testing
M-CGH = array-comparative genomic hybridization
Met-Fu = exploitation of metabolic alteration
Mol-Fu = exploitation of molecular finding
X-seq = brute force sequencing

Annexe C

Enrichissement en termes GO des clusters de gènes DILX

Terme GO	P-Value	Cluster
GO :0006259 DNA metabolic process	$6{,}86.10^{-3}$	1
GO :0006396 RNA processing	$8{,}51.10^{-3}$	1
GO :0016574 histone ubiquitination	$1{,}45.10^{-2}$	1
GO :0006281 DNA repair	$2{,}13.10^{-2}$	1
GO :0006974 response to DNA damage stimulus	$3{,}54.10^{-2}$	1
GO :0009451 RNA modification	$3{,}84.10^{-2}$	1
GO :0043414 biopolymer methylation	$5{,}47.10^{-2}$	1
GO :0051276 chromosome organization	$5{,}70.10^{-2}$	1
GO :0032259 methylation	$6{,}01.10^{-2}$	1
GO :0006364 rRNA processing	$7{,}23.10^{-2}$	1
GO :0033554 cellular response to stress	$7{,}49.10^{-2}$	1
GO :0016072 rRNA metabolic process	$7{,}54.10^{-2}$	1
GO :0006730 one-carbon metabolic process	$8{,}74.10^{-2}$	1
GO :0016567 protein ubiquitination	$9{,}26.10^{-2}$	1
GO :0016570 histone modification	$9{,}49.10^{-2}$	1
GO :0042254 ribosome biogenesis	$9{,}49.10^{-2}$	1
GO :0016569 covalent chromatin modification	$9{,}78.10^{-2}$	1
GO :0006350 transcription	$2{,}94.10^{-4}$	2
GO :0045449 regulation of transcription	$9{,}84.10^{-4}$	2
GO :0006355 regulation of transcription, DNA-dependent	$7{,}57.10^{-2}$	2
GO :0051252 regulation of RNA metabolic process	$8{,}00.10^{-2}$	2
GO :0055085 transmembrane transport	$1{,}77.10^{-3}$	4
GO :0006695 cholesterol biosynthetic process	$1{,}92.10^{-3}$	5
GO :0016126 sterol biosynthetic process	$2{,}59.10^{-3}$	5
GO :0006694 steroid biosynthetic process	$6{,}28.10^{-3}$	5
GO :0008203 cholesterol metabolic process	$6{,}80.10^{-3}$	5
GO :0016125 sterol metabolic process	$7{,}47.10^{-3}$	5
GO :0008202 steroid metabolic process	$1{,}49.10^{-2}$	5
GO :0008610 lipid biosynthetic process	$2{,}39.10^{-2}$	5
GO :0007242 intracellular signaling cascade	$9{,}83.10^{-3}$	6

Terme GO	P-Value	Cluster
GO :0007243 protein kinase cascade	2,90.10^{-2}	6
GO :0007610 behavior	4,49.10^{-2}	6
GO :0050877 neurological system process	5,31.10^{-2}	6
GO :0007611 learning or memory	7,91.10^{-2}	6
GO :0015031 protein transport	5,63.10^{-2}	7
GO :0045184 establishment of protein localization	5,68.10^{-2}	7
GO :0008104 protein localization	6,52.10^{-2}	7
GO :0006096 glycolysis	1,38.10^{-2}	13
GO :0006007 glucose catabolic process	1,70.10^{-2}	13
GO :0019320 hexose catabolic process	2,02.10^{-2}	13
GO :0046365 monosaccharide catabolic process	2,08.10^{-2}	13
GO :0046164 alcohol catabolic process	2,37.10^{-2}	13
GO :0044275 cellular carbohydrate catabolic process	2,49.10^{-2}	13
GO :0006796 phosphate metabolic process	2,81.10^{-2}	13
GO :0006793 phosphorus metabolic process	2,81.10^{-2}	13
GO :0016052 carbohydrate catabolic process	3,18.10^{-2}	13
GO :0006006 glucose metabolic process	4,45.10^{-2}	13
GO :0019318 hexose metabolic process	5,56.10^{-2}	13
GO :0005996 monosaccharide metabolic process	6,41.10^{-2}	13
GO :0006091 generation of precursor metabolites and energy	8,94.10^{-2}	13
GO :0046578 regulation of Ras protein signal transduction	7,13.10^{-5}	15
GO :0051056 regulation of small GTPase mediated signal transduction	1,23.10^{-4}	15
GO :0035023 regulation of Rho protein signal transduction	7,80.10^{-4}	15
GO :0008624 induction of apoptosis by extracellular signals	9,97.10^{-4}	15
GO :0030036 actin cytoskeleton organization	3,99.10^{-3}	15
GO :0030029 actin filament-based process	4,52.10^{-3}	15
GO :0006917 induction of apoptosis	7,86.10^{-3}	15
GO :0012502 induction of programmed cell death	7,90.10^{-3}	15
GO :0043065 positive regulation of apoptosis	1,39.10^{-2}	15
GO :0043068 positive regulation of programmed cell death	1,41.10^{-2}	15
GO :0010942 positive regulation of cell death	1,42.10^{-2}	15
GO :0007010 cytoskeleton organization	1,43.10^{-2}	15
GO :0006915 apoptosis	2,63.10^{-2}	15
GO :0012501 programmed cell death	2,71.10^{-2}	15
GO :0030031 cell projection assembly	3,63.10^{-2}	15
GO :0008219 cell death	3,67.10^{-2}	15
GO :0016265 death	3,72.10^{-2}	15
GO :0042981 regulation of apoptosis	4,51.10^{-2}	15
GO :0043067 regulation of programmed cell death	4,59.10^{-2}	15
GO :0010941 regulation of cell death	4,62.10^{-2}	15
GO :0016568 chromatin modification	2,03.10^{-2}	16
GO :0006325 chromatin organization	2,79.10^{-2}	16
GO :0051276 chromosome organization	3,59.10^{-2}	16
GO :0006576 biogenic amine metabolic process	3,54,10^{06}	24
GO :0006575 cellular amino acid derivative metabolic process	1,78.10^{-5}	24
GO :0042417 dopamine metabolic process	2,07.10^{-5}	24
GO :0006584 catecholamine metabolic process	6,10.10^{-5}	24
GO :0034311 diol metabolic process	6,10.10^{-5}	24

Terme GO	P-Value	Cluster
GO :0009712 catechol metabolic process	$6{,}10.10^{-5}$	24
GO :0018958 phenol metabolic process	$6{,}47.10^{-5}$	24
GO :0044271 nitrogen compound biosynthetic process	$1{,}33.10^{-4}$	24
GO :0046100 hypoxanthine metabolic process	$1{,}11.10^{-3}$	24
GO :0006164 purine nucleotide biosynthetic process	$1{,}16.10^{-3}$	24
GO :0006163 purine nucleotide metabolic process	$1{,}83.10^{-3}$	24
GO :0009165 nucleotide biosynthetic process	$1{,}83.10^{-3}$	24
GO :0034654 nucleobase, nucleoside, nucleotide and nucleic acid biosynthetic process	$1{,}97.10^{-3}$	24
GO :0034404 nucleobase, nucleoside and nucleotide biosynthetic process	$1{,}97.10^{-3}$	24
GO :0009113 purine base biosynthetic process	$3{,}32.10^{-3}$	24
GO :0006144 purine base metabolic process	$5{,}16.10^{-3}$	24
GO :0046112 nucleobase biosynthetic process	$5{,}16.10^{-3}$	24
GO :0009112 nucleobase metabolic process	$8{,}47.10^{-3}$	24
GO :0051289 protein homotetramerization	$9{,}57.10^{-3}$	24
GO :0007610 behavior	$1{,}12.10^{-2}$	24
GO :0021954 central nervous system neuron development	$1{,}18.10^{-2}$	24
GO :0051262 protein tetramerization	$1{,}43.10^{-2}$	24
GO :0046148 pigment biosynthetic process	$1{,}43.10^{-2}$	24
GO :0021953 central nervous system neuron differentiation	$1{,}47.10^{-2}$	24
GO :0042440 pigment metabolic process	$1{,}65.10^{-2}$	24
GO :0009116 nucleoside metabolic process	$2{,}34.10^{-2}$	24
GO :0051260 protein homooligomerization	$3{,}46.10^{-2}$	24
GO :0009152 purine ribonucleotide biosynthetic process	$4{,}25.10^{-2}$	24
GO :0009260 ribonucleotide biosynthetic process	$4{,}50.10^{-2}$	24
GO :0009150 purine ribonucleotide metabolic process	$5{,}00.10^{-2}$	24
GO :0009259 ribonucleotide metabolic process	$5{,}32.10^{-2}$	24
GO :0030900 forebrain development	$5{,}49.10^{-2}$	24
GO :0051259 protein oligomerization	$6{,}27.10^{-2}$	24
GO :0046649 lymphocyte activation	$7{,}14.10^{-2}$	24
GO :0048812 neuron projection morphogenesis	$7{,}63.10^{-2}$	24
GO :0045321 leukocyte activation	$8{,}63.10^{-2}$	24
GO :0048858 cell projection morphogenesis	$8{,}73.10^{-2}$	24
GO :0032990 cell part morphogenesis	$9{,}11.10^{-2}$	24
GO :0031175 neuron projection development	$9{,}11.10^{-2}$	24
GO :0007156 homophilic cell adhesion	$9{,}68.10^{-3}$	26
GO :0016337 cell-cell adhesion	$2{,}04.10^{-2}$	26
GO :0007155 cell adhesion	$5{,}17.10^{-2}$	26
GO :0022610 biological adhesion	$5{,}18.10^{-2}$	26
GO :0006406 mRNA export from nucleus	$2{,}44.10^{-3}$	34
GO :0006405 RNA export from nucleus	$3{,}03.10^{-3}$	34
GO :0051168 nuclear export	$4{,}44.10^{-3}$	34
GO :0051028 mRNA transport	$6{,}43.10^{-3}$	34
GO :0051236 establishment of RNA localization	$7{,}17.10^{-3}$	34
GO :0050658 RNA transport	$7{,}17.10^{-3}$	34
GO :0050657 nucleic acid transport	$7{,}17.10^{-3}$	34
GO :0006403 RNA localization	$7{,}39.10^{-3}$	34
GO :0015931 nucleobase, nucleoside, nucleotide and nucleic acid transport	$8{,}35.10^{-3}$	34
GO :0006913 nucleocytoplasmic transport	$1{,}15.10^{-2}$	34

Terme GO	P-Value	Cluster
GO :0051169 nuclear transport	1,17.10^{-2}	34
GO :0046907 intracellular transport	4,86.10^{-2}	34
GO :0000226 microtubule cytoskeleton organization	1,09.10^{-2}	39
GO :0007017 microtubule-based process	1,87.10^{-2}	39
GO :0007010 cytoskeleton organization	3,22.10^{-2}	39

TABLE C.1 – Enrichissement en termes GO ,calculé par l'outil DAVID pour les 17 clusters de gènes DILX regroupés par IntelliGO.

Annexe D

Recherche de gènes candidats pour Aicardi en amont du test biologique

D.1 Le syndrome d'Aicardi

Le syndrome d'Aicardi a été décrit pour la première fois en 1965 par le docteur Jean Aicardi suite à l'observation de filles présentant des spasmes en flexion (Aicardi *et al.*, 1965). En 1969 une seconde étude portant sur un plus grand nombre de cas lui a permis de proposer une triade de signes cliniques pour le diagnostic du syndrome d'Aicardi. Ainsi, ce syndrome était défini par une agénésie du corps calleux, des spasmes infantiles et des lacunes choriorétiniennes (Aicardi *et al.*, 1969). D'autres malformations de l'œil et du squelette peuvent aussi être aussi observées chez ces filles (Aicardi *et al.*, 1969, Donnenfeld *et al.*, 1989, Sutton *et al.*, 2005).

Le syndrome d'Aicardi n'est observé que chez les filles et de façon exceptionnelle chez les garçons présentant par ailleurs un syndrome Klinefelter (47, XXY). Le seul cas exceptionnel de deux sœurs atteintes du syndrome d'Aicardi plaide pour une origine génétique du syndrome (Molina *et al.*, 1989). L'hypothèse d'une néo mutation dominante liée au chromosome X expliquerait l'atteinte exclusive des filles. Un effet létal de cette mutation chez les garçons est avancé. Le gène responsable de ce syndrome n'étant toujours pas connu, nous avons décidé de mettre en place une approche différentielle similaire à celle de Vissers *et al.* (2010).

D.2 Premier séquençage

Afin de déterminer quel gène du chromosome X est muté dans le syndrome, l'exome de ce chromosome a été séquencé pour trois patientes (A, B et C) et les parents de C. Ainsi, le gène responsable de la maladie devrait présenter une mutation chez les trois patientes et pas chez les parents. Les données issues de l'analyse des résultats de séquençage ont été stockées dans NetworkDB, dont le schéma a été légèrement modifié en ajoutant les informations sur les patients et les mutations observées (Figure D.1). En plus des informations issues du séquençage, les interactions protéine-protéine, les réseaux biologiques et les termes GO des protéines mutées (et de leurs

interactants directs) sont stockées dans la base de données (Table D.1). Une variation détectée dans un gène est considérée comme une mutation si elle confirmée par au moins 4 lectures, qu'elle est couverte par au moins 25% des lectures à cette position et quelle est absentes des bases de données dbSNP et HapMap. Elle doit également ne pas être une variation synonyme.

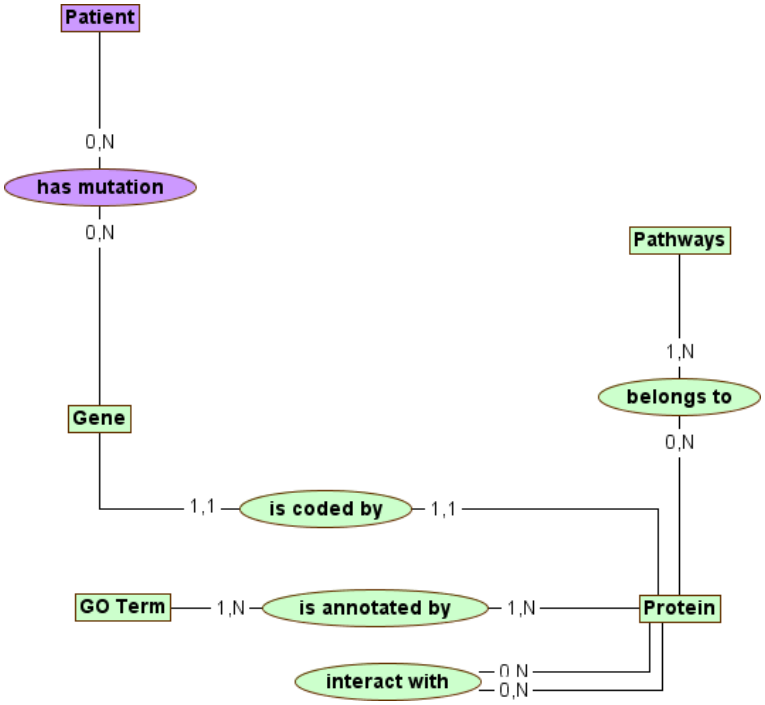


FIGURE D.1 – Schéma entité-association de NetworkDB adapté pour l'analyse des résultats de séquençage.

Table	Nombre d'éléments
Gene	111
Protein	464
Protein_protein_interaction	473
Pathway	398
GO_term	1990

TABLE D.1 – Statistiques de l'entrepôt NetworkDB.

Trois gènes, *DMD*, *VCX2* et *HDAC6*, ont été identifiés comme étant mutés chez les 3 patientes et non mutés chez les parents. DMD code pour la Dystrophine, qui est déjà connue pour être responsable de la Dystrophie musculaire de Duchenne (MIM : 310200), il s'agit donc d'un mauvais candidat pour le syndrome d'Aicardi. VCX2 est une protéine assez peu décrite (aucun terme GO, aucune interaction) mais dont une protéine de la même famille, VCXA, est exprimée dans le cerveau et qui est associée à une déficience intellectuelle liée à l'X (MIM : 300533, (Jiao *et al.*, 2009)). HDAC6 est une histone déacetylase, impliquée dans la régulation de la transcription via la modification de la conformation de la chromatine. Malheureusement, après reséquençage de ces trois gènes chez les filles et leurs parents, il s'est avéré que soit les variations étaient héritées mais non détectées chez les parents lors du premier séquençage, soit qu'il s'agissait d'erreurs du séquençage haut débit.

Aucun gène du chromosome X identifié comme muté chez les 3 patientes ne permet donc de

déterminer le gène candidat pour le syndrome d'Aicardi. La liste des gènes des 27 gènes mutés chez la patiente C et pas chez ses parents a été vérifiée expérimentalement (Table D.2) et malheureusement toutes les mutations se sont avérées être des faux positifs ou héritées des parents.

Gène	Patientes	Dérégulé dans le syndrome d'Aicardi	Expression dans le cerveau
ACRC	c	oui	oui
AR	c	non	oui
ARHGEF6	c	non	oui
ARHGEF7	c	non	oui
BMP15	c	non	non
BRWD3	c	non	non
CHIC1	bc	non	oui
DOCK11	ac	non	oui
FAM47A	c	non	non
FLNA	bc	non	oui
HUWE1	ac	non	oui
KDM5C	c	non	oui
L1CAM	c	non	oui
MAP3K15	c	non	non
MCF2	c	non	oui
PHEX	c	non	non
PHKA2	c	non	oui
PLXNA3	c	non	oui
PNCK	c	non	oui
PRICKLE3	c	non	oui
RPL36A	c	non	oui
TRO	ac	non	oui
TSPYL2	c	non	oui
UPF3B	ac	non	oui
ZMAT1	ac	non	oui
ZRSR2	c	non	oui

TABLE D.2 – Liste des gènes mutés chez la patiente C et non mutés chez ses parents. Les données de dérégulation sont celles obtenues par Yilmaz, S. (2007). Les données d'expression proviennent la base de données GeneCards.

En parallèle à la vérification biologique des mutations observées, j'ai cherché à établir la liste des gènes candidats potentiels. Comme proposé par Yilmaz *et al.* (2009), un gène candidat pour Aicardi peut être un interactant d'un gène muté. Il est possible d'étendre cette définition aux gènes d'un même réseau biologique. Ainsi en visualisant les informations contenues dans NetworkDB, on peut espérer trouver un bon candidat, à savoir un gène interagissant avec différentes protéines mutées chez les patientes ou un réseau biologique commun à différentes protéines mutées chez les 3 patientes.

Ainsi, en cherchant la liste des protéines qui interagissent avec des protéines mutées chez les 3 patientes, on obtient un liste de 23 interactants (Table D.3). Ces 23 protéines permettent de relier 22 protéines mutées chez A, (A et C), B, (B et C), C et (A, B et C). Elles sont ainsi des candidats potentiels pour le syndrome d'Aicardi.

Si on recherche un réseau biologique candidat plutôt qu'un gène à partir de ces données de

Nom
ATG2A protein
5'-AMP-activated protein kinase subunit beta-2
Protein NipSnap homolog 2
Adapter molecule crk
5'-AMP-activated protein kinase subunit gamma-1
Beta-arrestin-1
Battenin
Growth factor receptor-bound protein 2
Tectonin beta-propeller repeat-containing protein 1
Structure-specific endonuclease subunit SLX4
Beclin-1
Autophagy-related protein 16-1
Ras association domain-containing protein 5
Cysteine protease ATG4C
Serine/threonine-protein kinase ULK2
RB1-inducible coiled-coil protein 1
Microtubule-associated proteins 1A/1B light chain 3B
Gamma-aminobutyric acid receptor-associated protein-like 1
Calcium/calmodulin-dependent protein kinase kinase 2
DET1- and DDB1-associated protein 1
WD repeat domain phosphoinositide-interacting protein 2
Autophagy protein 5
UV radiation resistance-associated gene protein

TABLE D.3 – Liste des 23 protéines interagissant avec des protéines mutées chez les 3 patientes.

séquençage, il est important de ne sélectionner que les réseaux biologiques permettant de relier des protéines mutées chez les 3 patientes (Figure D.3). On remarque ainsi, que seuls 3 réseaux biologiques répondent à ce critère : le réseau "mRNA Splicing - Major Pathway", le réseau "RNA transport" et le réseau "Ubiquitin mediated proteolysis". Ces trois réseaux biologiques semblent donc de bon candidats pour le syndrome d'Aicardi, d'autant plus qu'ils appartiennent aux processus fréquemment associées aux DILX (section 4.3).

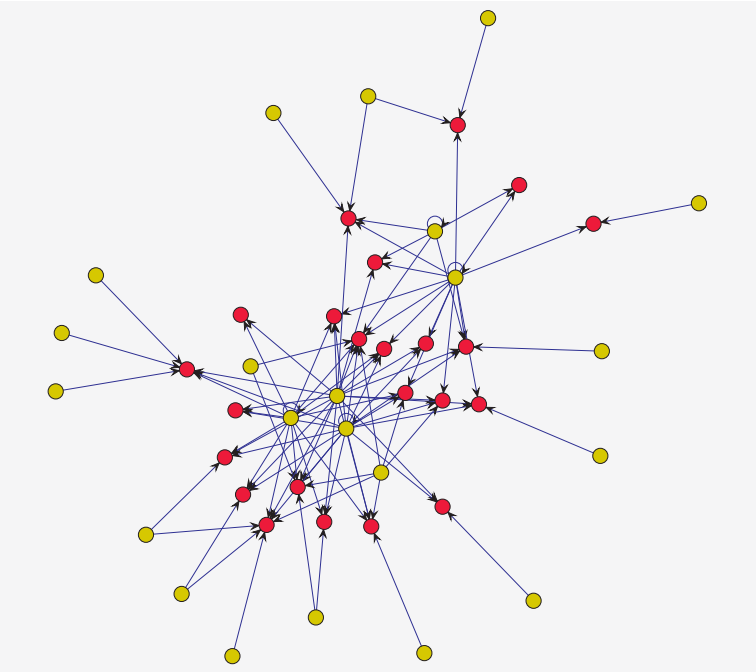


FIGURE D.2 – Réseau d'interaction des protéines présentant une "mutation" chez au moins une des patientes A, B et C (la "mutation" doit également être absente des parents de C). Les 22 protéines mutées sont représentées en jaune et leurs 23 interactants en rouge. Seules les protéines mutées qui sont reliées à d'autres protéines mutées via un interactant commun (ou une interaction directe) sont affichées

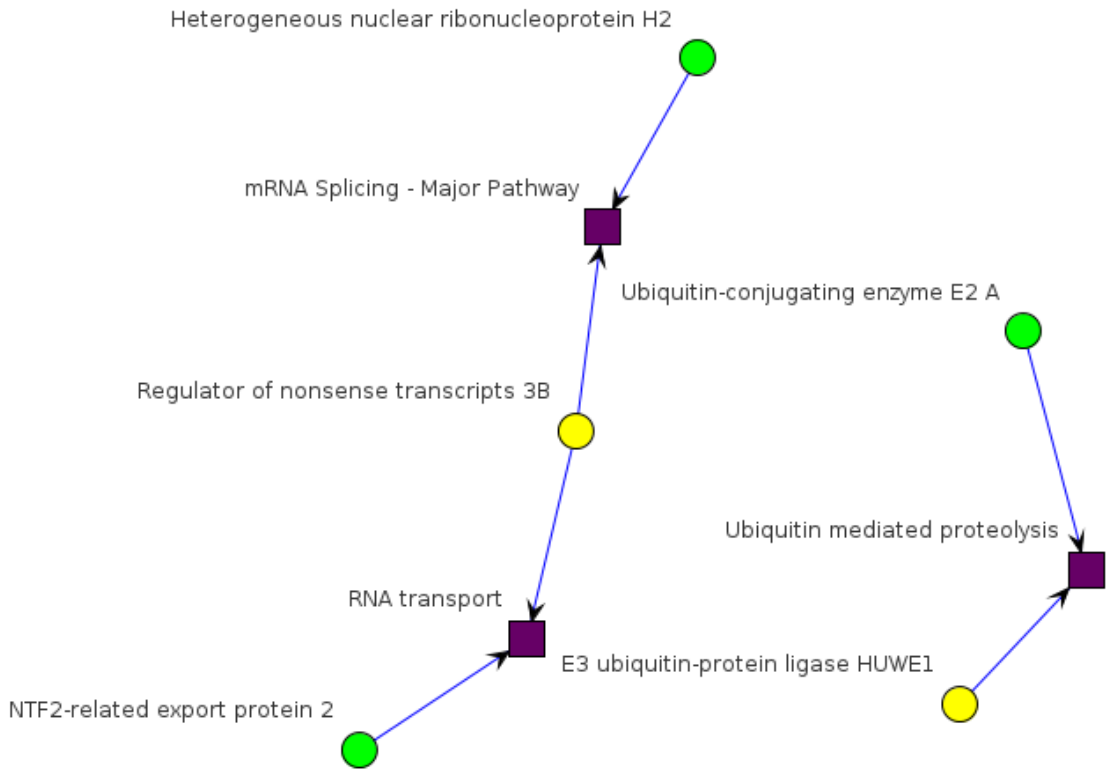


FIGURE D.3 – Graphe étendu des processus biologiques (carré violet) permettant de relier des protéines mutées (rond) chez les 3 patientes. Les protéines mutées chez les patientes A et C sont en jaune et les protéine mutés chez la patiente B sont en vert.

D.3 Second séquençage

D.3.1 Analyse initiale

Les résultats du premier séquençage n'ayant donné aucun résultat probant, une seconde campagne de séquençage a été menée. Cette fois ci, les exomes de 5 trios fille-parents ont été séquencés avec une couverture moyenne minimale de 50X (chaque position séquencée l'a été en moyenne 50x) afin de diminuer les risques de faux positifs. Afin de ne pas interférer avec les résultats du premier séquençage les nouvelles données ont été stockées dans une nouvelle version de NetworkDB. Parmi les résultats, le gène *FAM104B* est le seul pour lequel les 5 filles ont une mutation (de type insertion) qui n'est pas retrouvée identique chez leur parents. Cependant, l'analyse détaillée de ces insertions a montré des incohérences. On peut ainsi remarquer sur la figure D.4 que les insertions en amont de l'exon 3 observée chez le père et la mère n'ont pas été transmises à leur fille. De plus, la fille présente une insertion en amont de l'exon 2 hérité de son père plus grande que l'insertion observée chez lui. Enfin, toutes les insertions observées (dans les 5 trios) correspondent à des duplications de l'exon suivant/précédent. Après vérification, toutes ces insertions ce sont révélées être de faux positifs et il semblerait que cela provienne d'erreur lors de l'alignement. La qualité de l'alignement ayant mise en doute, une nouvelle étape d'alignement et d'analyse de variant à été réalisée avec un autre pipeline d'alignement au lieu de CASAVA.

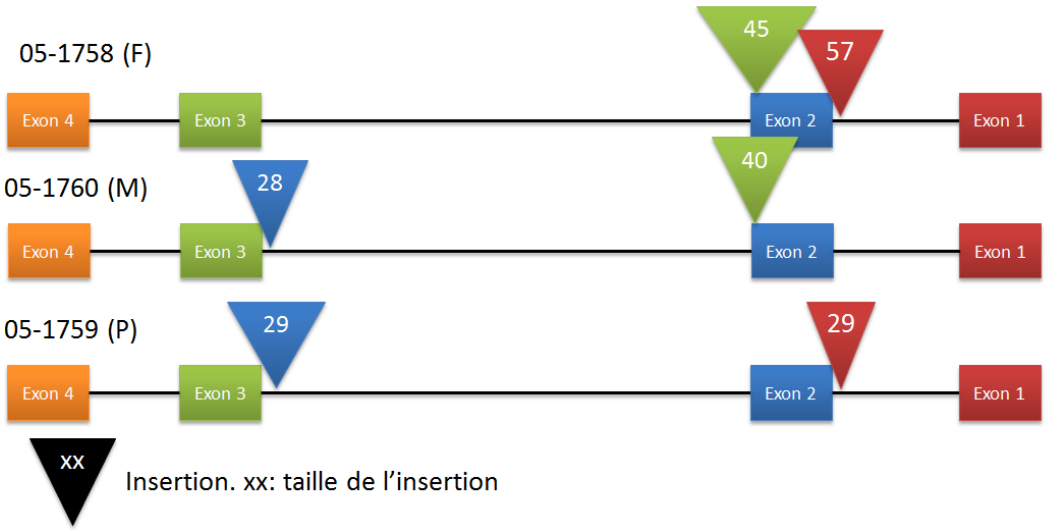


FIGURE D.4 – Exemple d'insertions observées dans le gène *FAM104B* chez un trio. La couleur est celle de l'exon qui possède la même séquence que la séquence insérée (dans tous les cas, la séquence insérée correspond au début d'un exon). 05-1758 est l'identifiant de la fille, 05-1760 est l'identifiant de la mère et 05-1759 correspond au père.

D.3.2 Analyses à la suite du réalignement des sequences

Le réalignement des séquences ainsi que la détection des variations a été effectuée par le Dr Hélène Dauchel de l'équipe TIBS (Traitement de l'information en Biologie Santé) au Laboratoire

d'Informatique Traitement de l'Information et des Systèmes (LITIS) de l'université de Rouen. Le pipeline développé pour la détection de variations nommé DGVar (Detection of Genetics Variations) est basé sur six outils libres et très utilisés internationalement, FastQC, BWA, SAMtools, IGV, puis Annovar et VEP, permettant une intégration dans l'outil d'interprétation EVA (Exome Variation Analyser) développé par l'équipe TIBS (Coutant *et al.*, 2012).

D.3.2.1 Recherche des gènes présentant des variations de novo chez plusieurs filles

D.3.2.1.1 Analyse via l'interface EVA : La recherche de variations *de novo* se fait à partir du filtre trio de l'interface EVA. Ce filtre permet de rechercher les variations présentes chez un individu et absentes chez d'autres personnes. Contrairement à ce que le nom du filtre sous-entend, il est possible de rechercher les variations qui sont observées chez un individu et qui sont absentes chez plus de 2 personnes. Pour chaque fille, j'ai ainsi pu filtrer les variations qui sont absentes des 10 parents. En regroupant les variations par gène, on extrait les gènes présentant des variations inconnues *de novo* chez plusieurs filles. Un total de 346 gènes mutés chez au moins 2 filles est obtenu.

Parmi ces gènes, 2 sont mutés chez les 5 filles : CAMTA1 et BCL6B. Pour le CAMTA1, les variations observées chez 4 filles sont homozygotes. Cette observation étant très peu probable dans le cas d'une mutation *de novo*, ce gène n'est donc vraiment « muté » (hétérozygote) que chez une seule fille. Les variations du second gène, BCL6B, sont un artéfact de l'alignement. En effet, toutes les variations observées sont des insertions d'un triplet AGC au sein d'une zone où ce triplet est répété. Les parents possèdent également une insertion d'un triplet AGC dans cette région mais le programme d'alignement place l'insertion à une position différente. Ceci entraine la détection d'une variation *de novo* chez la fille alors que la variation est héritée. Ainsi, aucune des 5 filles ne possède de variations *de novo* dans ce gène.

De façon similaire, l'analyse des gènes présentant des variations *de novo* chez 4 des 5 filles est décevante. En effet, un certain nombre des variations retenues sont annotées comme douteuses (c'est-à-dire ayant un score de qualité inférieur à 5), d'autres variations sont homozygotes et enfin certaines sont annotées comme proche des sites d'épissage alors qu'elles sont à plus de 50 bases du site d'épissage le plus proche. Ainsi en vérifiant manuellement les variations observées, on obtient les statistiques décrites dans la table D.4.

Ces erreurs (sites d'épissage) et le fait que la version distante d'EVA soit difficile à manipuler (requêtes pouvant prendre plusieurs minutes pour s'exécuter), m'ont poussé à utiliser directement les fichiers issus du pipeline DGVar en attendant que la version locale d'EVA soit disponible.

D.3.2.1.2 Analyse des fichiers d'annotations issues du pipeline DGVar : Les variations retenues sont des variations hétérozygotes non héritées des parents, elles sont également absentes des parents des autres filles et non référencées dans dbSNP. De plus, la position de la variation doit être couverte au minimum par 4 reads et le score de la variation doit être supérieur ou égal à 20. Le choix de ces paramètres est issu des analyses statistiques issues de l'alignement qui ont montrées que le score médian est de 25 et la profondeur médiane est 3X. Ces paramètres sont comparables à ceux utilisés par Riviere *et al.* (2012) pour identifier des mutations *de novo* responsables du syn-

Gène	# filles	#filles corrigé
CAMTA1	5	1
BCL6B	5	0
EFCAB4B	4	0
ZFHX3	4	0
GALNT9	4	0
RYR3	4	2
COL24A1	4	1
MAP1S	4	1
FAM171B	4	0
POU6F2	4	0
TBP	4	0

TABLE D.4 – Extraits des gènes mutés chez plusieurs filles selon EVA. La correction du nombre de fille est réalisée en éliminant les erreurs d'alignement dans les régions répétées, les variations douteuses (score de qualité inférieur à 5) et en vérifiant les variations annotées comme proches des site d'épissage.

drome Baraitser-Winter. Les variations doivent également être situées sur un exon et ne pas être de type synonyme afin d'être retenues. A partir de ces critères, une liste de 10 gènes présentant des variations chez plusieurs filles est obtenue (Table D.5). Au total, 155 autres gènes présentent des variations retenues mais chez uniquement une seule fille.

Gène	Chromosome	# filles
MUC4	3	4
SYT3	19	2
KIAA0284	14	2
BAIAP2L2	22	2
CBX4	17	2
DNAH11	7	2
ZNHIT2	11	2
SIRT3	11	2
ARHGEF19	1	2
NOTCH1	9	2

TABLE D.5 – Liste des gènes présentant des variations chez plusieurs filles

La première observation faite à partir de cette liste est qu'aucun gène ne présente de variations *de novo* chez les 5 filles, le meilleur « score » est de 4 filles pour le gène MUC4. Le mode d'apparition de la maladie faisant penser à une liaison au chromosome X, il est également étonnant de ne retrouver aucun gène présent sur ce chromosome parmi cette liste.

Le gène MUC4 est particulièrement intéressant à étudier car 4 des filles présentent des variations *de novo* au sein de ce gène. La cinquième fille possède également des variations de novo dans ce gène mais ces variations ont un score de qualité inférieur à 20 et ne sont donc pas retenues. La table D.6 récapitule des variations observées pour les patientes F2 à F5. La Patiente F4 présente 13 variations *de novo* affectant l'exon 2 ce qui parait très élevé ; la patiente F3 présente 2 variations dans l'exon 2, les patientes F2 et F5 ne présentent qu'une mutation affectant toutes deux l'exon 2. Cependant, en vérifiant les variations observées avec Genome Browser, je me suis aperçu que 7 d'entre elles, et notamment l'unique variation de la fille 2, correspondait à des variations répertoriées dans dbSNP. Le pipeline DGVar n'a pas annoté ces variations car il utilise une version plus ancienne

de dbSNP pour détecter les variations connues.

Position1	Position2	Score	A	C	G	T	Used	Ref	Alt	AA	Fille	Exon
195 506 294	195 506 294	32.0	0	11	0	8	19	T	C	N/D	F4	2
195 506 302	195 506 302	36.3	0	0	4	16	20	G	T	P/H	F4	2
195 506 542	195 506 542	74.0	0	0	13	14	27	G	T	P/H	F4	2
195 506 597	195 506 597	54.0	41	0	90	0	131	G	A	P/S	F5	2
195 509 918	195 509 918	135.0	0	39	61	0	100	G	C	H/D	F4	2
195 509 954	195 509 954	145.0	0	39	17	0	56	G	C	L/V	F4	2
195 510 062	195 510 062	107.0	0	31	33	0	64	G	C	H/D	F4	2
195 510 073	195 510 073	112.0	37	0	108	0	145	G	A	A/V	F3	2
195 510 082	195 510 082	24.0	31	0	43	0	74	G	A	P/L	F4	2
195 510 767	195 510 767	56.0	93	0	160	0	253	G	A	P/S	F4	2
195 510 900	195 510 900	143.0	0	0	159	121	280	G	T	D/E	F4	2
195 510 943	195 510 943	186.0	0	0	145	133	278	G	T	S/Y	F4	2
195 511 076	195 511 076	34.1	4	0	0	1	5	T	A	T/S	F2	2
195 511 465	195 511 465	41.0	72	0	70	0	142	G	A	A/V	F4	2
195 511 690	195 511 690	56.0	0	235	512	0	747	G	C	T/S	F4	2
195 512 480	195 512 480	97.0	94	0	77	0	171	A	G	S/P	F3	2
195 515 122	195 515 122	188.0	0	108	219	0	327	G	C	T/S	F4	2

TABLE D.6 – Variations retenues pour le gène *MUC4*. Les variations en jaune correspondent à des SNPs recensés dans dbSNP mais détectées sur la version 37.4 du génome humain (la version 37.1 est utilisée par DGVar).

D.3.2.1.3 Hypothèse de 2 gènes complémentaires : Parmi les mutations affectant seulement une ou deux filles il nous a semblé intéressant de regrouper des groupes de 2 gènes dont les mutations de novo sont réparties chez les cinq patientes sans chevauchement, c'est-à-dire que chaque patiente n'est affectée au minimum et au maximum que pour l'un de ces gènes.

Méthodologie utilisée : Cette recherche se fait en décrivant chaque gène par un vecteur à 5 dimensions correspondant aux 5 filles. Si une fille présente une mutation dans un gène, la valeur du vecteur gène sera de 1 pour la dimension correspondant à la fille. Ainsi, dans le tableau D.7, les filles F1 et F3 sont toutes les 2 mutées dans le gène A. Il suffit ensuite de rechercher les gènes complémentaires.

Gène	Filles				
	F1	F2	F3	F4	F5
A	1	0	0	0	1
B	0	1	1	1	0
C	1	1	1	1	0

TABLE D.7 – Exemples de représentation de gènes mutés

Deux gènes sont considérés comme complémentaires, si à eux deux, ils couvrent les 5 filles et qu'aucune des filles ne présente de mutation dans les 2 gènes. Ainsi, en suivant les exemples du tableau 4, les gènes A et B sont complémentaires tandis que les A et C ne sont pas puisqu'ils tous deux mutés chez la fille F1.

Afin de faciliter la recherche des couples complémentaires, j'ai utilisé la kappa statistique qui permet de calculer un score de similarité entre deux vecteurs :

$$\kappa(A, B) = \frac{P_o - P_e}{1 - P_e} \tag{D.1}$$

Où, P_o est la probabilité d'accord entre A et B,

$$P_o = \frac{|A \cap B|}{k} \tag{D.2}$$

P_e est la probabilité d'accord aléatoire entre A et B,

$$P_e = \frac{|A_1| \times |B_1| + |A_0| \times |B_0|}{k^2} \tag{D.3}$$

A_0 est l'ensemble de filles non mutées dans le gène A,

B_0 est l'ensemble de filles non mutées dans le gène B,

A_1 est l'ensemble de filles mutées dans le gène A,

B_1 est l'ensemble de filles mutées dans le gène B et

k est le nombre de filles.

Le score calculé par la kappa statistique varie entre -1 et 1. Un score de 1 correspond à deux gènes mutés chez exactement les 2 mêmes filles et un score de -1 correspond à deux gènes complémentaires.

Ainsi, en calculant le score de similarité selon la kappa statistique pour tous les couples de gènes mutés et en ne sélectionnant que les couples obtenant un score de -1, on obtient les couples de gènes complémentaires.

Résultats : En utilisant les valeurs de filtres utilisées précédemment (profondeur minimale$\geq$4 et score $\geq$20) on obtient 1 couple complémentaire : *MUC4* (F3, F4 et F5) - *DNAH11* (F1 et F2). Quelques informations ont été collectées sur les gènes *MUC4* et *DNAH11* et permettent de renforcer l'intérêt pour l'étude et la validation de ce couple de gène comme couple candidat.

Le gène *MUC4* code la protéine "Mucin-4" qui n'est associée à aucun processus biologique KEGG et à aucune interaction connue dans IntAct. Elle est annotée par les termes GO BP "O-glycan processing", "cell-matrix adhesion" et "post-translational protein modification".

DNAH11 est à l'origine de la protéine "Dynein heavy chain 11, axonemal". Comme MUC4, il n'existe que peu d'information sur cette protéine dans les bases de données. En effet, elle n'intervient dans aucun processus biologique KEGG et dans aucune interaction IntAct. Elle est cependant annotée par les termes GO suivants : "ATP catabolic process", "ciliary or flagellar motility", "determination of left/right symmetry' et "microtubule-based movement".

Ces deux gènes semblent liés à la migration cellulaire. En effet, *MUC4* est associé à l'adhésion de la cellule à la matrice et *DNAH11* à la motilité ciliaire. La migration cellulaire ayant été associée

au syndrome d'Aicardi (Yilmaz *et al.*, 2009), ce couple de gène forme un bon candidat pour ce syndrome.

D.3.2.1.4 Hypothèse de 3 gènes complémentaires : Cette hypothèse consiste à étudier toutes les combinaisons de trois gènes (G2a, G2b, G1) affectant 2, 2 et 1 patientes parmi les cinq, de telle sorte que la réunion des ensembles de patientes affectées soit égale à l'ensemble des cinq patientes, et que les intersections des ensembles de patientes affectées soit toujours nulle. L'hypothèse sous-jacente est que le phénotype Aicardi pourrait survenir lorsqu'au moins l'un de ces trois gènes est affecté, ce qui pourrait s'expliquer par des interactions fortes entre ces trois gènes pour la réalisation d'une fonction cellulaire. Cette hypothèse audacieuse représente la dernière chance de tirer quelque chose des données de séquençage d'exome avant de se tourner vers d'autres types d'analyse comme le séquençage complet des génomes ou les analyses de type RNAseq.

Au total, 9 couples G2a, G2b soutiennent cette hypothèse (Table D.8. Ces couples sont complétés par par l'ensemble des gènes G1 qui sont spécifiques d'une seule fille. Les patientes F1, F2, F3, F4 et F5 ont respectivement 25, 26, 27, 45 et 32 gènes qui leurs sont spécifiques.

G2a	G2b	G1
ARHGEF19	CBX4	Gènes F5
ARHGEF19	SYT3	Gènes F4
ARHGEF19	NOTCH1	Gènes F4
BAIAP2L2	SYT3	Gènes F3
BAIAP2L2	NOTCH1	Gènes F3
SIRT3	SYT3	Gènes F3
SIRT3	NOTCH1	Gènes F3
ZNHIT2	CBX4	Gènes F3
KIAA0284	CBX4	Gènes F3

TABLE D.8 – Combinaisons de trois gènes (G2a, G2b, G1) affectant 2, 2 et 1 patientes parmi les cinq.

Couple *ARHGEF19-CBX4* : ARHGEF19 ne possède aucune interaction connue dans IntAct et n'intervient dans aucun processus biologique KEGG. Ses termes GO BP sont "positive regulation of Rho GTPase activity", "regulation of actin cytoskeleton organization" et "wound healing".

CBX4 est annoté par les termes GO BP "chromatin modification", "negative regulation of apoptotic process", "negative regulation of transcription from RNA polymerase II promoter", "protein sumoylation" et"transcription, DNA-dependent'". Les interactions de CBX4 sont présentées en figure D.5. CBX4 n'intervient dans aucun processus biologique KEGG.

Du fait qu'ARHGEF19 ne possède aucune interaction, il est difficile de relier ces deux protéines. Néanmoins, leurs termes GO suggèrent que l'une entre elle est liée à la régulation du cytosquelette d'actine et la seconde est intervient dans la modification de la chromatine. Ces deux fonctions étant fréquemment associées aux déficiences intellectuelles, ce couple de gène est un bon candidat pour le syndrome d'Aicardi.

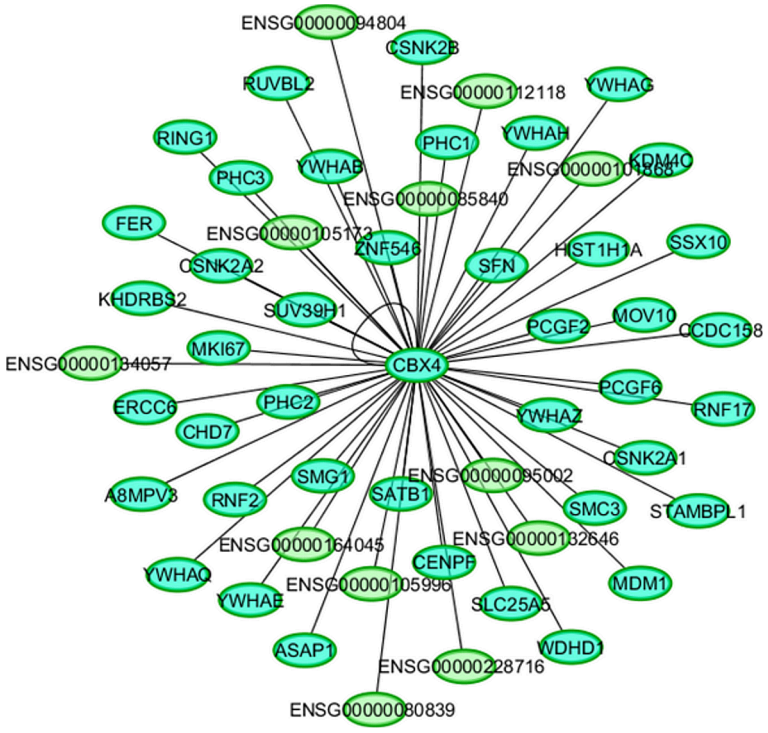


FIGURE D.5 – Interactions de CBX4

Couple *ARHGEF19-SYT3* : Il n'y a aucune information (pathway, interactions, termes GO BP) sur SYT3 dans les bases de données. Étant donnée l'absence d'informations sur SYT3, il est impossible de faire un lien entre les 2 protéines et la maladie.

Couple *ARHGEF19-NOTCH1* : NOTCH1 appartient aux processus biologiques KEGG "Dorso-ventral axis formation", "Notch signaling pathway" et "Prion diseases". Ses interactions sont présentées par la figure D.6. Cette protéines est annotées par les 95 termes GO BP suivants :

- ▷ "Notch receptor processing"
- ▷ "Notch signaling involved in heart development"
- ▷ "Notch signaling pathway involved in regulation of secondary heart field cardioblast proliferation"
- ▷ "anagen"
- ▷ "aortic valve morphogenesis"
- ▷ "arterial endothelial cell differentiation"
- ▷ "atrioventricular node development"
- ▷ "atrioventricular valve morphogenesis"
- ▷ "auditory receptor cell fate commitment"
- ▷ "axonogenesis"
- ▷ "branching morphogenesis of an epithelial tube"
- ▷ "cardiac chamber formation"
- ▷ "cardiac left ventricle morphogenesis"
- ▷ "cardiac muscle cell proliferation"
- ▷ "cardiac right atrium morphogenesis"
- ▷ "cardiac right ventricle formation"
- ▷ "cardiac vascular smooth muscle cell development"
- ▷ "cell fate specification"
- ▷ "cell migration involved in endocardial cushion formation"
- ▷ "cellular response to follicle-stimulating hormone stimulus"
- ▷ "cellular response to vascular endothelial

growth factor stimulus”
- ▷ “collecting duct development”
- ▷ “compartment pattern specification”
- ▷ “coronary artery morphogenesis”
- ▷ “coronary vein morphogenesis”
- ▷ “distal tubule development”
- ▷ “embryonic hindlimb morphogenesis”
- ▷ “endocardial cell differentiation”
- ▷ “endocardium morphogenesis”
- ▷ “endoderm development”
- ▷ “epithelial to mesenchymal transition involved in endocardial cushion formation”
- ▷ “forebrain development”
- ▷ “foregut morphogenesis”
- ▷ “glial cell differentiation”
- ▷ “glomerular mesangial cell development”
- ▷ “growth involved in heart morphogenesis”
- ▷ “hair follicle morphogenesis”
- ▷ “heart looping”
- ▷ “humoral immune response”
- ▷ “immune response”
- ▷ “in utero embryonic development”
- ▷ “inflammatory response to antigenic stimulus”
- ▷ “interleukin-4 secretion”
- ▷ “keratinocyte differentiation”
- ▷ “left/right axis specification”
- ▷ “liver development”
- ▷ “lung development”
- ▷ “mitral valve formation”
- ▷ “negative regulation of BMP signaling pathway”
- ▷ “negative regulation of anoikis”
- ▷ “negative regulation of calcium ion-dependent exocytosis”
- ▷ “negative regulation of canonical Wnt receptor signaling pathway”
- ▷ “negative regulation of cell migration involved in sprouting angiogenesis”
- ▷ “negative regulation of cell-substrate adhesion”
- ▷ “negative regulation of endothelial cell chemotaxis”
- ▷ “negative regulation of glial cell proliferation”
- ▷ “negative regulation of myoblast differentiation”
- ▷ “negative regulation of myotube differentiation”
- ▷ “negative regulation of oligodendrocyte differentiation”
- ▷ “negative regulation of ossification”
- ▷ “negative regulation of osteoblast differentiation”
- ▷ “negative regulation of photoreceptor cell differentiation”
- ▷ “negative regulation of pro-B cell differentiation”
- ▷ “negative regulation of stem cell differentiation”
- ▷ “negative regulation of transcription from RNA polymerase II promoter”
- ▷ “neural tube development”
- ▷ “neuronal stem cell maintenance”
- ▷ “pericardium morphogenesis”
- ▷ “positive regulation of BMP signaling pathway”
- ▷ “positive regulation of JAK-STAT cascade”
- ▷ “positive regulation of apoptotic process”
- ▷ “positive regulation of astrocyte differentiation”
- ▷ “positive regulation of cardiac muscle cell proliferation”
- ▷ “positive regulation of cell migration”
- ▷ “positive regulation of epithelial cell proliferation”
- ▷ “positive regulation of epithelial to mesenchymal transition”
- ▷ “positive regulation of keratinocyte differentiation”
- ▷ “positive regulation of transcription from RNA polymerase II promoter in response to hypoxia”
- ▷ “positive regulation of transcription of Notch receptor target”
- ▷ “prostate gland epithelium morphogene-

sis”
- ▷ “pulmonary valve morphogenesis”
- ▷ “regulation of epithelial cell proliferation involved in prostate gland development”
- ▷ “regulation of extracellular matrix assembly”
- ▷ “regulation of somitogenesis”
- ▷ “regulation of transcription from RNA polymerase II promoter involved in myocardial precursor cell differentiation”
- ▷ “response to muramyl dipeptide”
- ▷ “secretory columnal luminar epithelial cell differentiation involved in prostate glandular acinus development”
- ▷ “somatic stem cell division”
- ▷ “sprouting angiogenesis”
- ▷ “transcription initiation from RNA polymerase II promoter”
- ▷ “tube formation”
- ▷ “vasculogenesis involved in coronary vascular morphogenesis”
- ▷ “venous endothelial cell differentiation”
- ▷ “ventricular septum morphogenesis”
- ▷ “ventricular trabecula myocardium morphogenesis”

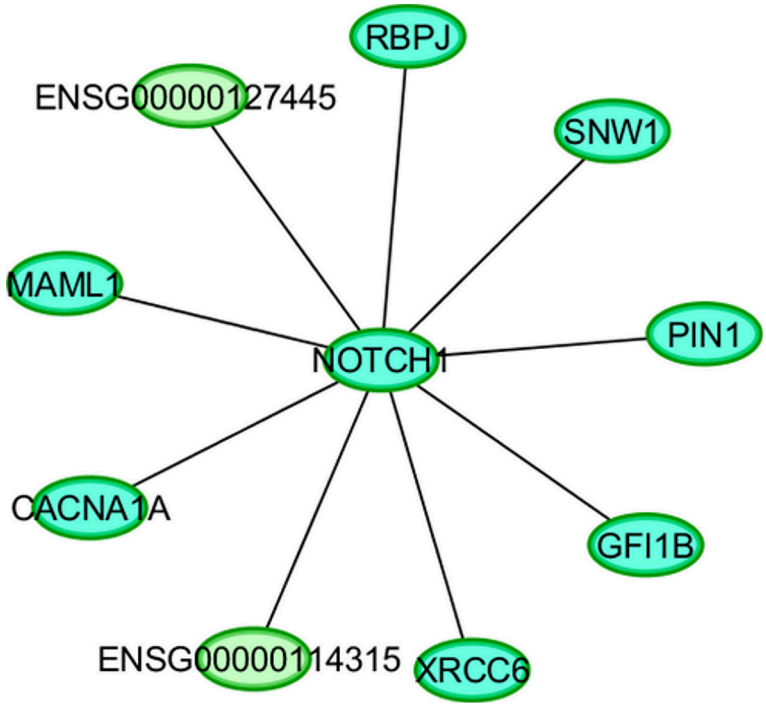


FIGURE D.6 – Interactions de NOTCH1

Il n’y a pas de lien évident entre les 2 protéines. NOTCH1 semble avoir un rôle particulièrement important dans le développement embryonnaire. Et ARHGEF19 est plutôt associée à la régulation du cytosquelette. De fait, ce couple de gène forme un assez bon couple candidat.

Couple *BAIAP2L2-SYT3* : BAIAP2L2 est annoté par les termes GO BP suivant : “cellular membrane organization”, “filopodium assembly” et “signal transduction”. BAIAP2L2 ne possède aucune interaction connue dans IntAct et n’intervient dans aucun pathway KEGG.

Étant donnée l’absence d’informations sur SYT3, il est impossible de faire un lien entre les 2 protéines et la maladie. Néanmoins, le fait que BAIAP2L2 intervienne dans l’assemblage des

filopodium est intéressant puisque les filopodium sont notamment impliqués dans la formation des axones.

Couple *BAIAP2L2-NOTCH1* : Ces 2 protéines semblent impliquées dans les processus de migrations cellulaire et forment donc un bon couple candidat.

Couple *SIRT3-SYT3* : SIRT3 est annotée par les termes GO BP “aerobic respiration”, “peptidyl-lysine deacetylation” et “protein ADP-ribosylation”. Cette protéine n’intervient dans aucun pathway KEGG. Ses interactions sont présentées par la figure D.7

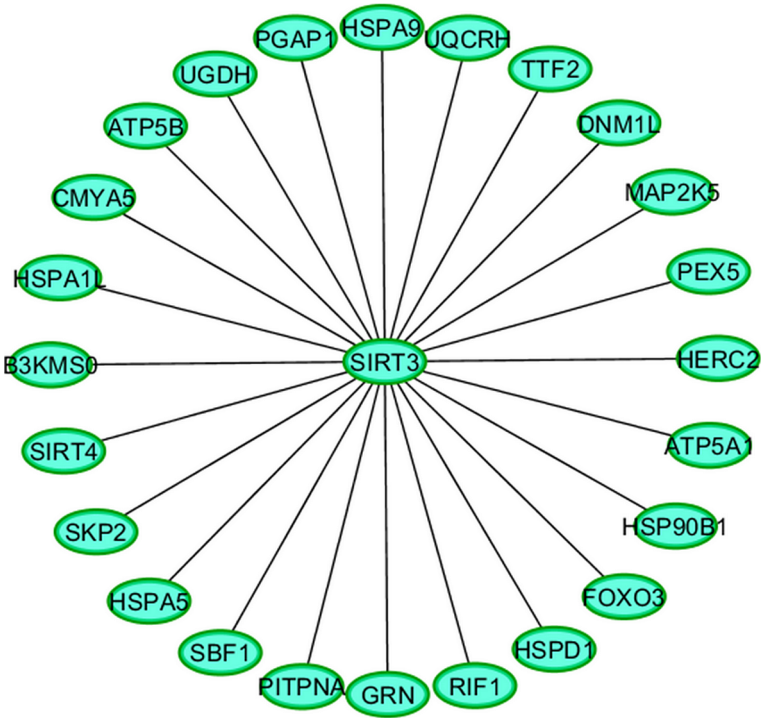


FIGURE D.7 – Interactions de SIRT3

Étant donnée l’absence d’informations sur SYT3, il est impossible de faire un lien entre les 2 protéines et la maladie.

Couple *SIRT3-NOTCH1* : Ces 2 protéines ne partagent pas d’interaction IntAct et les processus cellulaires dans lesquels elles interviennent sont différents. De plus les termes GO BP de SIRT3 ne semblent pas liés au syndrome d’Aicardi.

Couple *ZNHIT2-CBX4* : Il n’y aucune information (termes GO BP, interactions et pathway) sur ZNHIT2 dans les bases de données interrogées. Étant donnée l’absence d’informations sur ZNHIT2, il est impossible de faire un lien entre les 2 protéines et la maladie. Néanmoins, CBX4 est associé aux modification de la chromatine qui est un des mécanismes fréquements associés aux DILX.

Couple *KIAA0284-CBX4* : KIAA0284 n'est annotée par aucun terme GO et n'intervient dans aucun pathway KEGG. Sa seule interaction connue est avec PRKAA2, une Acetyl-CoA carboxylase kinase.

Ces 2 protéines ne possèdent donc pas d'interaction IntAct en commun et l'absence d'annotation de KIAA0284 empêche de faire un lien entre les 2 protéines et la maladie.

D.3.2.2 Conclusion

L'analyse des variations obtenues après réalignement des 15 exomes (5 trios fille malades- parents) a permis de détecter un couple de gènes complémentaires, composé de *MUC4* et *DNAH11*, qui permet de regrouper les 5 filles étudiées. Par ailleurs, l'analyse de ces données a également permis d'étudier les combinaisons de trois gènes (G2a, G2b, G1) affectant 2, 2 et 1 patientes parmi les cinq. Ainsi, 9 couples G2a et G2b ont pu être identifiés parmi lesquels certains sont de bon candidats car leurs fonctions sont similaires aux fonctions fréquemment associées aux DILX.

Bibliographie

Aerts,S., Lambrechts,D., Maity,S., Van Loo,P., Coessens,B., De Smet,F., Tranchevent,L.C., De Moor,B., Marynen,P., Hassan,B., Carmeliet,P. et Moreau,Y. (2006) Gene prioritization through genomic data fusion. *Nat. Biotechnol.,* **24** (5), 537–544.

Agapito,G., Guzzi,P.H. et Cannataro,M. (2013) Visualization of protein interaction networks : problems and solutions. *BMC Bioinformatics,* **14 Suppl 1**, S1.

Aicardi,J., Chevrie,J.J. et Rousselie,F. (1969) Spasma-in-flexion syndrome, callosal agenesis, chorioretinal abnormalities. *Arch. Fr. Pediatr.,* **26** (10), 1103–1120.

Aicardi,J., Lefebvre,J. et Lerique-Koechlin,A. (1965) A new syndrome : Spasms in flexion, callosal agenesis, ocular abnormalities. *Electroencephalogr. Clin. Neurophysiol.,* **19**, 609–610.

Alberts,B. (1998) The cell as a collection of protein machines : preparing the next generation of molecular biologists. *Cell,* **92** (3), 291–294.

Andrade,M.A., Perez-Iratxeta,C. et Ponting,C.P. (2001) Protein repeats : structures, functions, and evolution. *J. Struct. Biol.,* **134** (2-3), 117–131.

Apic,G., Ignjatovic,T., Boyer,S. et Russell,R.B. (2005) Illuminating drug discovery with biological pathways. *FEBS Lett.,* **579** (8), 1872–1877.

Atias,N. et Sharan,R. (2011) An algorithmic framework for predicting side effects of drugs. *J. Comput. Biol.,* **18** (3), 207–218.

Bakir-Gungor,B. et Sezerman,O.U. (2011) A new methodology to associate SNPs with human diseases according to their pathway related context. *PLoS ONE,* **6** (10), e26277.

Bandyopadhyay,S., Kelley,R. et Ideker,T. (2006) Discovering regulated networks during HIV-1 latency and reactivation. In *Pac Symp Biocomput* pp. 354–366.

Barabasi,A.L., Gulbahce,N. et Loscalzo,J. (2011) Network medicine : a network-based approach to human disease. *Nat. Rev. Genet.,* **12** (1), 56–68.

Becker,E., Robisson,B., Chapple,C.E., Guenoche,A. et Brun,C. (2012) Multifunctional proteins revealed by overlapping clustering in protein interaction network. *Bioinformatics,* **28** (1), 84–90.

Bell,L., Chowdhary,R., Liu,J.S., Niu,X. et Zhang,J. (2011) Integrated bio-entity network : a system for biological knowledge discovery. *PLoS ONE,* **6** (6), e21474.

Benabderrahmane,S., Smail-Tabbone,M., Poch,O., Napoli,A. et Devignes,M.D. (2010) IntelliGO : a new vector-based semantic similarity measure including annotation origin. *BMC Bioinformatics,* **11**, 588.

Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. et Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.,* **28** (1), 235–242.

Bhavani,S., Nagargadde,A., Thawani,A., Sridhar,V. et Chandra,N. (2006) Substructure-based support vector machine classifiers for prediction of adverse effects in diverse classes of drugs. *Journal of Chemical Information and Modeling,* **46** (6), 2478–2486.

Bizer,C., Heath,T. et Berners-Lee,T. (2009) Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems (IJSWIS),* **5** (3), 1–22.

Bolognesi,M.L. (2013) Polypharmacology in a single drug : multitarget drugs. *Curr. Med. Chem.,* .

Brandes,U., Eiglsperger,M., Herman,I., Himsolt,M. et Marshall,M.S. (2001) Graphml progress report. In *Graph Drawing* pp. 501–512.

Cami,A., Arnold,A., Manzi,S. et Reis,B. (2011) Predicting adverse drug events using pharmacological network models. *Sci Transl Med,* **3** (114), 114ra127.

Campillos,M., Kuhn,M., Gavin,A.C., Jensen,L.J. et Bork,P. (2008) Drug target identification using side-effect similarity. *Science,* **321** (5886), 263–266.

Carbon,S., Ireland,A., Mungall,C.J., Shu,S., Marshall,B., Lewis,S., Ireland,A., Lomax,J., Carbon,S., Mungall,C., Hitz,B., Balakrishnan,R., Dolan,M., Wood,V., Hong,E. et Gaudet,P. (2009) AmiGO : online access to ontology and annotation data. *Bioinformatics,* **25** (2), 288–289.

Chen,W., Stambolian,D., Edwards,A.O., Branham,K.E., Othman,M., Jakobsdottir,J., Tosakulwong,N., Pericak-Vance,M.A., Campochiaro,P.A., Klein,M.L., Tan,P.L., Conley,Y.P., Kanda,A., Kopplin,L., Li,Y., Augustaitis,K.J., Karoukis,A.J., Scott,W.K., Agarwal,A., Kovach,J.L., Schwartz,S.G., Postel,E.A., Brooks,M., Baratz,K.H., Brown,W.L., Brucker,A.J., Orlin,A., Brown,G., Ho,A., Regillo,C., Donoso,L., Tian,L., Kaderli,B., Hadley,D., Hagstrom,S.A., Peachey,N.S., Klein,R., Klein,B.E., Gotoh,N., Yamashiro,K., Ferris Iii,F., Fagerness,J.A., Reynolds,R., Farrer,L.A., Kim,I.K., Miller,J.W., Corton,M., Carracedo,A., Sanchez-Salorio,M., Pugh,E.W., Doheny,K.F., Brion,M., Deangelis,M.M., Weeks,D.E., Zack,D.J., Chew,E.Y., Heckenlively,J.R., Yoshimura,N., Iyengar,S.K., Francis,P.J., Katsanis,N., Seddon,J.M., Haines,J.L., Gorin,M.B., Abecasis,G.R., Swaroop,A., Johnson,R.N., Ai,E., McDonald,H.R., Stolarczuk,M., Pavan,P.R., Billiris,K.K., Iyer,M., Menosky,M.M., Pautler,S.E., Millard,S.M., Hubbard,B., Aaberg,T., DuBois,L., Lyon,A., Anderson-Nelson,S., Jampol,L.M., Weinberg,D.V., Munana,A., Rozenbajgier,Z., Orth,D., Cohen,J., MacCumber,M., Figliulo,C., Porcz,L., Folk,J., Boldt,H.C., Russell,S.R., Ivins,R., Hinz,C.J., Barr,C.C., Bloom,S., Jaegers,K., Kritchman,B., Whittington,G., Heier,J., Frederick,A.R., Morley,M.G., Topping,T., Davis,H.L., Bressler,S.B., Bressler,N.M., Doll,W., Trese,M., Capone,A., Garretson,B.R., Hassan,T.S., Ruby,A.J., Ostentoski,T., McCannel,C.A., Ruszcyzyk,M.J., Grand,G., Blinder,K., Holekamp,N.M., Joseph,D.P., Shah,G., Nobel,G.S., Antoszyk,A.N., Browning,D.J., Stallings,A.H., Singerman,L.J., Miller,D., Novak,M., Pendergast,S., Zegarra,H., Schura,S.A.,

Smith-Brewer,S., Davidorf,F.H., Chambers,R., Chorich,L., Salerno,J., Dreyer,R.F., Ma,C., Kopfer,M.R., Klein,M.L., Wilson,D.J., Nolte,S.K., Grunwald,J.E., Brucker,A.J., Dunaief,J., Fine,S.L., Maguire,A.M., Stoltz,R.A., McRay,M.N., Fish,G.E., Anand,R., Spencer,R., Arnwine,J., Chandra,S.R., Altaweel,M., Blodi,B., Gottlieb,J., Ip,M., Nork,T.M., Perry-Ramond,J., Fine,S.L., Maguire,M.G., Brightwell-Arnold,M., Harkins,S., Peskin,E., Ying,G.S. et Kurinij,N. (2010) Genetic variants near TIMP3 and high-density lipoprotein-associated loci influence susceptibility to age-related macular degeneration. *Proc. Natl. Acad. Sci. U.S.A.,* **107** (16), 7401–7406.

Chen,Y., Zhu,J., Lum,P.Y., Yang,X., Pinto,S., MacNeil,D.J., Zhang,C., Lamb,J., Edwards,S., Sieberts,S.K., Leonardson,A., Castellini,L.W., Wang,S., Champy,M.F., Zhang,B., Emilsson,V., Doss,S., Ghazalpour,A., Horvath,S., Drake,T.A., Lusis,A.J. et Schadt,E.E. (2008) Variations in DNA elucidate molecular networks that cause disease. *Nature,* **452** (7186), 429–435.

Cheng,F., Liu,C., Jiang,J., Lu,W., Li,W., Liu,G., Zhou,W., Huang,J. et Tang,Y. (2012) Prediction of drug-target interactions and drug repositioning via network-based inference. *PLoS Comput Biol,* **8** (5), e1002503.

Ciofani,M., Madar,A., Galan,C., Sellars,M., Mace,K., Pauli,F., Agarwal,A., Huang,W., Parkurst,C.N., Muratet,M., Newberry,K.M., Meadows,S., Greenfield,A., Yang,Y., Jain,P., Kirigin,F.K., Birchmeier,C., Wagner,E.F., Murphy,K.M., Myers,R.M., Bonneau,R. et Littman,D.R. (2012) A validated regulatory network for Th17 cell specification. *Cell,* **151** (2), 289–303.

Copley,R.R., Doerks,T., Letunic,I. et Bork,P. (2002) Protein domain analysis in the era of complete genomes. *FEBS Lett.,* **513** (1), 129–134.

Coutant,S., Cabot,C., Lefebvre,A., Leonard,M., Prieur-Gaston,E., Campion,D., Lecroq,T. et Dauchel,H. (2012) EVA : Exome Variation Analyzer, an efficient and versatile tool for filtering strategies in medical genomics. *BMC Bioinformatics,* **13 Suppl 14**, S9.

Curtis,R.E., Yin,J., Kinnaird,P. et Xing,E.P. (2012) Finding genome-transcriptome-phenome association with structured association mapping and visualization in genamap. pp. 327–338.

Demir,E., Cary,M.P., Paley,S., Fukuda,K., Lemer,C., Vastrik,I., Wu,G., D'Eustachio,P., Schaefer,C., Luciano,J., Schacherer,F., Martinez-Flores,I., Hu,Z., Jimenez-Jacinto,V., Joshi-Tope,G., Kandasamy,K., Lopez-Fuentes,A.C., Mi,H., Pichler,E., Rodchenkov,I., Splendiani,A., Tkachev,S., Zucker,J., Gopinath,G., Rajasimha,H., Ramakrishnan,R., Shah,I., Syed,M., Anwar,N., Babur,O., Blinov,M., Brauner,E., Corwin,D., Donaldson,S., Gibbons,F., Goldberg,R., Hornbeck,P., Luna,A., Murray-Rust,P., Neumann,E., Ruebenacker,O., Reubenacker,O., Samwald,M., van Iersel,M., Wimalaratne,S., Allen,K., Braun,B., Whirl-Carrillo,M., Cheung,K.H., Dahlquist,K., Finney,A., Gillespie,M., Glass,E., Gong,L., Haw,R., Honig,M., Hubaut,O., Kane,D., Krupa,S., Kutmon,M., Leonard,J., Marks,D., Merberg,D., Petri,V., Pico,A., Ravenscroft,D., Ren,L., Shah,N., Sunshine,M., Tang,R., Whaley,R., Letovksy,S., Buetow,K.H., Rzhetsky,A., Schachter,V., Sobral,B.S., Dogrusoz,U., McWeeney,S., Aladjem,M., Birney,E., Collado-Vides,J., Goto,S., Hucka,M., Le Novere,N., Maltsev,N., Pandey,A., Thomas,P., Wingender,E., Karp,P.D., Sander,C. et Bader,G.D. (2010) The BioPAX community standard for pathway data sharing. *Nat. Biotechnol.,* **28** (9), 935–942.

Derumeaux,G., Ernande,L., Serusclat,A., Servan,E., Bruckert,E., Rousset,H., Senn,S., Van Gaal,L., Picandet,B., Gavini,F. et Moulin,P. (2012) Echocardiographic evidence for valvular toxicity of benfluorex : a double-blind randomised trial in patients with type 2 diabetes mellitus. *PLoS ONE,* **7** (6), e38273.

Donnenfeld,A.E., Packer,R.J., Zackai,E.H., Chee,C.M., Sellinger,B. et Emanuel,B.S. (1989) Clinical, cytogenetic, and pedigree findings in 18 cases of Aicardi syndrome. *Am. J. Med. Genet.,* **32** (4), 461–467.

Emes,R.D. et Ponting,C.P. (2001) A new sequence motif linking lissencephaly, Treacher Collins and oral-facial-digital type 1 syndromes, microtubule dynamics and cell migration. *Hum. Mol. Genet.,* **10** (24), 2813–2820.

Finney,A. et Hucka,M. (2003) Systems biology markup language : Level 2 and beyond. *Biochem. Soc. Trans.,* **31** (Pt 6), 1472–1473.

Frachon,I., Etienne,Y., Jobic,Y., Le Gal,G., Humbert,M. et Leroyer,C. (2010) Benfluorex and unexplained valvular heart disease : a case-control study. *PLoS ONE,* **5** (4), e10128.

Freudenberg,J. et Propping,P. (2002) A similarity-based method for genome-wide prediction of disease-relevant human genes. *Bioinformatics,* **18 Suppl 2**, S110–115.

Ganesan,P., Garcia-Molina,H. et Widom,J. (2003) Exploiting hierarchical domain structure to compute similarity. *ACM Trans. Inf. Syst.,* **21** (1), 64–93.

Goh,K.I. et Choi,I.G. (2012) Exploring the human diseasome : the human disease network. *Brief Funct Genomics,* **11** (6), 533–542.

Harpaz,R., Chase,H.S. et Friedman,C. (2010) Mining multi-item drug adverse effect associations in spontaneous reporting systems. *BMC Bioinformatics,* **11 Suppl 9**, S7.

Hopkins,A.L. et Groom,C.R. (2002) The druggable genome. *Nat Rev Drug Discov,* **1** (9), 727–730.

Huang,d.a.W., Sherman,B.T. et Lempicki,R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc,* **4** (1), 44–57.

Huang,L.C., Wu,X. et Chen,J.Y. (2011) Predicting adverse side effects of drugs. *BMC Genomics,* **12 Suppl 5**, S11.

Ideker,T., Ozier,O., Schwikowski,B. et Siegel,A.F. (2002) Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics,* **18 Suppl 1**, S233–240.

International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature,* **431** (7011), 931–945.

Ito,T., Chiba,T., Ozawa,R., Yoshida,M., Hattori,M. et Sakaki,Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. U.S.A.,* **98** (8), 4569–4574.

Jeong,H., Mason,S.P., Barabasi,A.L. et Oltvai,Z.N. (2001) Lethality and centrality in protein networks. *Nature,* **411** (6833), 41–42.

Jiao,X., Chen,H., Chen,J., Herrup,K., Firestein,B.L. et Kiledjian,M. (2009) Modulation of neuritogenesis by a protein implicated in X-linked mental retardation. *J. Neurosci.,* **29** (40), 12419–12427.

Joshi-Tope,G., Gillespie,M., Vastrik,I., D'Eustachio,P., Schmidt,E., de Bono,B., Jassal,B., Gopinath,G.R., Wu,G.R., Matthews,L., Lewis,S., Birney,E. et Stein,L. (2005) Reactome : a knowledgebase of biological pathways. *Nucleic Acids Res.,* **33** (Database issue), D428–432.

Kajiwara,K., Berson,E.L. et Dryja,T.P. (1994) Digenic retinitis pigmentosa due to mutations at the unlinked peripherin/RDS and ROM1 loci. *Science,* **264** (5165), 1604–1608.

Karaboga,A.S., Petronin,F., Marchetti,G., Souchet,M. et Maigret,B. (2013) Benchmarking of HPCC : A novel 3D molecular representation combining shape and pharmacophoric descriptors for efficient molecular similarity assessments. *J. Mol. Graph. Model.,* **41**, 20–30.

Kerrien,S., Aranda,B., Breuza,L., Bridge,A., Broackes-Carter,F., Chen,C., Duesbury,M., Dumousseau,M., Feuermann,M., Hinz,U., Jandrasits,C., Jimenez,R.C., Khadake,J., Mahadevan,U., Masson,P., Pedruzzi,I., Pfeiffenberger,E., Porras,P., Raghunath,A., Roechert,B., Orchard,S. et Hermjakob,H. (2012) The IntAct molecular interaction database in 2012. *Nucleic Acids Res.,* **40** (Database issue), D841–846.

Knobbe,A., Crémilleux,B., Fürnkranz,J. et Scholz,M. (2008). From local patterns to global models : the lego approach to data mining.

Knox,C., Law,V., Jewison,T., Liu,P., Ly,S., Frolkis,A., Pon,A., Banco,K., Mak,C., Neveu,V., Djoumbou,Y., Eisner,R., Guo,A.C. et Wishart,D.S. (2011) DrugBank 3.0 : a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.,* **39** (Database issue), D1035–1041.

Kohler,J., Baumbach,J., Taubert,J., Specht,M., Skusa,A., Ruegg,A., Rawlings,C., Verrier,P. et Philippi,S. (2006) Graph-based analysis and visualization of experimental results with ONDEX. *Bioinformatics,* **22** (11), 1383–1390.

Kuhn,M., Campillos,M., Letunic,I., Jensen,L.J. et Bork,P. (2010) A side effect resource to capture phenotypic effects of drugs. *Mol. Syst. Biol.,* **6**, 343.

Laget,S., Joulie,M., Le Masson,F., Sasai,N., Christians,E., Pradhan,S., Roberts,R.J. et Defossez,P.A. (2010) The human proteins MBD5 and MBD6 associate with heterochromatin but they do not bind methylated DNA. *PLoS ONE,* **5** (8), e11982.

Laguri,C., Duband-Goulet,I., Friedrich,N., Axt,M., Belin,P., Callebaut,I., Gilquin,B., Zinn-Justin,S. et Couprie,J. (2008) Human mismatch repair protein MSH6 contains a PWWP domain that targets double stranded DNA. *Biochemistry,* **47** (23), 6199–6207.

Lamb,J., Crawford,E.D., Peck,D., Modell,J.W., Blat,I.C., Wrobel,M.J., Lerner,J., Brunet,J.P., Subramanian,A., Ross,K.N., Reich,M., Hieronymus,H., Wei,G., Armstrong,S.A., Haggarty,S.J., Clemons,P.A., Wei,R., Carr,S.A., Lander,E.S. et Golub,T.R. (2006) The Connectivity Map : using

gene-expression signatures to connect small molecules, genes, and disease. *Science,* **313** (5795), 1929–1935.

Lee,S., Lee,K.H., Song,M. et Lee,D. (2011) Building the process-drug-side effect network to discover the relationship between biological processes and side effects. *BMC Bioinformatics,* **12 Suppl 2**, S2.

Li,C., Donizelli,M., Rodriguez,N., Dharuri,H., Endler,L., Chelliah,V., Li,L., He,E., Henry,A., Stefan,M.I., Snoep,J.L., Hucka,M., Le Novère,N. et Laibe,C. (2010) BioModels Database : An enhanced, curated and annotated resource for published quantitative kinetic models. *BMC Systems Biology,* **4**, 92.

Li,Q. et Lai,L. (2007) Prediction of potential drug targets based on simple sequence properties. *BMC Bioinformatics,* **8**, 353.

Liu,M., Matheny,M.E., Hu,Y. et Xu,H. (2012*a*) Data mining methodologies for pharmacovigilance. *SIGKDD Explor. Newsl.,* **14** (1), 35–42.

Liu,M., Wu,Y., Chen,Y., Sun,J., Zhao,Z., Chen,X.W., Matheny,M.E. et Xu,H. (2012*b*) Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs. *J Am Med Inform Assoc,* **19** (e1), 28–35.

Ma'ayan,A., Jenkins,S.L., Goldfarb,J. et Iyengar,R. (2007) Network analysis of FDA approved drugs and their targets. *Mt. Sinai J. Med.,* **74** (1), 27–32.

Maere,S., Heymans,K. et Kuiper,M. (2005) BiNGO : a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics,* **21** (16), 3448–3449.

Martin,A., Ochagavia,M.E., Rabasa,L.C., Miranda,J., Fernandez-de Cossio,J. et Bringas,R. (2010) BisoGenet : a new tool for gene network building, visualization and analysis. *BMC Bioinformatics,* **11**, 91.

Martin,Y.C., Kofron,J.L. et Traphagen,L.M. (2002) Do structurally similar molecules have similar biological activity ? *J. Med. Chem.,* **45** (19), 4350–4358.

Mateja,A., Cierpicki,T., Paduch,M., Derewenda,Z.S. et Otlewski,J. (2006) The dimerization mechanism of LIS1 and its implication for proteins containing the LisH motif. *J. Mol. Biol.,* **357** (2), 621–631.

Merrill,G.H. (2008) The MedDRA paradox. In *AMIA Annu Symp Proc* pp. 470–474.

Molina,J.A., Mateos,F., Merino,M., Epifanio,J.L. et Gorrono,M. (1989) Aicardi syndrome in two sisters. *J. Pediatr.,* **115** (2), 282–283.

Muggleton,S., Srinivasan,A., King,R.D. et Sternberg,M.J.E. (1998) Biochemical knowledge discovery using inductive logic programming. In *Discovery Science* pp. 326–341.

Murzin,A.G., Brenner,S.E., Hubbard,T. et Chothia,C. (1995) SCOP – a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology,* **247** (4), 536–540.

Ndiaye,B., Bresso,E., Smaïl-Tabbone,M., Souchet,M. et Devignes,M.D. (2011). Modim : model-driven data integration for mining. Journées Ouvertes de Biologie, Informatique et Mathematiques 2011. www.pasteur.fr/ip/resource/filecenter/document/01s-00004f-0el/abstract-259.pdf.

Neale,B.M., Fagerness,J., Reynolds,R., Sobrin,L., Parker,M., Raychaudhuri,S., Tan,P.L., Oh,E.C., Merriam,J.E., Souied,E., Bernstein,P.S., Li,B., Frederick,J.M., Zhang,K., Brantley,M.A., Lee,A.Y., Zack,D.J., Campochiaro,B., Campochiaro,P., Ripke,S., Smith,R.T., Barile,G.R., Katsanis,N., Allikmets,R., Daly,M.J. et Seddon,J.M. (2010) Genome-wide association study of advanced age-related macular degeneration identifies a role of the hepatic lipase gene (LIPC). *Proc. Natl. Acad. Sci. U.S.A.,* **107** (16), 7395–7400.

Ogata,H., Goto,S., Fujibuchi,W. et Kanehisa,M. (1998) Computation with the KEGG pathway database. *BioSystems,* **47** (1-2), 119–128.

Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B. et Thornton,J.M. (1997) CATH - a hierarchic classification of protein domain structures. *Structure,* **5** (8), 1093–1108.

Overbeek,R., Fonstein,M., D'Souza,M., Pusch,G.D. et Maltsev,N. (1999) Use of contiguity on the chromosome to predict functional coupling. *In Silico Biol. (Gedrukt),* **1** (2), 93–108.

Page,D. et Craven,M. (2003) Biological applications of multi-relational data mining. In *SIGKDD Explorations* pp. 69–79.

Pang,K., Sheng,H. et Ma,X. (2010) Understanding gene essentiality by finely characterizing hubs in the yeast protein interaction network. *Biochem. Biophys. Res. Commun.,* **401** (1), 112–116.

Pauwels,E., Stoven,V. et Yamanishi,Y. (2011) Predicting drug side-effect profiles : a chemical fragment-based approach. *BMC Bioinformatics,* **12** (1), 169.

Pavlopoulos,G.A., Wegener,A.L. et Schneider,R. (2008) A survey of visualization tools for biological network analysis. *BioData mining,* **1**, 12.

Piro,R.M. et Di Cunto,F. (2012) Computational approaches to disease-gene prediction : rationale, classification and successes. *FEBS J.,* **279** (5), 678–696.

Puig,O., Caspary,F., Rigaut,G., Rutz,B., Bouveret,E., Bragado-Nilsson,E., Wilm,M. et Seraphin,B. (2001) The tandem affinity purification (TAP) method : a general procedure of protein complex purification. *Methods,* **24** (3), 218–229.

Pujol,A., Mosca,R., Farres,J. et Aloy,P. (2010) Unveiling the role of network and systems biology in drug discovery. *Trends Pharmacol. Sci.,* **31** (3), 115–123.

Quevillon,E., Silventoinen,V., Pillai,S., Harte,N., Mulder,N., Apweiler,R. et Lopez,R. (2005) InterProScan : protein domains identifier. *Nucleic Acids Res.,* **33** (Web Server issue), W116–120.

Riazati,D. et Thom,J.A. (2011) Matching star schemas. In *DEXA (2)* pp. 428–438.

Ricci,F., Staurenghi,G., Lepre,T., Missiroli,F., Zampatti,S., Cascella,R., Borgiani,P., Marsella,L.T., Eandi,C.M., Cusumano,A., Novelli,G. et Giardina,E. (2013) Haplotypes in IL-8 Gene Are Associated to Age-Related Macular Degeneration : A Case-Control Study. *PLoS ONE,* **8** (6), e66978.

Riviere,J.B., van Bon,B.W., Hoischen,A., Kholmanskikh,S.S., O'Roak,B.J., Gilissen,C., Gijsen,S., Sullivan,C.T., Christian,S.L., Abdul-Rahman,O.A., Atkin,J.F., Chassaing,N., Drouin-Garraud,V., Fry,A.E., Fryns,J.P., Gripp,K.W., Kempers,M., Kleefstra,T., Mancini,G.M., Nowaczyk,M.J., van Ravenswaaij-Arts,C.M., Roscioli,T., Marble,M., Rosenfeld,J.A., Siu,V.M., de Vries,B.B., Shendure,J., Verloes,A., Veltman,J.A., Brunner,H.G., Ross,M.E., Pilz,D.T. et Dobyns,W.B. (2012) De novo mutations in the actin genes ACTB and ACTG1 cause Baraitser-Winter syndrome. *Nat. Genet.,* **44** (4), 440–444.

Ropers,H.H. (2010) Genetics of early onset cognitive impairment. *Annu Rev Genomics Hum Genet,* **11**, 161–187.

Rual,J.F., Venkatesan,K., Hao,T., Hirozane-Kishikawa,T., Dricot,A., Li,N., Berriz,G.F., Gibbons,F.D., Dreze,M., Ayivi-Guedehoussou,N., Klitgord,N., Simon,C., Boxem,M., Milstein,S., Rosenberg,J., Goldberg,D.S., Zhang,L.V., Wong,S.L., Franklin,G., Li,S., Albala,J.S., Lim,J., Fraughton,C., Llamosas,E., Cevik,S., Bex,C., Lamesch,P., Sikorski,R.S., Vandenhaute,J., Zoghbi,H.Y., Smolyar,A., Bosak,S., Sequerra,R., Doucette-Stamm,L., Cusick,M.E., Hill,D.E., Roth,F.P. et Vidal,M. (2005) Towards a proteome-scale map of the human protein-protein interaction network. *Nature,* **437** (7062), 1173–1178.

Sakharkar,M.K., Li,P., Zhong,Z. et Sakharkar,K.R. (2008) Quantitative analysis on the characteristics of targets with FDA approved drugs. *Int. J. Biol. Sci.,* **4** (1), 15–22.

Santos,J.C., Nassif,H., Page,D., Muggleton,S.H. et E Sternberg,M.J. (2012) Automated identification of protein-ligand interaction features using Inductive Logic Programming : a hexose binding case study. *BMC Bioinformatics,* **13**, 162.

Schaefer,C.F., Anthony,K., Krupa,S., Buchoff,J., Day,M., Hannay,T. et Buetow,K.H. (2009) PID : the Pathway Interaction Database. *Nucleic Acids Res.,* **37** (Database issue), D674–679.

Schaffer,A.A. (2013) Digenic inheritance in medical genetics. *J. Med. Genet.,* .

Scheiber,J., Chen,B., Milik,M., Sukuru,S.C., Bender,A., Mikhailov,D., Whitebread,S., Hamon,J., Azzaoui,K., Urban,L., Glick,M., Davies,J.W. et Jenkins,J.L. (2009*a*) Gaining insight into off-target mediated effects of drug candidates with a comprehensive systems chemical biology analysis. *J Chem Inf Model,* **49** (2), 308–317.

Scheiber,J., Jenkins,J.L., Sukuru,S.C.K., Bender,A., Mikhailov,D., Milik,M., Azzaoui,K., Whitebread,S., Hamon,J., Urban,L., Glick,M. et Davies,J.W. (2009*b*) Mapping adverse drug reactions in chemical space. *Journal of Medicinal Chemistry,* **52** (9), 3103–3107.

Secrier,M., Pavlopoulos,G.A., Aerts,J. et Schneider,R. (2012) Arena3D : visualizing time-driven phenotypic differences in biological systems. *BMC Bioinformatics,* **13**, 45.

Smoot,M.E., Ono,K., Ruscheinski,J., Wang,P.L. et Ideker,T. (2011) Cytoscape 2.8 : new features for data integration and network visualization. *Bioinformatics (Oxford, England),* **27** (3), 431–432.

Stuart,J.M., Segal,E., Koller,D. et Kim,S.K. (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science,* **302** (5643), 249–255.

Sutton,V.R., Hopkins,B.J., Eble,T.N., Gambhir,N., Lewis,R.A. et Van den Veyver,I.B. (2005) Facial and physical features of Aicardi syndrome : infants to teenagers. *Am. J. Med. Genet. A,* **138A** (3), 254–258.

Takarabe,M., Kotera,M., Nishimura,Y., Goto,S. et Yamanishi,Y. (2012) Drug target prediction using adverse event report systems : a pharmacogenomic approach. *Bioinformatics,* **28** (18), i611–i618.

Tatonetti,N.P., Ye,P.P., Daneshjou,R. et Altman,R.B. (2012) Data-driven prediction of drug effects and interactions. *Sci Transl Med,* **4** (125), 125ra31.

Taubert,J., Sieren,K.P., Hindle,M., Hoekman,B., Winnenburg,R., Philippi,S., Rawlings,C.J. et Köhler,J. (2007) The oxl format for the exchange of integrated datasets. *J. Integrative Bioinformatics,* .

The UniProt Consortium (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.,* **40** (Database issue), D71–75.

Theocharidis,A., van Dongen,S., Enright,A.J. et Freeman,T.C. (2009) Network visualization and analysis of gene expression data using BioLayout Express(3D). *Nat Protoc,* **4** (10), 1535–1550.

Tranchevent,L.C., Barriot,R., Yu,S., Van Vooren,S., Van Loo,P., Coessens,B., De Moor,B., Aerts,S. et Moreau,Y. (2008) ENDEAVOUR update : a web resource for gene prioritization in multiple species. *Nucleic Acids Res.,* **36** (Web Server issue), W377–384.

van Bokhoven,H. (2011) Genetic and epigenetic networks in intellectual disabilities. *Annu. Rev. Genet.,* **45**, 81–104.

Venkatesan,K., Rual,J.F., Vazquez,A., Stelzl,U., Lemmens,I., Hirozane-Kishikawa,T., Hao,T., Zenkner,M., Xin,X., Goh,K.I., Yildirim,M.A., Simonis,N., Heinzmann,K., Gebreab,F., Sahalie,J.M., Cevik,S., Simon,C., de Smet,A.S., Dann,E., Smolyar,A., Vinayagam,A., Yu,H., Szeto,D., Borick,H., Dricot,A., Klitgord,N., Murray,R.R., Lin,C., Lalowski,M., Timm,J., Rau,K., Boone,C., Braun,P., Cusick,M.E., Roth,F.P., Hill,D.E., Tavernier,J., Wanker,E.E., Barabasi,A.L. et Vidal,M. (2009) An empirical framework for binary interactome mapping. *Nat. Methods,* **6** (1), 83–90.

Vissers,L.E., de Ligt,J., Gilissen,C., Janssen,I., Steehouwer,M., de Vries,P., van Lier,B., Arts,P., Wieskamp,N., del Rosario,M., van Bon,B.W., Hoischen,A., de Vries,B.B., Brunner,H.G. et Veltman,J.A. (2010) A de novo paradigm for mental retardation. *Nat. Genet.,* **42** (12), 1109–1112.

Vogt,I. et Mestres,J. (2010) Drug-target networks. *Molecular Informatics,* **29** (1-2), 10–14.

von Eichborn,J., Dunkel,M., Gohlke,B.O., Preissner,S.C., Hoffmann,M.F., Bauer,J.M., Armstrong,J.D., Schaefer,M.H., Andrade-Navarro,M.A., Le Novere,N., Croning,M.D., Grant,S.G., van Nierop,P., Smit,A.B. et Preissner,R. (2013) SynSysNet : integration of experimental data on synaptic protein-protein interactions with drug-target relations. *Nucleic Acids Res.,* **41** (Database issue), D834–840.

Wagner,A.H., Taylor,K.R., Deluca,A.P., Casavant,T.L., Mullins,R.F., Stone,E.M., Scheetz,T.E. et Braun,T.A. (2013) Prioritization of Retinal Disease Genes : An Integrative Approach. *Hum. Mutat.,* .

Walhout,A.J., Sordella,R., Lu,X., Hartley,J.L., Temple,G.F., Brasch,M.A., Thierry-Mieg,N. et Vidal,M. (2000) Protein interaction mapping in C. elegans using proteins involved in vulval development. *Science,* **287** (5450), 116–122.

Wang,Y., Reddy,B., Thompson,J., Wang,H., Noma,K., Yates,J.R. et Jia,S. (2009) Regulation of Set9-mediated H4K20 methylation by a PWWP domain protein. *Mol. Cell,* **33** (4), 428–437.

Wu,Z., Wang,Y. et Chen,L. (2013) Network-based drug repositioning. *Mol Biosyst,* .

Yamanishi,Y., Pauwels,E. et Kotera,M. (2012) Drug side-effect prediction based on the integration of chemical and biological spaces. *J Chem Inf Model,* **52** (12), 3284–3292.

Yanai,I., Derti,A. et DeLisi,C. (2001) Genes linked by fusion events are generally of the same functional category : a systematic analysis of 30 microbial genomes. *Proc. Natl. Acad. Sci. U.S.A.,* **98** (14), 7940–7945.

Yao,L. et Rzhetsky,A. (2008) Quantitative systems-level determinants of human genes targeted by successful drugs. *Genome Res.,* **18** (2), 206–213.

Yildirim,M.A., Goh,K.I., Cusick,M.E., Barabasi,A.L. et Vidal,M. (2007) Drug-target network. *Nat. Biotechnol.,* **25** (10), 1119–1126.

Yilmaz,S., Jonveaux,P., Bicep,C., Pierron,L., Smail-Tabbone,M. et Devignes,M.D. (2009) Gene-disease relationship discovery based on model-driven data integration and database view definition. *Bioinformatics,* **25** (2), 230–236.

Yilmaz, S. (2007). *Recherche de gènes candidats responsables du Syndrome d'Aicardi : Complémentarité des approches expérimentales et bioinformatiques*. PhD thesis, Université Henri Poincaré.

Zaoui,K., Benseddik,K., Daou,P., Salaun,D. et Badache,A. (2010) ErbB2 receptor controls microtubule capture by recruiting ACF7 to the plasma membrane of migrating cells. *Proc. Natl. Acad. Sci. U.S.A.,* **107** (43), 18517–18522.

Résumé

La compréhension des pathologies humaines et du mode d'action des médicaments passe par la prise en compte de multiples réseaux d'interactions entre biomolécules (appelés ici réseaux biologiques). Les recherches récentes sur les systèmes biologiques conduisent à produire de plus en plus de données, expérimentales ou inférées automatiquement, sur ces réseaux qui gouvernent les processus cellulaires. Cependant, l'hétérogénéité et la multiplicité croissantes de ces données rendent difficile l'intégration des réseaux biologiques dans les raisonnements des utilisateurs finaux que sont par exemple les cliniciens, confrontés à des phénotypes atypiques, ou les industriels de la pharmacie, désireux de comprendre les effets secondaires de certains médicaments. Dans cette thèse sont proposées des approches intégratives, guidées par les connaissances du domaine, et mettant en œuvre des techniques informatiques de gestion de données, de visualisation de graphes et de fouille de données, pour tenter de répondre au problème de l'exploitation insuffisante des données sur les réseaux biologiques dans la compréhension des phénotypes complexes associés aux maladies génétiques ou des effets secondaires des médicaments.

Le fondement commun et structurant du travail de cette thèse consiste en un entrepôt de données générique, NetworkDB. Cet entrepôt repose sur un modèle de données intégrant les protéines, leurs annotations fonctionnelles, leurs domaines, leurs interactions les unes avec les autres et les pathways dans lesquelles elles interviennent. Pour les applications choisies, ce modèle de données est étendu par les entités appropriées. L'entrepôt est peuplé de manière semi-automatique à partir de bases de données publiques telles qu'UniProt et en utilisant un système d'adaptateurs paramétrables développé au laboratoire (MODIM).

La première approche a pour but de faciliter l'accès aux réseaux biologiques pour l'étude de l'étiologie des maladies génétiques, notamment celles qui entraînent une déficience intellectuelle. L'entrepôt NetworkDB a été étendu aux gènes et à leurs mutations, et couplé à un système de visualisation de graphes interactifs (ONDEX). Ceci a permis aux cliniciens du Laboratoire de Génétique du CHU de Nancy de mieux comprendre certains phénotypes associés à des mutations dans des gènes peu connus. J'ai pu ensuite identifier et caractériser des sous-réseaux de gènes à partir de l'exploration systématique des réseaux associés à une centaine de gènes impliqués dans des déficiences intellectuelles liées à l'X.

La seconde approche concerne l'exploitation des réseaux biologiques pour la découverte de connaissances, en vue de mieux comprendre les effets secondaires des médicaments. Ici, l'entrepôt NetworkDB a été étendu aux médicaments, à leurs catégories, structures et effets secondaires, en lien avec les protéines cibles de ces médicaments. Une étape indispensable a consisté à regrouper en cluster de termes les effets secondaires similaires. Ces clusters de termes ont été associés aux médicaments à partir des données de la base SIDER. Ensuite, des profils combinant plusieurs clusters de termes partagés par les mêmes médicaments ont été extraits de NetworkDB. J'ai alors appliqué une méthode de fouille de données relationnelles pour caractériser ces profils. Les résultats se présentent sous forme de règles permettant de décrire quelles propriétés des médicaments et de leurs cibles sont préférentiellement associées à tel ou tel profil d'effets secondaires. Ces règles ont été proposées à l'entreprise Harmonic Pharma, partenaire industriel du contrat CIFRE support de cette thèse et spécialiste du repositionnement moléculaire, afin de lui permettre d'anticiper précocement l'apparition d'effets secondaires graves. Au-delà des applications traitées, les approches développées dans cette thèse contribuent à répondre au besoin croissant d'appréhender les phénomènes biologiques de façon intégrative, dans une logique de système biologique.

Mots-clés: réseaux d'interactions, représentation des connaissances, compréhension des effets secondaires, relations génotype-phénotype

Abstract

The understanding of human diseases and drug mechanisms requires today to take into account molecular interaction networks. Recent studies on biological systems are producing increasing amounts of data using experimental or computational methods. However, complexity and heterogeneity of these datasets make it difficult to exploit them adequately for understanding atypical phenotypes or drug side-effects. This thesis presents two knowledge-based integrative approaches that combine data management, graph visualization and data mining techniques in order to improve our understanding of complex phenotypes associated with genetic diseases or drug side-effects.

The first contribution of this thesis is a generic data warehouse, NetworkDB. This warehouse is based on a data model incorporating proteins, their functional annotations, their domains, their interactions and the pathways in which they operate. For selected applications, this data model is extended by the appropriate entities. The warehouse is semi-automatically populated from public databases such as UniProt, Gene Ontology, InterPro and KEGG Pathways, using a generic adapter system developed in the laboratory (MODIM for Model-Driven Database Integration for Mining).

The second contribution of this thesis is an approach based on graph visualization. This approach aims to facilitate access to biological networks in order to study genetic disease etiology, including X-linked intellectual disability (XLID). The NetworkDB warehouse was extended with appropriate sets of genes and their mutations and coupled with a visualization system for interactive graphs (ONDEX). This allowed the clinicians from the Genetics Laboratory at the University Hospital of Nancy to formulate new hypotheses about the association of mutations in certain genes with complex phenotypes. Moreover, I identified and characterized meaningful sub-networks of genes through systematic exploration of all molecular networks involving the genes associated with XLID.

The third contribution of this thesis concerns the use of biological networks for knowledge discovery, in order to better understand drug side-effects. Here, the NetworkDB warehouse was extended with a set of drugs, their categories, structures, side effects (as listed in the SIDER database) and with their target proteins. An essential step here was to cluster together the terms describing similar side effects on the basis of MedDRA terminology. These term clusters were then associated with drugs and side-effect profiles, defined as patterns of term clusters shared by the same drugs, were extracted from that binary table. A relational learning procedure, based on inductive logic programming, was set up in order to characterize these profiles. The resulting rules indicate which properties of drugs and their targets preferentially associate with a particular side-effect profile. These rules have been proposed, through a user-friendly interface, to Harmonic Pharma, the industrial partner of this thesis, expert in molecular repositioning, to help them anticipate early onset of serious side effects.

The approaches developed in this thesis constitute an answer to the growing need of clinicians and biologists of exploiting information about molecular networks when addressing complex dedicated questions from a systems biology point of view.

Keywords: Interaction networks, Knowledge representation, Understanding of drug side-effects, Genotype-phenotype relationships