HAL
archives-ouvertes.fr

Apport des ontologies de domaine pour l'extraction de connaissances à partir de données biomédicales

Gabin Personeni

▶ **To cite this version:**

Gabin Personeni. Apport des ontologies de domaine pour l'extraction de connaissances à partir de données biomédicales. Apprentissage [cs.LG]. Université de Lorraine, 2018. Français. NNT : 2018LORR0235 . tel-01925461

HAL Id: tel-01925461
https://hal.inria.fr/tel-01925461

Submitted on 16 Nov 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITÉ DE LORRAINE

École doctorale IAEM Lorraine

Apport des ontologies de domaine pour l'extraction de connaissances à partir de données biomédicales

THÈSE

présentée et soutenue publiquement le 09 Novembre 2018

pour l'obtention du

Doctorat de l'Université de Lorraine

(mention informatique)

par

Gabin Personeni

Composition du jury

Rapporteurs :	Olivier Dameron	Maître de Conférences, Université de Rennes 1
	Céline Rouveirol	Professeur, Université Paris 13
Examinateurs :	Jérôme Azé	Professeur, Université de Montpellier
	Anne Boyer	Professeur, Université de Lorraine
	Adrien Coulet	Maître de Conférences, Université de Lorraine
	Marie-Dominique Devignes	Chargée de Rercherches, CNRS
Invités :	Michel Dumontier	Distinguished Professor, Maastricht University
	Malika Smaïl-Tabbone	Maître de conférences, Université de Lorraine

Laboratoire Lorrain de Recherche en Informatique et ses Applications — UMR 7503

Mis en page avec la classe thesul.

Remerciements

Je tiens d'abord à remercier les rapporteurs et examinateurs pour avoir accepté de participer au jury de cette thèse.

Je tiens à remercier Marie-Dominique Devignes, chargée de recherches au CNRS et Adrien Coulet, maître de conférences à l'Université de Lorraine pour leur encadrement et leur soutien tout au long de cette thèse.

Je remercie Dave Ritchie pour m'avoir accueilli dans l'équipe CAPSID. Je tiens également à remercier tous les membres et anciens membres de l'équipe pour leur accueil.

Je remercie Michel Dumontier pour m'avoir accueilli dans son équipe à l'Université de Stanford et ainsi permis de réaliser une partie substantielle des travaux présentés dans cette thèse.

Je remercie le professeur Phillipe Jonveaux et Céline Bonnet pour avoir apporté leur expertise sur les déficiences intellectuelles aux différents travaux présentés dans cette thèse.

Je remercie Malika Smaïl-Tabbone et Emmanuel Bresso pour leur aide au cours de mes différents travaux de recherche.

Je remercie Jean Lieber et Alice Hermann pour m'avoir permis de découvrir le monde de la recherche en m'accueillant comme stagiaire dans l'équipe Orpailleur.

Finalement, je tiens à remercier camarades, amis ou collègues qui m'ont soutenu et aidé pendant ces quelques années.

Table des matières

Chapitre 2
Structures de patrons et ontologies pour la découverte d'associations entre effets indésirables de médicaments

Chapitre 3
Similarités sémantiques pour la classification de maladies à partir d'un diseasome

Table des figures

Liste des tableaux

Introduction

L'extraction de connaissances à partir de bases de données (ou KDD pour *Knowledge Discovery from Databases*) repose sur un processus en trois étapes : préparation des données, fouille de données et interprétation des résultats ; qui peuvent être réitérées en fonction du problème posé et des résultats obtenus. A chaque étape se posent non seulement les problèmes du choix parmi une très grande variété de méthodes et de la façon dont les connaissances du domaine préexistantes peuvent être utilisées pour guider l'ensemble du processus. La problématique de cette thèse s'intéresse en particulier à la mise en œuvre de ces connaissances de domaine dans l'étape de fouille de données.

Dans ce contexte, le Web sémantique propose un ensemble de standards et d'outils pour la formalisation et l'interopérabilité de connaissances partagées sur le Web. L'existence de ces standards a permis la mise à disposition de vastes ensembles de connaissances, représentées formellement sous la forme d'ontologies. Notamment, de nombreuses initiatives dans la communauté biomédicale ont suivi ces standards pour publier de nombreuses ontologies biomédicales, en offrant des liens entre elles et vers de nombreuses sources de données biomédicales. Les données et ontologies biomédicales ainsi associées constituent de nos jours un ensemble de connaissances complexes, massives, évolutives, hétérogènes et interconnectées, dont l'analyse est porteuse de grands enjeux en santé du point de vue de la prévention, des diagnostics, ou du suivi des traitements.

On s'intéressera dans cette thèse aux moyens d'utiliser ces ontologies biomédicales pour étendre les possibilités du processus de fouille de données. Parce que les données biomédicales sont hétérogènes et complexes, il peut être nécessaire de considérer des connaissances provenant de plusieurs ontologies dans ce processus. On proposera alors dans cette thèse des méthodes de fouille permettant de faire cohabiter les connaissances de plusieurs ontologies de domaine, afin de pouvoir associer une sémantique à l'ensemble des données et d'en tirer parti.

Les travaux de cette thèse concernent dans un premier temps l'extraction de connaissances à partir de données patients pour la découverte de co-occurences de événements indésirables médicamenteux. Ici se pose la problématique de la comparaison des événements complexes que sont ces événements indésirables, dans le but d'extraire des règles d'association plus générales. En effet, chaque événement indésirable peut être causée par plusieurs médicaments et présenter plusieurs phénotypes. On utilisera d'abord des ontologies biomédicales pour permettre la comparaison de médicaments ou de phénotypes deux à deux. On cherchera ensuite à étendre cette comparaison à des événements indésirables médicamenteux plus complexes, présentés par différents patients, en utilisant les ontologies pour tenir compte de la similarité de différents élements entrant en jeu dans deux réactions indésirables.

Cette étude utilise les structures de patrons, une extension de l'analyse formelle de concepts, afin d'extraire des règles d'association entre des réactions indésirables à certains médicaments ou

certaines classes de médicaments chez des groupes de patients. Il est donc nécessaire de proposer, de façon originale, des opérateurs de comparaison pouvant utiliser une ontologie de médicaments ou une ontologie de phénotypes pour permettre une comparaison sémantique des événements indésirables. Finalement, on définira un opérateur faisant cohabiter ces deux types d'ontologies. L'apport des bioontologies dans ce processus de KDD sera alors étudié sur deux ensembles de réactions indésirables aux médicaments, collectés à partir de dossiers patients électroniques ou d'un registre de réactions autodéclarées. Ainsi, la première contribution de cette thèse révèle comment une ontologie de médicaments et une ontologie de phénotypes cohabitent avec un ensemble de données patients pour améliorer un processus d'extraction de règles d'association. Différents opérateurs de comparaison exploitant chacun davantage de connaissances biomédicales seront proposés et comparés.

Dans un second temps, on utilisera une méthode numérique fondée sur des mesures de similarité sémantique pour la classification de déficiences intellectuelles génétiques. Ces maladies sont hétérogènes et peuvent être caractériséesà la fois par leurs phénotypes que de leur gènes responsables. Ce travail reproduit une méthodologie proposée récemment dans une étude sur l'ensemble des maladies humaines, et qui exploite une ontologie de phénotypes [Hoehndorf et al., 2015]. On étendra alors et applique cette méthodologie pour l'étude des déficiences intellectuelles, en tenant compte en plus de leur caractère génétique en considérant non seulement une ontologies de phénotypes, mais aussi les annotations fonctionnelles de leur gènes responsables. On s'intéresse alors à l'utilisation de mesures de similarité sémantiques pour comparer des maladies en utilisant plusieurs ontologies permettant de décrire les aspects phénotypiques et génétiques de ces maladies. On étudiera deux mesures de similarité utilisant des méthodes de calcul différentes, que l'on utilisera avec différentes combinaisons d'ontologies phénotypiques et géniques. On cherchera en particulier à quantifier l'influence que les différentes connaissances de domaine ont sur la capacité de classification de ces mesures, et comment ces connaissances peuvent coopérer au sein de telles méthodes numériques.

Une troisième étude s'intéresse à l'aspect "données" du Web sémantique dans le but de caractériser des gènes responsables de déficiences intellectuelles. On utilisera ici les Données Ouvertes Liées, (ou LOD pour *Linked Open Data*), qui sont de vastes ensembles de données proposées dans des formats intéropérables et liées à des ontologies de domaine du Web sémantique. Tout comme pour les ontologies biomédicales, de récentes initiatives de la communauté biomédicales ont permis la mise à disposition de tels jeux de données biomédicales, partiellement interconnectés et liés aux ontologies. Ces LOD représentent donc un objet d'étude idéal pour étudier la contribution des ontologies dans un processus de fouille de données. On s'intéressera également dans cette étude à l'utilisation des technologies du Web sémantique pour la sélection et l'intégration des données au cours du processus de KDD.

L'étape de fouille de données est instanciée dans cette étude par une méthodologie fondée sur la Programmation Logique Inductive (PLI), qui s'avère adaptée pour traiter des données relationnelles comme les LOD, en prenant en compte leurs relations avec les ontologies, notamment en mettant en œuvre des mécanismes de raisonnement. La PLI permet d'extraire d'un ensemble d'exemples positifs et négatifs un ensemble de règles logiques, ou théorie, caractérisant les exemples positifs, ici, les gènes responsables de déficiences intellectuelles. La PLI permet notamment de générer ces théories en prenant en considération des connaissances de domaine. Elles représentent des modèles à la fois prédictifs et descriptifs, et l'on cherchera à montrer comment

l'utilisation des ontologies liées au LOD améliorent ces deux qualités des théories obtenues par la PLI. Pour cela on testera différentes contraintes imposées sur le raisonnement effectué sur les connaissances de domaines.

Le point commun des travaux présentés dans cette thèse est d'utiliser les connaissances formalisées dans des ontologies du Web sémantique au sein de plusieurs processus variés de fouille de données biomédicales. Ils comparent des expériences intégrant plusieurs ontologies biomédicales et capacités de raisonnement sur celles-ci, de manière à en quantifier l'apport dans le processus de fouille. Ce document est organisé en cinq chapitres décrits ci-après.

Le Chapitre 1 présente un état de l'art détaillant les technologies du Web sémantique mises en œuvre ici, ainsi que les jeux de données et bioontologies utilisés dans les différentes expériences. On y présente également différentes techniques de fouille de données qui pourront être appliquées à des données liées à des ontologies. Différentes applications biologiques y sont ensuite décrites.

Le Chapitre 2 propose une approche utilisant les structures de patrons permettant de comparer des événements indésirables médicamenteux, et d'extraire des règles d'association entre ces événements à partir de données patients. On y propose différents opérateurs de comparaison mettant en œuvre un nombre incrémental d'ontologies de domaine.

Le Chapitre 3 présente une méthode fondée sur des mesures de similiarité sémantiques pour la classification de différentes classes de déficiences intellectuelles. On y présente différentes expériences exploitant chacune différentes combinaisons d'ontologies.

Le Chapitre 4 propose quant à lui une méthodologie pour la sélection, l'intégration et la fouille de LOD avec des ontologies, appliquée à la caractérisation des déficiences intellectuelles. On y compare différents degrés de raisonnement sur les connaissances formalisées dans les ontologies.

Finalement, le Chapitre 5 présente une synthèse des différents résultats, compare les qualités des différentes approches et propose des perspectives pour leur amélioration.

1

État de l'art

Sommaire

Ce Chapitre présente l'état de l'art de la découverte de connaissances biomédicale utilisant des ontologies de domaine. Il introduira dans un premier temps le web sémantique et ses ontologies. Seront ensuite présentées des méthodes de fouille de données et de leur capacité à utiliser des ontologies pour la découverte de connaissances. Finalement, plusieurs applications de ces méthodes au domaine biomédical seront détaillées.

1.1 Web sémantique et représentation de connaissances

Le réseau Internet permet de publier, de partager et d'accéder à une grande quantité de données, notamment via le Web. L'ensemble de ces données représente une opportunité pour les méthodes d'intelligence artificielle et de découverte de connaissances. Cependant, la plupart de ces données sont présentées dans des formats hétérogènes, et destinées à un accès par des utilisateurs humains. Pour répondre à cette problématique, le Web sémantique, dont les principes ont été énoncés dans [Berners-Lee et al., 2001], propose un ensemble de standards permettant de représenter des données et des connaissances de manière à les rendre interopérables et manipulables par des humains et des machines.

1.1.1 Ontologies

En informatique, une ontologie est une conceptualisation, ou représentation formelle des connaissances d'un domaine particulier pouvant être utilisée dans le cadre de raisonnement automatique [Grimm et al., 2011]. Cette définition d'une ontologie implique plusieurs propriétés [Gómez-Pérez et al., 2006] :

— Les ontologies doivent préférablement être exprimées dans un langage de représentation de connaissances formel, avec une sémantique clairement définie, tel que les logiques de description [Baader, 2003]. Le *Web Ontology Language* (OWL) est un des langages fréquemment utilisé pour encoder les ontologies du Web sémantique [W3C OWL Working Group, 2012].
— Une ontologie doit refléter et se limiter à représenter un consensus sur les connaissances d'un domaine.
— Les connaissances y sont exprimées explicitement de manière à les rendre accessibles à des processus de raisonnement automatiques. En particulier, il est nécessaire d'y rendre explicites les connaissances qui sont considérées comme évidentes par les créateurs de l'ontologie.
— Les concepts décrits dans l'ontologie doivent correspondre à des éléments humainement compréhensibles, en opposition à un modèle résultant d'un apprentissage automatique utilisant des approches numériques (*e.g.* machines à vecteurs de support, réseaux de neurones, etc.)

Cette définition d'une ontologie en tant que conceptualisation laisse cependant grande place à l'interprétation sur la façon d'encoder et l'expressivité d'une ontologie. Ainsi, on peut imaginer plusieurs niveaux de formalisme pour les ontologies, comme illustré en Figure 1.1.

A gauche sur la Figure 1.1, du côté le moins expressif, un simple vocabulaire contrôlé de termes, par exemple une liste de noms de médicaments, peut dans certains cas être considéré comme une ontologie naïve. Un glossaire associerait à chaque terme une définition en langage naturel, permettant de lui donner du sens, mais supportant difficilement le raisonnement automatique. Un thesaurus apporte une relation d'équivalence entre des termes synonymes. Il est également possible de trouver des connaissances organisées dans une hiérarchie informelle, comme par exemple dans des catégories et sous-catégories de produits sur un site marchand. La sémantique d'une telle hiérarchie peut cependant être variable, et donc peu utilisable pour du raisonnement.

Une ontologie formelle propose, au-delà d'un vocabulaire contrôlé de termes et leur équivalences, une hiérarchisation des concepts correspondants. Il existe alors une relation formelle appelée subsomption entre les concepts. On aura par exemple une relation de subsomption entre le concept `Animal` et `Chat` telle que `Animal` subsume `Chat` (qu'on notera `Chat rdfs:subClassOf`

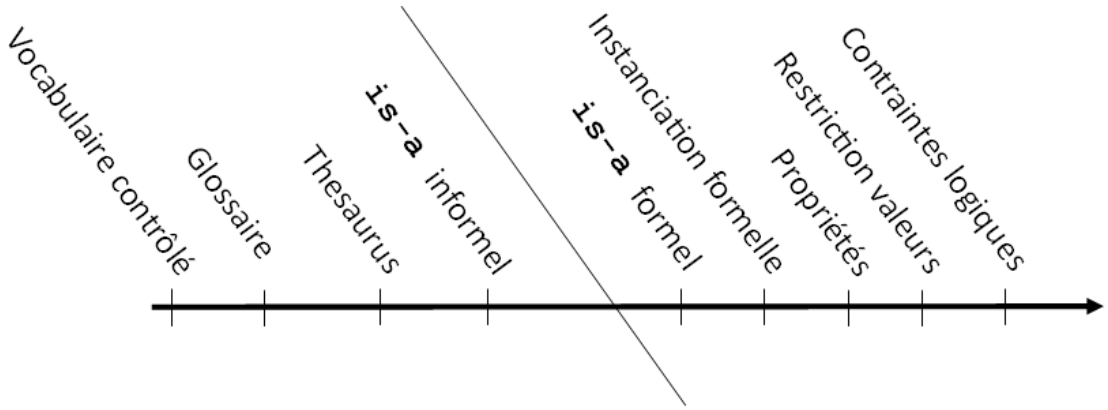


FIGURE 1.1 – Spectre représentant différents types d'ontologies — des ontologies informelles, aux ontologies formelles exprimées via des logiques de description, d'après [McGuinness, 2002]

`Animal` ou `Chat` $\sqsubseteq$ `Animal`). On pourra égamelement y trouver la notion d'instanciation, c'est-à-dire, une relation associant un objet à un concept, en tant qu'instance (qu'on notera `Felix rdf:type Chat` ou `Chat`($Felix$)). On pourra éventuellement y définir des propriétés : d'autres types de relations entre objets ou concepts ; et définir des contraintes logiques diverses (*e.g.* restrictions de domaine ou codomaine des propriétés, définition de concepts comme disjoints, etc.).

Plus une ontologie est décrite dans un langage de logique de descriptions expressif, plus il sera possible d'y représenter des connaissances complexes. En revanche, une plus grande expressivité requiert des algorithmes de raisonnement de plus grande complexité pour être expolitée. Ainsi, les ontologies sont souvent exprimées avec des contraintes logiques simples pour modérer le temps de calcul requis pour le raisonnement automatique.

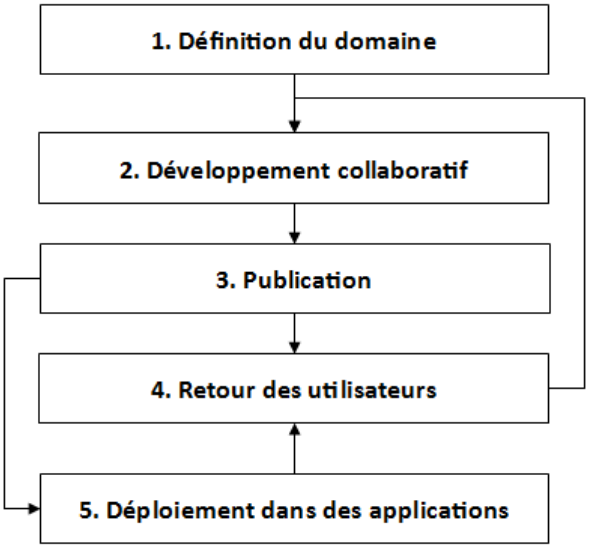


FIGURE 1.2 – Le cycle de vie d'une ontologie, d'après [Noy et al., 2010]

Par définition, les ontologies reflètent les connaissances actuelles d'un domaine, elles sont donc amenées à évoluer dans le temps et avec l'état de ces connaissances. Le cycle de vie d'une ontologie [Noy et al., 2010] s'organise alors de la manière suivante, comme illustré en Figure 1.2 :

1. les besoins auxquels doivent répondre l'ontologie sont définis, de manière à clairement délimiter le domaine de l'ontologie ;
2. elle est construite, par exemple à l'aide d'un éditeur d'ontologies (comme notamment *Protégé* [Knublauch et al., 2004]), en créant des liens avec d'autres ontologies préexistantes ;
3. une première version de l'ontologie est publiée et partagée en ligne ;
4. les retours des utilisateurs sont pris en compte pour permettre la création et la publication d'une nouvelle version de l'ontologie ;
5. l'ontologie peut être intégrée à des applications.

Structure Une ontologie s'organise en un graphe possédant plusieurs composants [Euzenat et al., 2007], notamment :

— Les individus qui sont les objets de l'ontologie
— Les classes ou concepts de l'ontologie, qui sont instanciés par un ensemble d'individus
— Les propriétés qui décrivent les relations entre individus et/ou classes, ou les attributs propres des individus et classes
— Les axiomes qui sont des assertions vérifiées dans le domaine de l'ontologie

Les concepts d'une ontologie sont organisés dans une hiérarchie par la relation de subsomption. Cette hierarchie forme un graphe dirigé acyclique, et peut être exploitée par des mesures de similarité sémantique ou des méthodes de raisonnement. On notera que la relation de subsomption est transitive, c'est-à-dire que pour tout concepts x, y et z, $x \sqsubseteq y$ et $y \sqsubseteq z$ impliquent $x \sqsubseteq z$. On appelera alors descendant de x tout concept w tel que $w \sqsubseteq x$, et on appelera réciproquement x un ancêtre de w.

La position relative de deux concepts dans la hiérarchie peut permettre de mesurer leur similarité. Un outil fréquemment utilisé pour cela est le plus petit ancêtre commun (LCA – *lowest common ancestor*) [Bender et al., 2005]. Si la structure de la hiérarchie est un arbre, alors il existe un LCA unique pour chaque paire de concepts.

Définition 1. *Soit un arbre T, pour tout x et y nœuds de T, le plus petit ancêtre commun de x et y, noté $LCA(x, y)$ est l'unique nœud vérifiant :*

— *x et y sont des descendants de $LCA(x, y)$ (chaque nœud est considéré comme son propre ancêtre et descendant), c'est-à-dire, $LCA(x, y)$ est un ancêtre commun à x et y*
— *la profondeur du nœud $LCA(x, y)$ est maximale parmi les ancêtres communs à x et y*

Cette définition du LCA peut être étendue à des hiérarchies dont la structure forme un graphe dirigé acyclique [Aït-Kaci et al., 1989, Bender et al., 2005]. Etant donné la nature de la relation de subsomption, le graphe de la hiérarchie ne peut pas contenir de cycle : les relations de subsomption de l'ontologie forment alors toujours un graphe dirigé acyclique (un arbre en étant un cas particulier). Dans tous les cas où la hiérarchie n'est pas un arbre, l'existence et l'unicité du LCA n'est pas garantie.

Définition 2. *Soient un graphe dirigé acyclique G, et x et y deux nœuds de G. Soit $G_{x,y}$ le sous-graphe formé par les ancêtres communs à x et y. L'ensemble des LCA des x et y est constitué des nœuds de $G_{x,y}$ de degré sortant 0 selon la relation de subsomption.*

De manière équivalente, pour deux concepts x et y d'une ontologie $\mathcal{O}$:

$$LCA(x, y) = \{e_i \in \mathcal{O} \mid (x \sqsubseteq e_i) \wedge (y \sqsubseteq e_i) \wedge \nexists e_j \in \mathcal{O}.\ (x \sqsubseteq e_j) \wedge (y \sqsubseteq e_j) \wedge (e_j \sqsubseteq e_i)\}$$

Le LCA est notamment utilisée pour définir des mesures de similarité sémantique permettant de comparer des concepts ou des ensembles de concepts. Ainsi, on pourra par exemple estimer que deux concepts sont très similaires si leur LCA est très spécifique. Des mesures de similarité sémantique utilisant le LCA seront présentées dans la Section 1.2.3.

1.1.2 Ontologies biomédicales

Les travaux présentés dans cette thèse s'intéressent à l'utilisation d'une ou plusieurs ontologies dans le processus de découverte de connaissances. Cette section présente les différentes ontologies utilisées dans ces travaux.

Gene Ontology Le consortium Gene Ontology (GO) [Ashburner et al., 2000] propose un vocabulaire pour la description des fonctions biologiques des gènes. Ce vocabulaire a pour but de permettre la réutilisation des connaissances sur des gènes similaires entre différentes espèces. Ce vocabulaire est contenu dans trois ontologies :

- *molecular function* (MF) qui décrit la fonction moléculaire des produits d'un gène, par exemple la catalyse d'une réaction ou le transport d'une molécule particulière.
- *biological process* (BP) qui décrit les processus biologiques dans lesquels un gène peut être impliqué. Un processus biologique représente souvent une transformation chimique ou physique dans l'organisme, réalisée par une ou plusieurs successions de fonctions moléculaires.
- *cellular component* (CC) qui décrit dans quel composant de la cellule le produit d'un gène est actif.

Les concepts de chaque ontologie sont organisés selon une hiérarchie structurée par la relation `is-a`. Chaque concept peut posséder plusieurs parents, ainsi la structure de chacune des trois ontologies GO forme un graphe dirigé acyclique. D'autres relations coexistent avec la relation de subsomption comme `PartOf`, notamment dans l'ontologie *cellular component*.

Ainsi, de nombreux gènes parmi plusieurs espèces ont été annotés par les termes des trois ontologies GO. Ces annotations sont mises à disposition via la base de données Gene Ontology Annotation (GOA) [Barrell et al., 2009]. Une grande partie de ces annotations sont assignées automatiquement aux gènes, en utilisant des liens entre différentes sources de données biologiques pour inférer les fonctions d'un gène ou de ses produits. A chaque annotation est associée le type de preuve qui a été utilisée pour la produire. Ces annotations peuvent être réparties en cinq catégories de par leur origine :

- *Experimental Evidence* – inférée expérimentalement
- *Computational Analysis* – inférée par analyse *in silico*
- *Author Statement* – annotation extraite d'un article
- *Curatorial Statement* – annotation manuelle ne correspondant à aucune catégorie précedente
- *Inferred from Electronic Annotation* – annotation automatique

Human Phenotype Ontology Human Phenotype Ontology (HPO) [Köhler et al., 2013, Robinson et al., 2008] est une ontologie construite dans le but de répertorier et de hiérarchiser les phénotypes anormaux présentés par les maladies génetiques. Ces phénotypes anormaux sont des symptômes de maladies extraits de la base de données *Online Mendelian Inheritance in Man* (OMIM) [Hamosh et al., 2005]. OMIM associe à une maladie génetique un ensemble de symptômes extraits de textes et classés par organe. Cette classification par organe ne permet cependant pas de juger de la similarité de deux symptômes affectant des organes différents. De

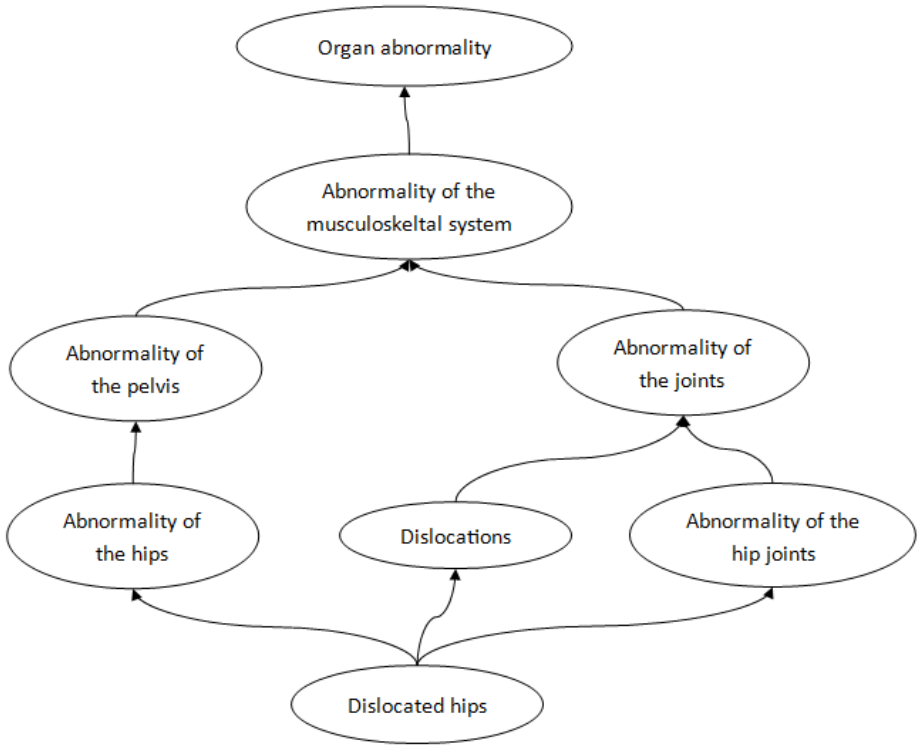


FIGURE 1.3 – Ancêtres du terme HPO *Dislocated hips*. Les arcs représentent la relation `is-a`, tels que $x \longrightarrow y$ représente x `is-a` y

même, il n'est pas possible de comparer des symptômes touchant le même organe. Une autre limite des données proposées par OMIM est le manque d'un vocabulaire contrôlé : ainsi plusieurs expressions peuvent être utilisées pour décrire un même symptôme, comme par exemple *amyotrophie généralisée* et *atrophie musculaire généralisée*.

L'ontologie HPO répond à ces limites en proposant un vocabulaire contrôlé de concepts organisés dans un graphe dirigé acyclique par des relations `is-a`. Un concept de l'ontologie peut avoir plusieurs parents, permettant de considérer différents aspects des phénotypes : par exemple, le phénotype *Dislocated hips* (voir Figure 1.3) possède deux parents indiquant la localisation anatomique du phénotype (*Abnormality of the hips*, *Abnormality of the hip joints*) et un parent indiquant le type d'anomalie (*Dislocations*).

Monarch Disease Ontology Monarch Disease Ontology (MonDO) [Mungall et al., 2017] propose d'unifier les données phénotypiques provenant de différentes espèces, dans le but de faciliter la réutilisation des connaissances sur les maladies dans des modèles animaux pour mieux comprendre les maladies analogues chez l'homme. Alors que seulement 51% des gènes humains codant pour une protéine possèdent des annotations phénotypiques, MonDO propose d'augmenter cette couverture à 89% en considérant les annotations des gènes orthologues [1] chez d'autres espèces. Ainsi MonDO intègre plusieurs sources de connaissances biomédicales, notamment OMIM et HPO, en combinant les ontologies par un algorithme bayésien [Mungall et al., 2016], utilisant les liens existant entre les différentes ontologies pour les unifier.

1. Gènes codant pour une fonction similaire dans des espèces différentes.

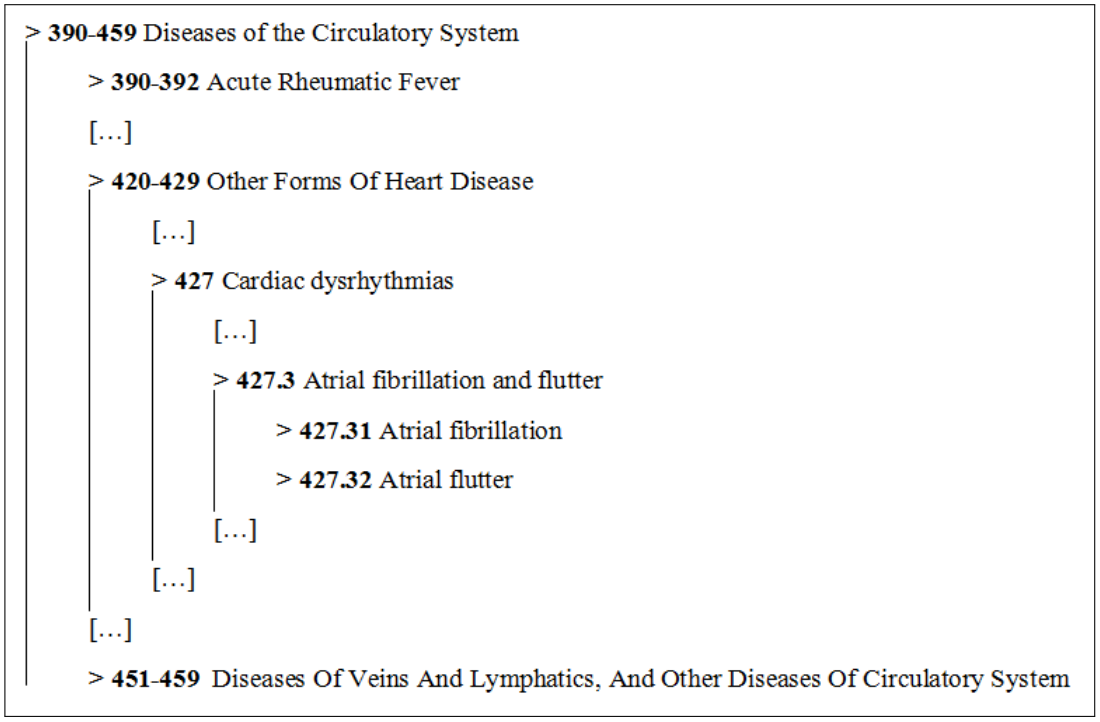


FIGURE 1.4 – Hiérarchie des concepts *Atrial fibrillation* et *Atrial flutter* extraite de ICD-9-CM. Le code ICD-9-CM de chaque concept est indiqué en gras.

ICD-9-CM La Classification Internationale des Maladies (*International statistical classification of diseases and related health problems*, abréviation : ICD – *International Classification of Diseases*, [WHO, 2004]) est un standard proposé par l'Organisation Mondiale de la Santé, notamment utilisé pour :

— l'encodage des statistiques d'incidence et de mortalité des maladies ;
— l'annotation des diagnostics dans des dossiers patients ;
— la gestion de la facturation par les praticiens et les assureurs.

On s'intéressera ici à une version particulière de cette classification : ICD-9-CM (*International Classification of Diseases, 9th revision, Clinical Modification*) [AMA, 2004]. ICD-9-CM est notamment le système officiel pour l'annotation de diagnostics dans les hôpitaux du monde entier.

ICD-9-CM propose une hiérarchisation de maladies structurée en un arbre. Chaque concept y est identifié par un code, permettant notamment de le situer au sein de la hiérarchie, comme illustré en Figure 1.4. Les codes des concepts des deux premiers niveaux les plus généraux de la hiérarchie sont des intervalles, tels que le code de chaque concept de premier niveau corresponde à l'union des intervalles des codes de ses enfants. Par exemple le concept **420-429** *Other forms of heart disease* est un enfant de **390-459** *Diseases of the Circulatory System*. Ces concepts correspondent à de larges catégories de maladies, et ne sont pas destinés à être utilisés pour encoder un diagnostic.

Les concepts du troisième niveau de l'ontologie sont identifiés par un nombre à 3 chiffres, compris dans l'intervalle correspondant au code de son parent. Les quatrième et cinquième niveaux reprennent chacun le code de niveau supérieur, suivi d'un chiffre supplémentaire (un point est ajouté pour séparer le troisième et quatrième chiffre du code).

Certaines limites existent à l'utilisation d'ICD-9-CM. Tout d'abord, la sémantique des noms de concepts n'est pas claire : par exemple, dans le concept *Atrial fibrillation and flutter*, le *and*

exprime un *ou* inclusif. On trouve également des concepts définis de manière peu précise, par exemple *Other Forms of Heart Disease*, qui décrit alors l'ensemble des maladies cardiaques autrement non décrites dans le reste de la classification, sans pour autant représenter une connaissance particulière sur cet ensemble de maladies [WHO, 2011]. D'autre part, la classification ICD-9-CM a d'abord été construite dans le but de faciliter la facturation, et ainsi peut proposer plusieurs codes pour une même maladie, ou bien ne pas proposer de concepts suffisamment précis pour certaines applications.

ATC La classification ATC (*Anatomical, Therapeutical, Chemical classification*) [WHOCC, 2013] organise les substances actives de médicaments dans une hiérarchie à cinq niveaux. Les différents niveaux peuvent correspondre :

- au groupe *Anatomique* de la substance, qui indique les organes ciblés par la substance (par exemple, le système nerveux) ;
- au groupe *Thérapeutique*, qui indique l'usage médical fait de la substance (par exemple, traitement contre le diabète) ;
- au groupe *pharmacologique*, qui indique les effets de la substance (par exemple, réduction de la glycémie) ;
- au groupe *Chimique* de la substance.

La hiérarchie ATC est organisée en un arbre à cinq niveaux, dont les feuilles correspondent aux substances actives de médicaments. À chacune de ces feuilles est attribué un code ATC unique, par exemple, la substance *paracetamol* a pour code N02BE01. Ce code permet de facilement situer la substance dans la hiérarchie, en effet, on peut le décomposer comme suit :

- la première lettre du code correspond au code premier niveau de la hiéarchie dont le nœud identifié par le code est le descendant. Dans l'exemple du paracetamol N02BE01, la lettre N permet d'identifier la classe anatomique correspondante : le système nerveux.
- Les deux chiffres suivants permettent d'identifier le deuxième niveau de la hiérarchie correspondant soit à un groupe thérapeutique ou pharmacologique. Dans l'exemple, cela correspond au groupe thérapeutique identifié par le code N02 : les analgésiques.
- Les deuxième et troisième lettres du code identifient respectivement aux troisième et quatrième niveaux de la hiérarchie, correspondant à des groupes thérapeutiques, pharmacologiques ou chimiques.
- Les deux derniers chiffres identifient la substance au cinquième niveau de la hiérarchie.

Une même substance pouvant affecter plusieurs organes ou présenter plusieurs applications thérapeutiques, une seule substance peut posséder plusieurs codes ATC. En effet, la structure en arbre de la classification ATC ne permet pas à un concept de posséder plusieurs parents. Pour les mêmes raisons, certains groupes chimiques peuvent apparaître dans plusieurs groupes anatomiques ou thérapeutiques. Ainsi, la sémantique exprimée par le nom d'un concept doit toujours être considérée dans le cadre des groupes anatomiques, thérapeutiques et pharmacologiques parents.

SNOMED CT L'ontologie SNOMED CT (*Clinical Terms*) [Donnelly, 2006] propose une riche terminologie de termes médicaux, regroupant symptômes, diagnostics, constats cliniques, causes, organes, substances pharmaceutiques, procédures médicales, etc. SNOMED CT étant exprimée dans un langage de logique de description, ses concepts peuvent être formellement définis à partir de relations à d'autres concepts. Par exemple, le concept *fracture du tibia* peut être défini comme un type de *fracture*, localisée au *tibia*.

Grâce à ces différents types de termes, SNOMED CT permet également une annotation dé-

taillée dans les rapports médicaux et dossiers patients électroniques, même lorsqu'un concept exact n'est pas pré-défini : en combinant par exemple un concept anatomique, un concept représentant une cause et un concept représentant un symptôme particulier.

1.1.3 Accès unifié aux ontologies biomédicales

Afin de faciliter l'accès aux ontologies biomédicales ainsi que leur interopérabilité, différents projets visent à rendre disponibles ces ontologies à travers un navigateur ou une API unique, accompagnées de liens entre leurs concepts. C'est notamment le cas du BioPortal et de l'UMLS, qui sont présentés dans cette section.

BioPortal Le BioPortal [Whetzel et al., 2011] est un portail Web développé par le *National Center for Biomedical Ontology* permettant l'accès à de nombreuses ontologies biomédicales. Le BioPortal propose notamment d'accéder à ces ontologies via une API REST [Masse, 2011], un point d'accès SPARQL [Prud et al., 2006] ou une interface graphique de navigation.

Le BioPortal hébergeait, en février 2018, 691 ontologies avec près de 9 millions de classes, contre 72 ontologies et 300 000 classes en 2008 [Noy et al., 2009]. On constate non seulement que le nombre d'ontologies biomédicales a presque décuplé en 10 ans, mais également que le nombre de classes moyen par ontologie a triplé sur cette même période.

Dans le cadre du cycle de vie des ontologies, le BioPortal met à disposition des utilisateurs des outils permettant de commenter les classes ou les ontologies ou de créer des liens (ou *mappings*) entre ces ontologies, rendant explicites les équivalences entre les classes de plusieurs ontologies. Ces commentaires sont alors pris en compte par les développeurs des ontologies afin d'en améliorer les futures versions. Les liens créés par les utilisateurs permettent l'intéropérabilité des ontologies, et facilitent l'intégration de données exprimées dans des vocabulaires différents.

Les ontologies prsentées précedemment sont toutes accessibles via le BioPortal.

UMLS L'*Unified Medical Language System* (UMLS) [Bodenreider, 2004] est un projet de l'*United States National Library of Medicine* (NLM) proposant à la communauté biomédicale un *Metathesaurus* (une collection d'ontologies), le *Semantic Network* (une ontologie de haut niveau) et des outils de traitement automatique du langage naturel.

Le *Metathesaurus* est un vaste ensemble de vocabulaires controlés médicaux intégrés. Il contient plus de 200 vocabulaires, dont notamment GO, HPO, ICD-9-CM, ATC et SNOMED CT présentés précedemment. Les termes synonymes dans ces différents vocabulaires sont fusionnés pour former les concepts de ce *Metathesaurus*. Chacun de ces concepts se voit alors attribuer un *Concept Unique Identifier* ou CUI, c'est-à-dire un code permettant de l'identifier de manière unique, et ce, même si il est défini dans plusieurs ontologies.

Le *Semantic Network*, l'ontologie de haut niveau de l'UMLS, organise les termes ou concepts dans une hiérarchie de haut niveau (phénotype, gène, maladie, etc.), et propose des types de relations entre ces concepts. La Figure 1.5 illustre la partie du *Semantic Network* liée au concept Organisme, ainsi que les différentes relations du *Semantic Network*.

Afin de mettre à profit le *Metathesaurus*, l'UMLS met à disposition des outils de traitement automatique du langage [McCray et al., 1994], permettant l'annotation de textes biomédicaux par des concepts du *Metathesaurus*, notamment le logiciel MetaMap [Aronson, 2001].

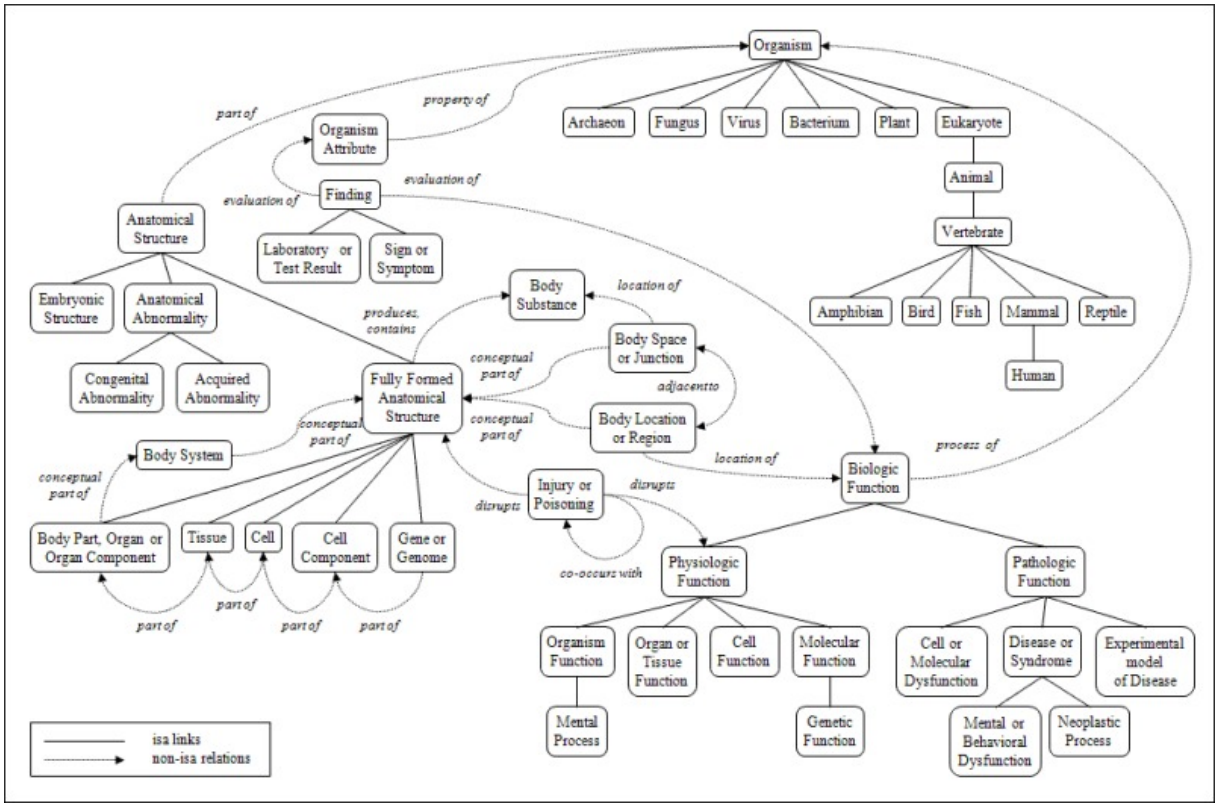


FIGURE 1.5 – Portion du *Semantic Network* de l'UMLS. Source : [US NLM, 2009]

Les ontologies du web sémantique décrivent des concepts formels permettant notamment le raisonnement automatique sur des instances de ces concepts. Pour que les connaissances de ces ontologies soient utiles, il est nécessaire de disposer de données instanciant les concepts y étant décrits : c'est justement ce que fournissent les Données Ouvertes et Liées, décrites dans la section suivante.

1.1.4 Données Ouvertes et Liées

Les Données Ouvertes Liées (ou *Linked Open Data* – LOD) font partie d'un effort communautaire pour la construction du Web sémantique [Berners-Lee, 2006, Bizer et al., 2009]. Les LOD sont disponibles sous la forme d'un grand nombre de jeux de données representés dans un format standard, partiellement connectés les uns aux autres ainsi qu'à des connaissances de domaine présente dans les ontologies du Web sémantique. Plus précisement, les LOD forment un ensemble d'assertions sur lesquels on peut opérer des inférences grâce aux connaissances de domaine qui peuvent y être liées.

Définitions Sir Tim Berners-Lee propose dans [Berners-Lee, 2006] plusieurs principes encadrant les LOD :

— Chaque resource est identifiée par un *Uniform Resource Identifier*[2] [Berners-Lee, 1994]

2. Les *URI* sont initialement nommés *Universal Resource Identifier* dans [Berners-Lee, 1994] avant que l'appelation *Uniform Resource Identifier* ne soit adoptée.

(URI).
— Chaque resource identifiée par un URI est décrite par un ensemble de triplets sous la forme *sujet-prédicat-objet*. Elle est ainsi liée à d'autres resources identifiées par leur URIs.
— Les données doivent être mises à disposition en ligne via le protocole HTTP, et sous une licence permettant la réutilisation et l'interopérabilité avec les autres sources de données.

Les triplets *sujet-prédicat-objet* constituant les données des LOD sont exprimés dans le *Resource Description Framework* (RDF) [Klyne et al., 2004, Cyganiak et al., 2014]. RDF est un modèle de données qui permet de représenter des données sous la forme de graphes dirigés, où chaque triplet *sujet-prédicat-objet* représente un arc *prédicat* reliant les nœuds *sujet* à *objet*, comme représenté en Figure 1.6.

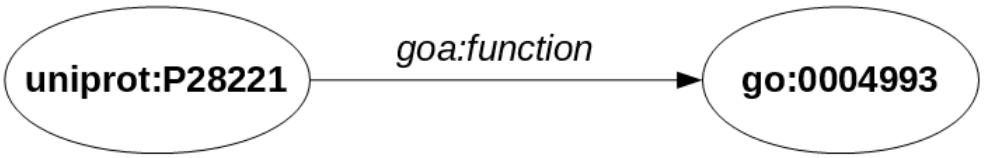


FIGURE 1.6 – Représentation en graphe du triplet `uniprot:P28221 goa:function go:0004993`, représentant l'annotation de la protéine P28211 par la fonction GO *serotonin receptor activity* (`go:0004993`). Ici les URIs utilisent une notation préfixée, telle que `uniprot:P28221` correspond à l'URI `http://bio2rdf.org/uniprot:P28221`, `goa:function` correspond à `http://bio2rdf.org/goa_vocabulary:function` et `go:0004993` correspond à `http://bio2rdf.org/go:0004993`.

RDF Schema étend RDF avec des concepts et relations normatifs, afin d'assurer l'interopérabilité des données exprimées en RDF. Ces triplets peuvent être stockés textuellement dans différents formats de fichiers, notamment RDF/XML [Gandon and Schreiber, 2014] et Turtle [Beckett and Berners-Lee, 2011], ou dans des *triplestore* : des systèmes de gestion de bases de données RDF, comme Virtuoso [Erling, 2012].

Le language de requête SPARQL (*SPARQL Protocol and RDF Query Language*) [Prud et al., 2006] permet d'interroger un ensemble de triplets RDF.

1.1.5 Sources de Données Ouvertes et Liées biomédicales

Les LODs connaissent une expansion rapide ces dernières années. Le site `http://lod-cloud.net` recensait en août 2014 la quantité de 570 sources de LOD (voir Figure 1.7), contre 1163 sources en 2017 (voir Figure 1.8). Par ailleurs, la quantité de sources de LOD biologiques a connu une progression encore plus impressionnante, passant de 63 sources en 2014 à 333 en 2017. On observe également sur la Figure 1.8 que ces sources biomédicales sont très fortement liées entre elles, les différentes initiatives d'unification des vocabulaires ayant certainement contribué à l'intégration entre elles de ces données.

Les LOD offrent de nouvelles opportunités pour le développement d'approches d'intégration de données et de découverte de connaissances. Cette récente disponibilité des LOD bénéficie particulièrement aux sciences du vivant, du fait que les données biologiques sont souvent proposées par de nombreuses sources, sans consensus universel sur une représentation unique des entités étudiées [Antezana et al., 2009].

L'intégration de ces données est un des premiers défis à relever afin de fouiller des données en provenance de plusieurs sources. De nombreux projets, tels que Bio2RDF [Belleau et al., 2008], Linked Open Drug Data [Samwald et al., 2011], PDBj [Kinjo et al., 2012] ou la plateforme EBI

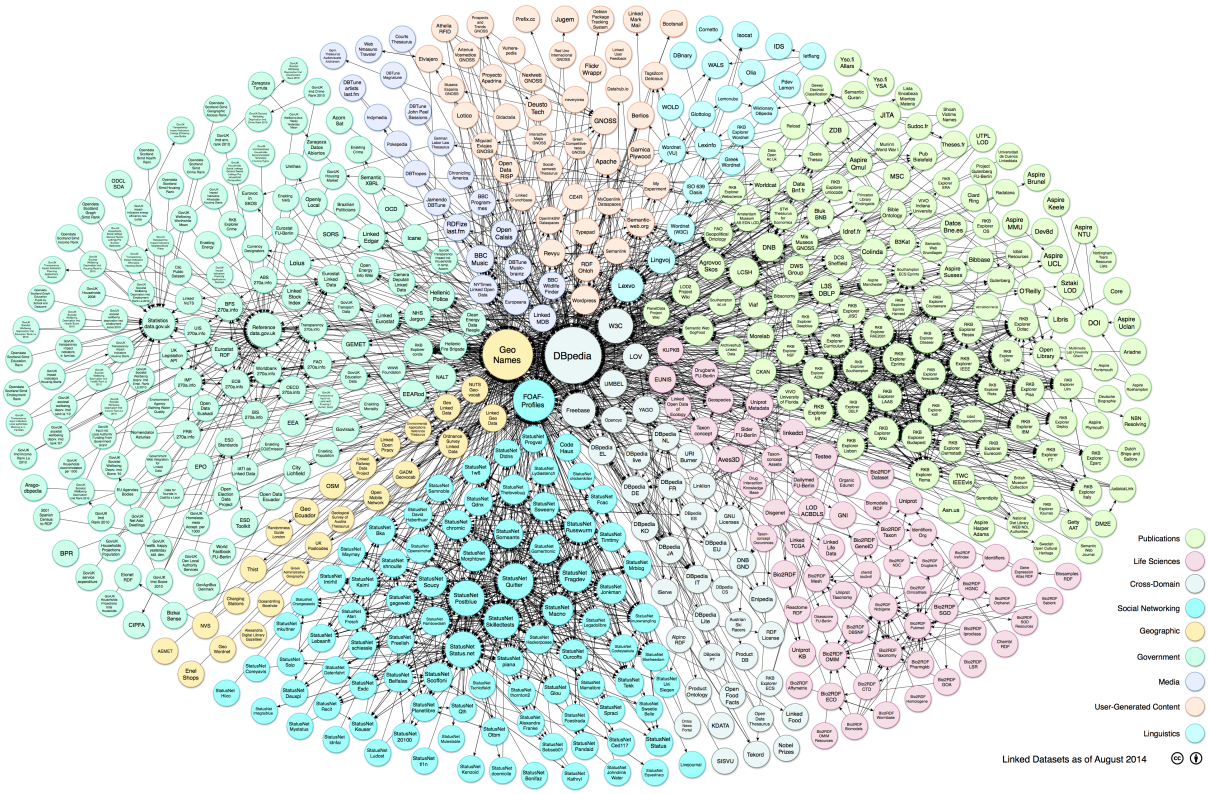


FIGURE 1.7 – Représentation de 570 sources de Données Ouvertes et Liées repertoriées en août 2014, dont en rose 63 sources de données biologiques. Source : [Schmachtenberg et al., 2014]

[Jupp et al., 2014] contribuent à rendre disponible des données biomédicales sous forme de LOD afin de faciliter leur intégration. Ces initiatives ont pu créer une vaste collection de données biomédicales dans un format standard et disponible pour la fouille. Cependant ces différentes sources de LOD ne sont pas nécessairement liées entre elles.

Au delà de ces interconnections, les LODs peuvent également être connectées aux connaissances de domaine contenues dans des ontologies, comme Gene Ontology [Ashburner et al., 2000] ou Human Phenotype Ontology [Robinson et al., 2008, Köhler et al., 2013]. Ces liens aux ontologies et les mécanismes de raisonnement associés à de la fouille de données peuvent permettre la découverte de connaissances. C'est notamment le sujet de cette thèse que d'étudier les liens entre le raisonnement et la fouille de données : on décrira ici les différentes sources de LOD utilisées qui y sont condidérées.

Bio2RDF Bio2RDF [Belleau et al., 2008] est un projet visant à transformer des sources de données biologiques hétérogènes non-LOD en RDF. Ces données peuvent provenir de bases de données relationnelles, de fichiers XML, HTML ou texte.

Trois versions de Bio2RDF ont été publiées, la dernière mettant à disposition plus de 11 milliards de triplets répartis dans une trentaine de jeux de données [Dumontier et al., 2014]. A titre de comparaison, la seconde version proposait une vingtaine de jeux de données pour seulement 1 milliards de triplets [Callahan et al., 2013]. Ces données sont stockées dans des *triplestore* Virtuoso et interrogeables via des *endpoint* SPARQL. Par ailleurs, les scripts ayant permis de peupler le *triplestore* à partir de chaque source de données originale sont mis à disposition librement sur

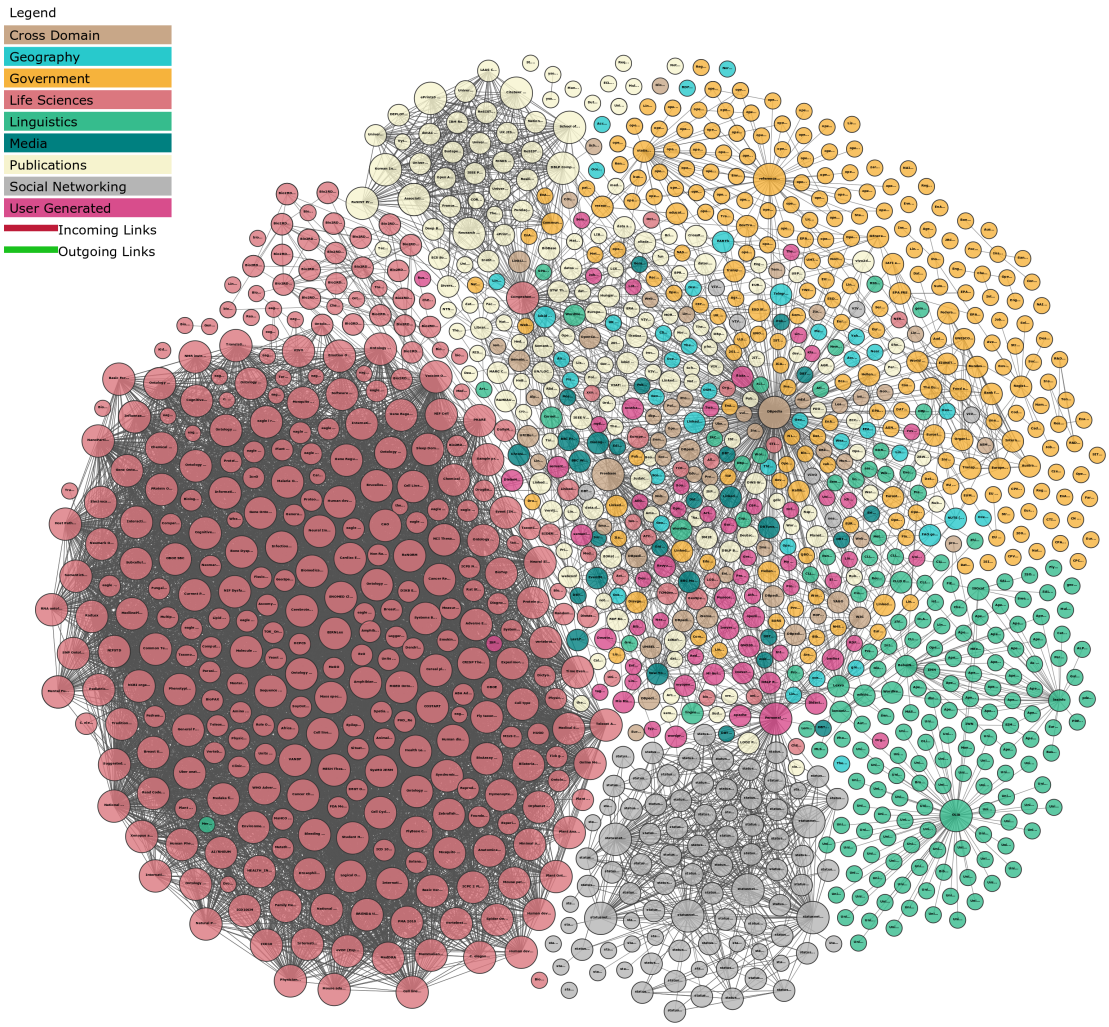


FIGURE 1.8 – Représentation de 1163 sources de Données Ouvertes et Liées repertoriées en août 2017, dont 333 sources de données biologiques. Source : [Abele et al., 2017]

la page GitHub du projet : `https://www.github.com/bio2rdf`.

On s'intéressera en particulier à six jeux de données de Bio2RDF, concernant les données sur les gènes et les protéines :

— OMIM [Hamosh et al., 2005], une collections de gènes humains associés à des maladies génétiques.

— KEGG (*Kyoto Encyclopedia of Genes and Genomes*) [Kanehisa and Goto, 2000], une base de connaissances construite dans le but de mieux comprendre les fonctions biologiques, en modélisant les gènes, les protéines et les réactions chimiques dans lesquelles elles sont impliquées. Ces réactions sont organisées en voies métaboliques ou *metabolic pathways*. KEGG intègre notamment des données sur les génes de 3500 espèces.

— InterPro [Hunter et al., 2008] est une base de données de familles, domaines et sites fonctionnels de protéines. Un domaine d'une protéine représente une partie de la séquence d'une protéine pouvant exister et fonctionner indépendamment du reste de la protéine.

Un domaine peut notamment correspondre à une fonction de la protéine.
— NCBI Gene [NCBI, 2005]
— GOA [Barrell et al., 2009], un ensemble d'annotations de gènes par des termes de Gene Ontology.
— iRefIndex [Razick et al., 2008], un ensemble d'interactions protéine-protéine.

UniProt UniProt [UniProt Consortium, 2016] est une base de connaissances sur les protéines, leur séquences et leurs annotations. UniProt contient plus de 60 millions de séquences de protéines dont 550000 annotées manuellement par des experts, dont l'ensemble des protéines humaines. UniProt propose ses données au format LOD via son propre *endpoint* SPARQL `https://sparql.uniprot.org/`.

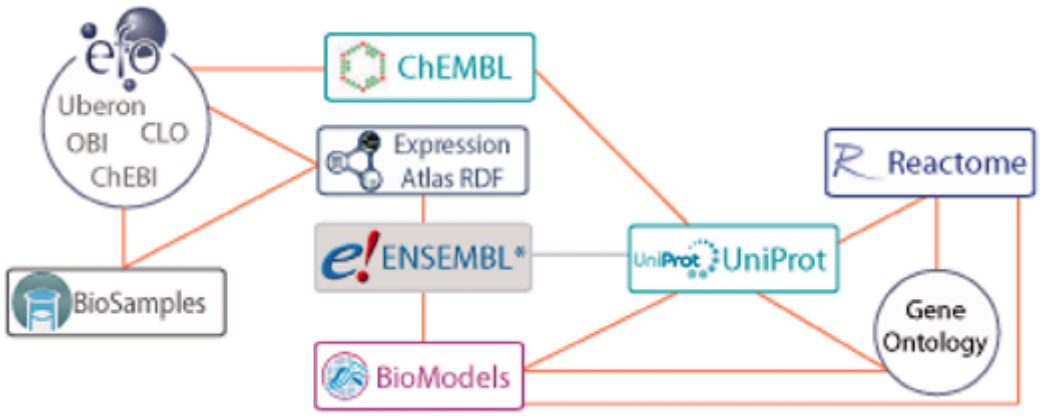


FIGURE 1.9 – Resources de la plateforme RDF de l'EBI (représentées par des rectangles) connectés à des ontologies (représentées par des cercles). Note : UniProt est indépendant de la plateforme EBI, cependant les données de la plateforme y sont fortement liées. Source : [Jupp et al., 2014].

Plateforme EBI RDF La plateforme RDF de l'EBI (*European Bioinformatics Institute*) [Jupp et al., 2014] met à disposition sous forme de LOD les différentes resources de l'EBI, illustrées en Figure 1.9. Cette plateforme propose notamment des ressources fortement liées à UniProt, référençant explicitement les URIs des LOD UniProt, comme Reactome, une base de données de voies métaboliques.

1.1.6 Ontologies et données ouvertes et liées

Afin de favoriser l'intégration entre ontologies et LOD, les ontologies peuvent être proposées au format RDF. Il est alors possible de faire des requêtes SPARQL sur les LOD en prenant en compte les connaissances des ontologies qui y sont liées. Par exemple, le BioPortal permet d'interroger ses ontologies via un endpoint SPARQL : `http://sparql.bioontology.org/`.

On peut considérer une distinction entre LOD et ontologies telle que les LOD représentent des informations sur des instances de concepts décrits par les ontologies. Cependant, cette distinction n'est pas toujours réalisée dans les connaissances biomédicales. En particulier, des ontologies comme GO ou HPO définissent des concepts abstraits (comme la fonction d'un gène) qui n'ont pas vocation à être instanciés, mais sont souvent utilisés comme annotations. On observera alors entre les données du LOD et les concepts des ontologies des relations autres que l'instanciation.

Par ailleurs, les LOD peuvent également exprimer des connaissances. On peut par exemple considérer qu'une base de données sur des interactions entre protéines comme des connaissances sur ces protéines : ici, une protéine ne représente pas un individu, mais une conceptualisation représentant un ensemble d'objets interchangeables. De même que l'interaction décrite entre deux protéines n'est pas une donnée résultant de l'observation de l'interaction de deux individus, mais une connaissance généralisable comme une relation entre deux classes d'individus.

L'avantage qu'ont les ontologies sur les LOD pour la représentation et l'exploitation des connaissances réside dans les capacités de raisonnement leur étant associées, notamment par l'utilisation de la hiérarchie de concepts. Il est en théorie possible d'utiliser ce raisonnement avec les données du LOD, cependant les grandes quantités de LOD rendent cette tâche de raisonnement difficilement généralisable à l'échelle des LOD.

1.2 Notions mathématiques

Cette section a pour objectif d'introduire les notions mathématiques qui seront utilisées dans les différents travaux présentés dans cette thèse.

1.2.1 Relations d'ordre et treillis

Les treillis sont des outils mathématiques qui peuvent notamment être utilisées pour comparer et organiser des données dans un processus de découverte de connaissances. Un treillis est une structure algébrique permettant d'organiser des objets munis d'une relation d'ordre [Birkhoff, 1949].

Définition 3. *Un ordre $\leq$ sur un ensemble E est une relation binaire entre les éléments de E, telle que, pour tout x, y et z éléments de E :*

- *$\leq$ est réflexif : $x \leq x$*
- *$\leq$ est anti-symétrique : $x \leq y$ et $y \leq x \Leftrightarrow x = y$*
- *$\leq$ est transitif : $x \leq y$ et $y \leq z \Rightarrow x \leq z$*

Afin de définir un treillis, il n'est pas nécessaire de disposer d'un ordre total, c'est-à-dire d'un ordre défini pour toute paire d'éléments de son domaine.

Définition 4. *Un ordre $\leq$ sur un ensemble E est dit total si et seulement si pour tout x et y éléments de E, x et y sont comparables, c'est-à-dire que l'on a $x \leq y$ ou $y \leq x$. Sinon, cet ordre est dit partiel.*

Un treillis peut alors être défini à partir d'un ensemble partiellement ordonné où chaque paire d'éléments admet une borne supérieure et inférieure. *A fortiori*, on peut définir un treillis à partir de tout ensemble totalement ordonné.

Définition 5. *Un treillis est un ensemble E muni de deux lois de composition internes*[3] *$\wedge$ (*meet*) et $\vee$ (*join*), telles que, pour tout x, y et z éléments de E :*

- *$\wedge$ et $\vee$ sont associatives : $x \wedge (y \wedge z) = (x \wedge y) \wedge z$*
- *$\wedge$ et $\vee$ sont commutatives : $x \wedge y = y \wedge x$*
- *$\wedge$ et $\vee$ vérifient la loi d'absorption : $x \wedge (x \vee y) = x \vee (x \wedge y) = x$*

3. Une loi de composition interne sur E est une opération binaire qui à deux éléments de E associe un élement de E.

On peut définir à partir des deux lois internes du treillis une relation d'ordre $\leq$ *telle que* $a \leq b \Leftrightarrow a \vee b = b$*. À l'inverse, on peut définir un treillis à partir d'un ensemble partiellement ordonné, c'est-à-dire, muni d'une relation d'ordre partielle sur ses éléments, tel que, pour tout* x *et* y *éléments de* E *:*

- x *et* y *admettent une borne supérieure* $\sup(x, y) \in E$
- x *et* y *admettent une bonne inférieure* $\inf(x, y) \in E$

Cette relation d'ordre permet alors de définir les lois internes du treillis telles que $x \wedge y = \inf(x, y)$ *et* $x \vee y = \sup(x, y)$*.*

Puisque chaque paire d'éléments du treillis admettent une borne inférieure et une borne supérieure, il existe une borne supérieure et une borne inférieure à l'ensemble de ses éléments.

1.2.2 Mesures de similarité et métriques

Une mesure de similarité est définie dans [Tversky, 1977] comme une fonction réelle permettant de comparer des objets représentés par des ensemble d'attributs. Les mesures de similarité proposent de quantifier la similarité de deux objets. A l'inverse, les métriques quantifient la dissimilarité de deux objets.

Définition 6. *Une métrique* d *sur un ensemble* E *est une fonction réelle* $d : E \times E \to \mathbb{R}^+$ *telle que, pour tout* x*,* y*, et* z *éléments de* E *:*

- d *vérifie l'identité des indiscernables :* $d(x, y) = 0 \Leftrightarrow x = y$
- d *est symétrique :* $d(x, y) = d(y, x)$
- d *vérifie l'inégalité triangulaire :* $d(x, z) \leq d(x, y) + d(y, z)$

Une fonction ne vérifiant que les deux premières propriétés est appelée semi-métrique.

On notera qu'on pourra dans la plupart des cas, pour une similarité s bornée par $[0, 1]$, définir une semi-métrique (et plus rarement une métrique) $d(x, y) = 1 - s(x, y)$.

Similarité de Dice La similarité de Dice [Dice, 1945] quantifie la similarité de deux ensembles, en comparant leurs cardinalités à celle de leur intersection.

Définition 7. *Soient deux ensembles* X *et* Y*, la similarité de Dice de ces deux ensembles est définie telle que :*

$$sim_{Dice}(X, Y) = \frac{2|X \cap Y|}{|X| + |Y|}$$

Similarité de Jaccard La similarité de Jaccard [Jaccard, 1901] quantifie la similarité de deux ensembles, en comparant la cardinalité de leur intersection avec celle de leur union.

Définition 8. *Soient deux ensembles* X *et* Y*, la similarité de Jaccard de ces deux ensembles est définie telle que :*

$$sim_{Jaccard}(X, Y) = \frac{|X \cap Y|}{|X \cup Y|}$$

Cette similarité peut être transformée en une métrique [Kosub, 2016] $d_{Jaccard}(X, Y) = 1 - sim_{Jaccard}(X, Y)$.

1.2.3 Similarité sémantique

Les mesures de similarité sémantique utilisent des ontologies afin de comparer des objets décrits par une classe ou un ensemble de classes de ces ontologies. Ces mesures s'appuient souvent la notion d'*Information Content* (IC) , qui quantifie l'information apportée par la présence d'une annotation.

Définition 9. *Information Content (Contenu d'Information) ou IC est une mesure de la surprise lors de l'échantillonage d'une variable aléatoire [Jones, 1979]. Soit e un événement avec une probabilité $P(e)$ non-nulle*[4]*. La quantité d'information contenue dans un message informant de l'occurence de e est :*

$$IC(e) = -log_2(P(e))$$

Dans notre cas, IC se définit par rapport à un corpus d'objets annotés à partir duquel il est possible de calculer la probabilité d'être annoté par une classe d'ontologie donnée. Les mesures de similarité fondées sur le contenu d'information nécessitent donc une ressource externe à l'ontologie pour leur calcul.

On notera que $IC(e)$ décroit quand $P(e)$ augmente, et inversement, et a pour limites 0 quand $P(e) = 1$ et $+\infty$ quand $P(e)$ approche de 0. Ainsi les annotations par des classes peu communes se voient attribuer une valeur d'IC plus élevée que des annotations communes.

Similarité sémantique entre termes

On s'intéressera d'abord aux similarités sémantiques permettant de comparer deux classes d'une ontologie. Ces similarités peuvent ensuite être utilisées pour comparer deux objets annotés par des classes d'ontologies en aggrégeant les similarités pair-à-pair entre leurs annotations.

Similarité de Resnik La similarité de Resnik est une mesure de similarité [Resnik, 1995] non bornée entre deux termes d'une ontologie. En particulier, elle quantifie la similarité de deux termes comme l'IC de leur LCA.

Définition 10. *Soient deux termes d'une ontologie t_1 et t_2, la similarité de Resnik entre ces deux termes est définie telle que :*

$$sim_{Resnik}(t_1, t_2) = IC(LCA(t_1, t_2))$$

Le LCA n'étant pas défini de manière unique dans une ontologie, on utilisera plutôt l'ancêtre avec le plus grand contenu d'information ou *Most Informative Common Ancestor* (MICA) [Resnik et al., 1999].

Définition 11. *Le* Most Informative Common Ancestor *(MICA) de deux classes t_1 et t_2 d'une ontologie est l'ancêtre commun à t_1 et t_2 maximisant IC, tel que :*

$$MICA(t_1, t_2) = t \mid t \in LCA(t_1, t_2) \wedge IC(t) \geq \max_{x \in LCA(t_1,t_2)} IC(x)$$

4. La fonction logarithme n'étant définie que pour les réels strictement positifs, IC n'est définie que sur l'ensemble des variables aléatoires de probabilité non-nulle.

Similarité de Lin La similarité de Lin [Lin et al., 1998] est fondée sur le même principe que la similarité de Resnik, mais est bornée par l'intervalle $[0, 1]$. Afin d'obtenir une fonction bornée, l'IC du LCA ou du MICA est normalisé par la somme des IC des termes comparés.

Définition 12. *Soient deux termes d'une ontologie t_1 et t_2, la similarité de Lin entre ces deux termes est définie telle que :*

$$sim_{Lin}(t_1, t_2) = \frac{2 \times IC(LCA(t_1, t_2))}{IC(t_1) + IC(t_2)}$$

Contrairement à la similarité de Resnik, cette similarité permet de garantir que $sim_{Lin}(t_1, t_2) = 1$ quand $t_1 = t_2$.

Simialrité de Wu-Palmer La similarité de Wu-Palmer [Wu and Palmer, 1994] est une similarité entre deux termes d'une ontologie. Cette similarité compare la profondeur de deux termes à celle de leur LCA.

Définition 13. *Soient deux termes d'une ontologie t_1 et t_2, la similarité de Wu-Palmer entre ces deux termes est définie telle que :*

$$sim_{Wu-Palmer}(t_1, t_2) = 2 \times \frac{Profondeur(LCA(t_1, t_2))}{Profondeur(t_1) + Profondeur(t_2)}$$

On note que cette mesure de similarité utilise la profondeur des différents termes plutôt que leur contenu d'information. Cependant, dans une ontologie, aucune sémantique formelle n'est associée à la profondeur d'un terme. Cette mesure de similarité peut donc être affectée par des modifications de l'ontologie qui ne changent pas la sémantique commune des termes comparés, mais ne nécessite pas de ressource externe à l'ontologie.

Aggrégation de similarités entre termes

Les similarités sémantiques comparant des termes peuvent être utilisées pour comparer des ensembles de termes à l'aide d'une fonction d'aggrégation.

Etant donnés deux ensembles de termes A et B et une similarité sim, on peut par exemple considérer la moyenne des similarités pair-à-pair entre ces termes [Lord et al., 2003] :

$$MoyennePairAPair(A, B) = \frac{\sum_{(t_i, t_j) \in A \times B} sim(t_i, t_j)}{|A \times B|}$$

Cette aggrégation a cependant l'inconvénient de ne pas respecter l'identité des indiscernables pour des ensembles de taille supérieure à 1. Par exemple, soient deux termes t_1 et t_2 et une similarité telles que $sim(t_1, t_1) = 1$, $sim(t_2, t_2) = 1$ et $sim(t_1, t_2) = 0$:

$$\begin{aligned} MoyennePairAPair(\{t_1, t_2\}, \{t_1, t_2\}) &= \frac{sim(t_1, t_1) + sim(t_1, t_2) + sim(t_2, t_1) + sim(t_2, t_2)}{4} \\ &= 0.5 \end{aligned}$$

Cet inconvénient provient du fait que l'on considère des similarités faibles (entre t_1 et t_2 dans l'exemple) alors qu'un terme beaucoup plus proche existe dans l'autre ensemble. Dans l'exemple ci-dessus, il semblerait suffisant de considérer la similarité de t_1 à lui-même et de t_2 à lui-même.

Pour remédier à ce problème, on peut utiliser une fonction d'aggrégation qui ne considère que les similarités les plus élevées [Sevilla et al., 2005] :

$$MaximumPairAPair(A,B) = \max_{(t_i,t_j)\in A\times B} sim(t_i,t_j)$$

Cela limite cependant l'information prise en compte par la similarité à une seule paire de termes, ce qui peut rendre cette méthode d'aggrégation peu précise pour comparer des grands ensembles de termes.

La fonction d'aggrégation *Best-Match Average* (BMA) [Schlicker et al., 2006, Couto et al., 2007] permet d'associer les avantages de ces deux approches, en considèrant pour chaque terme de A uniquement la meilleure similarité avec un terme de B (et inversement pour B et A de manière à ce que la similarité reste symmétrique) :

$$BMA(A,B) = \left(\frac{\forall_{t_i\in A}\max_{t_j\in B}(sim(t_i,t_j))}{|A|} + \frac{\forall_{t_i\in B}\max_{t_j\in A}(sim(t_i,t_j))}{|B|}\right)\times\frac{1}{2}$$

SimGIC

SimGIC [Pesquita et al., 2007] est une mesure de similarité sémantique permettant de comparer des objets, représentés par les ensembles de leurs annotations, étendus par les ancêtres de ces annotations dans l'ontologie. Elle utilise un principe proche de la similarité de Jaccard, et compare les termes et ancêtres communs à deux objets annotés. SimGIC ajoute à cela une pondération de chaque terme en fonction de son contenu d'information, de manière à donner une plus grande importance aux termes les plus spécifiques.

Définition 14. *Soient deux objets représentés sous la forme d'ensembles de termes A et B, la mesure de similarité simGIC est définie telle que :*

$$simGIC(A,B) = \frac{\sum_{t\in A\cap B} IC(t)}{\sum_{t\in A\cup B} IC(t)}$$

Similarités dans un espace vectoriel

On peut utiliser une représentation vectorielle d'objets annotés par des termes d'un vocabulaire pour permettre leur comparaison : deux objets sont d'autant plus similaires que l'angle entre les vecteurs les représentant est faible. Ainsi, on peut représenter un objet annoté par des termes d'un vocabulaire de termes $t_1,\ldots,t_n$ comme un vecteur $(a_1,\ldots,a_n)$, où $a_i = 1$ si cet objet est annoté par t_i, ou 0. On peut alors définir la mesure de similarité cosinus entre deux objets comme le cosinus de l'angle des vecteurs les représentant [Singhal et al., 2001].

Définition 15. *Soient $\overrightarrow{A}$ et $\overrightarrow{B}$ deux vecteurs de dimension n représentant deux objets annotés par les termes $t_1,\ldots,t_n$. La similarité cosinus entre $\overrightarrow{A}$ et $\overrightarrow{B}$ est égale à $\cos(\theta)$ où θ est l'angle entre $\overrightarrow{A}$ et $\overrightarrow{B}$, et est définie telle que :*

$$sim_{cosinus}(\overrightarrow{A},\overrightarrow{B}) = \frac{\overrightarrow{A}\cdot\overrightarrow{B}}{||\overrightarrow{A}||\times||\overrightarrow{B}||}$$

Le cosinus de l'angle entre deux vecteurs est de 1 pour un angle nul (deux vecteurs égaux) et de 0 pour deux vecteurs orthogonaux (aucune annotation en commun).

Pondération *tf-idf* *tf-idf* (*term frequency-inverse document frequency*) est utilisé en recherche d'information pour quantifier l'importance d'un mot dans un document. Il peut être utilisé pour donner un poids différent à chaque composante d'un vecteur d'annotations. Le calcul de *tf-idf* pour un terme t dans un document d est communément divisé en deux parties :

— *term frequency* : une pondération en fonction de la fréquence d'apparition de t dans d. Cette pondération peut notamment être binaire (1 si t apparaît dans d, 0 sinon) ou égale au nombre d'apparitions de t dans d normalisé ou non par la taille du document.

— *inverse document frequency* : une pondération similaire au contenu d'information, plus forte pour les termes apparaissant dans peu de documents. Soit D l'ensemble des documents :

$$\text{idf}(t) = -\log \frac{|\{d_i | d_i \in D \wedge t \text{ apparaît dans } d_i\}|}{|D|}$$

Ainsi la valeur de *tf-idf* pour un terme t et document d est égale à $\text{tf-idf}(t, d) = \text{tf}(t, d) \times \text{idf}(t)$.

On peut utiliser *tf-idf* pour la comparaison d'objets annotés par un ensemble de termes en considérant un objet comme un document. Chaque objet peut alors être représenté par un vecteur $\overrightarrow{A} = (\text{tf-idf}(t_1, A), \ldots, \text{tf-idf}(t_n, A))$ et comparé à l'aide de la similarité cosinus.

IntelliGO IntelliGO est une mesure de similarité fondée sur un modèle d'espace vectoriel, permettant de comparer des objets annotés par des termes d'une ontologie [Benabderrahmane et al., 2010]. Contrairement à la similarité cosinus, IntelliGO considère que les différentes dimensions du vecteur, c'est-à-dire, les termes du vocabulaire d'annotation, ne sont pas indépendants.

Définition 16. *Soit $\{\overrightarrow{e_1}, \ldots, \overrightarrow{e_n}\}$ un ensemble de vecteurs correspondant aux termes $t_1, \ldots t_n$ du vocabulaire d'annotation. $\{\overrightarrow{e_1}, \ldots, \overrightarrow{e_n}\}$ est une famille génératrice de l'espace vectoriel des vecteurs objets. Ainsi tout vecteur objet $\overrightarrow{g}$ peut s'exprimer comme une combinaison linéaire des éléments de cette famille, tel que :*

$$\overrightarrow{g} = \sum_i \alpha_i \times \overrightarrow{e_i}$$

Ici, α_i correspond au coefficient de pondération de l'annotation t_i pour l'objet $\overrightarrow{g}$. IntelliGO propose une fonctionnalité supplémentaire par rapport aux autres mesures de similarités puisqu'il permet d'intégrer une pondération des termes en fonction de l'origine de leur annotation : fondée sur des données expérimentales, extraite d'une publication, inférence automatique, etc. En effet, IntelliGO a été développé pour calculer des similarités sémantiques entre gènes annotés par des classes de Gene Ontology.

Ainsi, IntelliGO permet de donner un poids différent à chaque annotation en fonction du type de preuve supportant celles-ci pour des gènes comparés. Si l'objet n'est pas annoté par un terme, le poids correspondant est alors égal à 0. Cette pondération vient s'ajouter à la pondération par le contenu d'information de l'annotation, définie dans IntelliGO comme *Inverse Annotation Frequency* ou IAF. On a alors :

$$\alpha_i = w(g, t_i) \times IC(t_i)$$

IntelliGO utilise une ontologie pour permettre la comparaison de termes proches. Les similarités fondées sur un espace vectoriel classique utilisent une famille de vecteurs génerateurs orthogonaux entre eux, ce qui présuppose de l'indépendance des annotations. Au contraire, IntelliGO définit un produit scalaire entre deux vecteurs générateurs, exprimé en fonction de la profondeur dans la hiérarchie des deux annotations correspondantes et de leur ancêtre commun le plus spécifique.

Définition 17. *Soient $\overrightarrow{e_i}$ et $\overrightarrow{e_j}$ deux vecteurs génerateurs correspondant respectivement aux termes t_i et t_j, IntelliGO définit le produit scalaire entre deux tels vecteurs comme :*

$$\overrightarrow{e_i} * \overrightarrow{e_j} = 2 \times \frac{Profondeur(LCA(t_i, t_j))}{Profondeur(t_i) + Profondeur(t_j)}$$

où LCA est l'ancêtre commun de profondeur maximale.

IntelliGO définit ensuite un produit scalaire généralisé pour comparer les vecteurs d'annotations.

Définition 18. *Soient $\overrightarrow{g} = \sum_i \alpha_i \times \overrightarrow{e_i}$ et $\overrightarrow{h} = \sum_j \beta_j \times \overrightarrow{e_j}$ deux vecteurs de termes, IntelliGO définit le produit scalaire entre deux tels vecteurs comme :*

$$\overrightarrow{g} * \overrightarrow{h} = \sum_{i,j} \alpha_i \times \beta_j \times \overrightarrow{e_i} * \overrightarrow{e_j}$$

Finalement, IntelliGO définit la similarité entre deux objets annotés par des classes d'ontologies représentés par les vecteurs $\overrightarrow{g}$ et $\overrightarrow{h}$ de la même manière que la similarité cosinus.

Définition 19. *Soient $\overrightarrow{g} = \sum_i \alpha_i \times \overrightarrow{e_i}$ et $\overrightarrow{h} = \sum_j \beta_j \times \overrightarrow{e_j}$ deux vecteurs de termes, IntelliGO définit le similarité entre ces deux vecteurs comme :*

$$sim_{IntelliGO}(\overrightarrow{g}, \overrightarrow{h}) = \frac{\overrightarrow{g} * \overrightarrow{h}}{\sqrt{\overrightarrow{g} * \overrightarrow{g}}\sqrt{\overrightarrow{h} * \overrightarrow{h}}}$$

Les mesures de similarité présentées ici proposent différentes manières de comparer des termes, ou des objets annotés par des termes d'une ontologie. La Table 1.1 présente un résumé de ces différentes mesures, ainsi que leur différentes propriétés et les informations qu'elles considèrent. On notera qu'aucune de ces caractéristiques ne permet de juger de la performance de ces mesures de similarité, mais peuvent guider le choix d'une de ces mesures pour une application précise.

TABLE 1.1 – Comparaison des différentes mesures de similarités présentées dans cette section.

- Bornée : indique si la similarité est bornée par l'intervalle $[0, 1]$
- Identité : indique si la similarité vérifie l'identité des indiscernables. Le symbole $\Leftrightarrow$ dénote que $x = y \Leftrightarrow sim(x, y) = 1$. Le symbole $\Rightarrow$ dénote que $x = y \Rightarrow sim(x, y) = 1$.
- CA : indique si la similarité prend en compte les ancêtres communs (pas uniquement le ou les LCA) des termes comparés
- LCA : indique si la similarité prend en compte le LCA ou MICA des termes comparés
- IC : indique si la similarité utilise *Information Content*
- Profondeur : indique si la similarité prend en compte la profondeur des termes comparés et de leur LCA
- $\triangle$: indique si il existe une métrique définie par cette similarité vérifiant l'inégalité triangulaire

(1) Les aggrégations de similarités pair-à-pair héritent cette propriété de la similarité de termes
(2) Les ensembles comparés peuvent contenir soit les annotations les plus spécifiques, soit les annotations avec leurs ancêtres
(3) L'aggrégation maximum pair-à-pair vérifie $x = y \Rightarrow sim(x, y)$ si sa similarité de termes le vérifie également, mais ne peux pas vérifier $x = y \Leftrightarrow sim(x, y) = 1$
(4) Les vecteurs de termes comparés peuvent être pondérés par l'IC de chaque terme
L'ensemble des mesures de similarités présentées ici sont symmétriques.

Similarité	**Bornée**	**Identité**	**CA**	**LCA**	**IC**	**Profondeur**	$\triangle$
Dice	✓	$\Leftrightarrow$	✓				
Jaccard	✓	$\Leftrightarrow$	✓				✓
Resnik				✓	✓		
Lin	✓	$\Rightarrow$		✓	✓		
Wu-Palmer	✓	$\Leftrightarrow$		✓		✓	
Moyenne p.-à-p.	(1)		(2)	(1)	(1)	(1)	(1)
Maximum p.-à-p.	(1)	(3)	(2)	(1)	(1)	(1)	
BMA	(1)	(1)	(2)	(1)	(1)	(1)	
SimGIC	✓	$\Leftrightarrow$	✓		✓		✓
cosinus	✓	$\Leftrightarrow$	(2)		(4)		
IntelliGO	✓	$\Leftrightarrow$	✓	✓	✓	✓	

1.3 Fouille de données

La fouille de données est un processus d'extraction automatique de régularités dans un ensemble de données dans le but d'en extraire une connaissance [Fayyad et al., 1996]. On s'intéressera ici à plusieurs paradigmes de fouille :

— l'extraction de règles d'association, notamment avec l'analyse formelle de concepts et les structures de patrons ;
— le clustering permettant de partitionner un ensemble d'objets sur la base de leur similarités ;
— la programmation logique inductive permettant d'extraire des implications logiques entre des faits sur des exemples à décrire.

1.3.1 Extraction de règles d'association

L'extraction de règles d'association [Agrawal et al., 1993] est une méthode pour la découverte d'éléments fréquemment associés dans un jeu de donnés. Classiquement, l'extraction de règles d'association s'effectue sur un ensemble de transactions, représentées par des ensembles d'objets.

Une règles d'association est composée de deux ensembles d'objets L et R, et est notée $L \rightarrow R$. Une telle règle est interprétée comme « quand les objets L sont présents dans une transaction, alors les objets R le sont aussi ». Ces règles expriment ainsi une co-occurence d'objets, sans présumer d'une relation causale ou temporelle entre eux.

Une règle d'association peut être évaluée via différentes mesures, notamment la confiance et le support. La confiance d'une règle est la proportion de transaction contenant L contenant également R. Le support d'une règle est le nombre de transactions contenant L et R[5]. Par exemple, si une règle $\{A, B\} \rightarrow \{C\}$ a une confiance de 5/7 et un support de 5, alors C est présent dans 5 des 7 transactions où A et B sont présents, et A, B et C sont présents ensemble dans 5 transactions.

Plusieurs algorithmes pour l'extraction de règles d'association existent, comme par exemple l'agorithme Apriori, fondé sur les ensembles d'objets fréquents[Agrawal et al., 1994]. De tels ensembles fréquents peuvent notamment être identifiés par la construction d'un treillis d'ensembles d'éléments, défini par l'inclusion [Pasquier et al., 1999]. Dans les sections suivantes sont présentées l'Analyse Formelle de Concepts et son extension les structures de patrons, qui peuvent être utilisées pour l'extraction de règles d'association [Luxenburger, 1991, Lakhal and Stumme, 2005].

1.3.2 Analyse Formelle de Concepts

L'Analyse Formelle de Concepts (ou *FCA – Formal Concept Analysis*) [Ganter and Wille, 1997] est un cadre mathématique fondé sur la théorie des treillis pouvant être utilisé pour l'analyse de données et la découverte de connaissances. La FCA utilise des données représentées sous la forme d'un contexte formel, où des objets sont associés à des attributs par une relation binaire appelée relation d'incidence.

Définition 20. *Un contexte formel est un triplet (G, M, I) où G est un ensemble d'objets, M est un ensemble d'attributs et I est une relation binaire entre les objets de G et les attributs de M. Ainsi, pour un objet g et un attribut m, gIm indique que l'objet g possède l'attribut m.*

5. On peut également calculer le support relativement au nombre total de transactions, comme la proportion des transactions contenant L et R.

La Table 1.2 présente un exemple de contexte formel décrivant les différentes propriétés de quelques mesures de similarité. Dans cet exemple, les objets du contexte formel sont G = {Dice, Jaccard, Resnik, Wu_Palmer, SimGIC, IntelliGO} et les attributs sont M = {bornée, identité, ca, lca, ic, profondeur, it}.

TABLE 1.2 – Exemple de contexte formel pouvant être utilisé pour l'extraction de règles d'association en FCA, ayant pour objets les mesures de similarités décrites en Table 1.1 (l'attribut it correspondant à la la propriété "vérifie l'inégalité triangulaire").

	bornée	**identité**	**ca**	**lca**	**ic**	**profondeur**	**it**
Dice	×	×	×				
Jaccard	×	×	×				×
Resnik				×	×		
Wu_Palmer	×	×		×		×	
SimGIC	×	×	×		×		×
IntelliGO	×	×	×	×	×	×	

Il existe, entre les objets G et les attributs M du contexte formel, une correspondance de Galois .

Définition 21. *Une correspondance de Galois est un couple de fonctions (m_1, m_2) ayant respectivement pour domaine deux ensembles partiellement ordonnés $(P, \leq_P)$ et $(Q, \leq_Q)$ telles que pour $p \in P$ et $q \in Q$:*

- $m_1 : P \to Q$
- $m_2 : Q \to P$
- $q \leq_Q m_1(p) \Leftrightarrow p \leq_P m_2(q)$ *dans le cas d'une correspondances de Galois antitone*
- $m_1(p) \leq_Q q \Leftrightarrow p \leq_P m_2(q)$ *dans le cas d'une correspondances de Galois isotone*

Cette correspondance est établie via deux opérateurs de dérivation qui associent à un ensemble d'objets l'ensemble des attributs correspondants, et inversement.

Définition 22. *Pour un contexte formel (G, M, I), la correspondance de Galois entre $(2^G, \subseteq)$ et $(2^M, \subseteq)$ est définie, pour tout $A \subseteq G$ et $B \subseteq M$, par les opérateurs de dérivation notés $'$:*

$$A' = \{m \in M \mid gIm \text{ pour tout } g \in A\}$$

$$B' = \{g \in G \mid gIm \text{ pour tout } m \in B\}$$

Le contexte formel peut être organisé sous la forme d'un treillis de Galois de concepts formels, c'est-à-dire, une structure hiérarchique dans laquelle chaque nœud représente un concept formel, c'est-à-dire un ensemble d'objets partageant un ensemble d'attributs.

Définition 23. *Un concept formel est un couple (A, B), où A est un ensemble d'objets $A \subseteq G$, B est un ensemble d'attributs $B \subseteq M$, tels que $A' = B$; $B' = A$. On appelle alors A l'extension du concept et B son intention.*

Les concepts formels sont ordonnés par un ordre partiel défini sur l'inclusion de leur extensions, ou dualement sur l'inclusion de leur intentions.

Définition 24. *Soit un ordre partiel $\leq$ entre les concepts, tel que $(A_1, B_1) \leq (A_2, B_2)$ si et seulement si $A_1 \subseteq A_2$ et $B_2 \subseteq B_1$.*

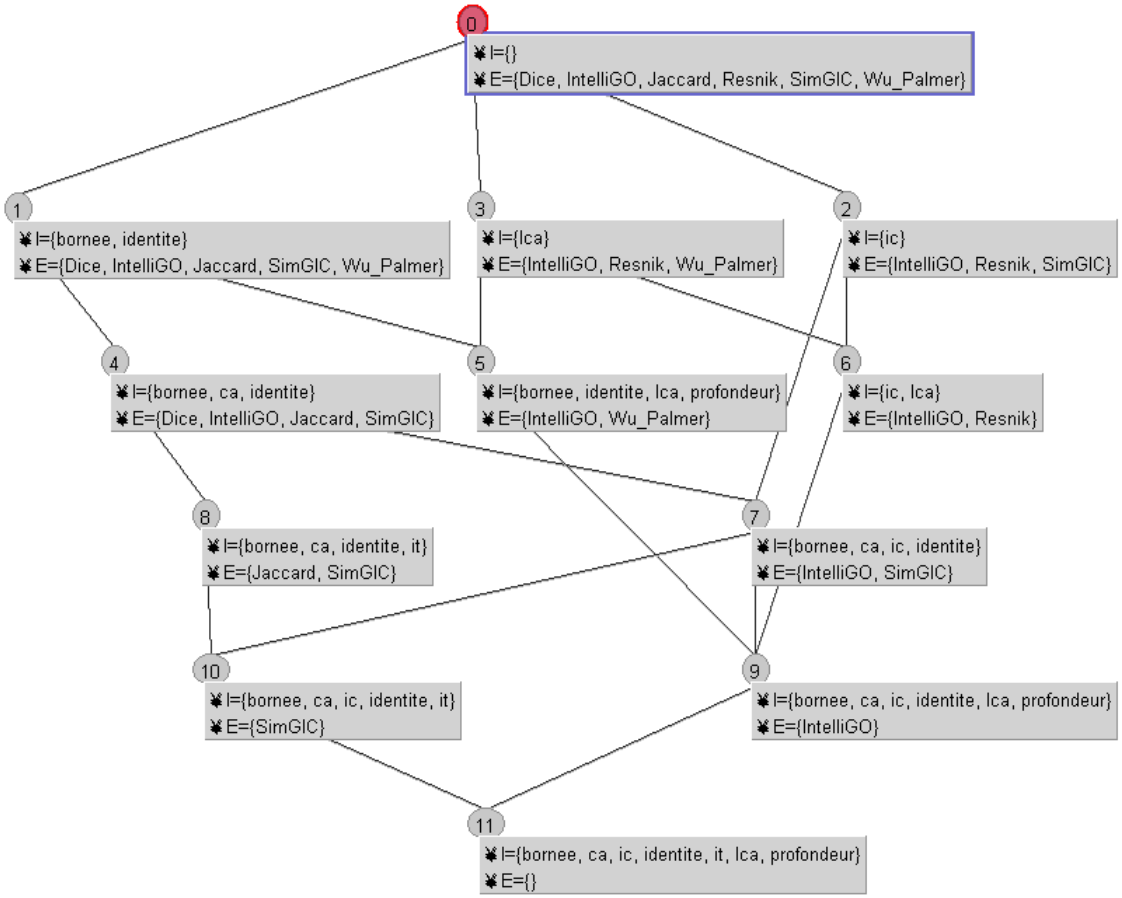


FIGURE 1.10 – Treillis de concepts généré à partir du contexte formel de la Table 1.2 en utilisant le logiciel Galicia [Valtchev et al., 2003]. Pour chacun des concepts, I est l'intention du concept et E son extension. Les arcs entre les concepts dénotent de bas en haut l'ordre partiel entre les concepts.

Ainsi, le contexte formel présenté en Table 1.2 peut être représenté sous la forme du treillis de concepts en Figure 1.10. Un tel treillis permet de faciliter l'extraction de règles d'association entre les attributs de M, en exposant comme intention de concepts des sous-ensembles d'attributs fréquents, et maximaux pour les objets qu'ils décrivent. Par exemple, pour le concept numéroté 2 sur la Figure 1.10, l'ensemble d'attributs {ic} est l'ensemble maximal d'attributs décrivant les objets {IntelliGO, Resnik, SimGIC}. Luxenburger [Luxenburger, 1991] propose une méthode simple pour l'extraction de règles d'association à partir d'un tel treillis. Soient deux concepts (A', A) et (B', B) directement liés dans un treillis de concepts tels que $(B', B) \leq (A', A)$, on peut extraire la règle d'association $A \rightarrow B \setminus A$ de support $|B'|$ et de confiance $|B'|/|A'|$.

A partir des concepts numérotés 2 et 6, on pourra par exemple extraire la règle d'association {ic} $\rightarrow$ {lca} de support 2 et de confiance 2/3.

Algorithmes De nombreux algorithmes existent pour la construction du treillis de concept en FCA. Certains de ces algorithmes considèrent l'ensemble des objets du contexte afin de produire des concepts du treillis [Bordat, 1986], tandis que d'autres algorithmes ajoutent les objets au

treillis un par un, en effectuant les modifications nécessaires. On s'intéressera ici à ce deuxième type d'algorithme, dits incrémentaux, en particulier aux algorithmes *AddIntent* [Van Der Merwe et al., 2004, Kourie et al., 2009] et *FastAddIntent* [Zou et al., 2015] .

Ces deux algorithmes explorent le treillis de concepts pour chaque objet, en partant de l'élement minimal du treillis pour trouver des concepts dits générateurs, à partir desquels les nouveaux concepts seront créés. Ces deux algorithmes reposent sur deux fonctions : `AddIntent` : $2^M \times Concepts \times Treillis \rightarrow Treillis$ et `GetMaximalConcept` : $2^M \times Concepts \times Treillis \rightarrow Concepts$. La fonction `GetMaximalConcept` prend en paramètre une intention i, un concept (A, B) et un treillis L. `GetMaximalConcept` explore le treillis à partir de (A, B) pour trouver le concept maximal $(C, D) \leq (A, B)$ avec $D \geq i$. Cette fonction permet de trouver les concepts générateurs du concept d'intention i. La fonction `AddIntent` prend en paramètre une intention i, un concept (A, B) et un treillis L et modifie L de manière à ce que L possède un concept d'intention i, en utilisant `GetMaximalConcept`.

Pour générer le treillis, la fonction `AddIntent` est appelée pour chaque objet $g \in G$, avec comme paramètres g', le concept minimal et le treillis à construire. Ces deux algorithmes présentent l'avantage d'être bien plus rapides que les autres algorithmes existants [Van Der Merwe et al., 2004, Zou et al., 2015]. De plus, les différentes implémentations des fonctions `AddIntent` et `GetMaximalConcept` sont interchangeables [6].

1.3.3 Structures de patrons

Les structures de patrons [Ganter and Kuznetsov, 2001] généralisent l'analyse formelle de concepts de manière à l'appliquer à des objets dotés de descriptions non seulement binaires, mais de nature plus complexe, comme par exemple des ensembles, graphes ou intervalles [Kaytoue et al., 2015]. En particulier, les structures de patrons ont été utilisées pour tirer parti des connaissances biomédicales pour fouiller des données annotées par des termes liés à des ontologies [Coulet et al., 2013].

Définition 25. *Une structure de patrons est un triplet $(G, (\mathcal{D}, \sqcap), \delta)$ où :*

- *G est un ensemble d'objets*
- *$\mathcal{D}$ est un ensemble de descriptions,*
- *$\sqcap$ est un opérateur de comparaison de descriptions tel que, pour deux descriptions X et Y dans $\mathcal{D}$, $X \sqcap Y$ est la similarité de X et Y. $X \sqcap Y$ est une description plus générale décrivant X et Y. Cet opérateur définit un ordre partiel $\leq_\sqcap$ sur les éléments de $\mathcal{D}$ tel que $X \leq_\sqcap Y \Leftrightarrow X \sqcap Y = X$.*
- *δ est une fonction qui associe une description à un objet.*

Il existe des similitudes entre un contexte formel (G, M, I) et une structure de patrons $(G, (\mathcal{D}, \sqcap), \delta)$. Notamment, il est possible d'exprimer un contexte formel binaire sous la forme d'une structure de patrons où les descriptions sont des ensembles d'attributs et où l'opérateur de comparaison $\sqcap$ est défini comme l'intersection ensembliste.

Définition 26. *Pour une structure de patrons, l'opérateur de dérivation $.^\square$ définit une correspondance de Galois entre $(2^G, \subseteq)$ et $(\mathcal{D}, \leq_\sqcap)$, telle que :*

$$A^\square = \prod_{g \in A} \delta(g) \text{ pour un ensemble d'objets } A \subseteq G$$

$$d^\square = \{g \in G \mid d \leq_\sqcap \delta(g)\} \text{ pour une description } d \in \mathcal{D}$$

6. Dans [Zou et al., 2015], `GetMaximalConcept` est nommée `GetClosureConcept` et `AddIntent` est nommée `FastAddIntent`.

Intuitivement, $A^{\square}$ est la description la plus spécifique (maximale selon $\leq_{\sqcap}$) pour l'ensemble d'objets A, et $d^{\square}$ est l'ensemble de tous les objets dont la description est plus spécifique ou égale à d. Une structure de patrons peut être représentée sous la forme d'un treillis de Galois de concepts formels.

Définition 27. *Un concept formel d'une structure de patrons est un couple (A, d), où A est un ensemble d'objets $A \subseteq G$ et d une description $d \in \mathcal{D}$ et $A^{\square} = d$; $d^{\square} = A$. On appelle alors A l'extension du concept et d son intention.*

Les concepts formels sont ordonnés par un ordre partiel défini sur l'inclusion de leur extensions, ou dualement sur l'ordre $\leq_{\sqcap}$ des descriptions de leur intention.

Définition 28. *Soit un ordre partiel $\leq$ entre les concepts, tel que $(A_1, d_1) \leq (A_2, d_2)$ si et seulement si $A_1 \subseteq A_2$ et $d_2 \leq_{\sqcap} d_1$.*

Le treillis d'une structure de patrons est alors défini par l'ensemble partiellement ordonné des concepts formels.

1.3.4 Clustering

Le clustering est un processus de fouille de données non supervisé, dont l'objectif est de former, à partir d'un ensemble d'objets, des groupes, ou *clusters*, d'objets similaires entre eux. La similarité entre les objets peut être exprimée à l'aide d'une fonction de similarité où d'une semi-métrique [Rokach and Maimon, 2005].

Clustering hiérarchique Le clustering hiérarchique regroupe un ensemble d'algorithmes permettant d'organiser des clusters dans un dendrogramme, c'est-à-dire, un arbre où l'ensemble des objets constitue le nœud racine, et où les enfants $e_1, \ldots, e_n$ d'un nœud N sont une partition de N. Chacune des feuilles de l'arbre représente un cluster d'un seul objet.

Il existe deux types d'algorithmes de clustering hiérarchiques :

— Les algorithmes par agglomération (approche *bottom-up*) qui construisent le dendrogramme à partir des feuilles. Initialement, chaque objet forme un cluster, et l'algorithme joint les deux clusters les plus proches jusqu'à n'en obtenir plus qu'un. Chaque fusion entre deux clusters est représentée dans le dendrogramme par la création d'un nœud parent à ces deux clusters. Il existe plusieurs méthodes pour évaluer la distance entre deux clusters à partir des distances pair-à-pair de leurs éléments, notamment en considérant le minimum [Sibson, 1973], maximum [Defays, 1977] ou la moyenne des différences pair-à-pair.

— Les algorithmes par division (approche *top-down*) qui construisent le dendrogramme à partir de la racine. Un seul cluster initial contient tout les objets, et l'algorithme divise le cluster le plus grand jusqu'à obtenir uniquement des clusters de taille 1 [Kaufman and Rousseeuw, 2009].

Les algorithmes de clustering hiérarchique permettent d'obtenir une partition de l'ensemble d'objets de départ en opérant une coupe dans le dendrogramme [Langfelder et al., 2007]. Cette coupe peut être faite sur plusieurs critères, notamment avec un seuil de similarité entre les clusters, ou en utilisant différents indices d'évaluation de clustering.

Clustering par k-moyennes et k-médoïdes Les algorithmes de clustering de type k-moyennes [MacQueen et al., 1967, Lloyd, 1982] et k-médoïdes [Kaufman and Rousseeuw, 1987] permettent de former un partitionnement d'un ensemble de d'objets en k clusters organisés autour d'un "centre".

Initialement, k centres de clusters sont choisis arbitrairement et chaque objet est associé au cluster dont le centre est le plus proche. Le centre de chaque cluster est ensuite recalculé, puis les objets sont de nouveau associés au cluster de centre le plus proche. Cette opération est répétée jusqu'à ce qu'un critère de convergence soit atteint, ou que le nombre maximal d'itérations soit dépassé.

Les méthodes des k-moyennes et k-médoïdes diffèrent sur la façon de calculer le centre de chaque cluster. La méthode des k-moyennes utilise un centre "artificiel" comme centre de cluster, correspondant à la moyenne des objets du clusters. Cette méthode est adéquate pour le clustering d'objets représentés sous forme de vecteurs numériques. Cependant, il n'est pas possible de l'utiliser pour des objets sur lesquels le calcul de la moyenne n'est pas possible, comme par exemple des gènes annotés par des termes GO. Dans ce cas, la méthode des k-médoïdes peut être utilisée : à la place de calculer le centre du cluster par la moyenne, un de ces objets est choisi comme centre du cluster. Ce point, dit médoïde, est sélectionné parmi les objets du cluster de manière à minimiser la somme des distances des points du cluster avec le médoïde.

Clustering flou Le clustering flou est une forme de clustering dans lequel chacun des objets peut appartenir à plusieurs clusters, avec différents degrés d'appartenance. Le clustering flou peut être plus approprié pour résoudre des problèmes de fouille de données où un objet peut appartenir à plusieurs groupes.

L'algorithme *Fuzzy c-means* [Dunn, 1973] est une version "floue" de k-means. Aux objets sont attribués un certain degré d'appartenance à chaque cluster, ce degré décroit avec leur distance au centre de ce cluster.

L'agorithme FANNY [Kaufman and Rousseeuw, 2009] est un autre algorithme de clustering flou, qui recherche les degrés d'appartenance de chaque objet à k clusters, de manière à minimizer la fonction :

$$\sum_{v=1}^{k} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} u_{i,v}^{r} \times u_{j,v}^{r} \times d_{i,j}}{2 \times \sum_{i=1}^{n} u_{j,v}^{r}}$$

où $u_{i,j}$ est le coefficient d'appartenance de l'objet i au cluster j, n est le nombre de d'objets, r est l'exposant d'appartenance et $d_{i,j}$ est la dissimilarité de i et j. L'exposant d'appartenance est un paramètre spécifique à FANNY, un grand exposant d'appartenance augmentant le caractère flou des clusters.

Pureté d'un clustering La pureté d'un clustering est une mesure d'évaluation d'un clustering par rapport à une classfication de référence [Larson, 2010].

Définition 29. *Soient une partition $\Omega = \omega_1, \ldots, \omega_n$ de N objets répartis en k clusters et un ensemble de classes $C = c_1, \ldots, c_J$. La pureté de Ω par rapport à C est définie comme :*

$$\text{pureté}(\Omega, C) = \frac{1}{N} \sum_{i=1}^{k} \max_{j} |\omega_i \cap c_j|$$

La pureté assigne à chaque cluster la classe la plus proche (c'est-à-dire dont l'intersection avec ce cluster est de cardinalité maximale) et calcule la proportion d'objets bien classés.

Indice de Rand L'indice de Rand [Rand, 1971] est une mesure de la similarité de deux partitions. Ils peut alors être utilisé pour comparer le résultat d'un clustering à un partitionnement ou une classification de référence.

Définition 30. *Soient deux partitions X et Y d'un même ensemble d'objets E, on défini l'indice de Rand R de ces deux partitions comme :*

$$R = \frac{a+b}{\binom{|E|}{2}}$$

où :
- *a est le nombre de paires (non-ordonnées) d'éléments de E qui sont dans le même cluster de X et dans le même cluster de Y ;*
- *b est le nombre de paires (non-ordonnées) d'éléments de E qui sont dans deux clusters différents de X et dans deux clusters différents de Y.*

Intuitivement, l'indice de Rand représente la proportion (de 0 à 1) de paires d'éléments de E pour lesquels X et Y sont en accord sur leur appartenance ou non au même cluster.

L'indice de Rand accorde la même importance aux paires correctement affectées à un même cluster (a), et aux paires correctement affectées à des clusters différents (b). Cela peut poser problème puisque dans de nombreux cas (notamment quand le nombre de clusters est élevé), une partition choisie de manière aléatoire aura une forte valeur de b. Afin de quantifier la différence entre une bonne partition et une partition aléatoire, on peut utiliser l'indice de Rand ajusté [Hubert and Arabie, 1985].

Définition 31. *Soient deux partitions X et Y d'un même ensemble d'objets E, on défini l'indice de Rand ajusté ARI (*Adjusted Rand Index*) de ces deux partitions comme :*

$$ARI = \frac{R - E(R)}{\max(R) - E(R)}$$

où :
- *R est l'indice de Rand de X et Y ;*
- *$E(R)$ est l'espérance de R ;*
- *$\max(R)$ est la valeur maximale de R.*

L'indice de Rand ajusté compare effectivement l'indice de Rand d'une partition avec l'indice de Rand attendu pour une partition aléatoire. Il existe notamment plusieurs manières équivalentes de le calculer sans avoir à génerer de partition aléatoire [Warrens, 2008].

1.3.5 Programmation Logique Inductive

La Programmation Logique Inductive (PLI) permet l'apprentissage de définitions de concepts (ici, des ensembles d'objets) sur la base d'observations [Muggleton, 1991]. Ces observations se présentent sous la forme d'un ensemble d'exemples positifs et d'un ensemble d'exemples négatifs, et peuvent être accompagnées de connaissances de domaine. A partir de ces observations et connaissances, la finalité de la PLI est d'induire un ensemble de règles, ou théorie, décrivant un maximum d'exemples positifs, et un minimum d'exemple négatifs.

Dans la plupart des systèmes d'ILP, les données sur les exemples, les connaissances de domaine et les théories générées sont représentées en logique du premier ordre, notamment grâce au langage Prolog [Clocksin and Mellish, 2003]. Ces connaissances et les règles de la théorie sont exprimées sous la forme de clauses de Horn : une disjonction de littéraux dont un seul est positif, par exemple $q \vee \neg p_1 \vee \ldots \vee \neg p_n$. Une clause de Horn peut également s'écrire sous la forme d'une implication équivalente $(p_1 \wedge \ldots \wedge p_n) \Rightarrow q$. En Prolog, une telle clause s'écrit sous la forme

`q :- p1, ...,pn`, où `q` forme la «tête» de la clause, et `p1, ...,pn` forme le « corps » de la clause. Une telle clause peut s'interpréter de la manière suivante : « si les conditions `p1, ...,pn` sont vraies, alors q est vraie ». Une clause dont le corps est vide est appelée un fait, puisque la vérité de la tête de la clause n'est pas conditionnée.

Les données sur les exemples incluent des descriptions des examples sous la forme de faits, construits à prédicats n-aires et de littéraux, tels que

`protein_mf`('gpaC','receptor binding').

exprimant que le fait que le terme Gene Ontology 'receptor binding' est une des fonctions moléculaire de la protéine 'gpaC'.

Les connaissances du domaine incluent un ensemble de faits et de règles ne référant à aucun exemple particulier mais exprimant des connaissances sur les éléments utilisés pour décrire les exemples. Par exemple le fait

`subClassOf`('insulin receptor binding','receptor binding').

exprime qu'il existe une relation sémantique *is-a* entre deux termes de Gene Ontology. On peut joindre à ces connaissances des règles d'inférence permettant, par exemple, d'exprimer la transitivité du prédicat `subClassOf` [7] :

`subClassOf(X,Z) :- subClassOf(X,Y), subClassOf(Y,Z).`

Une théorie T est un ensemble de règles couvrant le plus possible d'exemples positifs pour le moins possible d'exemples négatifs, c'est-à-dire des clauses dont le corps est vérifié par le plus possible d'exemples positifs pour le moins possible d'exemples négatifs. La tête de chaque règle représente le concept à apprendre tandis que le corps de la règle propose une description induite de ce concept, fondée sur la généralisation de plusieurs exemples. Par exemple, une règle obtenue lors d'un apprentissage sur des gènes responsables d'une maladie peut prendre la forme :

`is_responsible(X) :- gene_protein(X,Y), protein_mf(Y,'receptor binding').`

exprimant que si un gène `X` produit une protéine `Y` et que `Y` a pour fonction moléculaire 'receptor binding', alors le gène `X` est responsable de la maladie étudiée.

La recherche des règles de la théorie s'effectue dans l'espace des clauses, liées par leurs relations de généralisation et de spécialisation [Muggleton and Raedt, 1994]. Cet espace de clauses est trop conséquent pour en permettre une exploration exhaustive, il est donc nécessaire d'utiliser des méthodes heuristiques pour permettre la découverte de clauses couvrant le plus d'exemples positifs possible en ignorant un maximum de clauses couvrant trop d'exemples négatifs. Pour cela, il peut être nécessaire de définir des biais d'apprentissage qui définissent le type de règles désiré et d'ajuster la stratégie de recherche des règles.

Aleph Aleph est un programme Prolog implémentant un algorithme de PLI [Srinivasan, 2007]. Cet algorithme peut être décrit en quatre étapes :

1. Un exemple positif à généraliser est sélectionné. Le programme s'arrête s'il ne reste aucun exemple positif à généraliser.
2. Générer la clause la plus spécifique décrivant l'exemple sélectionné. Cette clause est générée à partir des faits relatifs à l'exemple sélectionné et aux connaissances du domaine. Cette clause est appelée *bottom clause*.

7. Dans la syntaxe Prolog, les variables sont identifiées par un nom commençant par une majuscule.

3. Explorer l'ensemble des clauses plus générales que la précédente. Les clauses plus générales peuvent être générées de deux manières :
   — en supprimant un terme de la clause,
   — en remplaçant un littéral par une variable.
   On cherche dans cet ensemble de clauses celle qui est vraie pour le plus d'exemples positifs et le moins d'exemples négatifs. Aleph propose plusieurs méthodes pour explorer cet ensemble de clauses et les évaluer en fonction de leur couverture des exemples.
4. En fonction des paramètres du programme, une clause est sélectionnée et ajoutée à la théorie. L'exemple sélectionné à l'étape 1 est retiré de l'ensemble des exemples à généraliser. Les autres exemples vérifiant la clause sélectionnée peuvent également être retirés. L'algorithme recommence à l'étape 1 pour les exemples restants.

Aleph requiert la définition de biais d'apprentissage définissant :
— le prédicat à prédire par la théorie, c'est-à-dire le prédicat apparaissant dans la tête de chaque règle de la théorie,
— le type d'argument (variable ou littéral) autorisé dans les prédicats du corps des règles. Par exemple, on pourra imposer que le second argument du prédicat `subClassOf` soit toujours un littéral.

Aleph propose également de nombreux paramètres pour guider la construction d'une théorie. Notamment, la fonction d'évaluation des clauses généralisées peut être modifiée. Par défaut, Aleph utilise la différence du nombre d'exemples positifs vérifiant la clause avec le nombre d'exemples négatifs la vérifiant. Il est également possible de spécifier un nombre minium d'exemples positifs pour qu'une clause soit ajoutée à la théorie, permettant d'exclure les clauses qui ne décrivent qu'un seul ou peu d'exemples.

La généralisation d'une clause peut également être contrainte via un paramètre limitant le nombre ou la proportion d'exemples négatifs vérifiant une clause, ou faux positifs. Par défaut cette limite est de 0. Cependant, il n'est pas toujours possible d'apprendre une théorie qui couvre correctement les exemples positifs en ne tolérant aucun faux positif, notamment lorsque les données sont bruitées. Il est alors souvent nécessaire de permettre un faible nombre de ces exemples négatifs. Une limite de faux positifs faible permet toutefois de réduire le nombre de généralisations à effectuer et d'améliorer la spécificité des règles.

1.4 Apport des ontologies dans la découverte de connaissances

Les ontologies sous leurs différentes formes : vocabulaire contrôlé, hiérarchie de termes, logiques de description, etc. permettent de faciliter et d'améliorer les différentes étapes du processus de découverte de connaissances. On s'intéressera ici en particulier aux applications des ontologies pour le domaine biomédical.

Les différents rôles des ontologies dans ce processus sont décrits par [Rubin et al., 2007] comme :
— l'interrogation de données biomédicales hétérogènes ;
— l'échange de données entre applications ;
— l'intégration des données ;
— la fouille de texte ;
— le raisonnement automatique.

On détaillera ici deux axes principaux pour l'utilisation des ontologies dans la découverte de connaissances : d'une part l'interopérabilité et l'intégration de données, d'autre part l'utilisation de raisonnement dans la découverte de connaissances.

1.4.1 Interopérabilité et intégration de données

L'intégration de sources de données hétérogènes est une tâche complexe, notamment lorsqu'une même entité est représentée de différentes manières dans une ou plusieurs sources [Köpcke and Rahm, 2010]. Une ontologie, en tant que vocabulaire contrôlé, permet de représenter ces entités de manière identique dans différents ensembles de données. L'ontologie peut également proposer des relations de synonymie, permettant de déterminer les correspondances non-explicites entre entités dans les données.

La fouille de texte bénéficie grandement des ontologies en utilisant une terminologie liant les termes dans le texte aux concepts d'une ontologie [Spasic et al., 2005, McCray et al., 1994, Aronson, 2001]. Les ontologies peuvent notamment être utilisées pour représenter cette terminologie. Des systèmes de récupération d'information peuvent également exploiter la sémantique de l'ontologie pour interroger une base de textes, notamment en utilisant des contraintes sur les classes d'objets à considérer et leurs relations [Müller et al., 2004] : on pourra par exemple récupérer toutes les instances de type *Gène* liées par une relation de type *régule*.

L'intégration de données est également facilitée par la présence de liens entre différentes ontologies. De nombreuses initiatives de la communauté biomédicale visent à la création de ces liens, comme notamment le BioPortal [Whetzel et al., 2011], ou à la création d'ontologies de haut-niveau unifiant les ontologies entre elles, comme le *Semantic Network* de l'UMLS [Bodenreider, 2004] ou l'ontologie MonDO [Mungall et al., 2017]. Les différentes sources biomédicales peuvent également être annotées par les termes de ces ontologies [Jonquet et al., 2011], ou rendues disponibles au format LOD et directement liées [Dumontier et al., 2014].

Le paradigme *Ontology-Based Data Access* (OBDA — accès aux données fondé sur les ontologies) [Calvanese et al., 2007, Poggi et al., 2008, De Giacomo et al., 2018] permet l'interrogation de nombreuses sources de données via les ontologies et du raisonnement. Ici, l'ontologie sert de vue globale sur un ensemble de sources de données : un système OBDA permet de transformer une requête sur cette vue globale en requête sur chaque source de données, à l'aide de correspondances définies entre le schéma de données global et les schémas locaux. Cette approche est de type *médiateur*, c'est-à-dire qu'elle permet de traiter plusieurs ensembles de données comme un seul, sans avoir à modifier les données elles-mêmes. Ce type de système a l'avantage de pouvoir prendre en compte des changements dans les sources de données sans avoir à effectuer un traitement sur ces données, tant que le schéma des données de change pas (ce qui n'est pas le cas par exemple d'une transformation en LOD d'une base de données : la transformation doit être effectuée pour chaque nouvelle version de la base de données).

1.4.2 Raisonnement dans la découverte de connaissances

Les mécanismes de raisonnement automatique fournis par les ontologies peuvent être utilisées dans la découverte de connaissances. Cette contribution est néanmoins limitée par la forte complexité de ces mécanismes. Les méthodes de fouille de données peuvent néanmoins être appliquées sur des données liées à des concepts d'ontologies, souvent en proposant des moyens d'exploiter la hiérarchie des concepts.

Par exemple, les différents niveaux de la hiérarchie des concepts peuvent être considérés pour extraire des règles d'association d'un ensemble d'objets annotés par des classes d'ontologie [Manda et al., 2012]. Ces objets peuvent également être comparés à l'aide de mesures de similarité sémantique pour y identifier des sous-groupes.

Analyse Formelle de Concepts Les ontologies et la FCA sont deux manières de modeliser et d'organiser des concepts. Cimiano *et al.* [Cimiano et al., 2004] décrivent un cycle d'interactions entre la FCA et les ontologies :

1. La FCA permet d'aider à la construction ou à l'intégration d'ontologies [Stumme and Maedche, 2001].
2. La FCA permet l'analyse et la visualisation des connaissances apportées par une ontologie [Alam et al., 2015].
3. Finalement, les ontologies peuvent êtres utilisées dans le processus de FCA. Il est alors nécessaire de représenter les connaissances de l'ontologie, ou un sous-ensemble de celles-ci, dans un context formel. Cimiano *et al.* proposent de définir les attributs du context formel comme correspondant à des définitions de concepts d'une ontologie, exprimées en logique de descriptions. Un système de raisonnement pourra ainsi pour un ensemble d'objets déterminer leur appartenance à chacun des concepts et construire un contexte formel binaire.

Les structures de patrons sont notamment adaptées pour l'utilisation d'ontologies en FCA, puisqu'elles permettent d'intégrer un processus de raisonnement dans l'opérateur de comparaison de descriptions. En particulier on peut définir un opérateur de comparaison exploitant la hiérarchie `is-a` d'une ontologie pour comparer des objets décrits par un ensemble de classes d'ontologie [Coulet et al., 2013].

Programmation Logique Inductive Les logiques de descriptions sont un fragment de la logique du premier ordre [Baader et al., 2005]. Les mécanismes de raisonnement des ontologies peuvent donc naturellement s'intégrer dans un processus de PLI. De manière similaire à la FCA, la PLI peut permettre la découverte de nouvelles connaissances pour enrichir les ontologies, tout en étant capable d'exploiter les connaissances existantes.

Lisi propose un formalisme combinant les logiques de descriptions et les clauses de Horn pour permettre l'intégration d'ontologies dans le processus de PLI [Lisi, 2008], et d'exprimer une théorie en logique de descriptions. Ce formalisme permet de former des concepts ou de raffiner les définitions des concepts d'une ontologie à partir d'un ensemble d'exemples.

Galárraga *et al.* proposent un algorithme, AMIE, de PLI permettant l'extraction de règles d'association sous l'hypothèse du monde ouvert (c'est-à-dire en se basant sur le principe que les connaissances sur les exemples sont non-exhaustives) [Galárraga et al., 2013]. Cet algorithme a la particularité de fonctionner sans nécessiter d'exemples négatifs : en effet, les ontologies et LODs ne comportent pas de faits négatifs, cet algorithme permet de traiter des connaissances sans ressource externe identifiant des exemples négatifs. Cela est possible notamment grâce à l'existence de relations fonctionnelles ou inverse-fonctionnelles (ou considérées comme telles) : on peut dans de tels cas considérer l'information présente comme complète.

Un objectif de cette thèse est d'étudier l'apport des ontologies et du raisonnement associé pour la découverte de connaissances. On y décrira plusieurs méthodes, utilisant notamment les structures de patrons pour exploiter la hiérarchie de concepts de plusieurs ontologies dans un processus d'extraction de règles d'association et les mécanismes de raisonnements offerts par la PLI dans une tâche de classification.

1.5 Contexte biomédical et applications

1.5.1 Pharmacovigilance

La pharmacovigilance est l'étude des effets secondaires, notamment les effets indésirables, causés par l'utilisation de médicaments. Il y a en particulier un besoin de détecter et d'étudier les effets indésirables après la commercialisation d'un médicament. Cette veille est dans un premier temps assurée par les médecins. Les données ainsi collectées peuvent alors être étudiées afin de démontrer ou non l'association entre un médicament et un effet indésirable. Il existe plusieurs difficultés à ce type d'études. D'abord les effets indésirables d'un médicament peuvent se présenter différement, ou pas du tout, chez plusieurs patients. Ensuite certain effets indésirables peuvent être causés par une combinaison de médicaments pris ensemble, qui pris séparément n'aurait pas eu d'effet secondaire.

De nombreuses études s'intéressent à la détection de "signaux", c'est-à-dire d'éléments quantitatifs ou qualitatifs indiquant une association entre un médicament et un effet secondaire [Meyboom et al., 1997]. On s'intéressera en particulier aux méthodes permettant de détecter des signaux pour des classes ou des combinaisons de médicaments.

Représentation vectorielle de profils patients Roitmann *et al.* proposent une représentation vectorielle des profils médicaments/effets indésirables de patients [Roitmann et al., 2014]. Chaque patient est représenté par un vecteur où chaque dimension est un effet indésirable declaré chez le patient. Ici les symptômes sont considérés comme des attributs indépendants. Cette représentation est utilisée par des algorithmes de clustering afin de former des groupes de patients ayant un profil similaire. Dans chacun de ces clusters sont ensuite étudiés les symptômes et médicaments les plus représentés. Cette méthode permet la découverte d'associations entre un ou plusieurs médicaments et un effet indésirable, et l'observation que, dans des clusters particuliers, des effets indésirables sont souvent associés.

Cette méthode d'apprentissage non supervisé permet d'extraire des associations sans hypothèse *a priori*, offrant la possibilité de faire de la pharmacovigilance à grande échelle sans nécessiter de connaissances du domaine. Cependant elle ne prend pas en considération les similarités entre phénotypes ou entre médicaments. En particulier, lorsque plusieurs médicaments sont fortement représentés dans un cluster, cette méthode ne permet pas à elle seule de déterminer si c'est la combinaison de ces médicaments qui est responsable d'un effet indésirable, ou si ces médicaments sont similaires car ils causent le même effet secondaire.

Fouille d'effets de classe Winnenburg *et al.* [Winnenburg et al., 2015] utilisent des événements indésirables extraits de la littérature sous forme de paires médicament-phénotype pour explorer les relations entre médicaments, classes de médicaments et leur effets secondaires. Dans cette étude, la prévalence des effets secondaires est considérée tant au niveau d'un médicament que de sa classe ATC. Les auteurs mettent alors en évidence des associations entre une classe de médicament et des effets secondaires, et poursuivent leur investigation au niveau des médicaments de cette classe.

Dans certain cas, l'association avec l'effet secondaire est présente pour seulement une partie des médicaments de la classe. On ne peut alors pas considérer l'effet intrinsèque à cette classe de médicaments. En revanches, si l'association existe pour chacun des médicaments de la classe, alors cela permet d'imputer l'effet secondaire à la classe de médicaments elle-même, qu'on peut alors appeler effet de classe. Ces résultats montrent qu'il est possible de considérer les effets secondaires des médicaments à différents niveaux.

Détection de signaux par la FCA La FCA peut être utilisée pour la détection de signaux pour la pharmacovigilance, par la mise en évidence d'attributs fréquemment associés. Dans [Lillo-Le Louët et al., 2009, Villerd et al., 2010], les auteurs utilisent un sous-ensemble des rapports d'un système d'auto-déclaration d'effets secondaires pour construire un contexte formel binaire dont les attributs sont des médicaments, phénotype et données démographiques. Ils identifient ensuite des concepts dans le treillis généré possédant commes attributs au moins un médicament et un phénotype, et vérifient la significativité statistique de leur association dans la population décrite par les attributs démographiques du concept. Cette approche permet alors d'extraire des associations entre un ensemble de médicaments et un phénotype.

1.5.2 Médecine fondée sur les réseaux et *diseasomes*

La médecine fondée sur les réseaux, ou *network-based medicine*, repose sur la représentation de données biologiques sous forme de graphes pour la découverte de connaissances biomédicales [Barabási et al., 2011]. Ainsi, l'étude topographique de réseaux de maladies connectées entre elles par des gènes communs, des interactions protéines-protéines ou des similarités sémantiques pourraient mener à l'identification des causes de ces maladies, ou la réutilisation d'un médicament contre une nouvelle maladie.

On s'intéressera ici aux *diseasomes*, c'est-à-dire à des graphes de maladies dans lesquels les liens dénotent une similarité entre deux maladies. Différentes approches pour la construction d'un diseasome ont été proposées, fondées sur différents critères de similarité. On définira d'abord quelques notions de bases sur la théorie des graphes.

Définition 32. *Un graphe simple, G est défini comme un ensemble de nœuds V (*vertices*) et un ensemble d'arêtes E (*edges*) représentées sous la forme d'une paire de noeuds, tel que $E \subseteq V \times V$.*

Les diseasomes utilisent fréquemment des graphes dont les nœuds représentent deux types d'objets, comme par exemple des maladies et des gènes. Il est courant dans ce cas que ce type de graphe présente des arcs uniquement entre un nœud gène et un nœud maladie. Un tel graphe est alors appelé graphe bipartie.

Définition 33. *Un graphe $G = (V, E)$ est dit bipartie s'il existe deux ensembles disjoints V_1 et V_2 tels que $V_1 \cup V_2 = V$, et $E \cap (V_1 \times V_1 \cup V_2 \times V_2) = \varnothing$; c'est-à-dire, qu'il existe deux sous-ensembles disjoints complémentaires de V, tels qu'aucune arête ne lie deux éléments d'un même de ces sous-ensembles.*

The Human Disease Network Dans [Goh et al., 2007], Goh *et al.* proposent le *Human Disease Network*, un graphe représentant la proximité génétique entre maladies. Ils construisent d'abord, à partir d'une liste d'associations entre gènes et maladies extraite d'OMIM, un graphe bipartie ayant pour nœuds, d'une part les gènes, d'autre part les maladies, et où une arête représente l'association entre un gène et une maladie.

Ce graphe bipartie permet ensuite la construction du *Human Disease Network*, un graphe où les nœuds représentent des maladies, et où un arc (d_i, d_j) représente le fait que d_i et d_j partagent un gène responsable. Les auteurs observent dans ce graphe quelques nœuds fortement connectés formant des *hubs* représentant des classes de maladies, et ce, bien que le graphe soit généré sans connaissance *a priori* de ces classes. C'est le cas par exemple des différents cancers, fortement connectés entre eux par les nombreux gènes associés à de nombreux types de cancers.

Un *Disease Gene Network* est également construit à partir du graphe bipartie initial : les nœuds y représentent des gènes, et les arcs (g_i, g_j) représentent le fait que g_i et g_j sont responsables d'une même maladie. Ce graphe propose une vue complémentaire des données à celle du

Human Disease Network, et on y retrouve également certains *hubs*, par exemple celui des gènes responsables de cancers.

Les auteurs analysent la topologie des deux graphes obtenus, par rapport à des graphes générés à partir de données aléatoires. Ils observent alors que la topologie du *Human Disease Network* et *Gene Disease Network* diffèrent significativement de ce qui serait attendu au hasard ($p < 10^{-4}$). Le *Human Disease Network* possède également près de 8 fois plus de liens entre maladies d'une même classe que des graphes produits à partir de données aléatoires.

Phenotypic Disease Network Dans [Hidalgo et al., 2009], Hidalgo *et al.* proposent le *Phenotypic Disease Network* pour représenter la comorbidité de maladies, identifiées par des codes ICD-9-CM. Ils utilisent les données du système d'assurances fédéral américain Medicare [Lauderdale et al., 1993, Mitchell et al., 1994] pour identifier des maladies co-occurant fréquemment dans un échantillon de plus de 30 millions de patients. Le système Medicare étant avant tout destiné aux patients de plus de 65 ans, les auteurs notent que le *Phenotypic Disease Network* n'est pas représentatif de la population globale.

Afin de quantifier la comorbidité de deux maladies, les auteurs utilisent le coefficient de corrélation de Pearson (ϕ) [Pearson, 1895] et le Risque Relatif (RR) [Copeland et al., 1977]. Le Risque Relatif calculé entre deux maladies X et Y équivaut au ratio entre le risque d'avoir la maladie X si l'on a Y et le risque d'avoir la maladie X dans l'ensemble de la population étudiée [8]. Chacune de ces mesures de la comorbidité présente des biais. Le risque relatif surestime la comorbidité des maladies rares et sous-estime celle des maladies communes, tandis que le coefficient de corrélation de Pearson estime correctement la comorbidité de maladies de fréquences similaires mais sous-estime la comorbidité de maladies de fréquences très différentes [Hidalgo et al., 2009].

Ainsi, les auteurs construisent un *Phenotypic Disease Network* pour ces deux mesures, puisqu'elles présentent des avantages et inconvénients complémentaires. Ils constatent que les valeurs de risque relatif calculées sont similaires à celles obtenues par des études de suceptibilité génétique. Les deux graphes construits présentent certaines similarités : par exemple, la proximité entre *Poisoning by drugs & toxic substances* et *Psychoses* et *Neurotic disorders* (voir Figure 1.11). Par ailleurs, on peut observer dans le graphe une bonne répartition des maladies par rapport à leur classe ICD-9-CM de premier niveau.

Certains faits intéressants peuvent être extraits du *Phenotypic Disease Network* : la connectivité d'une maladie dans le graphe est corrélée avec sa mortalité. Les auteurs démontrent que cette corrélation n'est pas seulement due au fait que un plus grand nombre de diagnostics puisse être corrélé avec un plus grand risque de mortalité. Au contraire, le nombre de visites chez un praticien et le nombre de diagnostics sont négativement corrélés à la mortalité. Le *Phenotypic Disease Network* permet donc de mettre en évidence avec une haute comorbidité sont associées à une plus haute mortalité.

Human Symptoms-Disease Network Pour les praticiens, les symptômes sont les manifestations les plus visibles d'une maladie. [Zhou et al., 2014] proposent donc de construire un diseasome se basant sur la similarité des symptômes des maladies. Les symptômes sont extraits des métadonnées MeSH (*Medical Subject Headings*) de PubMed [Wheeler et al., 2007, HJ and G, 1994], un moteur de recherche bibliographique répertoriant la littérature biomédicale. Sur un échantillon de plus de 7 millions de documents, les auteurs extraient 147978 liens entre 4219 maladies et 322 symptômes. Chaque maladie est alors représentée par un vecteur de symptômes, en

8. On notera que le risque relatif est symmétrique : $P(X \mid Y)/P(X) = P(Y \mid X)/P(Y) = P(X \cap Y)/P(X)P(Y)$

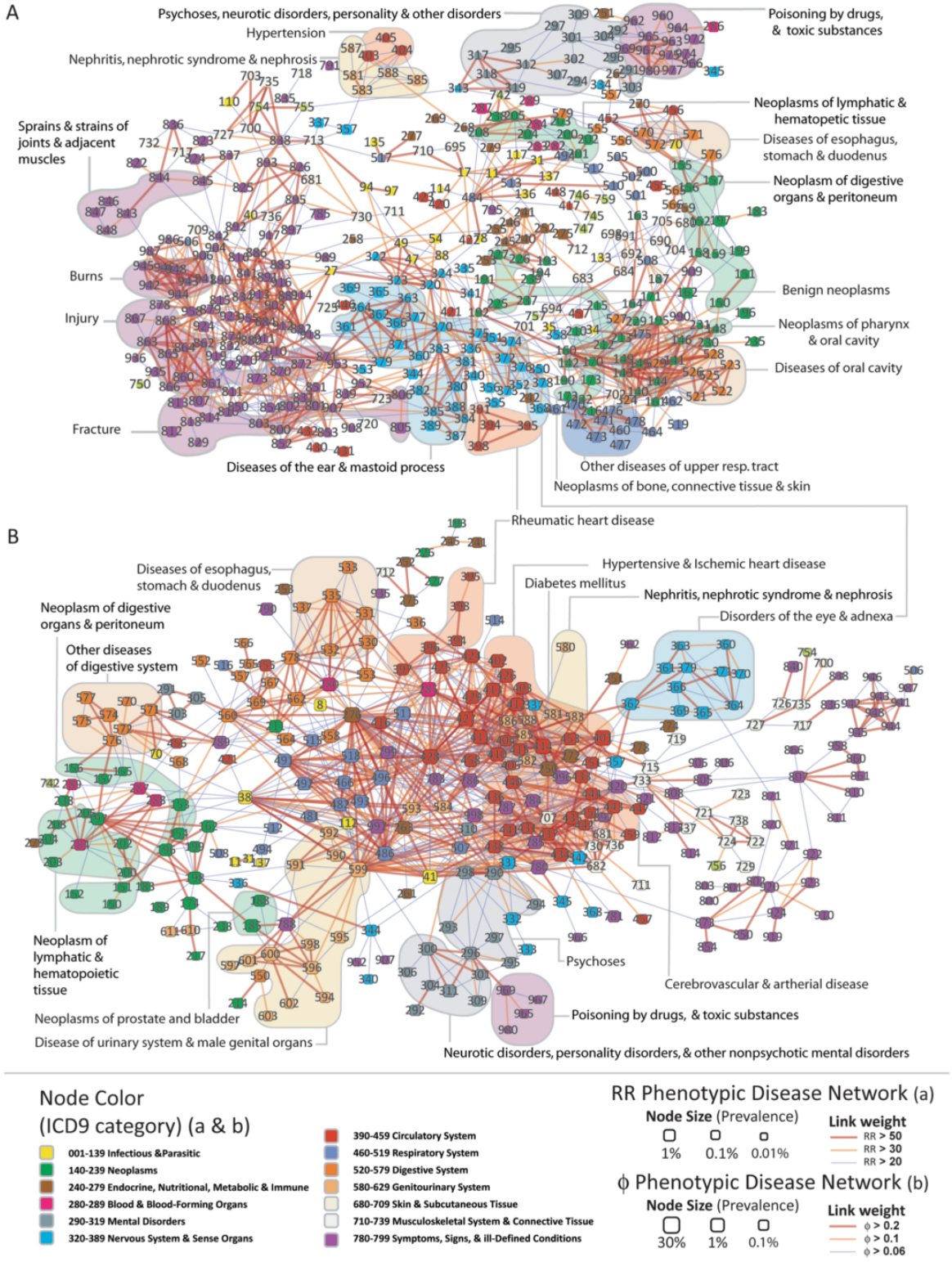


FIGURE 1.11 – Deux versions du *Phenotypic Disease Network*, le premier (A) construit à partir du Risque Relatif, le second (B) construit à partir du coefficient de corrélation de Pearson. Les couleurs de chaque nœud représentent une classe du premier niveau de la classification ICD-9-CM. Source : [Hidalgo et al., 2009]

utilisant une pondération *tf-idf*, et leurs similarités deux-à-deux sont calculées comme le cosinus de ces vecteurs.

Les auteurs construisent également un graphe liant entre eux des maladies sur la base des interactions entre les protéines produites par leur gènes responsables : deux maladies sont liées si un de leur gène produit une protéine interagissant avec une protéine d'un gène de l'autre maladie, ou s'il existe une troisième protéine avec laquelle les deux autres ont une interaction. *A fortiori*, deux maladies sont liées si elles partagent un gène en commun.

Les auteurs comparent ce graphe au graphe obtenu par le calcul des similarités phénotypiques. Ils observent une corrélation entre les interactions protéine-protéine des gènes de deux maladies et la similarité de leurs phénotypes. De même, il existe une corrélation entre les gènes partagés par deux maladies et leur similarité.

PhenUMA PhenUMA [Rodríguez-López et al., 2014] est une application web permettant la visualisation de réseaux de données biologiques mettant à profit les méthodes de similarité sémantique ainsi que les données sur les maladies génétiques et les interactions entre gènes.

Les auteurs calculent les similarités phénotypiques des différents gènes et maladies en appliquant la mesure de similarité de Resnik, aggrégée avec la fonction *Best-Match Average*, à des annotations de phénotypes exprimées dans le vocabulaire de l'ontologie HPO. Ils calculent également les similarités fonctionnelles des gènes par la même méthode en utilisant les annotations GO des gènes.

PhenUMA propose alors de visualiser et d'interroger plusieurs types de graphes : graphe de maladies, graphe de gènes, fondés respectivement sur la similarité des annotations HPO et GO, interactions des protéines de deux gènes, ou sur les gènes partagés par deux maladies.

Diseasome fondé sur l'interactome La *network-based medicine* peut également être utilisée pour évaluer la proximité de maladies et de médicaments. [Guney et al., 2016] proposent d'estimer pour des paires médicament-maladie, l'efficacité de ce médicament contre cette maladie. Pour cela, les auteurs proposent d'explorer un graphe représentant les associations entre gènes ou protéines[9] et maladies, les protéines ciblées par les médicaments (par des liens médicament-protéine) et les interactions entre les protéines.

Ce graphe permet la quantification de la proximité d'un médicament à une maladie : pour une paire médicament-maladie, on peut calculer la distance en nombre d'arcs dans le graphe entre une protéine ciblée par le médicament et une protéine impliquée dans la maladie. Les auteurs proposent 4 mesures de proximité entre un médicament et une maladie sur la base de cette distance :

- *closest* : moyenne des distances, pour chaque protéine cible du médicament, à la protéine de la maladie la plus proche ;
- *shortest* : moyenne des distances, entre chaque paire protéine cible et protéine de la maladie ;
- *kernel* : similaire à *shortest*, mais une pondération exponentielle donne plus d'importance dans le calcul de la moyenne aux chemins les plus courts ;
- *centre* : moyenne des distances, pour chaque protéine cible du médicament, au centre topologique des protéines de la maladie ;

Chaque distance calculée pour une paire médicament-maladie est ensuite comparée à une distribution aléatoire de la même distance calculée pour un ensemble de protéines cibles et un ensemble de protéines de la maladie pris au hasard (mais de même cardinalités que ceux

9. Dans le graphe, aucune distinction n'est faite entre protéine et gène.

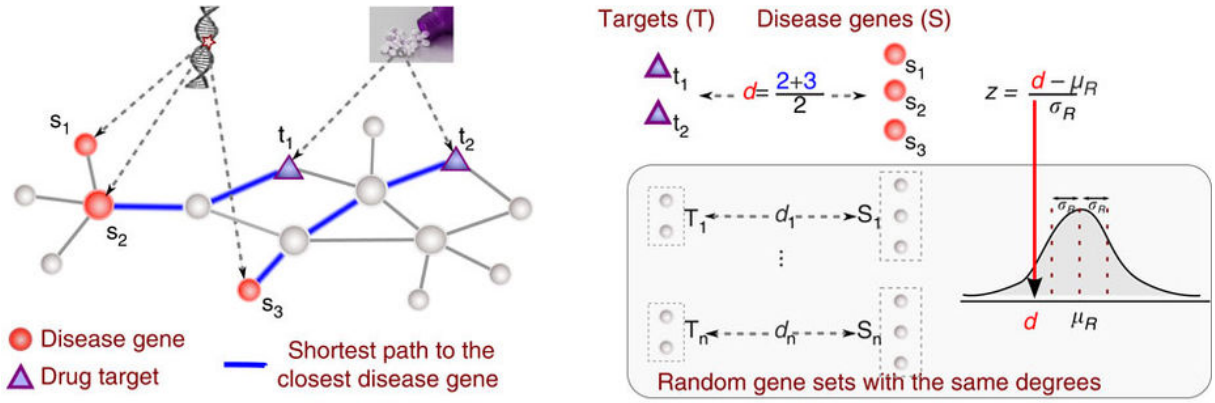


FIGURE 1.12 – Calcul de la proximité entre un médicament et une maladie avec la mesure *closest*. A gauche, le graphe des interactions de gènes ou protéines, où sont représentés les chemins entre chaque cible du médicament et la protéine de la maladie la plus proche. A droite, le calcul de la distribution de la distance pour des cibles et protéines de maladies randomisées, ainsi que le calcul final de la proximité médicament-maladie en nombre d'écarts-types. Source : [Guney et al., 2016]

correspondants à la paire médicament-maladie en question). En particulier, les auteurs calculent le nombre d'écarts-types de la distance calculée pour la paire par rapport à la distribution des distances randomisées. La Figure 1.12 illustre le calcul de la proximité entre un médicament et une maladie avec la mesure *closest*.

Les auteurs évaluent leur différentes mesures de proximité sur l'ensemble des paires médicament-maladie de leur jeu de données, en essayant de distinguer les paires "positives", c'est-à-dire correspondant à une indication connue, des autres considérées comme négatives. Les paires médicament-maladie sont alors classées en fonction de leur proximité relative à la proximité randomisée, permettant de calculer l'aire sous la courbe ROC. Les auteurs concluent que la mesure *closest* est la plus performante avec une aire sous la courbe de 0.66, ce qui semble indiquer qu'il n'est pas nécessaire pour un médicament d'avoir des cibles proches de toutes les protéines d'une maladie pour être efficace contre cette maladie. De plus, les paires associant un médicament et une maladie proches sont sur-représentées dans les essais cliniques ($p < 10^{-8}$). Ainsi, cette approche semble adéquate pour le classement ou la découverte de médicaments candidats pour des essais cliniques.

Diseasome fondé sur la similarité sémantique des phénotypes Un diseasome peut également être construit à partir d'une similarité sémantique. [Hoehndorf et al., 2015] proposent un diseasome construit à partir de plus de 6000 maladies. Chacune de ces maladies est associée à un ensemble de phénotypes extraits par fouille de texte sur plus de 5 millions de résumés d'articles MEDLINE [US NLM, 2018]. Ces phénotypes sont des classes de l'ontologie MonDO [Mungall et al., 2017], permettant la comparaison sémantique de ces maladies. Ces annotations de maladie ont été évaluées par comparaison aux phénotypes décrits dans OMIM, puis ont été mises à disposition librement [10].

Les auteurs utilisent la mesure de similarité sémantique SimGIC [Pesquita et al., 2007] et l'ontologie MonDO pour calculer la similarité des maladies. Les paires de maladies sont alors

10. Voir `http://aber-owl.net/aber-owl/diseasephenotypes/`

classées en fonction de leur similarité. Le diseasome est construit comme le graphe dont les nœuds sont les maladies et les arcs sont les 0.5% paires de maladies à la similarité la plus élevée.

Les auteurs conduisent ensuite trois évaluations des similarités calculées, et par extension du diseasome généré à partir de cette similarité. Ils évaluent d'abord le classement par similarité par rapport à un ensemble de 22 classes de maladies de haut niveau extrait de Disease Ontology [Schriml et al., 2011]. Pour chaque maladie d, le classement des autres maladies d_i par rapport à leur similarité à d est évalué par l'aire sous la courbe ROC, où une maladie d_i est considéré comme positive si elle partage une classe Disease Ontology avec d.

Ils évaluent ensuite le même classement par rapport aux médicaments partagés par des maladies. Pour l'ensemble des paires de maladies (d_1, d_2), cette paire est considérée positive si il existe un médicament indiqué pour traiter d_1 et d_2. Les indications de médicaments sont tirées de la base de données SIDER 2 [Kuhn et al., 2015]. Cette évaluation permet de conclure que pour une paire positive et une paire négative prises au hasard, la paire positive a 64.7% de chance d'avoir une similarité plus élevée.

Finalement, les auteurs utilisent les mesures de similarité pour produire un clustering des maladies, et le comparent aux classes Disease Ontology sélectionnées auparavant. Ce clustering comporte 198 clusters, et ils obtiennent alors une pureté moyenne par cluster de 0.575, et un indice de Rand de 0.828.

Ces différentes études sur les réseaux de maladies montrent l'existence de liens entre les différents aspects d'une maladie : gènes responsables, symptômes et traitements. Ainsi, ces *diseasomes* permettent la mise en évidence de la proximité de certains mécanismes des maladies étudiées, afin de parfaire les connaissances sur leur causes et de proposer des médicaments réutilisables d'une maladie proche à l'autre.

1.6 Conclusion

Cet état de l'art présente différentes méthodes permettant l'intégration de connaissances de domaine dans le processus de découverte de connaissances. On peut distinguer parmi ces méthodes les méthodes symboliques comme l'Analyse Formelle de Concepts et la Programmation Logique Inductive, et les méthodes numériques comme celles fondées sur une mesure de similarité sémantique. Les travaux présentés dans cette thèse se fondent sur ces différentes approches pour tirer parti des connaissances de domaine contenues dans les ontologies pour la découverte de connaissances biomédicales dans deux contextes principaux : la découverte d'associations entre événements indésirables médicamenteux et la caractérisation des déficiences intellectuelles, notamment par la mise en évidence de leurs gènes responsables.

L'Analyse Formelle de Concepts (FCA) est un premier outil capable de tenir compte des connaissances d'une ontologie. Le Chapitre 2 propose une approche fondée sur la FCA et les structures de patrons pour l'extraction d'associations dans des dossiers patients. Cette approche permet de comparer des représentations complexes d'événements indésirables médicamenteux à l'aide d'une ontologie de symptômes et d'une ontologie de médicaments. Le treillis construit permet ensuite l'extraction de règles d'association entre deux ou plus de ces événements. L'utilisation de descriptions complexes de ces événements, rendue possible grâce aux structures de patrons, jointe aux ontologies permet d'obtenir des règles expressives à différents niveaux de l'ontologies : on peut par exemple comparer des événements décrits par un ou plusieurs médicaments, voire une classe de médicaments.

La médecine fondée sur les réseaux représente une manière nouvelle d'explorer les interactions entre gènes, maladies et médicaments, et notamment de distinguer des groupes de maladies similaires. Le Chapitre 3 décrit la construction d'un diseasome en utilisant la mesure de similarité IntelliGO [Benabderrahmane et al., 2010] et l'ontologie MonDO [Mungall et al., 2017]. Le diseasome résultant et sa capacité à identifier des maladies partageant un gène responsable ou un traitement médicamenteux sont comparés au travail de Hoehndorf *et al.* [Hoehndorf et al., 2015]. Une étude des différents groupes de déficiences intellectuelles au sein de ce diseasome est également présentée.

La Programmation Logique Inductive (PLI) permet le raisonnement sur des ontologies liées au LOD à travers le formalisme de la logique du premier ordre. Le Chapitre 4 propose une application de la PLI aux gènes reponsables de déficiences intellectuelles, utilisant des LODs décrivant un ensemble de gènes et liées à Gene Ontology. Des LODs de multiples sources sont intégrées selon un modèle de données local permettant la fouille avec le programme de PLI Aleph. Cette approche permet d'obtenir une théorie caractérisant les gènes responsables de déficiences intellectuelles.

2

Structures de patrons et ontologies pour la découverte d'associations entre effets indésirables de médicaments

Sommaire

2.1 Problèmatique des associations entre Evénements Indésirables Médicamenteux

2.1.1 Définition et classification des EIM

Les Evénements Indésirables Médicamenteux (EIM) sont des manifestations d'un effet secondaire non-recherché d'un médicament administré à un patient chez qui il entraine une déterioration de son état de santé. Les EIM constituent une problématique majeure de santé publique. En effet les essais cliniques effectués pour la pré-autorisation de mise sur le marché d'un médicament ne peuvent permettre d'en prévenir les effets secondaires, le rôle de ces études étant d'abord d'évaluer le ratio bénéfice/risque d'une substance. Certains effet indésirables sont en effet trop rares pour se déclarer lors d'une étude sur un petit nombre de patients, ou peuvent se déclarer lorsque que le médicament est combiné avec une autre substance.

Les effets secondaires de médicaments sont particulièrement le domaine détude de la pharmacovigilance. On s'intéressera ici à la pharmacovigilance après mise sur le marché des médicaments, en particulier via des EIM extraits de Dossiers Médicaux Electroniques (DME), ou d'un système de rapport d'effets secondaires de médicaments.

Classification Les EIM furent d'abord classés en deux catégories [Rawlins and Thompson, 1977, Rawlins and Thompson, 1981] :

- Catégorie A (liés au dosage – *Augmented*) : les EIM liés au dosage d'un médicament sont assez communs (80% des EIM) et prévisibles, puisqu'il résultent d'une augmentation du l'effet primaire du médicament. Par exemple, un anticoagulant peut provoquer des saignements à trop forte dose. À la suite d'un tel EIM un ajustement de la dose prescrite peut être suffisant.
- Catégorie B (particulier – *Bizarre*) : ces EIM sont plus rares et sans relation avec les effets thérapeutiques du médicament. Les réactions allergiques à un médicament entrent dans cette catégorie. La nature de ces EIM les rends imprévisibles et hautement dangereux, et imposent l'arrêt du traitement.

Cette catégorisation des EIM a ensuite été étendue, et comporte maintenant six catégories, pour mieux décrire les EIM liés aux traitements prolongés [Grahame-Smith DG, 1984, Royer, 1997] ou à l'inefficacité du traitement [Hartigan-Go and Wong, 2000] :

- Catégorie C (liés au dosage et au temps – *Chronic*) : cette catégorie comporte les EIM causés par la dose cumulative d'un médicament sur la durée du traitement. À la suite d'un tel EIM un ajustement de la dose prescrite peut être suffisant.
- Catégorie D (liés au temps – *Delayed*) : cette catégorie comporte les EIM apparaissant un certain temps après le traitement.
- Catégorie E (liés à l'arrêt du traitement – *End of use*) : cette catégorie comporte les EIM se présentant peu après l'arrêt du traitement, notamment dans le cas des médicaments pouvant provoquer une addiction comme les opiacés.
- Catégorie F (échec du traitement – *Failure*) : cette catégorie décrit les effets insuffisants ou l'absence d'effet d'un traitement, du fait par exemple de son dosage ou de son interaction avec d'autres substances.

Cette classification n'est cependant pas suffisante pour décrire tous les EIM [Edwards and Aronson, 2000], et est amenée à évoluer avec les connaissances sur les mécanismes des EIM.

Conséquences Les façons de traiter un EIM sont tout aussi diverses que les EIM en eux-mêmes. L'EIM peut être traité par un ajustement de la dose du médicament, l'arrêt du traitement ou l'administration d'un nouveau traitement visant à traiter les symptomes de l'EIM si le traitement ne peut pas être interrompu. La modification du traitement du patient comporte cependant un risque supplémentaire d'EIM.

2.1.2 Recherche d'associations entre EIM

De nombreux travaux s'intéressent à la détection d'associations entre un médicament et un effet secondaire, via des systèmes de rapport d'événements indésirables[Sakaeda et al., 2013] ou de la fouille de DME [LePendu et al., 2013]. Ces études, en conjonction avec les essais cliniques, permettent notamment de connaître les effets secondaires qui peuvent se présenter lors de la prise d'un médicament donné.

Cependant, les EIM ne se présentent pas nécessairement de la même manière chez différents groupes de patients, et les causes de ces variations ne sont pas toujours connues. En effet, les causes d'un EIM peuvent être multiples et de différentes natures : génétiques, métaboliques, interactions avec d'autre substances, etc. On fait ici l'hypothèse que l'on peut fouiller des données patients dans le but de révéler que des sous-groupes de patients qui sont susceptibles aux effets secondaires d'un médicament peuvent présenter une susceptibilité aux effets secondaires d'un autre médicament. Dans de tels cas, plusieurs EIM, chacun causés par des médicaments différents pourraient se réveler comme fréquemment associés. De telles associations sont connues pour certains médicaments d'une même classe [Winnenburg et al., 2015], on cherchera ici à étudier ces associations entre EIM au-delà de ces classes. On propose dans ce chapitre une méthode permettant d'indentifier des EIM fréquemment associés dans des sous-groupes de patients.

Les manifestations d'EIM sont complexes et peuvent être reportées de différentes manières. En effet, les EIM ne sont pas limités au simple cas d'un médicament causant un phénotype, mais peuvent être l'association de plusieurs médicaments et phénotypes. De plus, ces médicaments et phénotypes peuvent être reportés dans des vocabulaires différents et avec des niveaux de détail variables. Ainsi, les ontologies biomédicales représentent une resource indispensable pour permettre de considérer les relations sémantiques entre les EIM. On utilisera l'Analyse Formelle de Concepts (FCA) et son extension, les structures de patrons, pour permettre la représentation d'EIM avec des termes d'ontologies biomédicales et leur comparaison sémantique. La FCA a déjà été utilisée pour comparer des EIMs avec un ou plusieurs médicaments et phénotypes dans le cadre de la détection de signaux, c'est-à-dire des associations entre médicaments et phénotypes [Lillo-Le Louët et al., 2009, Villerd et al., 2010]. On proposera ici une approche utilisant en plus des ontologies pour la comparaison d'EIM, décrits par des classes de médicaments et de phénotypes, grâce aux structures de patrons. De plus, cette approche permet l'extraction d'associations de plus haut niveau, entre les EIM eux-mêmes.

Cette approche est expérimentée sur deux jeux de données utilisant différentes ontologies biomédicales. Le premier jeu de données est un ensemble de DME desquels sont extraits des EIM. Le second est extrait d'un système de rapport d'événements indésirables, FAERS [FDA, 2016].

2.2 Comparaison d'événements indésirables

Dans cette section sont décrits trois opérateurs de comparaison d'EIM. Chacun d'entre eux définit une représentation différente des EIM d'un patient, ainsi qu'une structure de patrons associée, chacune faisant un plus grand usage des ontologies biomédicales. Le premier opérateur

n'utilise pas l'ontologie de médicaments ATC, et généralise les phénotypes représentés par leur code ICD-9-CM par la classe ICD-9-CM la plus générale les décrivant. Le second opérateur introduit l'ontologie ATC pour permettre des généralisations sur les médicaments, tandis que le troisième introduit la possibilité de généraliser sur plusieurs niveaux d'ICD-9-CM.

2.2.1 Formalisation d'un événement indésirable

Un Evénement Indésirable Médicamenteux (EIM) est un événement complexe qui peut impliquer plusieurs médicaments, et se manifester par plusieurs phénotypes. Un EIM peut ainsi être caractérisé par un ensemble de médicaments et un ensemble de phénotypes. Afin de faciliter la comparaison entre des EIM, on considérera des ensembles d'ingrédients actifs de médicaments, plutot que des ensembles de formes commerciales de médicaments. Ainsi, on utilisera ici le terme "médicament" pour désigner l'ingrédient actif d'un médicament.

Dans ce travail, on souhaite représenter un EIM comme une paire (D_i, P_i), où D_i est un ensemble de médicaments, et P_i est un ensemble de phénotypes.

TABLE 2.1 – Exemple de jeu de données contenant 3 patients avec chacun 2 EIM.
— ICD 599.8 : *other specified disorders of the urethra and urinary tract*
— ICD 599.9 : *unspecified disorders of the urethra and urinary tract*
— ICD 719.4 : *pain in joint*

Patient	EIM	
P1	({acetaminophen},{ICD 599.9})	({prednisone},{ICD 599.8})
P2	({prednisone},{ICD 599.8})	({prednisone},{ICD 719.4})
P3	({acetaminophen},{ICD 719.4})	({acetaminophen, prednisone},{ICD 599.9})

2.2.2 Premier opérateur de comparaison

Il est nécessaire de disposer d'une représentation qui permet la prise en compte de la similarité entre les EIM, c'est-à-dire qui ne considère pas les EIM comme des attributs indépendants. A cette fin, cette première expérience propose une représentation qui regroupe les EIM partageant une classe de phénotype, et définit un opérateur permettant la comparaison de leurs ensembles de médicaments. Un EIM est alors représenté comme un simple ensemble de médicaments, permettant la comparaison d'EIM deux à deux par intersection. Un patient sera alors représenté par des ensembles d'EIM, dans chaque classe de phénotype. On choisit ici de répartir les EIM en fonction de la classe de plus haut niveau de leur phénotype dans ICD-9-CM.

On définit ici la structure de patrons $(G, (\mathcal{D}_1, \sqcap_1), \delta_1)$. G est un ensemble de patients. Chaque description de $\mathcal{D}_1$ est un vecteur d'ensembles d'EIM, où chaque dimension représente une des classes du premier niveau d'ICD-9-CM. Chaque EIM est alors représenté par l'ensemble de médicaments qui y sont impliqués.

Par exemple, en considérant seulement les deux classes ICD de la Table 2.2 et les EIM décrits en Table 2.4 :

$$\delta_{1,\text{ICD 580-629}}(\text{P1}) = \{\{\text{prednisone}\}, \{\text{acetaminophen}\}\}$$
$$\delta_{1,\text{ICD 710-739}}(\text{P1}) = \varnothing$$

Ici, les EIM sont décomposés sur la base de leur phénotypes. On appellera sous-descriptions les ensembles de médicaments associés à une classe de premier niveau d'ICD pour représenter

TABLE 2.2 – Exemple de représentation des EIM de 3 patients pour la structure de patrons $(G, (\mathcal{D}_1, \sqcap_1), \delta_1)$, avec 2 classes ICD de premier niveau.
— ICD 580-629 : *diseases of the genitourinary system*
— ICD 710-739 : *diseases of the musculoskeletal system and connective tissue*

Patient	ICD 580-629 (*genitourinary system*)	ICD 710-739 (*musculoskeletal system*)
P1	{{prednisone}, {acetaminophen}}	$\varnothing$
P2	{{prednisone}}	{{prednisone}}
P3	{{prednisone, acetaminophen}}	{{acetaminophen}}

les EIM : un patient présente un phénotype de cette classe ICD après avoir pris les médicaments de cette sous-description. Dans l'exemple présenté dans la Table 2.2, le patient P1 a présenté un EIM avec un phénotype de la classe ICD 580-629 deux fois : une fois après la prise de *prednisone*, une autre fois après la prise d'*acetaminophen*. En revanche, le patient P3 a subi un EIM avec un phénotype de cette classe une seule fois, après la prise simultanée de *prednisone* et d'*acetaminophen*.

On définit une sous-description comme un ensemble de prescriptions, chaque prescription étant un ensemble de médicaments, tel qu'aucune des prescriptions n'est comparable aux autres par l'ordre partiel $\subseteq$. On définit alors l'opérateur de comparaison de sous-descriptions $\sqcap_1$.

Définition 34. *Soit $\sqcap_1$ un opérateur de comparaison d'ensembles d'ensembles de médicaments. On définit $\sqcap_1$ tel que, pour toute paire de tels ensembles X et Y :*

$$X \sqcap_1 Y = \max(\subseteq, \{x \cap y \mid (x, y) \in X \times Y\})$$

où $\max(\leq_i, S)$ est l'unique ensemble des éléments maximaux appartenant à un ensemble S pour un ordre partiel $\leq_i$.

Définition 35. *Soit S tout ensemble partiellement ordonné muni de l'ordre partiel $\leq_i$, on définit la fonction* max *telle que :*

$$\max(\leq_i, S) = \{s \mid \nexists x.(s \leq_i x)\}$$

La fonction max permet de conserver uniquement les ensembles les ensemble de médicaments les plus descriptifs, et ainsi d'éviter de représenter des données redondantes dans les descriptions. Par exemple, étant donné 4 médicaments $d_1, \ldots, d_4$:

$$\begin{aligned}
&\{\{d_1, d_2, d_3\}\} \sqcap_1 \{\{d_1, d_2\}, \{d_2, d_4\}\} \\
&= \max(\subseteq, \{\{d_1, d_2, d_3\} \cap \{d_1, d_2\}, \{d_1, d_2, d_3\} \cap \{d_2, d_4\}\}) \\
&= \max(\subseteq, \{\{d_1, d_2\}, \{d_2\}\}) \\
&= \{\{d_1, d_2\}\}
\end{aligned}$$

On ne conserve que $\{d_1, d_2\}$ puisque $\{d_2\} \subseteq \{d_1, d_2\}$ et que $\{d_1, d_2\}$ est le seul élément maximal selon $\subseteq$. En effet, l'information exprimée par $\{d_2\}$ – une prescription contenant le médicament d_2 – est plus générale que l'information exprimée par $\{d_1, d_2\}$ – une prescription contenant les médicaments d_1 et d_2.

Puisque chaque patient a une description de ses EIM pour chaque classe ICD, on peut généraliser l'opérateur de comparaison de ces descriptions à un vecteur de telles descriptions :

$$\begin{aligned}
\delta_1(\text{P1}) \sqcap_1 \delta_1(\text{P2}) &= \langle \delta_{1,1}(\text{P1}), \ldots, \delta_{1,n}(\text{P1}) \rangle \sqcap_1 \langle \delta_{1,1}(\text{P2}), \ldots, \delta_{1,n}(\text{P2}) \rangle \\
&= \langle \delta_{1,1}(\text{P1}) \sqcap_1 \delta_{1,1}(\text{P2}), \ldots, \delta_{1,n}(\text{P1}) \sqcap_1 \delta_{1,n}(\text{P2}) \rangle
\end{aligned}$$

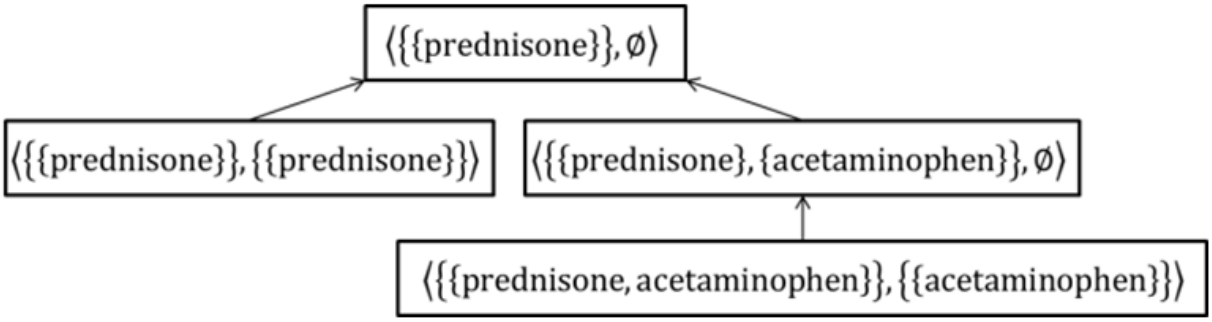


FIGURE 2.1 – Représentation des données de la Table 2.2 dans un semi-treillis construit à partir de la structure de patrons $(G, (\mathcal{D}_1, \sqcap_1), \delta_1)$, où les flèches représentent l'ordre partiel $\leq_{\sqcap_1}$ défini par $\sqcap_1$.

La Figure 2.1 présente le semi-treillis de descriptions associé à cette structure de patrons, construit à partir des descriptions en Table 2.2. L'opérateur de comparaison présenté ici est néanmoins limité puisqu'il ne permet pas de tenir compte de la similarité ni entre médicaments ni entre phénotypes.

2.2.3 Second opérateur de comparaison, utilisant une ontologie de médicaments

On souhaite étendre le premier opérateur de comparaison de manière à permettre l'extraction d'associations entre des classes de médicaments. On propose alors une structure de patrons permettant de prendre en compte une ontologie de médicaments. Chaque médicament y est représenté par une classe de l'ontologie, comme illustré en Table 2.3 avec un example utilisant l'ontologie ATC.

TABLE 2.3 – Exemple de représentation des EIM de patients pour $(G, (\mathcal{D}_2, \sqcap_2), \delta_2)$, avec 2 classes ICD de premier niveau.
- H02AA03 : *desoxycortone*
- H02AB07 : *prednisone*
- N02BE01 : *acetaminophen*

Patient	ICD 580-629 (genitourinary system)	ICD 710-739 (musculoskeletal system)
P1	{{H02AB07},{N02BE01}}	∅
P2	{{H02AB07}}	{{H02AB07}}
P3	{{H02AB07, N02BE01}}	{{N02BE01}}
P4	{{H02AA03}}	∅

On définit une seconde structure de patrons $(G, (\mathcal{D}_2, \sqcap_2), \delta_2)$ où les descriptions de $\mathcal{D}_2$ sont des vecteurs dont les dimensions sont des ensembles de prescriptions. Une prescription est représentée par un ensemble de classes d'une ontologie de médicaments, correspondant aux différents médicaments prescrits. On note que cette représentation permet de prendre en compte plusieurs classes pour un seul médicament, comme il peut en exister dans l'ontologie ATC par exemple.

Pour permettre la comparaison d'ensembles de classes d'une ontologie $\mathcal{O}$, on définit l'opérateur de comparaison $\sqcap_{\mathcal{O}}$.

Définition 36. *Soit $\mathcal{O}$ une ontologie définissant une hiérarchie de classes formant un ordre*

partiel $\sqsubseteq$. *On définit un opérateur de comparaison d'ensembles de classes de* $\mathcal{O}$, *noté* $\sqcap_{\mathcal{O}}$, *tel que, pour tout* x *et* y *deux ensemble de classes de* $\mathcal{O}$:

$$x \sqcap_{\mathcal{O}} y = max(\sqsubseteq, \{LCA(c_x, c_y) \mid (c_x, c_y) \in x \times y\})$$

où $LCA(c_x, c_y)$ *est l'ancêtre commun le plus spécifique de* c_x *et* c_y *dans* $\mathcal{O}$.

Ici, pour tout ensemble de classes C, $\max(\sqsubseteq, C)$ est le sous-ensemble des classes les plus spécifiques de C (qui n'ont pas de descendant dans C). Ainsi, $x \sqcap_{\mathcal{O}} y$ est le sous-ensemble des ancêtres les plus spécifiques communs aux classes de x et y.

A partir de $\sqcap_{\mathcal{O}}$ on définit l'ordre partiel $\leq_{\mathcal{O}}$, qui compare deux ensembles de classes d'une ontologie, x et y, tel que $x \leq_{\mathcal{O}} y \Leftrightarrow x \sqcap_{\mathcal{O}} y = x$. $x \leq_{\mathcal{O}} y$ exprime alors que y est un ensemble de classes de $\mathcal{O}$ plus spécifique que x.

On définit enfin l'opérateur de comparaison $\sqcap_2$ similairement à $\sqcap_1$, mais utilisant $\sqcap_{\mathcal{O}}$ pour comparer des ensembles de classes de médicaments plutot que la simple intersection $\cap$.

Définition 37. *Soit* $\sqcap_2$ *un opérateur de comparaison d'ensembles d'ensembles de médicaments représentés par des classes d'une ontologie* $\mathcal{O}$. *On définit* $\sqcap_2$ *tel que, pour toute paire de descriptions* X *et* Y *de* $\mathcal{D}_2$:

$$X \sqcap_2 Y = max(\leq_{\mathcal{O}}, \{x \sqcap_{\mathcal{O}} y \mid (x, y) \in X \times Y\})$$

Comme pour $\sqcap_1$, on généralisera $\sqcap_2$ à un vecteur de descriptions d'EIM en l'appliquant à chacune des dimensions des vecteurs comparés. La structure de patrons $(G, (\mathcal{D}_2, \sqcap_2), \delta_2)$ permet la généralisation d'EIM impliquant différents médicaments appartenant à une même classe de plus haut niveau. Par exemple, on peut comparer les patients P1 et P4 de l'exemple présenté en Table 2.3 et obtenir une description commune mettant en évidence un EIM décrit avec une classe de médicaments de plus haut niveau :

$$\begin{aligned}
\delta(\text{P1}) \sqcap_2 \delta(\text{P4}) &= \langle\{\{\text{H02AB07}\}, \{\text{N02BE01}\}\}, \varnothing\rangle \sqcap_2 \langle\{\{\text{H02AA03}\}\}, \varnothing\rangle \\
&= \langle \max(\leq_{\mathcal{O}}, \{\{\text{H02AB07}\} \sqcap_{\mathcal{O}} \{\text{H02AA03}\}, \\
&\qquad\qquad \{\text{N02BE01}\} \sqcap_{\mathcal{O}} \{\text{H02AA03}\}\}), \varnothing\rangle \\
&= \langle \max(\leq_{\mathcal{O}}, \{\{\text{H02A}\}, \{\top\}\}), \varnothing\rangle \\
&= \langle\{\{\text{H02A}\}\}, \varnothing\rangle
\end{aligned}$$

Ici, on utilise $\sqcap_{\mathcal{O}}$ pour comparer des ensembles de médicaments représentés par leur classe ATC. La comparaison de {H02AA03} (*desoxycortone*) et {H02AB07} (*prednisone*) produit leur ancêtre commun dans l'ontologie ATC : {H02A} (*corticosteroids for systemic use, plain*). On observe que {N02BE01} (*acetaminophen*) et {H02AA03} (*desoxycortone*) n'ont que la racine $\top$ de l'ontologie en commun, ainsi $\{\text{N02BE01}\} \sqcap_{\mathcal{O}} \{\text{H02AA03}\} = \{\top\}$. La fonction max le retire du résultat final, puisque $\{\top\}$ est redondant avec {H02A}, puisque $\{\top\} \leq_{\mathcal{O}} \{\text{H02A}\}$. Le vecteur $\langle\{\{\text{H02A}\}\}, \varnothing\rangle$ représente la généralisation la plus spécifique des descriptions des patients P1 et P4, et peut être lu comme : les médicaments de la classe H02A (*corticosteroids for systemic use, plain*) sont associés à un phénotype de la classe ICD *diseases of the genitourinary system* (580-629), et aucun médicament n'est associé à la classe ICD *diseases of musculoskeletal system and connective tissue* (710-739).

2.2.4 Troisième opérateur de comparaison, utilisant une ontologie de médicaments et une ontologie de phénotypes

On définit une structure de patrons permettant d'utiliser une ontologie de médicaments ainsi qu'une ontologie de phénotypes pour décrire et comparer les EIM. Cet opérateur peut s'appliquer à des données exprimées à l'aide de classes provenant de différentes ontologies de phénotypes. Cet opérateur est ici décrit de manière générique, et est utilisable avec n'importe quelle ontologie.

Dans notre cas, afin déviter une sur-généralisation des descriptions d'EIM, on exclut les 2 niveaux de la hiérarchie ICD les plus généraux, et les 3 plus généraux de SNOMED CT. La Table 2.4 présente un exemple de la représentation de données utilisées avec la structure de patrons décrite ici, exprimée avec les ontologies ATC et ICD.

Ici, un EIM est représenté comme un vecteur $\langle D_i, P_i \rangle$ à deux dimensions : un ensemble de médicaments D_i pour la première, associé à un ensemble de phénotypes P_i pour la deuxième. La description d'un patient est alors un ensemble de tels vecteurs.

TABLE 2.4 – Exemple de représentation des EIM de patients pour $(G, (\mathcal{D}_3, \sqcap_3), \delta_3)$.
— ICD 599.8 : *other specified disorders of the urethra and urinary tract*
— ICD 599.9 : *unspecified disorders of the urethra and urinary tract*
— ICD 719.4 : *pain in joint*

Patient	Description
P1	{⟨{H02AB07},{ICD 599.8}⟩, ⟨{N02BE01},{ICD 599.9}⟩}
P2	{⟨{H02AB07},{ICD 599.9}⟩, ⟨{H02AB07},{ICD 719.4}⟩}
P3	{⟨{H02AB07, N02BE01},{ICD 599.9}⟩, ⟨{N02BE01},{ICD 719.4}⟩}

On définit la structure de patrons $(G, (\mathcal{D}_3, \sqcap_3), \delta_3)$, où les descriptions de $\mathcal{D}_3$ sont des ensembles d'EIM. On définit d'abord un opérateur $\sqcap_{EIM}$ permettant la comparaison des représentations d'EIM deux à deux :

$$v_x \sqcap_{EIM} v_y = \langle D_x, P_x \rangle \sqcap_{EIM} \langle D_y, P_y \rangle$$
$$= \begin{cases} \langle D_x \sqcap_{\mathcal{O}} D_y, P_x \sqcap_{\mathcal{O}} P_y \rangle \text{ si chaque dimension contient une classe non racine,} \\ \langle \varnothing, \varnothing \rangle \text{ sinon.} \end{cases}$$

L'opérateur $\sqcap_{EIM}$ applique l'opérateur de comparaison $\sqcap_{\mathcal{O}}$ sur les deux dimensions du vecteur représentant l'EIM, en utilisant l'ontologie correspondant aux données. Si au moins une des deux dimensions du résultat devait être égale à $\varnothing$, alors le résultat est remplacé par $\langle \varnothing, \varnothing \rangle$ afin de l'ignorer dans les prochaines généralisations. En effet, on ne veut pas considérer la similarité de deux EIM si ils n'ont aucun médicament ou phénotype commun.

Finalement on définit l'opérateur de comparaison de notre structure de patrons $\sqcap$ tel que, pour toute paire de descriptions (X, Y) :

$$X \sqcap_3 Y = \max(\leq_{EIM}, \{v_x \sqcap_{EIM} v_y \mid (v_x, v_y) \in X \times Y\})$$

Comparé à $\sqcap_2$, $\sqcap_3$ introduit un niveau supplémentaire de comparaison avec $\sqcap_{EIM}$ qui permet la généralisation d'EIM en appliquant l'opérateur $\mathcal{O}$ à une ontologie de phénotypes.

2.2.5 Extraction et filtrage des règles d'association

Chacune des structures de patrons présentées dans cette section permettent la construction d'un treillis de concepts. On peut ensuite extraire de ces treillis des règles d'association en

utilisant la méthode de Luxenburger [Luxenburger, 1991] : pour deux concepts (A', A) et (B', B) tels que $(B', B) \leq (A', A)$ directement liés dans le treillis, on peut extraire la règle d'association $A \to B \setminus A$ de support $|B'|$ et de confiance $|B'|/|A'|$. On obtient alors des règles d'association de la forme $A \to B$.

Parmi les règles extraites, on souhaite ne conserver que les règles permettant d'identifier des associations entre des EIM différents. Par exemple, une règle ne décrit pas une telle association lorsque le ou les EIM de la partie droite ne sont qu'une spécialisation de celui ou ceux de la partie gauche (par exemple l'association $\{\langle\{d_1\}, \{p_1\}\rangle\} \to \{\langle\{d_1, d_2\}, \{p_1\}\rangle\}$ semble triviale dans le cadre ce cette étude).

On ne conserve alors que les règles possédant dans leur partie droite un EIM (D_R, P_R) telle que il n'existe aucun EIM (D_L, P_L) dans la partie gauche de la règle tel que D_R et D_L ou P_R et P_L soient comparable par $\leq_{\mathcal{O}}$. Cette condition permet de s'assurer que la partie droite de la règle introduise de nouvelles classes de médicaments et phénotypes, qui ne sont pas des généralisations ou spécialisations de celles de la partie gauche.

2.3 Traitement des données patient

Cette section présente les deux jeux de données desquels ont été extraites des règles d'association entre EIM. Le premier jeu de données a été extrait d'un ensemble de DME de patients diagnostiqués comme atteints par le Lupus Érythémateux Disséminé (LED), une grave maladie auto-immune. Ces patients sont fortement susceptibles aux EIM puisqu'ils prennent de nombreux traitements, contre le LED et les maladies opportunes [Vasudevan and Ginzler, 2009]. Le second jeu de données a été extrait du système américain de rapports d'EIM, FAERS (*Food & Drug Administration — FDA — Adverse Event Reporting System*).

2.3.1 STRIDE

Le premier jeu de données utilisé dans cette étude est un ensemble de 6869 DME de patients extrait de STRIDE, une entrepôt de DME de l'Hôpital de l'Université de Stanford [Lowe et al., 2009]. Ce jeu de données est anonyme et concerne uniquement des patients atteints de LED, suivis entre 2008 et 2014 ; 451000 visites hospitalières y sont documentées. Il y est renseigné les délais entre chaque visite, les diagnostics encodés par un code ICD-9-CM et les prescriptions sous la forme d'une liste d'ingrédients actifs representés par leur identifiants RxNorm. Ceux-ci sont mis en correspondance avec les classes ATC associées.

Les EIM sont extraits des dossiers patients comme décrit dans l'Algorithme 2.1. On établit d'abord une liste d'EIM candidats, en considérant le DME de chaque patient séparément. Un EIM candidat est une paire (D_i, P_i) dans laquelle le lien entre les médicaments de D_i et les phénotypes de P_i est à confirmer. Pour chaque paire de visites consécutives dans le DME, on extrait l'ensemble des médicaments D_i prescrits durant la première visite, et les phénotypes P_i diagnostiqués durant la seconde. On ne sélectionne que les paires de visites où l'intervalle entre deux visites consécutives est inférieur à 14 jours : on s'intéressera ici à des effets secondaires à plutôt court terme, en faisant l'hypothèse que les phénotypes P_i ont davantage de chance d'être causés par les médicaments D_i si la période séparant les deux visites est courte. De plus, la Table 2.5 illustre qu'augmenter cet intervalle n'augmente pas grandement le nombre de patients dans le jeu de données extrait des DME.

On obtient alors un ensemble d'EIM candidats sous la forme (D_i, P_i). On ne conserve dans P_i que les phénotypes reportés comme un effet secondaire d'au moins un médicament de D_i dans la base de données d'indications et d'effets secondaires SIDER 4.1 [Kuhn et al., 2015]. On ne

Data: Le DME d'un seul patient, sous la forme d'une séquence de visites hospitalières $\langle v_0, v_1, \dots v_n \rangle$
La base de données SIDER, qui permet d'associer à un médicament d l'ensemble de ses effets secondaires, que l'on notera $EffetsSecondaires(d)$
$IntervalleMax$, le nombre maximal de jours dans lesquels on recherche un EIM après la prescription d'un médicament
Result: Un ensemble d'EIM

$EIM = \varnothing$
for *chaque visite* v_i *dans la séquence d'événements du DME* **do**
  if *la visite* v_{i+1} *suivante est dans un délai de moins de* $IntervalleMax$ *jours* **then**
    $D_i \leftarrow$ ensemble des médicaments prescrits durant v_i
    $P_i \leftarrow$ ensemble des phénotypes diagnostiqués durant v_{i+1}
    $P_i \leftarrow P_i \cap \bigcup_{d \in D_i} EffetSecondaires(d)$
    if $P_i \neq \varnothing$ **then**
      $EIM \leftarrow EIM \cup \{(D_i, P_i)\}$
    end
  end
end
for *chaque EIM candidat* (D_i, P_i) *dans* EIM **do**
  if *il existe une visite* v_j *où est prescrit l'ensemble des médicaments* D_j *tel que* $D_i \subseteq D_j$ **then**
    $EIM \leftarrow EIM - (D_i, P_i)$
  end
end
return EIM

Algorithm 2.1: Algorithme d'extraction d'EIM du DME d'un patient

TABLE 2.5 – Nombre de patients présentant au moins 2 EIM sélectionnés par l'Algorithme 2.1 pour différentes valeurs d'$IntervalleMax$.

Intervalle (jours)	1	2	6	10	14	18	22	26	30
|Patients|	434	461	498	526	548	555	558	564	576
|EIM|	2,396	2,587	2,902	3,110	3,286	3,388	3,454	3,501	3,621

conserve dans le jeu de données extrait uniquement les candidats où $P_i \neq \varnothing$, c'est-à-dire pour lesquel au moins un effet secondaire est présent. De plus, on exclut un EIM candidat (D_i, P_i) s'il existe pour le même patient une visite où est prescrit l'ensemble de médicaments D_j tel que $D_i \subseteq D_j$. En effet, on considèrera que des prescriptions répétées du même ensemble de médicaments par un praticien comme indiquant une absence de nocivité pour le patient.

Dans les cas où plusieurs EIM possèdent des ensembles de médicaments comparables (par l'ordre partiel $\subseteq$), on conserve uniquement les EIM avec un ensemble de médicaments maximal. En effet, on cherche à trouver des associations entre différents EIM, on souhaite alors éviter de considérer plusieurs fois des ensembles de médicaments similaires. Finalement, on ne conserve que les patients avec au moins deux EIM, l'objectif étant d'identifier des associations entre EIM fréquemment associés. Après les différents filtres appliqués aux données, on obtient un total de 3286 EIM répartis entre 548 patients.

Dans ce jeu de données, l'ontologie de médicaments utilisée est ATC, tandis que l'ontologie de phénotypes est ICD-9-CM.

2.3.2 FAERS

FAERS (FDA (U.S. *Food & Drug Administration*) *Adverse Event Reporting System*) est le système de rapport d'EIM américain [FDA, 2016]. La FDA publie une base de données comportant les EIM reportés par les patients, professionnels de la santé et laboratoires pharmaceutiques aux États-Unis. Cette base de données est utilisée par la FDA pour la pharmacovigilance post-commercialisation des médicaments autorisés à la vente sur le territoire américain. Elle constitue une ressource importante pour la recherche en pharmacovigilance, notamment pour la détection d'EIM [Sakaeda et al., 2013] ou d'interactions indésirables entre médicaments [Kakar et al., 2016]. La diversité des sources des différents rapports d'EIM rend néanmoins FAERS difficile à utiliser sans un traitement des données préalable, les médicaments et phénotypes rapportés n'étant pas nécessairement décrits dans un vocabulaire normalisé.

Une ressource de 2016, AEOLUS [Banda et al., 2016] propose des outils pour faire correspondre des rapports FAERS aux vocabulaires RxNorm pour les médicaments et SNOMED CT pour les phénotypes. Nous avons utilisé ces outils pour reconstruire une base de données de rapports FAERS sur la période 2012 quatrième trimestre à 2016 second trimestre inclus, liée à RxNorm et SNOMED CT. On a ensuite mis en correspondance les identifiants RxNorm des médicaments à leur classes ATC.

Data: Un cas de FAERS, sous la forme d'une collection de rapports d'EIM
Result: Un ensemble d'EIM
$EIM = \varnothing$
for *chaque rapport* r_i *du cas* **do**
  $D_i \leftarrow$ ensemble des médicaments rapportés r_i
  $P_i \leftarrow$ ensemble des phénotypes diagnostiqués durant r_i
  $EIM \leftarrow EIM \cup \{(D_i, P_i)\}$
end
for *chaque EIM candidat* (D_i, P_i) *dans* EIM **do**
  if *il existe une rapport* r_j *où l'ensemble des médicaments* D_j *est tel que* $D_i \subseteq D_j$ **then**
    $EIM \leftarrow EIM - (D_i, P_i)$
  end
end
return EIM

Algorithm 2.2: Algorithme d'extraction d'EIM d'un cas FAERS

L'Algorithme 2.2 décrit le processus de séléction des EIM dans FAERS. Chaque rapport d'événement dans FAERS liste l'ensemble des médicaments pris avant l'EIM D_i et la liste de phénotypes caractérisant cet EIM P_i. Ainsi chaque rapport d'EIM peut être formalisé sous la forme (D_i, P_i). Ces rapports d'EIM sont groupés par "cas", un cas se constituant de plusieurs rapports d'EIM d'un même patient. Nous avons selectionné dans la base de données FAERS les cas présentant plusieurs EIM, en excluant les EIM où l'ensemble des médicaments est déjà inclus dans un autre EIM. Avec ces contraintes, nous avons pu extraire 570 cas de patients avec 2 ou plus EIM, pour un total de 1148 EIM.

Dans ce jeu de données, l'ontologie de médicaments utilisée est ATC, tandis que l'ontologie de phénotypes est SNOMED-CT.

2.4 Résultats

Cette section présente les résultats obtenus à partir des expériences d'extraction de règles d'association utilisant les différents opérateurs. Les première et seconde structures de patrons utilisant la structure en arbre d'ICD-9-CM pour simplifier la représentation des EIM, ne peuvent être appliquées que sur les données extraites des DME codés avec ICD-9-CM. La troisième structure de patrons permet d'utiliser n'importe quelle ontologie de phénotypes, et peut donc être appliquée aux données des DME et de FAERS.

On présente alors les résultats de quatre expériences : trois correspondant aux structures de patrons présentées précédemment sur les DME, et une utilisant la troisième structure de patrons avec le jeu de données extrait de FAERS.

2.4.1 Vue d'ensemble des résultats

Les structures de patrons présentées précedemment permettent de construire chacune un treillis de concepts duquel sont extraites des règles d'association. On choisit d'extraire uniquement des règles avec un support et une confiance d'au moins, respectivement, 5 et 0.75. La Table 2.6 présente quelques données quantitatives de ce processus.

TABLE 2.6 – Données quantitatives sur le processus d'extraction de règles d'association entre EIM pour les différentes expériences réalisées.

Expérience	1 (DME)	2 (DME)	3 (DME)	3 (FAERS)
Nombre de patients	548	548	548	570
Nombre d'EIM	3,286	3,286	3,286	1,148
Taille du treillis de concepts	1.9 million	2.3 million	2.5 million	22,700
Règles extraites	5 million	7 million	9 million	18,500
Règles extraites après filtrage	772	1,907	913	493
Règle avec un support de 8 ou plus	8	50	15	151
Support maximum	9	10	10	27

On observe d'abord que l'expérience sur FAERS produit un nombre de règles beaucoup moins important que celles avec les DME de STRIDE. On observe également que l'expérience utilisant les données de FAERS génère un treillis de concepts beaucoup plus petit que celles utilisant les données de DME (100 fois moins de concepts), et ce malgré un nombre de patients comparable. Cependant après filtrage des règles, leur nombre est du même ordre de grandeur (seulement 2 fois moins). Cette différence peut être expliquée par les différences entre les deux jeux de données : le jeu de données de DME ne concerne que des patients atteints de lupus érythémateux disséminé, tandis que le jeu de données de FAERS concerne une population plus générale. De plus, le plus grand nombre d'EIM extraits des DME tend à augmenter la similarité entre patients, augmentant alors la taille du treillis de concepts généré. L'ensemble de règles filtrées extraites de chaque expérience est disponible à l'adresse : `https://github.com/g-a-perso/ADE-associations/`.

2.4.2 Analyse statistique des associations entre EIM

Les Figures 2.2 et 2.3 proposent une vue quantitative des classes ATC associées par des règles d'association extraites dans la troisième expérience sur les DME. On a isolé chaque paire de classes ATC associées par une règle, c'est-à-dire où chacune des deux classes ATC apparaît d'un côté différent au sein d'une même règle d'association. On obtient ainsi un ensemble d'associations entre deux classes ATC. La Figure 2.2 donne la fréquence de ces associations tandis que la Figure 2.3 donne la différence de cette fréquence avec la fréquence qui serait attendue si ces associations étaient aléatoires.

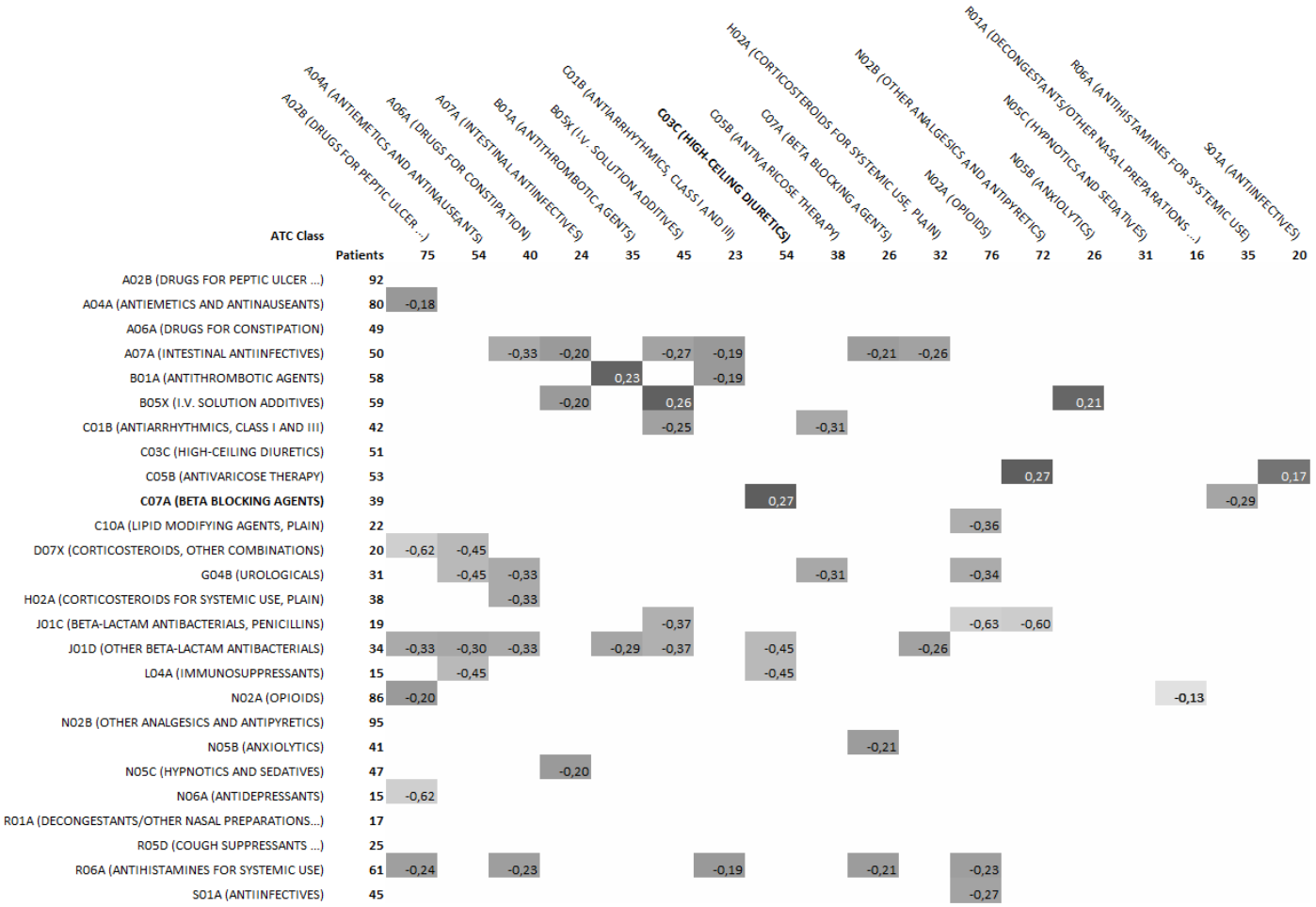
Pour chaque paire de classes de médicaments ATC (l, r), on cherche l'ensemble des règles extraites des DME et de FAERS de la forme $L \rightarrow R$ telles que l (ou un de ses descendants) apparaît dans L et r (ou un de ses descendants) apparaît dans R. On calcule ensuite le support de cet ensemble de règles comme étant le nombre de patients vérifiant au moins une de ces règles. On calcule ensuite pour chaque (l, r) le ratio entre *(i)* le support des règles telles que l apparaît dans L et r apparaît dans R, *(ii)* le support des règles telles que l apparaît dans L. Ces valeurs sont visibles en Figure 2.2. Ce ratio exprime la fréquence à laquelle les règles associent un EIM impliquant r à un EIM impliquant l. On notera que le total de ces ratios peut être plus grand que 1 dans chaque ligne puisque une règle peut associer plus que deux classes ATC, et qu'un patient peut vérifier plusieurs règles.

On utilise un test Z pour évaluer la significativité de l'écart obtenu au ratio attendu si les associations entre EIM étaient extraites aléatoirement. La Figure 2.3 présente les écarts au ratio attendu significatives (avec $p < 0.001$). Pour chaque classe ATC, le ratio attendu est calculé comme le support de l'ensemble des règles où cette classe apparaît dans la partie droite, divisé par le support de l'ensemble des règles (soit le nombre de patients décrits par une règle). Le test

FIGURE 2.2 – Distribution des associations des classes de médicaments trouvées dans la troisième expérience sur les DME. Sur la gauche, les classes ATC apparaissant dans la partie de gauche de règles d'association et le support (en nombre de patients décrits) de l'ensemble des règles correspondantes. En haut, les classes ATC apparaissant dans la partie droite de règles d'association et le support de l'ensemble des règles correspondantes. Les valeurs dans les cellules représentent, pour a la classe en ligne et b la classe en colonne le ratio entre *(i)* le support des règles $A \rightarrow B$ telles que a apparaît dans A et b apparaît dans B, *(ii)* le support des règles telles que l apparaît dans A. Par exemple, le support des règles ou *Beta-Blocking Agents* (C07A) est présent dans la partie gauche est 39 et le support du sous-ensemble de ces règles où *High-Ceiling Diuretics* (C03C) est présent dans la partie droite est 72 % (0.72) de 39.

ATC Class	Patients	A02B (DRUGS FOR PEPTIC ULCER ...)	A04A (ANTIEMETICS AND ANTINAUSEANTS)	A06A (DRUGS FOR CONSTIPATION)	A07A (INTESTINAL ANTIINFECTIVES)	B01A (ANTITHROMBOTIC AGENTS)	B05X (I.V. SOLUTION ADDITIVES)	C01B (ANTIARRHYTHMICS, CLASS I AND III)	C03C (HIGH-CEILING DIURETICS)	C05B (ANTIVARICOSE THERAPY)	C07A (BETA BLOCKING AGENTS)	H02A (CORTICOSTEROIDS FOR SYSTEMIC USE, PLAIN)	N02B (OTHER ANALGESICS AND ANTIPYRETICS)	N02A (OPIOIDS)	N05B (ANXIOLYTICS)	N05C (HYPNOTICS AND SEDATIVES)	R01A (DECONGESTANTS/OTHER NASAL PREPARATIONS ...)	R06A (ANTIHISTAMINES FOR SYSTEMIC USE)	S01A (ANTIINFECTIVES)
		75	54	40	24	35	45	23	54	38	26	32	76	72	26	31	16	35	20
A02B (DRUGS FOR PEPTIC ULCER ...)	92	0,59	0,40	0,23	0,16	0,28	0,36	0,20	0,43	0,29	0,12	0,25	0,61	0,61	0,26	0,23	0,17	0,28	0,18
A04A (ANTIEMETICS AND ANTINAUSEANTS)	80	0,44	0,40	0,24	0,16	0,30	0,36	0,06	0,48	0,19	0,19	0,18	0,54	0,56	0,24	0,28	0,18	0,15	0,11
A06A (DRUGS FOR CONSTIPATION)	49	0,69	0,43	0,51	0,10	0,41	0,45	0,10	0,45	0,31	0,35	0,20	0,61	0,59	0,27	0,35	0,22	0,22	0,16
A07A (INTESTINAL ANTIINFECTIVES)	50	0,52	0,24	0	0	0,16	0,10	0	0,40	0,22	0	0	0,50	0,54	0,12	0,18	0	0,28	0
B01A (ANTITHROMBOTIC AGENTS)	58	0,71	0,50	0,50	0,26	0,52	0,55	0	0,43	0,34	0,19	0,22	0,53	0,60	0,34	0,33	0,21	0,24	0,16
B05X (I.V. SOLUTION ADDITIVES)	59	0,58	0,54	0,36	0	0,41	0,63	0,08	0,56	0,19	0,20	0,14	0,73	0,61	0,42	0,27	0,20	0,15	0,17
C01B (ANTIARRHYTHMICS, CLASS I AND III)	42	0,43	0,29	0,12	0,14	0,36	0,12	0,12	0,36	0	0,17	0,21	0,48	0,64	0,21	0,43	0,14	0,31	0,14
C03C (HIGH-CEILING DIURETICS)	51	0,69	0,49	0,49	0,12	0,45	0,33	0,16	0,43	0,33	0,31	0,29	0,55	0,61	0,31	0,27	0,20	0,25	0,12
C05B (ANTIVARICOSE THERAPY)	53	0,62	0,43	0,28	0,11	0,36	0,32	0,09	0,40	0,47	0,13	0,23	0,57	0,87	0,28	0,28	0,13	0,43	0,34
C07A (BETA BLOCKING AGENTS)	39	0,46	0,36	0,21	0	0,26	0,41	0,18	0,72	0,15	0,36	0,13	0,56	0,49	0,15	0,15	0,21	0	0
C10A (LIPID MODIFYING AGENTS, PLAIN)	22	0,50	0,59	0,23	0	0,59	0,36	0	0,68	0,27	0,36	0	0,27	0,41	0,32	0	0	0,23	0
D07X (CORTICOSTEROIDS, OTHER COMBINATIONS)	20	0	0	0	0	0	0,25	0	0,35	0	0	0	0,55	0,25	0	0	0	0,25	0,25
G04B (UROLOGICALS)	31	0,58	0	0	0	0,16	0,32	0	0,32	0	0	0,16	0,29	0,32	0,16	0	0	0,16	0,16
H02A (CORTICOSTEROIDS FOR SYSTEMIC USE, PLAIN)	38	0,50	0,37	0	0	0,34	0,16	0	0,29	0,13	0	0,24	0,50	0,45	0,16	0,24	0,16	0,16	0
J01C (BETA-LACTAM ANTIBACTERIALS, PENICILLINS)	19	0,58	0,26	0	0	0,42	0	0	0,68	0	0	0	0	0	0	0	0	0	0
J01D (OTHER BETA-LACTAM ANTIBACTERIALS)	34	0,29	0,15	0	0,15	0	0	0,15	0	0,38	0	0	0,47	0,62	0,15	0,32	0	0,38	0,15
L04A (IMMUNOSUPPRESSANTS)	15	0,40	0	0	0,33	0	0	0,33	0	0	0	0	0,47	0,47	0	0	0	0	0
N02A (OPIOIDS)	86	0,42	0,41	0,27	0,16	0,29	0,30	0,19	0,36	0,19	0,17	0,22	0,51	0,56	0,22	0,28	0	0,20	0,08
N02B (OTHER ANALGESICS AND ANTIPYRETICS)	95	0,52	0,51	0,23	0,16	0,28	0,39	0,12	0,37	0,35	0,17	0,27	0,63	0,61	0,20	0,27	0,14	0,29	0,20
N05B (ANXIOLYTICS)	41	0,39	0,51	0,17	0,12	0,37	0,41	0	0,44	0,15	0	0,15	0,54	0,37	0,20	0,24	0,12	0,20	0,15
N05C (HYPNOTICS AND SEDATIVES)	47	0,62	0,51	0,38	0	0,40	0,40	0,15	0,49	0,15	0,23	0,28	0,62	0,51	0,32	0,38	0,11	0,26	0,17
N06A (ANTIDEPRESSANTS)	15	0	0,33	0	0	0,53	0,33	0	0,73	0	0	0	0,60	0,40	0	0	0	0,33	0
R01A (DECONGESTANTS/OTHER NASAL PREPARATIONS...)	17	0,76	0,76	0,41	0	0,76	0,29	0	0,71	0,59	0,41	0,35	0,47	0,76	0	0	0,53	0	0,29
R05D (COUGH SUPPRESSANTS ...)	25	0,56	0,60	0,48	0	0,56	0,56	0	0,36	0,28	0,20	0	0,48	0,64	0,36	0,36	0	0,40	0
R06A (ANTIHISTAMINES FOR SYSTEMIC USE)	61	0,38	0,28	0,10	0,18	0,13	0,44	0	0,34	0,23	0	0,16	0,39	0,56	0,10	0,13	0	0,34	0,11
S01A (ANTIINFECTIVES)	45	0,53	0,31	0,36	0,29	0,27	0,38	0	0,51	0,29	0,11	0,11	0,36	0,62	0,16	0,24	0,16	0,22	0,29

FIGURE 2.3 – Résultats des tests de significativité statistique de la distribution des associations entre EIM extraites dans la troisième expérience sur les DME. Le ratio dans chaque cellule de la Figure 2.2 a été comparé à sa valeur attendue en faisant l'hypothèse d'une distribution aléatoire des classes ATC dans la partie droite des règles d'association. Les cellules vides indiquent que la différence entre les ratios observés et attendus n'est pas significative ($p > 0.001$, test Z). Les autres cellules indiquent la différence entre les ratios observés et attendus, et que cette différence est significative ($p < 0.001$, test Z).



Z permet ensuite de trouver les ratios dont la différence à ce ratio attendu est signifivative à $p < 0.0001$.

On observe notamment quelques associations d'intérêt entre classes ATC (avec $p < 0.001$). Par exemple, on constate notamment que les EIM impliquant des agents bêta-bloquants (classe ATC C07A) sont fortement associés à des EIM impliquant des diurétiques de l'anse (classe ATC C03C). Ces deux classes de médicaments sont impliquées dans des thérapies contre l'hypertension, parfois en combinaison, ce qui peut expliquer la forte association entre des EIM causés par ces deux types de médicaments. Ainsi, il semble probable qu'un grand nombre de patients se voient prescrire des médicaments de ces deux classes. On observe également que les EIM impliquant des agents antithrombotiques (classe ATC B01A) sont associés à d'autres EIM impliquant d'autres médicaments de la même classe. Ainsi, il semble que l'approche proposée ici permet de révéler des associations significatives entres EIM causés par des médicaments d'une même classe ou de plusieurs classes différentes.

2.4.3 Exemples de règles d'association

La Table 2.7 présente les 15 règles d'association extraites lors de la troisième expérience sur les DME avec un support de 8 ou plus. Par exemple, on obtient la règle d'association suivante, avec un support de 10 et une confiance de 0.77 :

$$\{\langle\{\text{C08DB}\}, \{\text{ICD 428.0}\}\rangle\} \rightarrow \{\langle\{\text{A02B}\}, \{\text{ICD 427.31}\}\rangle\}$$

Cette règle exprime qu'environ 77% des patients qui présentent le phénotype *congestive heart failure* (ICD 428.0) après une prescription de *benzothiazepine derivatives* (C08DB) présentent également le phénotype *atrial fibrillation* (ICD 427.31) après prescription d'un médicament de la classe ATC *drug for peptic ulcer and gastro-esophageal reflux disease* (A02B). Cette règle est vérifiée pour 10 patients.

La Table 2.8 présente des exemples d'associations entre EIM obtenus dans les trois expériences sur les DME. Ici, on trouve dans ces trois expériences une association similaire à différents niveaux de généralisation. Dans l'intérêt de la lisibilité, les trois règles sont exprimées dans le formalisme de la troisième structure de patrons.

Dans cet exemple, on observe que la règle d'association de l'expérience 2 est plus générale que la règle de l'expérience 1 (R06A est un ancêtre de la doxylamine dans ATC). En effet, la seconde structure de patrons permettait la généralisation sur les classes de médicaments. Dans chaque expérience, la représentation des EIM permet de décrire l'implication de plusieurs médicaments ou classes de médicaments. Les règles d'association peuvent également associer plus que deux EIM, comme par exemple la règle de la troisième expérience.

2.4.4 Evaluation des règles extraites sur STRIDE

Le jeu de données extrait des DME de STRIDE représente uniquement les patients atteints de LED, soit une petite partie de la base de données STRIDE qui contient environ 2 millions de DME. On a évalué les 15 règles ayant le plus grand support (entre 8 et 10) parmi les règles extraites des DME, listées en Table 2.7. Chacune de ces règles a été transformée en une requête SQL permettant d'identifier les DME de STRIDE vérifiant chaque règle. La Table 2.7 donne le support de chaque règle dans le jeu de données original (S_1) et son support dans la base de données STRIDE complète (S_2). On observe dans que pour ces règles, le nombre de DME les vérifiant varie entre 33 à 326 pour une moyenne de 86.2 Cela illustre que des règles d'association extraites à partir de DME de patients atteints de LED peuvent être pertinentes en dehors du jeu de données initial.

TABLE 2.7 – Exemple de 15 règles d'association extraites des DME de STRIDE dans la troisième expérience et présentant les supports les plus élevés. S_1 est le support dans le jeu de données utilisé pour l'extraction. S_2 est le support dans la base de données STRIDE complète.

Rule	S_1	S_2
{⟨{Anilides}, {Thrombocytopenia, unsp.}⟩, ⟨{Antithrombotic agents}, {Thrombocytopenia, unsp.}⟩} → {⟨{Opioids}, {Anemia, unsp.}⟩}	9	326
{⟨{Serotonin (5HT3) antagonists}, {Thrombocytopenia, unsp.}⟩, ⟨{Anilides}, {Thrombocytopenia, unsp.}⟩, ⟨{Antithrombotic agents}, {Thrombocytopenia, unsp.}⟩} → {⟨{Opioids}, {Anemia, unsp.}⟩}	8	256
{⟨{Proton pump inhibitors}, {Thrombocytopenia, unsp.}⟩, ⟨{Antithrombotic agents}, {Thrombocytopenia, unsp.}⟩} → {⟨{Opioids}, {Anemia, unsp.}⟩, ⟨{Drugs for peptic ulcer and GORD}, {Anemia, unsp.}⟩}	9	176
{⟨{Proton pump inhibitors}, {Thrombocytopenia, unsp.}⟩, ⟨{Anilides}, {Thrombocytopenia, unsp.}⟩, ⟨{Antithrombotic agents}, {Thrombocytopenia, unsp.}⟩} → {⟨{Drugs for peptic ulcer and GORD}, {Anemia, unsp.}⟩, ⟨{Opioids}, {Anemia, unsp.}⟩}	8	157
{⟨{Benzothiazepine derivatives}, {Congestive heart failure, unsp.}⟩} → {⟨{Drugs for peptic ulcer and GORD}, {Atrial fibrillation}⟩}	10	129
{⟨{Drugs for peptic ulcer and GORD}, {Atrial fibrillation}⟩, ⟨{ACE inhibitors, plain}, {Atrial fibrillation}⟩, ⟨{Anilides}, {Atrial fibrillation}⟩} → {⟨{Serotonin (5HT3) antagonists}, {Heart failure}⟩, ⟨{Drugs for peptic ulcer and GORD}, {Congestive heart failure, unsp.}⟩}	8	66
{⟨{Serotonin (5HT3) antagonists}, {Atrial fibrillation}⟩, ⟨{Drugs for peptic ulcer and GORD}, {Atrial fibrillation}⟩, ⟨{ACE inhibitors, plain}, {Atrial fibrillation}⟩} → {⟨{Electrolyte solutions}, {Congestive heart failure, unsp.}⟩, ⟨{Osmotically acting laxatives}, {Heart failure}⟩}	8	64
{⟨{Proton pump inhibitors}, {Thrombocytopenia, unsp.}⟩, ⟨{Anilides}, {Thrombocytopenia, unsp.}⟩, ⟨{Glucocorticoids}, {Thrombocytopenia, unsp.}⟩} → {⟨{Opioids}, {Anemia, unsp.}⟩, ⟨{Drugs for peptic ulcer and GORD}, {Anemia, unsp.}⟩}	10	49
{⟨{Proton pump inhibitors}, {Congestive heart failure, unsp.}⟩, ⟨{Antithrombotic agents, Anilides, Opium alkaloids and derivatives}, {Heart failure}⟩, ⟨{Anilides}, {Congestive heart failure, unsp.}⟩, ⟨{Anxiolytics}, {Heart failure}⟩, ⟨{Electrolyte solutions}, {Congestive heart failure, unsp.}⟩} → {⟨{Opioids}, {Anemia, unsp.}⟩}	9	37
{⟨{Sulfonamides, plain}, {Congestive heart failure, unsp.}⟩, ⟨{Antithrombotic agents, Anilides, Opium alkaloids and derivatives}, {Heart failure}⟩, ⟨{Proton pump inhibitors}, {Congestive heart failure, unsp.}⟩, ⟨{Anxiolytics}, {Heart failure}⟩, ⟨{Anilides}, {Congestive heart failure, unsp.}⟩, ⟨{Electrolyte solutions}, {Congestive heart failure, unsp.}⟩, ⟨{Sulfonamides, plain, R05D}, {Heart failure}⟩} → {⟨{Opioids}, {Anemia, unsp.}⟩}	8	33

Rule	S_1	S_2
{⟨{Anilides, Opium alkaloids and derivatives, Proton pump inhibitors}, {Heart failure}⟩, ⟨{Anilides, Proton pump inhibitors}, {Congestive heart failure, unsp.}⟩, ⟨{Antithrombotic agents, Anilides, Opium alkaloids and derivatives}, {Heart failure}⟩, ⟨{Anxiolytics}, {Congestive heart failure, unsp.}⟩, ⟨{Electrolyte solutions}, {Congestive heart failure, unsp.}⟩} → {⟨{Opioids}, {Anemia, unsp.}⟩}	8	31
{⟨{Antithrombotic agents, Anilides, Opium alkaloids and derivatives, Serotonin (5HT3) antagonists}, {Heart failure}⟩, ⟨{Anilides, Serotonin (5HT3) antagonists}, {Congestive heart failure, unsp.}⟩, ⟨{Proton pump inhibitors}, {Congestive heart failure, unsp.}⟩, ⟨{Anxiolytics}, {Heart failure}⟩, ⟨{Electrolyte solutions}, {Congestive heart failure, unsp.}⟩} → {⟨{Opioids}, {Anemia, unsp.}⟩}	8	29
{⟨{Other antiinfectives}, {Congestive heart failure, unsp.}⟩, ⟨{Drugs for peptic ulcer and GORD}, {Congestive heart failure, unsp.}⟩, ⟨{Anilides, Glucocorticoids}, {Congestive heart failure, unsp.}⟩} → {⟨{Sulfonamides, plain}, {Anemia, unsp.}⟩}	9	26
{⟨{Other antiinfectives}, {Congestive heart failure, unsp.}⟩, ⟨{Antithrombotic agents}, {Heart failure}⟩, ⟨{Drugs for peptic ulcer and GORD}, {Congestive heart failure, unsp.}⟩, ⟨{Anilides, Glucocorticoids}, {Congestive heart failure, unsp.}⟩} → {⟨{Sulfonamides, plain}, {Anemia, unsp.}⟩}	8	25
{⟨{Antibiotics}, {Congestive heart failure, unsp.}⟩, ⟨{Proton pump inhibitors}, {Congestive heart failure, unsp.}⟩, ⟨{Other antiinfectives}, {Congestive heart failure, unsp.}⟩, ⟨{Electrolyte solutions}, {Congestive heart failure, unsp.}⟩, ⟨{Anilides, Glucocorticoids}, {Congestive heart failure, unsp.}⟩} → {⟨{Sulfonamides, plain}, {Anemia, unsp.}⟩}	8	16

TABLE 2.8 – Example de règles similaires à différents niveaux de généralisation dans les trois expériences sur les DME (1 : pas de généralisation, EIM regroupés par code ICD-9-CM de haut niveau, 2 : généralisation des médicaments avec ATC, 3 : généralisation avec ATC et ICD-9-CM).
- A02B : *drugs for peptic ulcer and gastro-oesophagal reflux disease*
- G04BE : *drugs used in erectile dysfunction*
- G04BE04 : *yohimbine*
- N06BC : *xanthine derivatives*
- N06BC01 : *caffeine*
- R06A : *antihistamines for systemic use*
- R06AA : *aminoalkyl ethers*
- R06AA09 : *doxylamine*
- ICD 280-289 : *diseases of the blood and blood-forming organs*
- ICD 285.9 : *anemia, unspecified*
- ICD 580-629 : *diseases of the genitourinary system*
- ICD 586 : *renal failure, unspecified*

Expérience	Règle	Support
1 (DME)	{⟨{yohimbine, doxylamine, vancomycin, caffeine}, {ICD 580-629}⟩} →{⟨{doxylamine, tocainide}, {ICD 280-289}⟩}	5
2 (DME)	{⟨{G04BE, N06BC}, {ICD 580-629}⟩} →{⟨{R06A}, {ICD 280-289}⟩}	9
3 (DME)	{⟨{G04BE, N06BC}, {ICD 586}⟩, ⟨{A02B, N06BC}, {ICD 586}⟩} →{⟨{R06AA}, {ICD 285.9}⟩}	5

2.5 Discussion

Ce chapitre explore une approche fondée sur les structures de patrons pour l'extraction d'EIM fréquemment associés dans des DME et FAERS. Les structures de patrons permettent de fouiller des descriptions détaillées d'EIM. Cependant il est nécessaire de considérer la complexité des algorithmes d'analyse formelle de concepts combinés avec les opérateurs de comparaison proposés. Les ontologies biomédicales permettent une meilleure comparaison des EIM d'un patient, notamment en étant capable de généraliser plusieurs EIM causés par des médicaments d'une même classe. Cela augmente néanmoins le nombre de concepts générés, puisque l'on augmente les possiblités de trouver une description générale commune à de nombreux patients. On observe en effet une augmentation de la taille du treillis de concept lorsque l'on utilise des descriptions plus détaillées et des opérateurs de comparaison faisant davantage usage des ontologies.

2.5.1 Interprétation des associations entres EIM

Une limitation des représentations d'EIM proposées est l'absence de temporalité entre les EIM. Cette limitation est justifiée par deux arguments. D'une part, l'ordre d'occurence d'EIM similaires peut varier entre différents patients. D'autre part, cet ordre peut être verifié dans les DME pour les cas d'intérêt. En se reportant aux DME, on peut par exemple constater dans certains cas que des EIM présents dans la partie gauche d'une règle d'association se produisent avant les EIM de la partie droite. Ainsi les règles d'association n'expriment qu'une association fréquente de ces EIM chez les patients, sans information de temporalité.

Dans les expériences utilisant les données extraites des DME, seuls les EIM exprimés dans une fenêtre de 14 jours après la prescription ont été considérés. Ainsi, seuls des EIM à court terme ont été considérés. La représentation des EIM pourrait être enrichie avec une information sur le délai entre la prescription et l'apparition du phénotype indésirable. Cela permettrait d'extraire des associations dans un jeu de données d'EIM à court ou long terme, tout en étant capable de discriminer au besoin entre ces différentes manifestations. Cela permettrait par exemple l'extraction d'associations entre un EIM à court terme et un EIM à long terme : la toxicité d'un médicament à court terme pourrait dans ce cas être utilisée comme un prédicteur de la toxicité à long terme d'un autre.

Une autre limitation de l'extraction de règles d'association est que les règles extraites ne permettent pas d'exposer de relation causale entre les EIM associés. Cependant, il semble plus approprié de rechercher une cause biologique commune à deux EIM associés par une règle, que de chercher une relation causale directe entre ces deux EIM.

2.5.2 Application à différents jeux de données et ontologies

Une grande quantité de règles d'association peut être extrait des treillis de concepts générés. Les règles sont filtrées automatiquement de manière à exclure les règles qui ne répondent pas à la problématique étudiée. L'approche proposée est flexible et est applicable à différents jeux de données, cependant, il est difficile de comparer les règles d'association extraites de jeux de données très différents et décrits par différentes ontologies. Ainsi, on a pu tester les règles selectionnées obtenues sur le jeu de données des DME de STRIDE sur la base de données STRIDE entière. Les résultats de ces tests, présentés en Table 2.7, indiquent que les règles extraites d'un ensemble de DME de patients diagnostiqués avec un LED peuvent s'appliquer à un ensemble plus général de patients. En effet, les patients atteints de LED sont susceptibles à de nombreuses occurrences d'EIM causés par la grande diversité de médicaments qui leur sont prescrits. Les DME de tels patients, utilisés conjointement avec des ontologies biomédicales peuvent permettre l'identification d'EIM fréquemment associés, même dans la population générale des patients.

On observe également une grande différence entre les résultats obtenus par les expériences sur les DME de STRIDE et sur les rapports de FAERS. Cela peut s'expliquer par la différence de nature des deux jeux de données : le jeu de données extrait de FAERS est constitué d'EIM rapporté d'abord par des patients, tandis que le jeu de données provenant des DME est constitué d'EIM automatiquement extraits de ces dossiers. Le jeu de données issu de STRIDE ne concerne que des patients diagnostiqués avec un LED, tandis que FAERS concerne la population générale, avec nécessairement une plus grande diversité des traitements et des phénotypes. Les patients présents dans STRIDE présentent également davantage d'EIM en moyenne, mais cela est attendu étant donné le traitement lourd prescrit contre le LED.

2.5.3 Conclusion

Les structures de patrons permettent d'avoir un processus d'extraction d'associations utilisant une représentation expressive des EIM. Cette représentation permet de considérer les multiples médicaments et phénotypes impliqués dans un tel événement. Les structures de patrons permettent également d'enrichir cette représentation avec plusieurs ontologies biomédicales, rendant ainsi possible la comparaison sémantique des EIM. A notre connaissance, cette approche est unique dans sa capacité à comparer des représentations détaillées d'EIM en considérant de multiples ontologies dans le but d'extraire des associations entre EIM fréquemment associés. Cette approche est également flexible et peut être appliquée à différents DME et systèmes de

rapport d'EIM liés à des ontologies biomédicales quelconques. La généricité de cette approche a été démontrée par son application à deux jeux de données, chacune lié à deux parmi trois ontologies biomédicales différentes.

Les règles d'association extraites dans ce travail pourraient servir comme base pour un système de recommandation. Par exemple, un tel système pourrait proposer une recommandation contre la prescription d'un médicament d à un patient donné si ce patient a déjà présenté un EIM associé par une règle à un autre EIM impliquant le médicament d. Il serait néanmoins nécessaire d'utiliser et d'évaluer une méthode plus précise pour l'extraction des EIM dans les DME.

Les médicaments impliqués dans des règles d'association pourraient être étudiés à la recherche de mécanismes communs pouvant causer des EIM fréquemment associés. Il pourrait alors être utile de classer les règles d'association en fonction de la gravité des phénotypes présentés par les EIM. La représentation des EIM choisie pourrait être étendue pour inclure davantage de propriétés des médicaments et phénotypes, comme par exemple les cibles des médicaments annotées par des classes GO. Cela permettrait notamment d'extraire des règles d'association tout en prenant en compte directement les mécanismes d'action des médicaments.

3

Similarités sémantiques pour la classification de maladies à partir d'un diseasome

Sommaire

3.1 Problème biomédical

Depuis quelques années prévaut l'hypothèse selon laquelle l'étude de réseaux de maladies, ou diseasomes, permettrait la découverte de nouvelles connaissances sur les mécanismes ou les traitements possibles de ces maladies [Barabási et al., 2011]. De nombreuses méthodes pour la construction d'un diseasome existent dans la littérature. Dans un tel réseau, on pourra par exemple relier entre elles les maladies partageant une propriété comme un gène responsable [Goh et al., 2007], des phénotypes [Hidalgo et al., 2009, Hoehndorf et al., 2015] ou rapprochées par une chaîne d'interactions entre les produits de leur gènes [Guney et al., 2016].

On détaille dans ce chapitre la construction et l'analyse d'un diseasome fondé sur la similarité sémantique des phénotypes de maladies. On s'intéressera en particulier au pouvoir explicatif du modèle représenté par le diseasome pour l'étude des maladies génétiques et notamment les déficiences intellectuelles, qui seront le sujet d'étude du Chapitre 4. Ces maladies sont nombreuses et hétérogènes, et peuvent être causées par différents gènes [Inlow and Restifo, 2004, Kaufman et al., 2010]. Une classification experte de ces maladies en 5 classes est presentée ici. On étudiera

alors la capacité de la mesure de similarité, et du diseasome obtenu, à isoler ces différentes classes de déficiences intellectuelles.

3.2 Matériel et méthodes

Dans une étude récente, [Hoehndorf et al., 2015] proposent un diseasome où les maladies sont connectées en fonction de leur similarité phénotypique. Les auteurs ont produit un jeu de données, extrait de la littérature, qui comprend plus de 6000 de maladies OMIM annotées par les classes de phénotypes de l'ontologie MonDO [Mungall et al., 2017]. Ainsi, une maladie est décrite par un ensemble de classes de cette ontologie. Ils utilisent ensuite la fonction de similarité sémantique SimGIC [Pesquita et al., 2007] pour comparer les maladies sur la base de leur phénotypes. On propose ici de construire un diseasome similaire en utilisant la fonction de similarité sémantique IntelliGO [Benabderrahmane et al., 2010], et de comparer les deux diseasomes dans capacité à caractériser les différentes classes de maladies, notamment les déficiences intellectuelles.

3.2.1 Données et ontologies

On dispose pour cette expérience du jeu de données d'annotations de 6220 maladies OMIM utilisé par [Hoehndorf et al., 2015]. Ce jeu de données est disponible à l'adresse `http://aber-owl.net/aber-owl/diseasephenotypes/` [1]. Il contient des association entre des maladies et des phénotypes extraites d'un ensemble de plus de 5 millions de résumés d'articles de MedLine [US NLM, 2018]. Les phénotypes de ces associations sont exprimés sous la forme de classes de l'ontologie MonDO, une ontologie unifiant différentes ontologies et jeux de données sur les maladies et leur phénotypes.

On utilisera trois autres jeux de données comme vérité terrain pour l'évaluation du diseasome produit. Parmi ces trois jeux de données, deux ont été mis à disposition par Hoehndorf *et al.*. Le premier jeu de données proposé par Hoehndorf *et al.* est extrait de SIDER [Kuhn et al., 2015], une base de données d'indications et d'effets secondaires de médicaments. A partir des indications de SIDER, une liste de paires de maladies partageant au moins un médicament est établie. Le second jeu de données proposé par Hoehndorf *et al.* est une classification de maladies extraite de Disease Ontology (DO), utilisant 22 classes de haut niveau de DO. Certaines maladies peuvent néanmois appartenir à plusieurs de ces classes. Finalement, on utilisera les relations entre gènes et maladies fournies par OMIM pour une évaluation qui s'intéressera aux maladies génétiques.

3.2.2 Construction du diseasome fondé sur la similarité phénotypique

On suit ici la méthode décrite dans [Hoehndorf et al., 2015] pour la construction d'un diseasome fondé sur la similarité sémantique des phénotypes. On utilise la mesure de similarité sémantique IntelliGO [Benabderrahmane et al., 2010], décrite en Section 1.2.3 du Chapitre 1, pour calculer la similarité sémantique de chaque paire de maladies du jeu de données. Pour le calcul de cette similarité, IntelliGO tient compte :

— des annotations par des classes MonDO de chaque maladie, fournies dans le jeu de données partagé par Hoehndorf *et al.*.

— de la fréquence d'annotation par chacune des classes de MonDO pour cet ensemble de maladies, de manière à accorder plus d'importance aux annotations moins fréquentes ou par des classes plus spécifiques,

1. Accédé au 28/08/2018.

— des relations `rdfs:subClassOf` et `owl:equivalentClass` dans la hiérarchie des classes de MonDO.

IntelliGO est initialement conçu pour fonctionner avec une hiérarchie de classes sous la forme d'un graphe dirigé acyclique et enraciné. Cependant, MonDO contient des liens d'équivalence entre les classes formant des cycles entre deux classes. Soit $G(V, E)$ le graphe dirigé des classes de MonDO tel que V est l'ensemble des classes et $E = \{(x, y) \in V^2 \mid x \sqsubseteq y \vee x \equiv y\}$, on définit un graphe dirigé acyclique représentant la hiérarchie de ses classes $G'(V', E')$ tel que :

— V' est l'ensemble des classes d'équivalence de V. On rassemble en un seul nœud de V' chaque ensemble de classes équivalentes. Ainsi, les nœuds V' sont une partition de V. On notera $[x]$ la classe d'équivalence d'une classe x, telle que $[x] = \{y \mid x \equiv y\}$.

— $E' = \{([x], [y]) \mid [x] \neq [y] \wedge (x, y) \in E\}$.

De manière similaire, on remplacera, pour chaque maladie, chacune de ses annotations par la classe d'équivalence correspondante. Ce graphe dirigé acyclique de classes extrait de MonDO permet à IntelliGO de calculer la similarité entre les classes, avant d'aggréger ces similarités classe à classe pour obtenir la similarité entre maladies. Cette aggrégation tient compte du contenu d'information (IC) de chaque classe d'équivalence MonDO, en donnant un poids pour l'aggrégation plus élevé aux classes les moins fréquentes.

On utilise IntelliGO pour construire une matrice de similarités deux à deux pour les 6220 maladies. Pour la construction du diseasome, on sélectionne un seuil de similarité correspondant au minimum des 0.5% plus hautes similarités. Chaque maladie est représentée dans le graphe du diseasome par un nœud tandis que les arcs représentent les similarités supérieures ou égales à ce seuil.

3.2.3 Evaluation du diseasome

Hoehndorf *et al.* proposent deux évaluations quantitatives de la capacité de leur diseasome à correctement rapprocher des maladies similaires. Une première évaluation compare les similarités entre maladies à une classification de maladies extraite de Disease Ontology [Schriml et al., 2011]. Une seconde évaluation compare ces similarités aux indications de médicaments partagées par les maladies selon SIDER. Comme l'on s'intéresse aux déficiences intellectuelles qui sont des maladies génétiques, on propose ici une troisième évaluation comparant ces similarités à une similarité fondée sur l'existence d'un gène responsables partagé par les deux maladies.

Ces différentes évaluations sont effectuées grâce à une analyse ROC (*Receiver Operating Characteristic*) qui permet de quantifier le pouvoir de prédiction d'un modèle de classification à différents niveaux de sensibilité. Ici, on considérera le classement des paires de maladies comme un modèle de classification dont on peut faire varier la sensibilité en classant comme positives toutes les paires dont la similarité dépasse un certain seuil. Chaque évaluation repose sur un critère selon lequel une paire de maladies est considérée comme faisant partie de la classe positive ou non : même classe DO, indication ou gène responsable en commun. Une courbe de ROC représente alors le taux de vrais positifs en fonction du taux de faux positifs. On peut ensuite calculer l'aire sous cette courbe (notée ROCAUC — *ROC Area Under Curve*) pour quantifier le pouvoir de prédiction du classement évalué. La valeur de la ROCAUC représente ici la probabilité qu'une paire de maladies choisie aléatoirement dans la classe positive ait une similarité plus élevée qu'une paire de maladies choisie aléatoirement dans la classe négative.

On notera ici $ROCAUC(C, P)$ la fonction permettant de calculer la ROCAUC pour un classement d'objets C et une classe positive définie par l'ensemble d'objets P.

Evaluation par une ontologie de maladies La première évaluation quantifie la capacité de la mesure de similarité à associer les paires de maladies ayant en commun une des 22 classes DO de haut niveau. On calcule ici, pour chacune des 6220 maladies du jeu de données, la similarité sémantique avec les 6219 autres maladies. On obtient alors un classement des paires de maladies croissant avec leur similarité phénotypique. Pour chaque maladie d, on calcule la ROCAUC sur la tâche de classification des paires de maladies partageant une classe DO de haut niveau avec d. On obtient ainsi une valeur de ROCAUC pour chaque maladie. On peut ensuite calculer la valeur moyenne des ROCAUC sur l'ensemble des maladies (comme décrit dans l'Algorithme 3.1), puis sur les maladies de chaque classe DO de haut niveau (comme décrit dans l'Algorithme 3.2), pour évaluer la capacité de classification globale de la mesure de similarité pour les classes DO.

Data: L'ensemble des maladies D, une mesure de similarité $sim : D \times D \to \mathbb{R}^+$
Result: ROCAUC moyen pour l'ensemble des maladies D

$ROCAUC_{moyen} = 0$
for *chaque maladie* $d \in D$ **do**
  $classement \leftarrow tri\_decroissant(D - \{d\} \mid x \geq y \Leftrightarrow sim(d,x) \geq sim(d,y))$
  $positifs \leftarrow \{x \in D \mid classesDO(d) \cap classesDO(x) \neq \varnothing\}$
  $ROCAUC_{moyen} \leftarrow ROCAUC_{moyen} + ROCAUC(classement, positifs)/|D|$
end
return $ROCAUC_{moyen}$

Algorithm 3.1: Algorithme d'évaluation d'une mesure de similarité par le test des classes DO communes pour l'ensemble des maladies

Data: L'ensemble des maladies D, une classe de haut niveau DO $\mathcal{C}$, une mesure de similarité $sim : D \times D \to \mathbb{R}^+$
Result: ROCAUC moyen pour l'ensemble des maladies de la classe C

$D_{\mathcal{C}} \leftarrow$ les maladies de D de classe $\mathcal{C}$
$ROCAUC_{moyen} = 0$
$positifs \leftarrow \{x \in D \mid \mathcal{C} \in classesDO(x)\}$
for *chaque maladie* $d \in D_{\mathcal{C}}$ **do**
  $classement \leftarrow tri\_decroissant(D - \{d\} \mid x \geq y \Leftrightarrow sim(d,x) \geq sim(d,y))$
  $ROCAUC_{moyen} \leftarrow ROCAUC_{moyen} + ROCAUC(classement, positifs)/|D_{\mathcal{C}}|$
end
return $ROCAUC_{moyen}$

Algorithm 3.2: Algorithme d'évaluation d'une mesure de similarité par le test des classes DO communes pour une classe DO donnée

Evaluation de la mesure par les médicaments communs La seconde évaluation quantifie la capacité de la mesure de similarité à détecter des maladies ayant au moins un médicament indiqué en commun. On ordonne ici l'ensemble des paires de maladies en fonction de leur similarité phénotypique. On calcule ensuite la ROCAUC pour ce classement de paires de maladie sur la tâche de classification des paires de maladies ayant au moins un médicament indiqué en commun d'après la base de données SIDER. Cette classe positive est constituée de 16174 paires de maladies distinctes avec une indication en commun.

Soit D l'ensemble des 6220 maladies et sim une mesure de similarité à évaluer. Soit C_{sim} le classement des paires de maladies $(x,y) \in D \times D \mid x \neq y$ par rapport à leur similarité $sim(x,y)$. On définit l'ensemble des paires positives pour cette évaluation $P_{medicaments}$ comme $\{(x,y) \in D \times$

$D \mid \exists m$ un médicament indiqué pour x et $y\}$. On calcule alors $ROCAUC(C_{sim}, P_{medicaments})$, correspondant à la probabilité de mieux classer par la similarité *sim* une paire de maladies pour lesquelles il existe un médicament indiqué commun par rapport à une paire de maladies sans un tel médicament.

Evaluation de la mesure de similarité appliquée aux maladies génétiques On propose une évaluation supplémentaire par rapport au travail de [Hoehndorf et al., 2015] visant à quantifier la capacité de la mesure de similarité à associer des maladies génétiques partageant un gène responsable en commun. On limite cette évaluation à l'ensemble des maladies possédant au moins un gène responsable connu selon OMIM. On ordonne ensuite l'ensemble des paires de ces maladies en fonction de leur similarité phénotypique et on calcule la ROCAUC pour ce classement sur la tâche de classification avec comme classe positive l'ensemble des paires de maladies partageant un gène responsable d'après OMIM.

On définit l'ensemble des paires positives pour cette évaulation P_{genes} comme $\{(x, y) \in D \times D \mid \exists g$ un gène responsable de x et $y\}$. On calcule alors $ROCAUC(C_{sim}, P_{genes})$, correspondant à la probabilité de mieux classer par la similarité *sim* une paire de maladies pour lesquelles il existe un gène responsable par rapport à une paire de maladies sans gène commun.

Significativité des valeurs de ROCAUC La significativité des valeurs de ROCAUC obtenues dans l'ensemble de ces évaluations peut être calculée. Il existe une correspondance entre la ROCAUC calculée pour un classement et le test U de Mann-Whitney ou test de Wilcoxon-Mann-Whitney [Wilcoxon, 1945, Mann and Whitney, 1947] permettant de tester si la distribution de deux groupes de valeurs est identique. Dans les évaluations présentées précédemment, on cherche en effet à évaluer la différence de distribution des valeurs de similarités entre un groupe positif et un groupe négatif.

La statistique U correspondant à une valeur de ROCAUC est définie telle que :

$$U = ROCAUC * n_1 * n_2$$

où n_1 et n_2 sont les effectifs des deux groupes comparés, ici les groupes positif et négatif. On peut considérer que cette statistique U suit une distribution normale. Cette distribution est centrée sur $m_U = \frac{1}{2} n_1 n_2$. Similairement la distribution des ROCAUC est centrée sur $\frac{1}{2}$. L'écart-type de cette distribution est défini tel que :

$$\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

permettant ensuite de calculer le nombre d'écarts types à la moyenne d'une statistique U comme

$$z_U = \frac{U - m_U}{\sigma_U}.$$

La valeur de z_U obtenue permet par un test Z de calculer un seuil au dessus duquel la statistique U et la valeur de ROCAUC sont significatives. En l'occurence, chacune des ROCAUC présentées dans la section suivante sont significatifs pour $p < 10^{-5}$, c'est-à-dire que la probabilité qu'un classement aléatoire des maladies obtienne une valeur de ROCAUC supérieure ou égale à celle trouvée est inférieure à 10^{-5}.

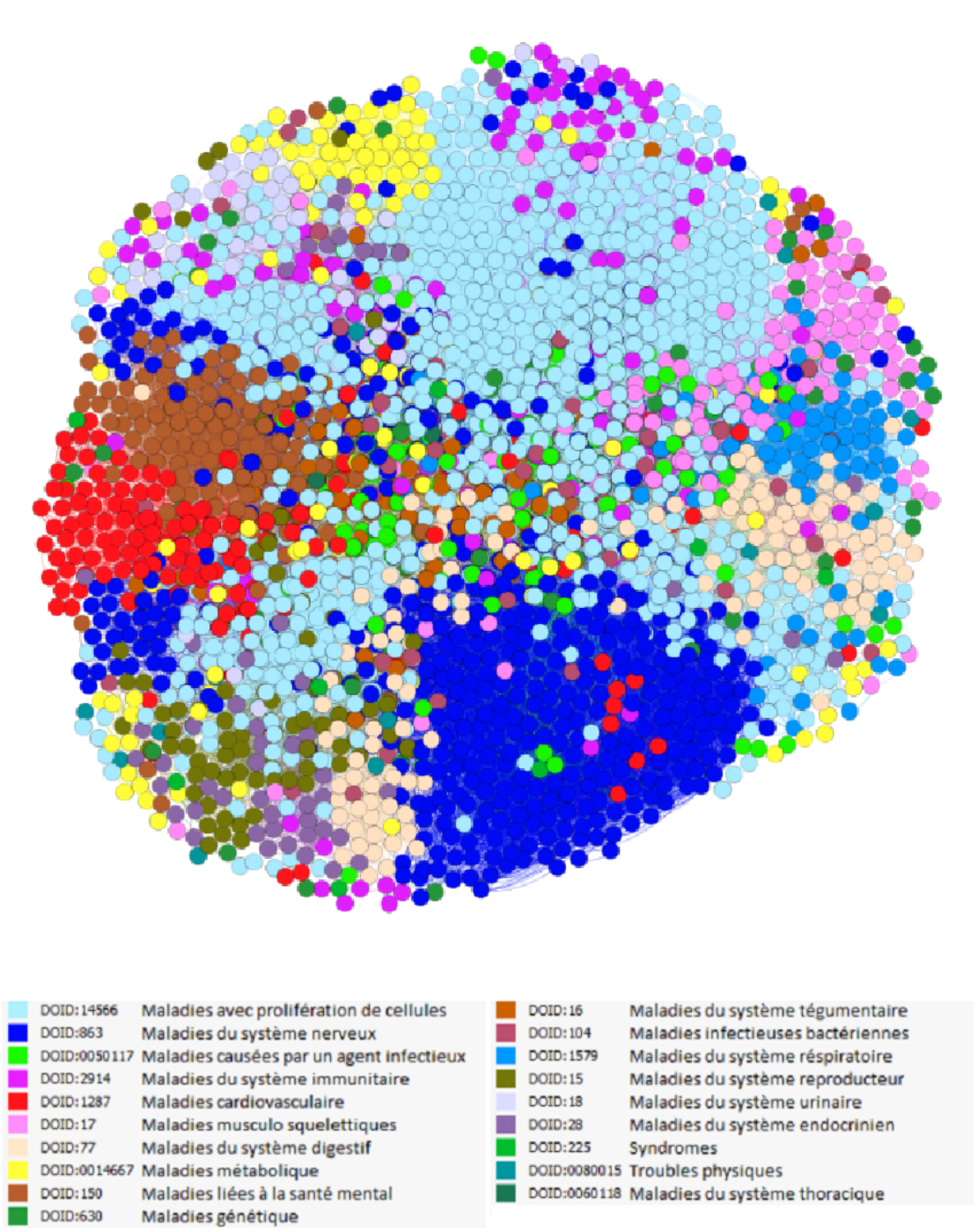


FIGURE 3.1 – Diseasome de 4773 maladies construit à partir de leur similarité phénotypique calculée avec la mesure IntelliGO. La visualisation du graphe est produite par le logiciel Gephi et l'algorithme Force Atlas 2 [Jacomy et al., 2014] qui regroupe spatialement les nœuds les plus connectés.

TABLE 3.1 – Evaluation de la classification des paires de maladies avec une classe DO commune pour les mesures de similarité SimGIC (par [Hoehndorf et al., 2015]) et IntelliGO. Pour chaque classe DO, la ROCAUC est calculée pour les deux mesures SimGIC et IntelliGO comme décrit dans l'Algorithme 3.2. Les résultats toutes maladies confondues présentés dans cette Table sont obtenus par la méthode de calcul décrite dans l'Algorithme 3.1.

Classe DO	SimGIC	IntelliGO	Différence
Toutes classes confondues	0.720	0.759	+0.039
Bacterial infectious disease (`DOID:104`, 168 maladies)	0.743	0.765	+0.022
Cardiovascular system disease (`DOID:1287`, 268 maladies)	0.720	0.745	+0.025
Disease by infectious agent (`DOID:0050117`, 393 maladies)	0.738	0.756	+0.018
Disease of cellular proliferation (`DOID:14566`, 1764 m.)	0.672	0.770	+0.098
Disease of mental health (`DOID:150`, 205 maladies)	0.783	0.875	+0.092
Disease of metabolism (`DOID:0014667`, 240 maladies)	0.733	0.784	+0.051
Endocrine system disease (`DOID:28`, 130 maladies)	0.711	0.761	+0.050
Fungal infectious disease (`DOID:1564`, 58 maladies)	0.864	0.857	−0.007
Gastrointestinal system disease (`DOID:77`, 281 maladies)	0.718	0.718	+0.000
Genetic disease (`DOID:630`, 171 maladies)	0.656	0.661	+0.005
Immune system disease (`DOID:2914`, 278 maladies)	0.730	0.761	+0.031
Integumentary system disease (`DOID:16`, 243 maladies)	0.743	0.826	+0.083
Musculoskeletal system disease (`DOID:17`, 314 maladies)	0.704	0.765	+0.061
Nervous system disease (`DOID:863`, 796 maladies)	0.712	0.649	−0.063
Parasitic infectious disease (`DOID:1398`, 103 maladies)	0.729	0.745	+0.016
Physical disorder (`DOID:0080015`, 57 maladies)	0.593	0.541	−0.052
Reproductive system disease (`DOID:15`, 585 maladies)	0.868	0.893	+0.025
Respiratory system disease (`DOID:1579`, 140 maladies)	0.868	0.890	+0.022
Syndrome (`DOID:225`, 47 maladies)	0.613	0.643	+0.030
Thoracic disease (`DOID:0060118`, 10 maladies)	0.659	0.757	+0.098
Urinary system disease (`DOID:18`, 149 maladies)	0.870	0.901	+0.031
Viral infectious disease (`DOID:934`, 97 maladies)	0.745	0.777	+0.032

3.3 Diseasome et résultats

Les similarités phénotypiques calculées par IntelliGO permettent la construction d'un diseasome des 6220 maladies du jeu de données de [Hoehndorf et al., 2015]. Le diseasome est obtenu en conservant comme arcs les 0.5% meilleures similarités et est présenté en Figure 3.1. Ce graphe possède une composante connexe de 4773 maladies liées par 70923 arcs.

3.3.1 Diseasome des 6220 maladies

On peut dans un premier temps observer une certaine homogénéité des classes de maladies parmi plusieurs zones du diseasome. On observe néanmoins que certaines classes de maladies sont divisées et réparties dans plusieurs parties du diseasome. C'est notamment le cas des cancers (`DOID:14566` en cyan sur la figure) : le diseasome confime l'hétérogénénité de cette classe de maladies.

On évalue la pertinence du regroupement des maladies observé dans ce diseasome en calculant la ROCAUC représentant la probabilité qu'une paire de maladies de même classe de haut niveau

dans DO possède une similarité plus élevée qu'une paire de maladie de classes DO différentes. Cette évaluation est conduite séparément pour chaque maladie, c'est-à-dire restreinte aux paires contenant cette maladie. Les résultats de ces évaluations, toutes classes confondues ou aggrégées par classe DO de haut niveau, sont présentés dans la Table 3.1.

Les résultats obtenus par l'utilisation de la similarité sémantique IntelliGO comparés à ceux de la mesure SimGIC, utilisée par [Hoehndorf et al., 2015], montrent que IntelliGO permet globalement de mieux grouper les maladies de même classe dans le diseasome. Sur l'ensemble des maladies, IntelliGO obtient une ROCAUC moyenne de 0.759 contre 0.720 pour SimGIC. On constate également que IntelliGO a une meilleure performance pour la majorité des classes de haut niveau de DO évaluées séparément (19 classes sur 22).

Diseasome pour la découverte d'indications médicamenteuses On évalue ensuite la qualité de prédiction du diseasome pour l'identification d'éventuels médicaments communs à plusieurs maladies. Pour cela, on calcule le ROCAUC représentant la probabilité que deux maladies pour lesquelles un même médicament est indiqué aient une plus forte similarité que deux maladies sans indication commune. On a 931 maladies dans le jeu de données qui sont associées à au moins un médicament indiqué d'après la base de données SIDER, pour 16174 paires de maladies ayant un médicament associé commun.

Ici, IntelliGO obtient une valeur de ROCAUC de 0.574 contre 0.648 pour SimGIC, comme présenté en Figure 3.2. Cependant ce test est effectué, comme dans les expériences de [Hoehndorf et al., 2015], sur l'ensemble des maladies, y compris celles qui ne sont associées à aucun médicament. En faisant le test uniquement sur les maladies associées à au moins un médicament dans SIDER, on obtient une valeur de ROCAUC de 0.593 pour IntelliGO et de 0.617 pour SimGIC, comme présenté en Figure 3.3. Ces résultats montrent qu'en se limitant aux maladies associées à au moins un médicament, on augmente l'efficacité de la classification de la mesure IntelliGO ($+0.019$) tout en diminuant celle de SimGIC (-0.031). Ceci suggère qu'il existe des biais différents dans ces deux méthodes de mesure de la similarité associés à la présence ou non de maladies sans traitement connu.

Pour essayer de comprendre l'origine de ce biais, on établit par une méthode similaire de calcul de la ROCAUC que SimGIC a une probabilité de 0.513 de mieux classer une paire de maladie si au moins une des maladies a au moins un médicament associé. Cette probabilité est de 0.545 lorsque les deux maladies ont au moins un médicament indiqué (commun ou non). On observe un effet inverse pour IntelliGO, puisque cette probabilité est de 0.478 quand une maladie a un médicament indiqué et de 0.482 quand les deux maladies ont un médicament indiqué. Effectuer le test restreint à l'ensemble de maladie ayant au moins un traitement dans SIDER atténue ce biais pour les deux mesures de similarité, puisque SimGIC qui favorisait dans le classement les maladies avec des médicaments voit son score diminuer de 0.031 tandis que IntelliGO qui avait le biais inverse voit son score augmenter de 0.019.

Diseasome pour la découverte de gènes responsables On complémente le processus d'évaluation du diseasome proposé par [Hoehndorf et al., 2015] en s'intéressant cette fois aux gènes responsables de maladies génétiques. Pour cela, on calcule le ROCAUC représentant la probabilité que deux maladies avec un gène responsable en commun aient une plus forte similarité que deux maladies sans gène commun. On a 651 maladies dans le jeu de données qui possèdent au moins un gène responsable d'après OMIM, pour 587 paires de maladies partageant une gène.

On obtient dans cette évaluation une ROCAUC de 0.771 pour IntelliGO et de 0.730 pour SimGIC, comme présenté en Figure 3.4.

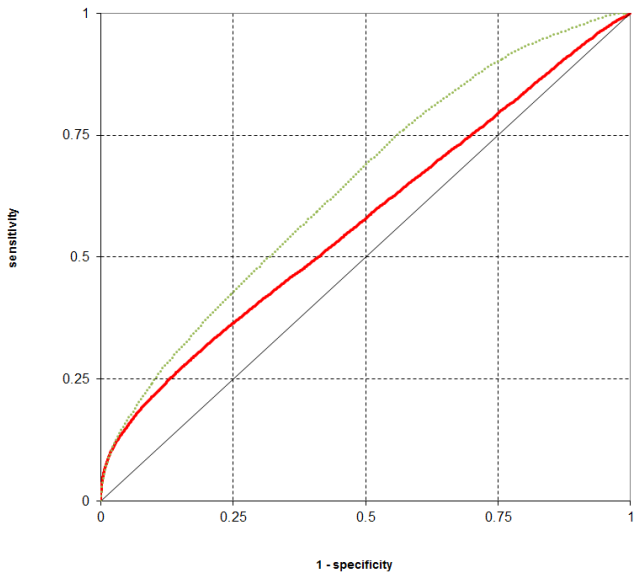


FIGURE 3.2 – Courbe de ROC représentant les performances de IntelliGO (trait plein rouge, ROCAUC = 0.574) et SimGIC (trait pointillé vert, ROCAUC = 0.648) sur la tâche de classification des maladies ayant au moins un médicament en commun, en utilisant l'ontologie MonDO

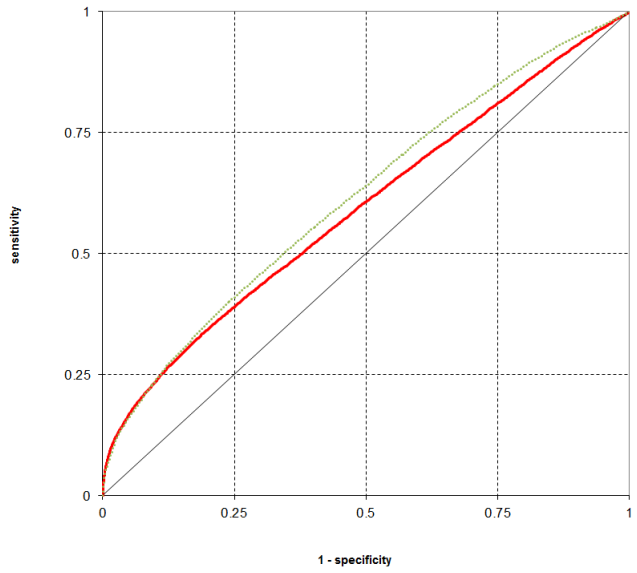


FIGURE 3.3 – Courbe de ROC représentant les performances de IntelliGO (trait plein rouge, ROCAUC = 0.593) et SimGIC (trait pointillé vert, ROCAUC = 0.617) sur la tâche de classification des maladies ayant au moins un médicament en commun, restreinte aux maladies ayant au moins une association à un médicament, en utilisant l'ontologie MonDO

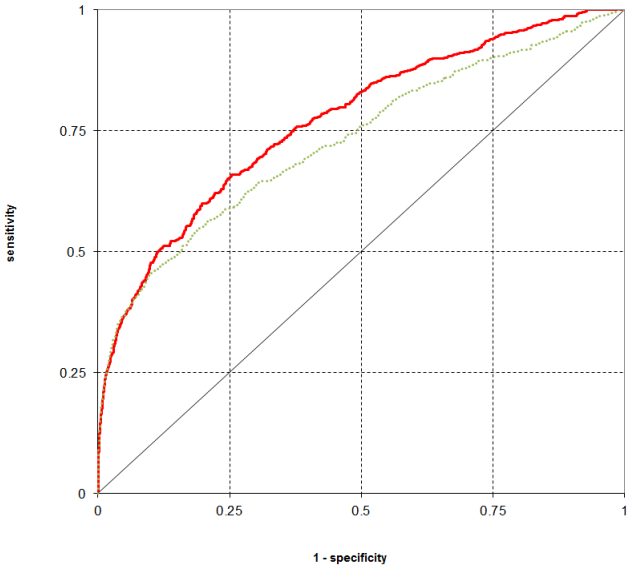


FIGURE 3.4 – Courbe de ROC représentant la performance IntelliGO (trait plein rouge, ROCAUC = 0.771) et SimGIC (trait pointillé vert, ROCAUC = 0.730) sur la tâche de classification des maladies partageant un gène responsable, en utilisant l'ontologie MonDO

3.3.2 Caractérisation de déficiences intellectuelles dans le diseasome

On souhaite s'intéresser à la place qu'occupent les déficiences intellectuelles dans le diseasome obtenu avec IntelliGO. La Figure 3.5 présente le sous-ensemble du diseasome des 6220 maladies, réduit aux déficiences intellectuelles.

On observe trois composantes connexes dans ce graphe où sont surreprésentées certaines classes DO. La composante numérotée 1 comporte 13 maladies dont 12 classées comme métaboliques dans DO. La composante 2 comporte 9 maladies dont 7 classées comme maladies du système nerveux. Enfin, la composante 3 comporte 26 maladies dont 13 classées comme génétiques dans DO.

Les évaluations précédentes montrent que IntelliGO est une mesure de similarité sémantique appropriée pour comparer des maladies sur la base de leur phénotypes. IntelliGO est notamment capable d'identifier certaines maladies partageant un gène responsable, ainsi, on propose d'utiliser IntelliGO pour classer l'ensemble hétérogène des déficiences intellectuelles dans plusieurs classes distinctes. Bien que le sous-ensemble du diseasome obtenu soit cohérent avec la classification extraite de DO, cette classification semble trop généraliste pour étudier plus en détail les différents types de déficiences intellectuelles. Il semble alors nécessaire de recourir à une classification experte des déficiences intellectuelles, davantage spécialisée qu'une ontologie de maladies généraliste. Cette classification, réalisée en collaboration avec un expert du domaine est présentée en Annexe A. Elle concerne 374 déficiences intellectuelles de la liste établie par [Gilissen et al., 2014], et classées en 5 catégories non disjointes : *(i)* maladies métaboliques, *(ii)* maladies de la neurogenèse, *(iii)* régulation, *(iv)* régulation de l'expression génétique, *(v)* maladies synaptiques. On étudie alors la capacité de classification sur ces 5 classes de déficiences intellectuelles de la méthode utilisant des similarités sémantiques pour comparer des maladies présentée précedemment.

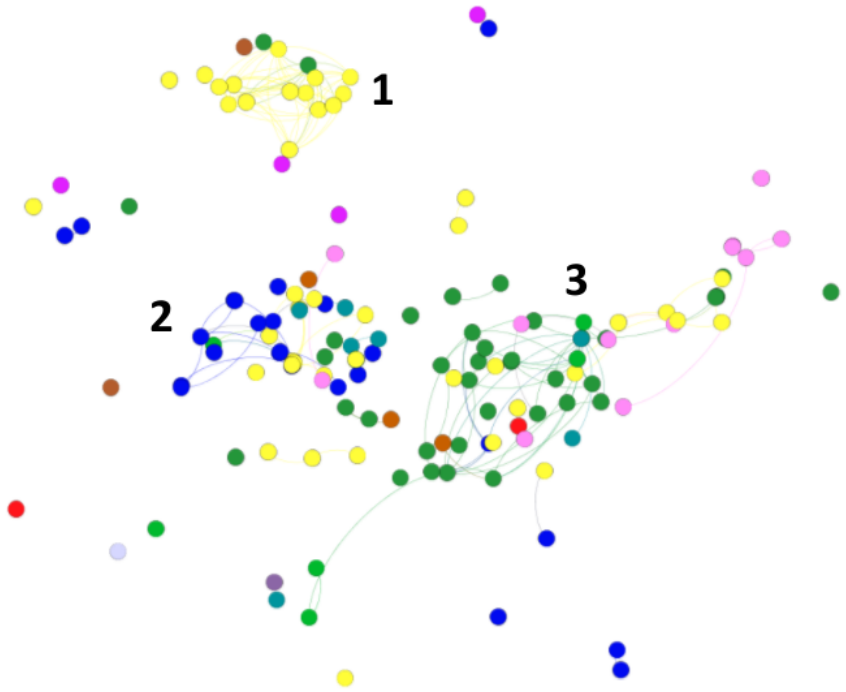


FIGURE 3.5 – Diseasome des déficiences intellectuelles extrait du diseasome des 6220 maladies, avec trois composantes connexes correspondant à trois classes DO de haut niveau : 1. maladies métaboliques (en jaune), 2. maladies du système nerveux (en bleu), 3. maladies génétiques (en vert)

3.3.3 Application à la caractérisation de 5 classes de déficiences intellectuelles

On applique la méthode décrite dans ce chapitre au jeu de données des 5 catégories de déficiences intellectuelles, disponible en Annexe A. Pour cet ensemble de maladies on dispose d'annotations de phénotypes exprimées sous forme de classes Human Phenotype Ontology (HPO). Puisque l'ensemble des maladies étudiées ici sont génétiques, on pourra également utiliser, en plus des phénotypes associés aux maladies, les annotations Gene Ontology (GO) de leur gènes responsables. Ces annotations sont réparties dans les trois aspects de GO : *Biological Component* (BP), *Cellular Component* (CC), *Molecular Function* (MF). On considérera chacun des aspects de GO comme une ontologie indépendante : en effet il n'existe aucun lien de subsomption entre ces trois aspects de GO, ce qui rend impossible la comparaison de termes GO appartenant à deux aspects différents.

On s'intéressera dans un premier temps aux résultats obtenus en utilisant chacune des quatre ontologies séparément, c'est-à-dire en calculant la similarité sémantique de deux maladies à partir des annotations d'une seule ontologie. On propose ensuite d'aggréger les similarités calculées sur les différentes ontologies pour obtenir une mesure de similarité sémantique qui considère plusieurs ontologies. Cette aggrégation est effectuée en calculant la moyenne des similarités calculées à partir de chaque ontologie.

On propose alors une mesure de similarité entre les déficiences intellectuelles pour chaque combinaison possible des 4 ontologies BP, CC, MF et HPO, et pour les deux mesures de similarité sémantique IntelliGO et SimGIC. On obtient alors 30 mesures de similarité entre les déficiences intellectuelles (15 à partir d'IntelliGO et 15 à partir de SimGIC, pour chaque sous-ensemble d'ontologies non vide de {BP, CC, MF, HPO}). Ces mesures sont ensuite évaluées en calculant une ROCAUC moyenne pour chacune des 5 classes de déficiences intellectuelles, comme décrit précedemment par l'Algorithme 3.2, puis de manière synthétique sur l'ensemble de ces 5 classes par la méthode décrite par l'Algorithme 3.1.

Résultats La Table 3.2 présente une partie des résultats de ces évaluations pour des similarités calculées à partir d'une sélection de combinaisons d'ontologies. Ces résultats montrent une performance variable des mesures de similarité IntelliGO et SimGIC en fonction des ontologies utilisées et des classes de maladies sur lesquelles sont effectuées les évaluations. En particulier, on peut constater qu'ajouter HPO aux ontologies prises en compte par une mesure de similarité a généralement un effet négatif ou nul sur la performance dans la plupart des tâches de classification, à l'exception de la tâche de classification de la classe Neurogenèse. On note également que les mesures utilisant uniquement HPO ont des performances globales très inférieures aux autres mesures utilisant seulement BP, CC ou MF. On peut alors supposer que, pour les quatre classes autres que Neurogenèse, HPO n'apporte pas plus d'information que de bruit.

A l'inverse, combiner deux ou trois des aspects de GO ensemble a un effet positif sur la performance globale des mesures de similarité. Cet effet est le plus visible pour la classe Régulation de l'Expression Génétique : la performance d'IntelliGO passe de 0.740 en utilisant seulement MF à 0.803 en utilisant les trois aspects de GO, tandis que celle de SimGIC passe de 0.905 en utilisant seulement BP à 0.936.

Si la combinaison des trois aspects de GO offre la meilleure performance globale, on peut néanmoins observer que ce n'est pas nécessairement le cas pour la classification pour une classe prise isolément :

— La classe Régulation est globalement mal prédite. En particulier, la performance sur cette classe est proche du hasard lorsque l'on utilise HPO seule.

— La classe Régulation de l'Expression Génétique est mieux prédite par SimGIC (0.936)

TABLE 3.2 – ROCAUC obtenues dans l'évaluation de SimGIC et IntelliGO sur la tâche de classification des maladies en 5 classes de déficiences intellectuelles (Régulation, Régulation de l'Expression Génétique, Métabolique, Synaptique, Neurogenèse, décrites en Annexe A) utilisant différentes ontologies (les trois aspects de GO : BP, CC, MF et l'ontologie de phénotypes HPO). Pour chaque classe de déficiences intellectuelles, la ROCAUC est calculée pour les deux mesures SimGIC et IntelliGO comme décrit dans l'Algorithme 3.2. Les résultats toutes classes confondues présentés dans cette Table sont obtenus par la méthode de calcul décrite dans l'Algorithme 3.1. Chaque cellule de la Table donne les résultats pour une tâche de classification donnée et un ensemble d'ontologies considérées pour les deux mesures IntelliGO (nombre au dessus) et SimGIC (nombre en dessous).
Les valeurs en gras dénotent dans chaque colonne un résultat maximum pour une tâche de classification donnée, pour l'une des deux similarités utilisées.

Classes	Rég.	Rég. EG	Métab.	Synap.	Neuro.	Toutes
Effectif	154	70	105	31	33	374
Seuil$_{(p<0.01)}$	0.5071	0.5135	0.5105	0.5331	0.5310	0.5060
BP	0.509	0.681	0.838	0.711	0.548	0.652
	0.523	0.905	0.785	0.758	0.568	0.691
CC	**0.532**	0.7186	0.681	0.723	0.563	0.627
	0.515	0.846	0.659	0.787	0.561	0.641
MF	0.526	0.740	0.749	0.628	0.610	0.642
	0.507	0.830	0.735	0.588	0.560	0.641
HPO	0.481	0.584	0.608	0.530	0.590	0.548
	0.502	0.574	0.574	0.559	0.594	0.548
3GO	0.526	**0.803**	**0.862**	0.732	0.595	**0.693**
	0.527	**0.936**	0.764	0.7911	0.581	**0.695**
HPO, 3GO	0.520	0.787	0.854	0.720	0.625	0.687
	0.524	0.933	0.748	0.768	0.628	0.691
BP, CC	0.512	0.750	0.844	**0.772**	0.547	0.673
	0.525	0.927	0.750	**0.819**	0.578	0.690
BP, MF	0.523	0.768	0.856	0.690	0.611	0.681
	0.522	0.916	**0.785**	0.711	0.571	0.688
CC, HPO	0.499	0.684	0.691	0.677	0.582	0.608
	0.514	0.817	0.662	0.757	**0.633**	0.640
MF, HPO	0.515	0.732	0.743	0.620	**0.657**	0.638
	0.513	0.813	0.643	0.625	0.616	0.622
Maximum	**0.532**	**0.803**	**0.862**	**0.772**	**0.657**	**0.693**
	0.527	**0.936**	**0.785**	**0.819**	**0.633**	**0.695**

en utilisant les trois aspects de GO. Dans ce cas, ajouter HPO a un effet négligeable. Le meilleur résultat pour IntelliGO (0.803) est également obtenu en utilisant les trois aspects de GO.

— La classe Métabolique est mieux prédite par IntelliGO (0.862) en utilisant à nouveau les trois aspects de GO. En revanche, SimGIC obtient sa meilleure performance pour cette classe en considérant seulement BP et MF.

— La classe Synaptique est mieux prédite lorsque l'on utilise uniquement les deux aspects BP et CC pour les deux mesures SimGIC (0.819) et IntelliGO (0.772).

— La classe Neurogenèse est la seule classe pour laquelle on obtient de meilleurs résultats en utilisant HPO avec un ou plusieurs aspects de GO. IntelliGO obtient un score de 0.657 en utilisant MF et HPO (pour 0.610 avec MF seul) tandis que SimGIC obtient un score de 0.633 en utilisant CC et HPO (pour 0.561 avec CC seul).

Ces résultats viennent confirmer l'hététogénéité des déficiences intellectuelles et les difficultés inhérentes à leur classification. On a dans un premier temps pu observer plusieurs groupes de ces maladies dans le diseasome phénotypique ; ici, on voit que selon la classe de déficiences intellectuelles, les ontologies GO ou HPO sont plus ou moins pertinentes pour la classification.

3.4 Discussion

Une première partie des travaux présentés dans ce Chapitre s'intéresse à la construction d'un diseasome phénotypique à l'aide d'une mesure de similarité sémantique. On a observé que la mesure de similarité sémantique IntelliGO, créée pour la comparaison de gènes annotés par des termes issus de l'ontologie GO, pouvait s'adapter à la comparaison de maladies munies d'annotations phénotypiques. En particulier, cette mesure obtient une performance supérieure à celle de l'état de l'art utilisée dans [Hoehndorf et al., 2015] pour la tâches de classification des classes DO. On introduit également un nouveau critère d'évaluation par rapport à ceux de l'état de l'art : une tâche de classification dont le but est d'identifier des paires de maladies partageant au moins un gène responsable, sur laquelle IntelliGO obtient aussi des performance supérieures. L'étude des déficiences intellectuelles au sein du diseasome phénotypique construit par IntelliGO révèle que mis à part un ou deux groupes assez homogènes, ces maladies sont très dispersées au sein de l'ensemble de plus de 6000 maladies étudié.

Ce constat nous a mené à recourir à une classification experte de ces déficiences intellectuelles en 5 classes. Dans une seconde partie du travail, on peut alors évaluer les mesures de similarités IntelliGO et SimGIC sur les tâches de classification de ces 5 classes, en utilisant différentes combinaisons d'ontologies. Les annotations de phénotype issues de l'ontologie HPO se révèlent insuffisantes pour résoudre ce problème, tandis que les annotations GO permettent d'obtenir de bonnes performances de classification. La combinaison de plusieurs des aspects de GO offre un effet positif sur la performance d'IntelliGO et de SimGIC.

On note néanmoins que la pertinence de l'utilisation d'une ontologie dépend fortement de la classe de maladies à prédire, et que prendre en considération un grand nomre d'ontologies d'ontologies peut parfois avoir un effet négatif sur la performance de classification. D'autres modes d'aggrégation de différentes mesures de similarités seraient à considér.

3.4.1 Limites et perspectives

Une limite de l'utilisation de mesures de similarité sémantique est que ces mesures considérent les annotations même lorsqu'elles ne sont pas pertinentes au problème. Cette approche requiert donc de sélectionner manuellement les données et ontologies à fournir à la mesure, ou d'itérer

empiriquement sur des sous-ensembles de ces données, à condition de disposer d'une verité terrain. De plus, l'ensemble des déficiences intellectuelles peuvent être étudiées à différents niveaux de granularité arbitrairement définis, comme par exemple via leur classe DO ou la classification experte que nous avons utlisée.

Les mesures de similarité sémantiques permettent la classification de certaines de ces classes, en revanche, elles ne permettent pas de justifier cette classification auprès d'un expert. Si les ontologies du web sémantiques sont destinées à être compréhensibles autant par les humains que par les machines, le traitement qui en est fait par les différentes mesures de similarité synthétise les connaissances sur les maladies en des données numériques non-exploitables par un utilisateur humain.

Le Chapitre 4 propose une méthodologie pour la caractérisation de gènes responsables de déficiences intellectuelles utilisant une méthode d'apprentissage supervisée : la Programmation Logique Inductive (PLI). On a constaté que l'apport des annotations phénotypiques était peu discriminant pour la différentiation des déficiences intellectuelles, mais que les différents aspects de GO avaient un apport positif sur la performance de classification : on s'intéressera donc à ces maladies au niveau de leur gènes responsables plutôt que de leurs phénotypes. La PLI permet de considérer les différents aspects de GO séparément ou en conjonction pour former des règles en logique du premier ordre caractérisant des sous-ensembles de gènes responsables. Ici, un des avantages de la PLI est sa capacité à sélectionner par induction les données pertinentes à sa tâche de classification : ainsi, il n'y a pas d'inconvénient à inclure des données non-discriminantes, autre que l'augmentation du temps d'exécution de l'algorithme. Les règles produites permettent la classification de gènes responsables, tout en caractérisant ces gènes de manière intelligible pour un expert.

3.4.2 Conclusion

Les méthodes utilisant des mesures de similarité sémantiques peuvent se montrer performantes dans de nombreuses tâches de classification. L'utilisation conjointe de plusieurs ontologies de domaine se révèle pertinente dans certains cas, mais nécessite une sélection préalable pour ne pas introduire de bruit dans la mesure de similarité. Si ces mesures sont capables de tenir compte de la sémantique de la hiérarchie de concepts des ontologies en entrée, elles ne proposent pas de résultats pouvant être analysés qualitativement par un expert du domaine.

On notera que les mesures de similarité IntelliGO et SimGIC obtiennent des résultats très différents sur les tâches de classification presentées ici, sans que l'une ne se montre constamment plus perfomante que l'autre. On aura d'ailleurs noté différents biais de ces mesures sur la tâche de classification des maladies partageant une indication médicamenteuse. En effet, les méthodes de calcul employées par ces deux mesures sont très différentes, ce qui pourraient expliquer ces résultats.

Néanmoins, les mesures de similarité sémantique semblent pouvoir proposer une représentation des maladies sous la forme d'un diseasome en concordance avec une classification humaine grâce aux ontologies de domaine. Ce diseasome permet d'explorer les interactions entre gènes, maladies et médicaments, notamment en permettant de prédire l'existence de gènes communs entre maladies. Le diseasome ainsi obtenu représente donc à sa façon la synthèse d'un grand ensemble de connaissances sur les maladies humaines, qui permet d'en avoir une vue d'ensemble. Afin de pouvoir étudier en particulier les déficiences intellectuelles, on présentera dans le Chapitre 4 une méthode utilisant les ontologies biomédicales pour caractériser plus en détails les gènes responsables de ces maladies.

4

Prise en compte des ontologies dans la Programmation Logique Inductive appliquée aux Données Ouvertes Liées

Sommaire

4.1 Problème biomédical

Le Chapitre 3 présente une approche fondée sur la similarité sémantique permettant de distinguer différentes classes de maladies. En s'intéressant en particulier aux déficiences intellectuelles, qui sont des maladies génétiques, on a observé que les annotations des gènes responsables sont importantes pour la caractérisation de ces maladies. En revanche, les annotations phénotypiques apportent peu d'information pour distinguer différents sous-groupes dans ces maladies.

Il nous semble alors pertinent de s'intéresser à ces maladies par l'intermédiaire des données disponibles sur les gènes qui leur sont associés. On propose ici une méthode utilisant des Données Ouvertes Liées (LOD) et une ontologie biomédicale : Gene Ontology (GO), pour la caractérisation

et la classification de gènes responsables pour des déficiences intellectuelles. On souhaite pour cela tirer partie des différentes intiatives récentes visant à rendre disponibles sous forme de LOD de nombreuses sources de données biologiques [Dumontier et al., 2014, Whetzel et al., 2011].

L'approche présentée ici propose, à partir d'une modélisation des données sur les gènes, de sélectionner, intégrer et fouiller des LOD, afin de proposer un modèle décrivant les gènes responsables de déficiences intellectuelles. Cette approche qui se veut explicite veut faciliter les échanges et la prise en considération des connaissances d'experts du domaine biomédical, tant sur la sélection des données que sur l'interprétation des résultats. On utilisera ici la Programmation Logique Inductive (PLI) pour produire un modèle des gènes responsables : en effet, la PLI permet d'apprendre un modèle caractérisant l'ensemble de ces objets, appelé théorie en PLI. Ce modèle est constitué de règles qui peuvent ensuite être évaluées d'une part comme tout modèle de prédiction ou éxaminées par un expert. On évaluera en particulier la contribution d'une part des LOD, d'autre part d'une ontologie de domaine, à la qualité de la caractérisation et de la prédiction du modèle.

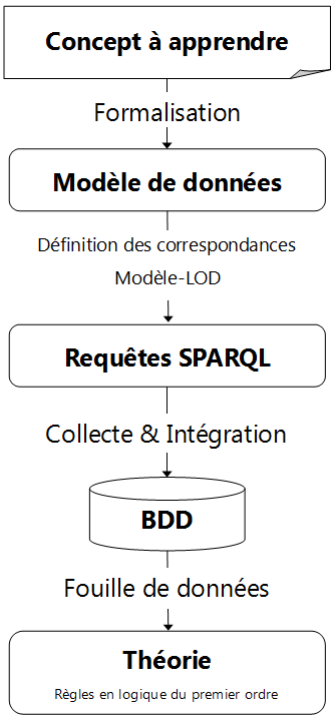


FIGURE 4.1 – Etapes de la méthode d'intégration et de fouille de LOD décrite dans ce Chapitre. On part d'un concept à apprendre : ici, on cherche à décrire les gènes responsables de déficiences intellectuelles. On formalise ensuite des connaissances expertes sur les données du problème pour établir un modèle de données à partir duquel on pourra effectuer la fouille. Ce modèle est alors mis en correspondance avec les entités et propriétés RDF des LOD, ce qui permet de construire les requêtes SPARQL permettant de collecter les données nécessaires à la fouille en instanciant le modèle établi. Ces données sont matéralisées dans une base de données de triplets qui seront utilisés dans le processus de fouille pour la production d'une théorie décrivant les gènes responsables de déficiences intellectuelles.

On propose ici une méthode d'intégration de données extraites des LOD en vue de permettre la fouille de données. Cette méthode est appliquée au problème biomédical de caractérisation des gènes responsables de déficiences intellectuelles et se compose de plusieurs étapes, comme illustré en Figure 4.1 :

— Les données du problème doivent d'abord être formalisées. Cette étape nécessite des connaissances expertes sur le domaine. Notamment, ce modèle peut être calqué sur les ontologies de domaine, ce qui facilitera l'intégration des données liées à ces ontologies. On choisit d'utiliser le modèle Entité-Association où les entités et associations reflètent respectivement les entités et propriétés RDF.
— Chaque entité et association du modèle est ensuite mise en correspondance avec les données des LOD d'un ou plusieurs jeux de données, en en donnant une définition exprimée en utilisant les éléments des LOD. Ces définitions permettront de construire des requêtes SPARQL pour la collecte des données et l'instanciation du modèle de données.
— Les données sont ensuite collectées en utilisant ces requêtes SPARQL. Durant cet étapes, les instances d'une même entité qui décrivent le même objet dans plusieurs jeux de données doivent être mis en correspondance fusionnées si nécessaire. Si possible, cette étape utilise les liens entre les différents jeux de données déjà présents dans les LOD.
— Finalement, on utilise la PLI pour générer à partir d'un ensemble d'exemples positifs et un ensemble d'exemples négatifs une théorie caractérisant les exemples positifs. Dans notre cas d'étude, la PLI nous permet de proposer une théorie caractérisant les gènes responsables de déficiences intellectuelles.

4.2 Intégration de données extraites des LOD

4.2.1 Modèle de données

Dans notre approche d'intégration de données extraites des LOD, la première étape est d'établir un modèle Entité-Association (EA) décrivant les entités à considérer pour une étude particulière. Le rôle de ce modèle est de présenter un modèle abstrait des données pertinentes à fouiller. Cette étape est réalisée avec un expert du domaine, sans nécessairement requérir de connaissances sur les données présentes dans les LOD ou leur structure. Il peut néanmoins s'inspirer de la structure des ontologies de domaine. Bien qu'un modèle EA est généralement constitué d'entités, d'associations et d'attributs, dans le cas présent seules les entités et les associations sont utilisées dans le modèle. Les associations n-aires ou associations comportant des attributs sont représentées par composition d'associations binaires suivant le mécanisme de réification. La Figure 4.2 présente le modéle EA défini pour notre étude sur les gènes responsables de déficiences intellectuelles.

4.2.2 Définition de correspondances avec les LOD

L'intégration des données extraites des LOD consiste dans un premier temps à établir des correspondances entre notre modèle EA et des types d'entités et de propriétés dans les LOD. Ces correspondances sont à définir entre chaque entité du modèle et un ou plusieurs types RDF des LOD ; et entre chaque association du modèle avec une propriété RDF des LOD. En effet, chaque jeu de données présent dans les LOD est susceptible d'utiliser un type distinct pour décrire une seule entité dans notre modèle. Par exemple, l'entité `Gene` du modèle correspond à deux types dans les LOD : le type `<http://bio2rdf.org/geneid_vocabulary:Gene>` et le type

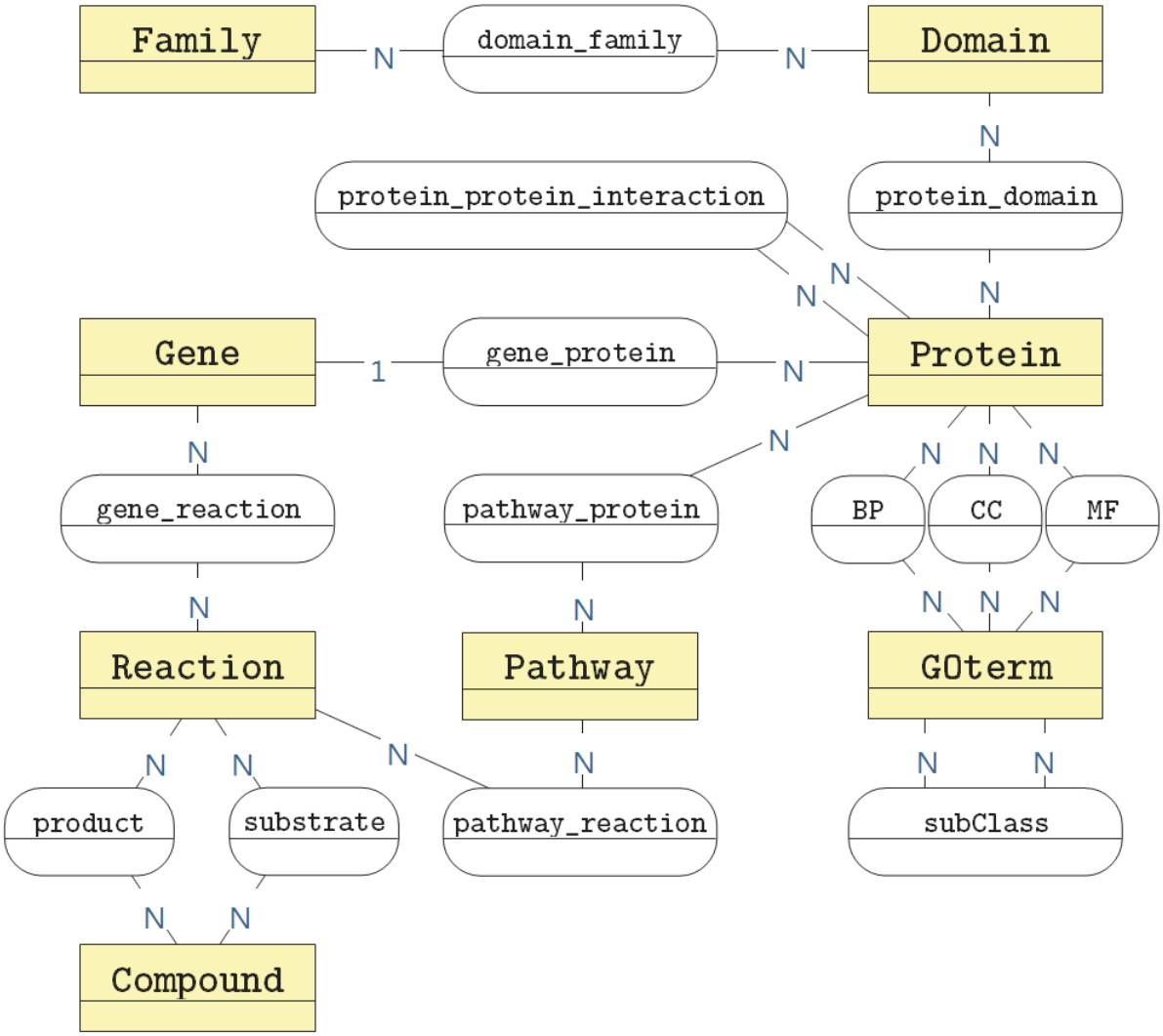


FIGURE 4.2 – Modèle Entité-Association (EA) des données sur les gènes responsables de déficiences intellectuelles. `BP`, `CC` et `MF` représentent les 3 types d'annotations GO, respectivement : *Biological Process*, *Cellular Component* et *Molecular Function*. Cette figure omet les données sur la localisation des gènes dans les chromosomes, qui est composée de 5 entités associées à `Gene` : `Chromosome`, `Chromosome_Arm`, `Chromosome_Region`, `Chromosome_Band`, `Chromosome_SubBand`, `Chromosome_SubSubBand`.

<http://bio2rdf.org/kegg_vocabulary:Gene>[1] utilisés respectivement dans deux jeux de données de Bio2RDF : NCBI Gene et KEGG.

Chaque entité est ensuite définie plus finement par une définition de concept pouvant être un type d'entité RDF, le domaine ou codomaine d'une propriété, la négation, l'union ou l'intersection d'une telle définition de concept. Cette définition devra pouvoir s'exprimer sous la forme d'une requête SPARQL. De manière similaire, les associations du modèle EA auront une définition les mettant en correspondance avec une ou plusieurs propriétés, compositions de propriétés (notée $\mathtt{p}_1 \circ \mathtt{p}_2$) ou propriétés inverses (notée $\mathtt{p}^-$). Par exemple, la relation `gene_reaction` entre un gène et une réaction (qui représente le fait qu'un gène produit une enzyme qui catalyse la réaction ou une protéine impliquée dans la réaction) peut être mise en correspondance avec $\mathtt{kegg:xGene}^- \circ \mathtt{kegg:xEnzyme}^-$ et $\mathtt{kegg:xGene}^- \circ \mathtt{kegg:xReaction}$, comme illustré en Figure 4.3. La requête SPARQL correspondant à la définition de le relation `gene_reaction` est quant à elle présentée en Figure 4.4. Les Tables 4.1 et 4.3 listent les entités et association de notre modèle EA ainsi que les jeux de données auxquels ils sont mis en correspondance.

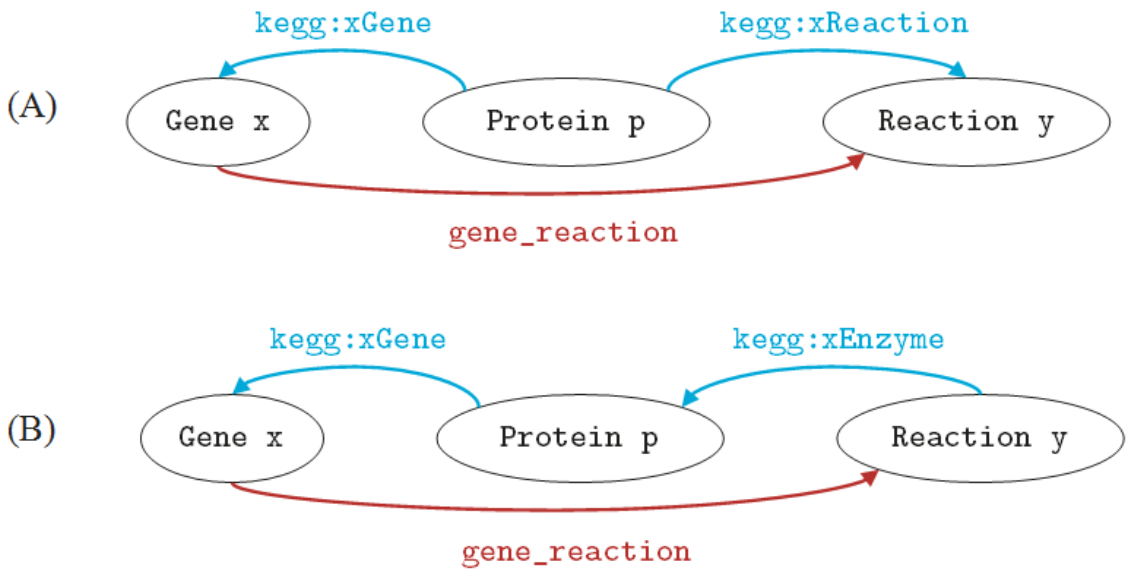


FIGURE 4.3 – Représentation des associations du modèle EA (en rouge) et des propriétés des LOD (en bleu) liant les entités `Gene` et `Reaction`. La propriété `kegg:xGene` lie une protéine au gène qui la produit, `kegg:xEnzyme` lie une réaction à ses enzymes et `kegg:xReaction` lie les protéines aux réactions dans lesquelles elles sont impliquées. L'association `gene_reaction`, définie dans notre modèle EA entre les entités `Gene` et `Reaction`, est ici mise en correspondance avec les compositions de propriétés : $\mathtt{kegg:xGene}^- \circ \mathtt{kegg:xReaction}$ et $\mathtt{kegg:xGene}^- \circ \mathtt{kegg:xEnzyme}^-$.

4.2.3 Sélection des exemples positifs et négatifs

Afin de pouvoir produire un concept des gènes responsables de déficiences intellectuelles via l'ILP, il est nécessaire de disposer de données sur un ensemble d'exemples positifs et un

1. On utilisera ensuite les préfixes `geneid:` et `kegg:` pour désigner respectivement les espaces de noms `http://bio2rdf.org/geneid_vocabulary:` et `http://bio2rdf.org/kegg_vocabulary:`. On écrira par exemple `kegg:Gene` pour désigner `<http://bio2rdf.org/kegg_vocabulary:Gene>`.

TABLE 4.1 – Définitions, sources de données (URL de l'endpoint SPARQL) et nombre d'individus pour chaque entité de notre modèle EA. Les entités marquées du symbole * sont définies uniquement par le domaine ou codomaine de leur associations.

Entité	Définitions et sources	Instances
Gene	geneid:Gene,kegg:Gene **Sources :** cu.gene.bio2rdf.org/sparql cu.kegg.bio2rdf.org/sparql	549
Protein*	**Source :** beta.sparql.uniprot.org/sparql	1257
Pathway	biopax3:Pathway **Sources :** cu.kegg.bio2rdf.org/sparql www.ebi.ac.uk/rdf/services/reactome/sparql	580
Reaction*	**Source :** cu.kegg.bio2rdf.org/sparql	433
Compound*	**Source :** cu.kegg.bio2rdf.org/sparql	628
GOterm	owl:Class **Sources :** cu.goa.bio2rdf.org/sparql sparql.bioontology.org/sparql	7770
Domain	interpro:Domain **Source :** cu.interpro.bio2rdf.org/sparql	262
Family	interpro:Family **Source :** cu.interpro.bio2rdf.org/sparql	781
	Total	12260

ensemble d'exemples négatifs. Dans ce travail, les exemples positifs définis par une liste de gènes responsables pour des déficiences intellectuelles selon l'état de l'art, tandis que les exemples négatifs sont des gènes qui ne causent pas, à notre connaissance, ce type de maladie. On dispose de 282 exemples positifs tirés d'une étude sur les gènes responsables de déficiences intellectuelles conduite par Inlow et Restifo [Inlow and Restifo, 2004]. On établit également une liste de 267 exemples négatifs, présentés en Annexe B, qui sont des gènes responsables de maladies non liées aux déficiences intellectuelles.

Pour établir cette liste de gènes négatifs, nous avons dans un premier temps sélectionné des groupes de phénotypes dans OMIM qui n'avaient pas de déficience intellectuelle comme symptôme. Parmi celles-ci, nous avons choisi et validé avec l'aide d'experts du domaine un ensemble de phénotypes distincts des déficiences intellectuelles. La Table 4.2 propose une liste de ces phénotypes. Les gènes responsables de ces phénotypes ont ensuite été extraits d'OMIM. L'ensemble des 267 exemples négatifs présentés en Annexe B constitue un échantillon représentatif de ces gènes, c'est-à-dire que la proportion de gènes responsables pour chaque phénotype est preservée dans notre échantillon par rapport à l'ensemble des gènes.

4.2.4 Collecte des triplets RDF

À partir du modèle EA et des définitions de chaque entité et association en correspondance avec les entités et propriétés des LOD, il est possible de construire de manière systématique des requêtes SPARQL permettant de récupérer les données des LOD qui instancient le modèle EA. Une requête SPARQL est construite pour chaque association, à partir de la définition de l'association et des définitions des entités qu'elle associe. Une telle requête récupère alors un

TABLE 4.2 – Liste de phénotypes distincts des déficiences intellectuelles à partir de laquelle les gènes négatifs ont été sélectionnés (signes cliniques OMIM).

Phénotype
deafness
retinitis pigmentosa
obesity
cataract
muscular dystrophy
myopathy
hemolytic anemia
anemia
complement component deficiency
osteoarthritis
ectodermal dyspolasia
thrombophilia

ensemble de paires d'URI, chaque paire correspondant à deux instances liées par l'association.

Par exemple, la requête présentée en Figure 4.4 permet de récupérer toutes les associations d'un gène à une réaction. Cette requête est construite à partir de la définition de `gene_reaction` (décrite en Figure 4.3), à laquelle on ajoute une contrainte de domaine et de codomaine, c'est-à-dire les définitions des entités `Gene` :`geneid:Gene` ⊔ `kegg:Gene` et `Reaction` : `kegg:Reaction`. Dans cette requête, la variable `?x` correspond à un `Gene` tandis que la variable `?y` correspond à une `Reaction`. Les paires de résultats (`?x`, `?y`) instancient l'association `gene_reaction`.

Une fois les requêtes construites, nous avons automatisé le processus de collecte de données. En partant d'une liste initiale d'instances de `Gene` (correspondant aux exemples positifs et négatifs), les requêtes sont automatiquement executées de manière à collecter les données concernant ces gènes. Pour restreindre la collecte aux données pertinentes, une clause `FILTER` est ajoutée à chaque requête avant son éxecution. Par exemple, dans la requête présentée en Figure 4.4, la clause `FILTER(?x = ...)` est ajoutée pour restreindre la collecte d'associations `gene_reaction` à nos gènes exemples.

Une fois les requêtes SPARQL executées, les données récupérées sont automatiquement stockées dans une base de donnés relationnelle, construite à partir du modèle EA. À chaque entité correspond une table dont les colonnes représentent les URIs de l'instance dans chaque jeu de données, et un identifiant local comme clé primaire. À chaque association correspond une table stockant les paires instanciant cette association, référençant les clés primaires des différentes entités. Le nombre d'instances collectées par cette méthode à partir de nos listes d'exemples positifs et négatifs est présenté dans les Tables 4.1 et 4.3. Cette méthode permettrait l'ajout de nouvelles entités ou associations à notre modèle de manière incrémentale, cependant, la mise à jour des données collectées n'est possible qu'en recollectant l'intégralité d'une entité ou d'une association.

4.2.5 Mise en correspondance des individus

Les instances de nos entités sont identifiées dans les LOD par des URI. Puisque qu'une entité du modèle EA peut correspondre à plusieurs entités RDF dans différents jeux de donnés, il est possible de collecter des données redondantes : on peut notamment collecter différentes URI

```
PREFIX kegg:<http://bio2rdf.org/kegg_vocabulary:>
PREFIX geneid:<http://bio2rdf.org/geneid_vocabulary:>
SELECT ?x ?y
WHERE
{
   {?x rdf:type geneid:Gene}
   UNION                                  } Mapping of Gene
   {?x rdf:type kegg:Gene}

   {?y rdf:type kegg:Reaction}            } Mapping of Reaction

   ?p kegg:xGene ?x.
   {?y kegg:xEnzyme ?p}                   } Mapping of
   UNION                                    gene_reaction
   {?p kegg:xReaction ?y}
}
```

FIGURE 4.4 – Requête SPARQL construite pour instancier l'association `gene_reaction`, à partir des définitions de `Gene`, `Reaction` et `gene_reaction`.

désignant un même objet. Par exemple, les URI `geneid:5091` et `kegg:hsa:5091` désignent le même gènes dans deux jeux de données distinct : le gène nommé *pyruvate carboxylase* ayant pour Gene ID 5091. Des liens d'équivalence peuvent exister entre les URI de deux jeux de données, idéalement formellement exprimés par la propriété `owl:sameAs`. On remarque cependant que ces références entre jeux de données sont exprimés par des propriétés avec une sémantique plus faible, comme par exemple `rdfs:seeAlso`, `skos:relatedMatch` ou une propriété propre à un jeu de données.

Pour les entités mises en correspondance avec plusieurs jeux de données de LOD, il est nécessaire de proposer une façon automatique de résoudre l'identité des instances. Pour notre étude, nous avons utilisé différents moyens en fonction des cas présents :

— Utiliser, lorsque qu'ils sont disponibles dans les LOD, des liens exprimant l'équivalence entre différents URI codant pour un même objet. Ces liens sont matérialisés par des propriétés avec une sémantique formelle, comme `owl:sameAs` ou sont propres à un jeu de données particulier, comme `rdfs:seeAlso`.

— Utiliser d'autres données des LOD associées aux individus pour tester l'identité de deux instances :

  ◦ Les URI eux-mêmes peuvent parfois contenir suffisamment d'information pour affirmer que deux individus sont identiques. Par exemples, dans certains jeux de données, les URI désignant des gènes contiennent un NCBI Gene ID : le gène humain avec le Gene ID 5091 est représenté par l'URI `<http://bio2rdf.org/geneid:5091>` dans Bio2RDF NCBI Gene, et `<http://bio2rdf.org/kegg_vocabulary:hsa:5091>` dans Bio2RDF KEGG. Un lien d'équivalence entre ces deux URI peut alors être établi sur la base du Gene ID et d'expressions régulières.

TABLE 4.3 – Sources et nombre d'instances pour chaque association du modèle EA

Association	Endpoint SPARQL	Instances
gene_protein	beta.sparql.uniprot.org/sparql	819
gene_reaction	cu.kegg.bio2rdf.org/sparql	500
pp_interaction	cu.irefindex.bio2rdf.org/sparql	742
pathway_protein	www.ebi.ac.uk/rdf/services/reactome/sparql	767
protein_domain	cu.interpro.bio2rdf.org/sparql	262
pathway_reaction	cu.interpro.bio2rdf.org/sparql	706
substrate	cu.kegg.bio2rdf.org/sparql	938
product	cu.kegg.bio2rdf.org/sparql	960
protein_bp	cu.goa.bio2rdf.org/sparql	10242
protein_cc	cu.goa.bio2rdf.org/sparql	4358
protein_mf	cu.goa.bio2rdf.org/sparql	4063
subClass	sparql.bioontology.org/sparql	12779
domain_family	cu.interpro.bio2rdf.org/sparql	1238
gene_chromosome	cu.gene.bio2rdf.org/sparql	538
gene_chromosome_arm	cu.gene.bio2rdf.org/sparql	538
gene_chromosome_region	cu.gene.bio2rdf.org/sparql	538
gene_chromosome_band	cu.gene.bio2rdf.org/sparql	538
gene_chromosome_subband	cu.gene.bio2rdf.org/sparql	311
gene_chromosome_subsubband	cu.gene.bio2rdf.org/sparql	63
	Total	40900

- Les individus peuvent être associés avec des littéraux qui peuvent les identifier dans plusieurs jeux de données, par exemple le symbole de gène HGNC, qui est utilié pour identifier les gènes dans les jeux de données Bio2RDF-NCBI Gene et Bio2RDF-OMIM.

En utilisant ces méthodes, et étant donné un URI dans un jeu de données particulier, on peut trouver un URI correspondant dans un autre jeu de données. Ces méthodes sont néanmoins à adapter à chaque cas où l'on cherche à mettre en correspondance des URI provenant de différents jeux de données. Cette étape d'identification d'individus peut s'effectuer par des requêtes SPARQL récupérant des triplets d'équivalence entre individus. Les individus identiques sont stockés dans la base de données comme une seule entrée regroupant leur différentes URI. Quand cela est possible, on peut remplacer cette étape par le pré ou post-traitement d'une requête récupérant les instances d'une relation (par exemple, en transformant une URI récupérée `<http://bio2rdf.org/geneid:5091>` en `<http://bio2rdf.org/kegg_vocabulary:hsa:5091>`).

4.3 Programmation Logique Inductive avec des ontologies

L'étape de fouille des données issues des LOD est effectuée en utilisant la PLI [Muggleton, 1991] pour apprendre une conceptualisation des gènes responsables de déficiences intellectuelles. La PLI permet l'apprentissage de définition de concepts sur la base d'observations. Ces observations se présentent sous la forme d'un ensemble d'exemples positifs et d'un ensemble d'exemples négatifs, et peuvent être accompagnées de connaissances de domaine. A partir de ces observations et connaissances, la PLI permet d'induire un ensemble de règles, ou théorie, décrivant un

maximum d'exemples positifs, et un minimum d'exemple négatifs. Ici, on dispose de l'ensemble des triplets intégrés selon notre modèle EA décrivant les gènes ou exemples positifs et négatifs, associés ou non à des déficiences intellectuelles.

Les expériences ont été effectuées avec le programme Aleph [Srinivasan, 2007] en utilisant les paramètres suivants :

— *rule size* = 6, le nombre maximal de termes dans une règle ;
— *minpos* = 5, le nombre minimum d'exemples positifs qu'une règle doit couvrir ;
— *noise* = 3, le nombre maximum d'exemples négatifs qu'une règle peut couvrir ;
— *minacc* = 0.85, le ratio minimum d'exemples positifs parmi les exemples couverts par une règle.

Le paramètre *noise* permet d'obtenir une théorie tolérant quelques exceptions (exemples négatifs), ce qui est une nécessité lorsque l'on manipule des données potentiellement bruitées telles que notre ensemble de données intégrées à partir des LOD.

On propose ici 5 expériences de fouille utilisant notre jeu de données intégrées. Chaque expérience utilise un niveau de raisonnement variable sur la hiérarchie des termes GO. La théorie résultante de chaque expérience de fouille est utilisée à la fois dans un but descriptif et prédictif. La description donnée des gènes responsables de déficiences intellectulles est évaluée qualitativement, tandis que le pouvoir de prédiction des règles exprimées en logique du première ordre est évalué par un processus de validation croisée. A cette fin, des workflows KNIME [Berthold et al., 2009], une plateforme permettant le traitement et la fouille de données, avec une implémentation d'Aleph ont été utilisés [Grisoni et al., 2013], pour effectuer la validation croisée en *leave-one-out*[2], comme illustré en Figure 4.5. Un gène est prédit comme étant responsable d'une déficience intellectuelle si et seulement si il est couvert par au moins une règle de la théorie évalueée.

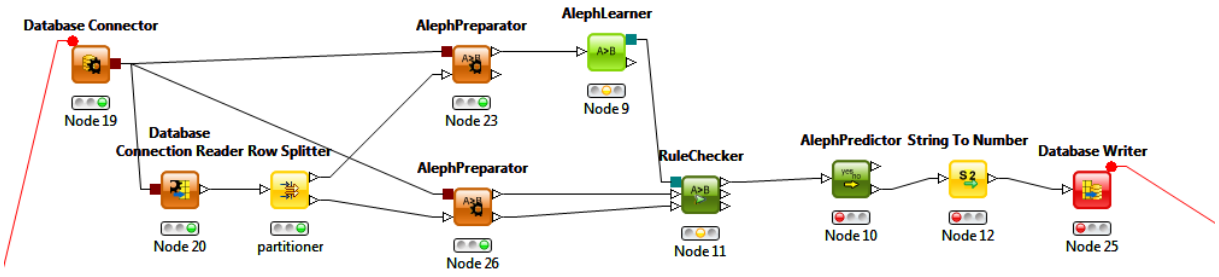


FIGURE 4.5 – Représentation du Workflow KNIME pour un pli (ou *fold*) de validation croisée, utilisant une version adaptée des nœuds de PLI : AlephPreparator, AlephLearner, RuleChecker et AlephPredictor [Grisoni et al., 2013].

La première expérience, $G1$, s'applique aux gènes munis des connaissances sur leur protéines, réactions, voies biologiques, etc. ainsi que leur annotations GO, plus les classes parentes au premier degré de ces annotations, dénotés explicitement par la propriété `rdfs:subClassOf`. On notera ici cette relation de parenté directe entre termes GO $\texttt{subClassOf}_1$.

2. Une validation croisée utilisant autant de pli (ou *folds*) qu'il y a d'exemples. Ici, pour chaque exemple, on apprendra une théorie sur l'ensemble des autres exemples, qu'on évaluera ensuite sur ce seul exemple.

Le logiciel Aleph permet de prendre en compte des règles d'inférence exprimées en logique du premier ordre. Ces règles permettent de raisonner en considérant à la fois les données et connaissances liées aux exemples avant d'effectuer l'étape d'apprentissage, enrichissant ainsi les descriptions de chaque exemple. Afin d'évaluer la contribution des connaissances contenues dans les ontologies de domaine, on propose 3 autres expériences permettant de considérer la transitivité de la relation `subClassOf`, à différents niveaux de généralisation. Pour permettre à Aleph d'effectuer n étapes des généralisation, on lui fournit $2n$ règles d'inférence comme suit :

Une règle pour chaque $i \in [2, n]$ exprimant la transitivité de `subClassOf` au i-ième degré :

$\texttt{subClassOf}_i(\texttt{X},\texttt{Z})$ `:-` $\texttt{subClassOf}_{i-1}(\texttt{X},\texttt{Y})$, $\texttt{subClassOf}_1(\texttt{Y},\texttt{Z})$.

Une règle pour chaque $i \in [1, n]$:

$\texttt{subClassOf}(\texttt{X},\texttt{Y})$ `:-` $\texttt{subClassOf}_i(\texttt{X},\texttt{Y})$.

Une règle exprimant la réflexivité de `subClassOf` :

`subClassOf(X,X) :- goterm(X).`

Ici, l'expérience de fouille a été effectuée pour n variant de 1 à 4. On notera les expériences correspondantes $G1$, $G2$, $G3$ et $G4$, utilisant chacune respectivement jusqu'à 1, 2, 3 ou 4 étapes de généralisation. L'examination des 4 théories produites révèle que les règles produites contiennent pour la plupart des prédicats liés à des termes GO. En revanches, les prédicats référant aux voies biologiques ou interactions entre protéines sont peu présents. Cela peut s'expliquer par le fait que les annotations GO sont très nombreuses dans notre jeu de données, relativement aux autres entités. On propose alors une cinquième expérience, notée $no - GO$ pour analyser l'apport des prédicats non liés à GO, et contextualiser l'apport de l'ontologie GO.

4.4 Résultats

Les 5 expériences de fouille détaillées précedemment permettent de produire 5 théories décrivant les gènes responsables de déficiences intellectuelles. L'intégralité de ces théories est disponible en Annexe C.

TABLE 4.4 – Métriques sur les théories produites par les 5 expériences : nombres de règles (#Règles), nombre moyen d'exemples positifs couverts par une règle (#Ex. pos. moyen), nombre maximum d'exemples positifs couverts par une règle (#Ex. pos. max.), nombre minimum d'exemples positifs couverts par une règle (#Ex. pos. min.).

Expérience	#Règles	#Ex. pos. moyen	#Ex. pos. max.	#Ex. pos. min.
$no - GO$	11	8.4	15	5
$G1$	22	14	35	6
$G2$	19	15.5	38	6
$G3$	18	15.1	39	6
$G4$	16	16.2	42	5

La Table 4.4 présente quelques métriques sur les différentes théories permettant de constater les effets du raisonnement sur GO sur les règles obtenues. Avec GO, on constate que le plus on permet de généralisations, le moins on obtient de règles. Nénamoins, les règles obtenues sont couvrent davantage d'exemples en moyenne. Par rapport à l'expérience $no - GO$, le nombre de

TABLE 4.5 – Résultats de prédiction pour les 5 expériences, par validation croisée *leave-one-out*. VP/FP : Vrai/Faux Positifs, VN/FN : Vrai/Faux Négatifs, Sens. : Sensibilité, Spec. : Spécificité, Prc : Précision.

Expérience	VP	FP	VN	FN	Sens.(%)	Spec.(%)	Prc.(%)
$no - GO$	75	**15**	**252**	207	26.6	**94.4**	59.6
$G1$	135	50	217	147	47.9	81.3	64.1
$G2$	157	52	215	125	55.7	80.5	67.8
$G3$	157	49	218	125	55.7	81.7	68.3
$G4$	**161**	45	222	**121**	**57.1**	83.1	**69.8**

règles dans la théorie double lorsque l'on ajoute des termes GO, et le nombre moyen d'exemples couverts augmente de 8.4 à 14. Cela indique que les termes GO ont un rôle très positif pour l'apprentissage du concept des gènes responsables de déficiences intellectuelles. Lorsque l'on augmente le nombre d'étapes de généralisation de 1 à 4, le nombre de règles diminue de 22 à 16, tandis que le nombre moyen d'exemples positifs couverts augmente légèrement de 14 à 16.2. Ces résultats confirment que avec l'addition d'étapes de raisonnement, les théories tendent à devenir plus compactes avec moins de règles, chacune couvrant plus d'exemples. Il est cependant nécessaire de mesurer le pouvoir de prédiction de chacune de ces théories, c'est-à-dire dans quelle mesure ces règles permettent de prédire si un gène donné est responsable de déficiences intellectuelles.

4.4.1 Analyse quantitative des théories sur une tâche de classification

On évalue le résultat de chaque expérience de fouille sur le critère du pouvoir de prédiction par validation croisée, où un gène est prédit comme étant responsable uniquement s'il est couvert par au moins une règle de la théorie évalueée. La Table 4.5 présente les résultats de la validation croisée *leave-one-out* des expériences d'apprentissage $no - GO$, et $G1$ à $G4$.

Les résultats montrent que sans utiliser d'annotations GO ($no - GO$), la précision de la prédiction est plutôt faible (59.6%), avec une très haute specificité mais une faible sensibilité. Utiliser des termes GO augmente la qualité de la prédiction, jusqu'à une précision de 69.8% lorsque que l'on permet jusqu'à 4 étapes de généralisation. On observe notamment qu'augmenter le nombre maximum d'étapes de généralisation permet d'augmenter la précision de la prédiction : au-delà des simples annotations GO, les connaissances de domaine proposées par les ontologies permettent une meilleure caractérisation des déficiences intellectuelles.

4.4.2 Analyse qualitative des théories

Au-delà de la capacité de prédiction des théories, on souhaite analyser leur capacité descriptive de ces théories, c'est-à-dire, leur caractérisation des gènes responsables de déficiences intellectuelles. La Table 4.6 présente les règles obtenues dans l'expérience $no - GO$: en l'absence d'annotations GO, on observe plusieurs règles présentant des prédicats concernant la localisation chromosomique des gènes, tels que les règles 4 et 5 pointant les chromoses 1 et X comme porteurs potentiels de gènes responsables pour les déficiences intellectuelles. Effectivement, il existe de nombreuses déficiences intellectuelles liées au chromosome X [Bresso, 2013]. De plus, la règle 8 pointe vers une localisation plus précise sur le chromosome 22. D'autres règles contiennent le

TABLE 4.6 – Corps des règles de la théorie $no-GO$, suivis, respectivement, du nombre d'exemples positifs et du nombre d'exemples négatifs couverts par chaque règle. La tête de chaque règle est `is_responsible(A)`.

#	Corps des règles		
1	`gene_in_reaction(A, 'Ubiquinol+Acceptor⇔Ubiquinone+Reduced Acceptor')`.	7	0
2	`gene_in_reaction(A, B)`, `gene_protein(A, C)`, `pp_interaction(C,D)`, `pp_interaction(D, C)`.	6	0
3	`gene_in_reaction(A, B)`, `gene_protein(A, C)`, `pp_interaction(C, P30480)`.	7	0
4	`gene_in_reaction(A, B)`, `gene_ch(A, '1')`	14	0
5	`gene_in_reaction(A, B)`, `gene_ch(A, x)`	15	2
6	`gene_in_pathway(A, 'Alanine and aspartate metabolism')`.	6	1
7	`gene_in_pathway(A, 'Valine, leucine and isoleucine degradation')`.	11	1
8	`gene_chromosome_band(A, '22q13')`.	6	0
9	`gene_in_pathway(A, 'N-Glycan biosynthesis')`.	8	0
10	`gene_in_pathway(A, 'Formation of TC-NER repair complex')`.	5	0
11	`gene_in_pathway(A, 'Glycosaminoglycan degradation')`.	8	0

prédicat `gene_in_pathway` (règles 6, 7, 9, 10, 11) dans lesquelles on trouve des voies métaboliques impliquées dans le métabolisme de la cellule. En effet, les maladies métaboliques génétiques sont connues pour être une cause importante de déficiences intellectuelles [van Karnebeek and Stockler, 2012].

Lorsque l'on permet au processus d'apprentissage d'utiliser les annotations et l'ontologie GO (expériences $G1$ à $G4$), le répertoire des termes GO apparaissant dans les règles, soit comme l'annotation directe d'une protéine ou un ancêtre commun, varie avec le nombre de généralisation permises par nos règles d'inférences. Les théories résultantes de ces expériences sont disponibles en Annexe C. Sur l'ensemble des expériences, on dénombre respectivement 47, 7 et 14 termes GO distincts pour chacun des trois aspects *Biological Process* (BP), *Molecular Function* (MF) et *Cellular Component* (CC). Parmi les termes BP, on peut à nouveau observer des termes décrivant des processus métaboliques de la cellule, mais aussi des termes liés aux méchanismes d'expression génetique et du développement du système nerveux. Ces observations sont cohérentes avec la caractérisation des déficiences intellectuelles souhaitée. Certaines règles peuvent également combiner des termes des plusieurs aspects de GO, comme la règle 16 de la théorie $G4$ présentant les prédicats `protein_bp` et `protein_mf` :

```
is_responsible(A) :- gene_protein(A,B), protein_mf(B,C),
                     subClassOf(C,'ion binding'),
                     protein_bp(B,'carbohydrate metabolic process').
```

Cette règle illustre que la PLI est capable de prendre en compte plusieurs ontologies, ici les différents aspects indépendants de GO, tout en opérant à différents niveaux de généralisation. En particulier, cette règle exprime qu'un gène `A` est responsable d'une déficience intellectuelle s'il code pour une protéine `B` annotée par les termes GO *carbohydrate metabolic process* et par un terme `C` descendant de *ion binding*.

L'impact de l'ajout de règles d'inférence peut être illustré par des règles présentant le prédicat `subClassOf`. En particulier, on s'intéressera à une règle référençant des termes descendants de *organonitrogen compound metabolic process* dans les quatres théories $G1$ à $G4$:

— La règle 3 de la théorie $G1$ (et règle 7 dans $G2$, 23 exemples positifs couverts) :

`is_responsible(A) :- gene_protein(A,B), protein_bp(B,C),`
`protein_bp(C,` 'organonitrogen compound **metabolic** process')`.`

— La règle 4 des théories $G3$ et $G4$ (respectivement 39 et 42 exemples positifs couverts) :

`is_responsible(A) :- gene_in_reaction(A,B), gene_protein(A,C), protein_bp(C,D),`
`protein_bp(C,` 'organonitrogen compound **catabolic** process')`.`

On observe d'abord que le terme GO présent dans les théories $G3$ et $G4$, *organonitrogen compound* ***catabolic*** *process*, est plus spécifique que celui présent dans les théories $G1$ et $G2$, *organonitrogen compound* ***metabolic*** *process*. On notera néanmoins que des règles identiques dans différentes théories peuvent avoir une couverture différents, puisque l'on doit les considérer dépendamment des règles d'inférence utilisées pour créer cette théorie. Ici, permettre davantage d'étapes de généralisation permet d'augmenter la couverture de cette règle, tout en obtenant une caractérisation plus spéficique.

4.5 Discussion

L'approche présentée ici permet, à partir d'une conceptualisation d'un problème, l'intégration et la fouille de données extraites des LOD, en vue de proposer un modèle pouvant prédire et caractériser une classe d'objets. Cette approche est ici appliquée sur une tâche de caractérisation des gènes responsables des déficiences intellectuelles, tirant parti de la mise à disposition sous forme de LOD de nombreux jeux de données biologiques.

4.5.1 Intégration de Données Ouvertes Liées

On notera d'abord que l'étape de sélection et d'intégration requiert à ce jour un travail manuel afin d'établir des correspondances entre différents jeux de données. En particulier, on propose ici un modèle entité-association que l'on va mettre en correspondance avec les types et propriétés présentes dans les LOD. Ce travail est néanmoins nécessaire puisque les différents jeux de données n'utilisent pas un vocabulaire unifié, ni au niveau des classes, ni au niveau des individus : on aura par exemple à mettre en correspondance les deux types :

— `<http://bio2rdf.org/geneid_vocabulary:Gene>`,

— `<http://bio2rdf.org/kegg_vocabulary:Gene>`.

Idéalement, on souhaiterait que la plupart des jeux de données biologiques fassent réference à une ontologie biologique de haut niveau, définissant des concepts tels que gène ou protéine et leur relations, ou trouver un lien d'équivalence entre les types des différents jeux de données. En outre, il est nécessaire de disposer d'un modèle de données de haut niveau pour faciliter la communication avec des experts du domaine. Ainsi, au sein de notre approche, le modèle EA a un rôle à la fois dans l'intégration des données et dans la communication avec les experts.

4.5.2 Programmation Logique Inductive pour la fouille avec des ontologies de domaine

Au-delà de la mise en œuvre en elle-même, les résultats tant de prédiction que de caractérisation semblent indiquer que la PLI est bien adaptée pour la fouille de donnés extraites des LOD et intégrées. De plus, on observe dans les théories que les connaissances des différents aspects

de GO sont exploitées dans le processus d'apprentissage. Le grand nombre d'annotations GO dans le jeu de données intégrées conduit à une majorité de prédicats référant aux termes GO dans les théories $G1$ à $G4$. On constate pourtant grâce à la théorie $no-GO$ que les autres prédicats possède un bon pouvoir prédictif sur une partie des gènes. La prévalence des prédicats associés à GO pourrait être réduite en limitant le nombre de termes GO sur le base des codes d'évidence associés aux annotations GO. Par ailleurs, on peut envisager la création d'une théorie globale combinant les règles de la théorie $no-GO$ et celles des théories $G1$ à $G4$ [Berthold et al., 2007, Knobbe et al., 2008].

Un avantage que la PLI présente, par rapport aux méthodes fondées sur les similarités sémantiques détaillées dans le Chapitre 3, est la prise en compte des différents aspects de GO uniquement lorsque nécessaire : on observe dans une même théorie des règles présentant des termes d'un seul ou plusieurs de ces aspects. Ainsi, il n'est pas nécessaire de répéter les expériences avec différentes combinaisons des trois aspects de GO pour optimiser le résultat. Un autre avantage de la PLI est qu'elle permet d'obtenir une théorie sous forme de règles de logique du premier ordre, permettant à un expert du domaine de comprendre le modèle et d'en expliquer les prédictions. Les théories obtenues présentent une forte spécificité, limitant le nombre de faux positifs, en dépit d'une sensibilité plus faible. Cela pourrait être expliqué par la variabilité des quantité et qualité des données disponibles sur chacun des gènes. Chaque règle caractérise un sous-ensemble significatif des gènes positifs (16 en moyenne pour la meilleure théorie), croissant avec le nombre de généralisations permises par les règles d'inférence. Ces théories pourraient ensuite être utilisés pour proposer des gènes candidats aux experts.

4.5.3 Conclusion

Les LOD mettent à disposition une vaste collection de données biologiques liées à des ontologies. Les LOD restent néanmoins fragmentés en différents jeux de données, qui nécessitent la création de nouveaux liens pour permettre leur interopérabilité. Néanmoins, le format unique des LOD et les liens existants permette la proposition de correspondances et en font une resource opportune pour la fouille de données dans le domaine biomédical.

La PLI est un outil efficace pour la fouille de LOD, de part l'utilisation de la logique du premier ordre permettant d'exprimer tant les données des LOD que les connaissances des ontologies. La PLI se révèle plus appropriée que les méthodes fondées sur les similarités sémantiques pour la prise en compte de plusieurs ontologies. En particulier, cette méthode ne requiert pas de sélection préalable des ontologies, ici les différents aspects de GO, à appliquer pour classer un ensemble de gènes.

Les différentes règles et théories produites par la PLI illustrent, tout comme les expériences décrites dans le Chapitre 3, que les déficiences intellectuelles forment un ensemble hétérogène de maladies. D'une part, le Chapitre 3 montre que chaque aspect de GO et contribue à la prédiction à la classe de ces maladies de manière différente selon la classe. D'autre part, les expériences de PLI décrites ici produisent de nombreuses règles chacune caractérisant un petit sous-ensemble de ces maladies.

Une grande partie du travail présenté ici consiste à l'établissement de correspondances entre types, propriétés et individus des jeux de données des LOD. On peut envisager que, si cette étape n'était plus nécessaire, un processus de PLI puisse être appliqué directement sur les LOD, permettant de raisonner partiellement avec les ontologies liées aux données.

5

Conclusions et perspectives

Les représentations formelles de connaissances représentent une opportunité pour le domaine de l'intelligence artificielle et notamment la fouille de données. Les technologies du Web sémantique proposent des standards pour formaliser et partager des connaissances de domaine et ont ainsi permis la mise à disposition de ces connaissances en grandes quantités, notamment dans le domaine biomédical. On distingue au sein du Web sémantique deux composantes principales : les données ouvertes et liées (LOD) et les ontologies. Les LOD sont un ensemble de jeux de données proposés de manière à être interopérables, grâce au langage standard RDF, aux URI et aux liens entre eux. De leur côté, les ontologies du Web sémantique permettent une interprétation formelle et du raisonnement sur les LOD.

Cette thèse décrit trois études utilisant les connaissances formalisées par des ontologies du Web sémantique dans un processus de fouille de données biomédicales. Dans chacun de ces travaux, l'apport des ontologies est évalué grâce à plusieurs expériences utilisant chacune différentes ontologies ou différents degrés de raisonnement. On y présente notamment une série d'expériences intégrant plusieurs sources de données et de connaissances associées au sein d'un même processus de fouille de données. L'ensemble de ces travaux illustrent l'apport des ontologies dans deux problématiques biomédicales distinctes :

- la découverte d'associations entre Evénements Indésirables Médicamenteux (EIM) à partir de données patients ;
- la classification des déficiences intellectuelles en utilisant des connaissances phénotypiques ou génétiques.

Dans ces différents travaux, les expériences utilisent une ou plusieurs des ontologies suivantes dans le processus de fouille de données :

- ATC, une ontologie de médicaments,
- ICD-9-CM, une ontologie de maladies,
- SNOMED CT, une ontologie de termes médicaux,
- *Gene Ontology* (GO), trois ontologies ou aspects indépendants décrivant les fonctions biologiques des gènes : *Biological Process* (BP), *Cellular Component* (CC), *Molecular Function* (MF),
- *Human Phenotype Ontology* (HPO), une ontologie de phénotypes humains,
- *Monarch Disease Ontology* (MonDO), une ontologie de phénotypes intégrant notamment HPO et OMIM.

Dans le Chapitre 2, j'ai proposé une approche utilisant des ontologies biomédicales pour permettre la comparaison sémantique d'EIM, afin d'extraire des règles d'association entre ces événements. Cette méthode est fondée sur les structures de patrons, une extension de l'analyse formelle

de concepts, pour laquelle j'ai proposé trois opérateurs de comparaison, permettant d'utiliser, respectivement, 0, 1 ou 2 ontologies. En particulier, le troisième opérateur combine une ontologie de phénotypes et une ontologie de médicaments, permettant de trouver des associations au niveau de classes de haut niveau de médicaments ou de phénotypes. Comparativement aux autres opérateurs exploitant moins d'ontologies, cet opérateur permet de trouver des règles d'association avec un meilleur support. Une analyse qualitative de ces règles permet de constater que les deux ontologies ont bien été utilisées pour généraliser des classes de médicaments ou de phénotypes à différents niveaux lors de la généralisation.

Le Chapitre 3 présente une méthode numérique pour utiliser les ontologies dans un processus de fouille de données. Cette méthode est fondée sur les similarités sémantiques. On y reproduit d'abord une expérience décrite par [Hoehndorf et al., 2015], où une similarité sémantique, SimGIC, est utilisée avec l'ontologie MonDO pour la construction d'un graphe de maladies ou diseasome. Ce diseasome est ensuite évalué sur plusieurs tâches de classification, dont le but est la mise en évidence de maladies partageant une classe Disease Ontology dans un cas, une indication médicamenteuse dans l'autre. Dans ce Chapitre, l'approche a été étendue avec l'utilisation de la mesure de similarité sémantique IntelliGO, ainsi qu'une évaluation complémentaire concernant la classification de maladies partageant un gène responsable. J'ai mis en évidence différents biais entre les similarités SimGIC et IntelliGO, notamment concernant les maladies pour lesquelles un médicament est connu dans la base de données d'indications SIDER. J'ai également réalisé une nouvelle expérience utilisant conjointement les ontologies HPO et GO pour classifier des déficiences intellectuelles dans 5 classes définies par des experts du domaine. On utilise alors les deux similarités sémantiques SimGIC et IntelliGO avec différentes combinaisons d'ontologies parmi HPO et les trois aspects de GO. On observe que, globalement, considérer les 4 ontologies (HPO et les trois aspects de GO) permet d'améliorer la performance sur la tâche de classification. Cependant, pour certaines classes de déficiences intellectuelles, ne considérer qu'un sous-ensemble de ces ontologies permet d'améliorer la performance. Ainsi, il existe des cas où ajouter davantage de connaissances du domaine a un effet négatif sur le processus de fouille de données. On constate ici une limite de ce type de méthode lorsque que l'on considère des données non pertinentes pour un problème donné.

Le Chapitre 4 propose une approche permettant d'utiliser les trois aspects de GO pour la caractérisation de gènes responsables de déficiences intellectuelles. Cette méthode utilise la Programmation Logique Inductive (PLI) qui, à partir d'exemples de gènes responsables décrits par des données biologiques exprimées en logique du premier ordre, construit une théorie, c'est-à-dire un ensemble de règles en logique du premier ordre, décrivant ces gènes responsables. Plusieurs expériences ont été réalisées utilisant des règles d'inférences permettant de raisonner sur les hiérarchies de concepts de GO, chacune utilisant un degré de raisonnement différent, c'est-à-dire une prise en considération d'une partie plus ou moins grande de la hiérarchie de concepts. J'ai également réalisé une expérience sans utiliser de termes GO. On montre alors qu'un degré de raisonnement supérieur améliore la caractérisation des gènes responsables, d'une part en comparant qualitativement les théories obtenues et, d'autre part, en les évaluant sur une tâche de classification. De plus, on observe que la PLI est capable de considérer simultanément les trois aspects de GO pour produire des règles caractérisant ces gènes responsables.

Discussion

Les trois contributions présentées dans cette thèse se révèlent complémentaires puisqu'elles permettent chacune l'utilisation d'ontologies de domaine dans un processus de fouille avec une

méthode différente. Les structures de patrons permettent l'extraction de règles d'association de manière non supervisée. Ainsi, cette méthode est tout à fait appropriée pour la découverte d'associations entre EIM sans apriori sur les médicaments ou phénotypes impliqués. Les mesures de similarité sémantique permettent de quantifier la proximité de maladies annotées par des classes d'ontologies, et peuvent être utilisées dans un processus de clustering ou pour ordonnancer les maladies. Les méthodes fondées sur les mesures de similarité sémantique sont également non supervisées. Cependant, parce que ces méthodes n'effectuent pas un raisonnement formel sur les connaissances de l'ontologie, il est nécessaire d'évaluer le modèle obtenu en se réferant à une réalité terrain pré-établie, comme une classification de maladies. La PLI permet quant à elle d'apprendre une théorie décrivant un ensemble d'objets donné, apprenant ainsi à décrire et à prédire une classe de manière supervisée.

On distinguera parmi les différentes méthodes mises en œuvre dans cette thèse deux méthodes symboliques, les structures de patrons et la PLI, et une méthode numérique de par l'utilisation de similarités sémantiques. Les deux méthodes symboliques permettent, à partir d'un ensemble de données liées à des ontologies, d'extraire un modèle interprétable par un expert du domaine. En cela, ces méthodes conservent un des avantages des ontologies, qui sont interprétables à la fois par une machine et un humain. Dans ces deux méthodes, on constate que l'ajout de davantage d'ontologies ou de capacités de raisonnement permet de générer des modèles plus détaillés.

En revanche, une limite de la méthode numérique fondée sur les similarités sémantiques est que, bien qu'elle considère la sémantique associée à la hiérarchie de concepts de l'ontologie, on ne peut pas considérer qu'elle mette en œuvre un mécanisme de raisonnement formel sur les connaissances de l'ontologie. Notamment, il n'est pas possible d'associer une sémantique formelle au modèle résultant de l'apprentissage, ici, le diseasome. Aussi, pour l'interprétation d'un tel diseasome par un expert, il pourra être nécessaire de mettre en œuvre des méthodes symboliques qui permettent d'expliquer la similarité entre deux maladies. A cette fin, une approche similaire à celle décrite pour les opérateurs de comparaison présentés dans le Chapitre 2 permettrait de comparer les annotations de deux maladies. De fait, il apparaît que les méthodes symboliques sont capables de proposer des modèles interprétables par des experts, ce qui représente un grand avantage dans le cadre de potentielles applications biologiques.

Une autre limite posée par les mesures de similarité sémantiques est la difficulté de leur évaluation objective. On a notamment observé que les deux mesures SimGIC et IntelliGO ont des écarts de performance très variables sur différentes tâches. De même, leur performance semble varier en fonction des ontologies utilisées. Ainsi, le choix de la meilleure mesure de similarité sémantique est dépendant de la tâche et des connaissances à considérer. Ces différences de perfomance peuvent s'expliquer du fait que ces deux mesures de similarité ont des méthodes de calcul différentes. Notamment, SimGIC est une mesure fondée sur les nœuds tandis qu'IntelliGO est fondé sur les arcs du graphe de l'ontologie. De plus IntelliGO considère une similarité entre les concepts de l'ontologie deux-à-deux, tandis que SimGIC compare uniquement des ensembles de concepts. On note également que SimGIC respecte l'inégalité triangulaire, ce qui n'est pas le cas d'IntelliGO. La pertinence de cette propriété peut dépendre de la construction de l'ontologie. Par exemple, soient trois concepts A, B et C tels que $A \sqsubseteq B$ et $A \sqsubseteq C$: on peut supposer de la proximité entre A et B et entre A et C de fait de leur relation de subsomption. En revanche, il n'y a pas de proximité nécessaire entre B et C. Notamment, dans une ontologie comme HPO, deux parents d'un concept peuvent représenter différentes perspectives. Par exemple, le concept *Dislocated hips* est enfant des deux concepts *Dislocations* et *Abnormality of the hips*. Ici, une mesure respectant l'inégalité triangulaire devrait considérer les deux derniers concepts comme similaires du fait qu'ils aient un concept enfant commun. Il peut alors être nécessaire, lors du choix d'une mesure de similarité, de considérer les propriétés mathématiques de la mesure dans

le contexte de l'ontologie à exploiter.

Les définitions de SimGIC et IntelliGO, ainsi que les différents biais de SimGIC et IntelliGO mis en évidence dans les expériences du Chapitre 3, laissent supposer que ces mesures exploitent des élements des ontologies auxquels aucune sémantique formelle n'est associée, comme par exemple le nombre d'annotations d'une maladie ou la profondeur dans la hiérarchie d'un concept. Cela rend ces mesures susceptibles à des mises à jour de l'ontologie, même lorsque cette mise à jour ne changent pas la sémantique des concepts comparés. Par exemple, les similarités calculées par SimGIC seraient affectées par l'ajout ou la suppression d'un concept redondant dans l'ontologie. Similairement, IntelliGO produirait des résultats différents si l'on considérait que tout concept de l'ontologie est enfant de `owl:Thing`. Ces biais peuvent poser des problèmes dans le contexte de l'utilisation des ontologies du Web sémantique, qui peuvent être liées entre elles ou présenter des concepts redondants. Il pourrait alors être pertinent de concevoir des mesures de similarité sémantiques qui soient le plus insensibles possibles aux éléments non sémantiques des ontologies.

Les travaux de cette thèse explorent des méthodologies visant à faire coopérer plusieurs ontologies. De telles méthodes semblent nécessaires pour tirer parti au mieux des connaissances et données publiées via le Web sémantique. En particulier, la PLI permet d'intégrer avec un formalisme unique, la logique du premier ordre, les LOD et les ontologies qui y sont liées. On notera néanmoins que, dans notre cas d'étude des gènes responsables de déficiences intellectuelles, les données extraites et intégrées des LOD peuvent s'apparenter à des connaissances biomédicales. Par exemple, les prédicats décrivant les protéines produites par un gène ou d'une réaction décrivent une connaissance générale sur ces gènes, protéines ou réactions. Par ailleurs, une protéine définie dans les LOD comme une instance d'une classe *Protéine* ne représente pas un individu particulier, mais une conceptualisation représentant un ensemble de protéines interchangeables. Ainsi, on peut trouver dans les LOD comme dans les ontologies des éléments de l'ordre des connaissances biomédicales. Cela pourrait poser problème pour la représentation de données et l'apprentissage automatique avec les LOD, s'il on voulait y inclure des données patients qui ne seraient pas au même niveau d'abstraction que des connaissances d'ordre plus général.

Les LOD et ontologies du Web sémantique sont proposées dans un format intéropérable et partiellement intégrées par des liens entre les jeux de données. Cependant, comme on a pu le remarquer pendant les expériences d'intégration et de fouille de LOD, une phase manuelle de découverte, de sélection et de mise en correspondance des jeux de données peut être nécessaire pour permettre de les exploiter. On a en particulier proposé un modèle de données de haut niveau restreint à notre cas d'étude, en correspondance avec les différents jeux de données des LOD biomédicales utilisés. Ce modèle permet de faciliter l'intégration de ces données pour la fouille de données par la PLI. Le modèle proposé ici propose des définitions d'entités et de relations fondées sur les types d'entités et propriétées existantes dans les LOD, pour permettre l'intégration dans une base de données relationelle, dans un but d'interopérabilité avec le logiciel de PLI Aleph. Une approche similaire peut néanmoins être appliquée pour la création d'ontologies de haut niveau au sein du Web sémantique lui-même. L'extension de la couverture d'un tel modèle à l'échelle de l'ensemble des ontologies biomédicales est néanmoins limitée par plusieurs facteurs. En particulier, différents usages des données imposent différentes contraintes sur la construction d'une ontologie. En effet, différentes ontologies peuvent proposer différentes perspectives sur les connaissances d'un domaine. Par exemple, l'ontologie ICD-9-CM répartit les maladies en plusieurs classes orientées anatomiquement, et de manière à permettre la facturation des soins. En revanche, l'ontologie HPO, dont les classes peuvent avoir plusieurs parents, peut proposer plusieurs perspectives sur un phénotype, avec une classe parente décrivant la localisation anatomique du phénotype et une autre classe parente décrivant le type de symptôme. De

plus, différentes applications peuvent également demander différents niveaux de granularité dans les ontologies. L'intégration de données expérimentales, ou de données patients comme celles de rapports anonymes d'effets secondaires pourrait également servir à enrichir les LOD et à étendre les possibilités des méthodes de fouille de LOD. Des principes de publication et de partage des jeux de données ont été récemment définis pour faciliter leur réutilisation et leur interopérabilité [Wilkinson et al., 2016]. Des initiatives comme le BioPortal [Whetzel et al., 2011] proposent également une vaste collection ontologies enrichie par des liens proposés par la communauté, favorisant l'intégration de plusieurs ontologies. D'autres, comme BioSchemas [Gray et al., 2017] proposent une ontologie biomédicale de haut niveau permettant l'interopérabilité de nombreuses sources de données.

Perspectives

Les différentes méthodes présentées dans cette thèse permettent la considération de nouvelles sources de données et de connaissances pour la fouille de données. On propose ici différentes manières d'étendre ces méthodes pour permettre d'y prendre en compte davantage de données et connaissances.

Les structures de patrons utilisées pour l'extraction d'associations d'EIM pourraient notamment être étendues pour considérer davantage de connaissances sur les médicaments et phénotypes représentés, comme par exemple les fonction biologiques affectées par un médicament, ou pouvant être à l'origine d'un phénotype. Les structures de patrons présentées permettent de considérer la similarité entre EIM en tenant compte d'ontologies biomédicales, cependant, elles ne peuvent pas considérer les relations entre médicaments et phénotypes. En augmentant la représentation des EIM proposée avec des données sur les mécanismes biologiques des différents médicaments et phénotypes, cela permettrait de mettre en évidence les mécanismes des EIM associés, et, éventuellement d'obtenir des associations plus générales et indépendantes d'une classification particulière de médicaments ou de phénotypes.

Dans l'étude présentée au Chapitre 2, seuls des EIM à court terme ont été considérés. La représentation des EIM pourrait donc être enrichie en tenant compte du délai entre la prescription du médicament et l'apparition du phénotype indésirable. Cela permettrait d'extraire des associations dans un jeu de données d'EIM à court ou long terme, tout en étant capable de discriminer au besoin entre ces différentes manifestations. Eventuellement, cela pourrait permettre d'utiliser un EIM à court terme pour prédire l'apparition d'un EIM sur le long terme.

D'autre données pourraient être incluses au-delà des EIM dans la description de chaque patient, comme les antécédants médicaux du patient ou les maladies pour lesquelles il est actuellement traité. En effet, au-delà des interactions entre médicaments qui peuvent être considérées par la représentation des EIM proposée, il est également nécessaire de prendre en compte les interactions entre maladies et médicaments [Goldberg et al., 1996]. Ce type de données pourrait notamment faciliter l'interprétation des associations obtenues.

De telles règles d'association extraites sur un plus grand ensemble de patients pourraient servir comme base pour un système de recommandation, en permettant de mettre en évidence des contre-indications possibles. En effet, il a été démontré que des systèmes d'aide à la décision peuvent réduire les EIM liés à la prescription de médicaments [Schedlbauer et al., 2009].

L'utilisation de la PLI pour fouiller des données extraites des LOD, et accompagnées d'ontologies biomédicales a permis de génerer un modèle décrivant les gènes responsables de déficiences

intellectuelles. La méthodologie proposée ici requiert néanmoins la sélection et l'intégration des LOD en utilisant un modèle de données abstrait instancié et mis en correspondance avec les LOD. Cependant, ce modèle pourrait être remplacé par une ou plusieurs ontologies biomédicales de haut niveau, formalisant l'ensemble des connaissances sur les gènes étudiés et dont les concepts seraient instanciés par des LOD. Cela permettrait de réduire considérablement le travail nécessaire pour la sélection et l'intégration des données. Ainsi, les différentes classes et propriétés des LOD et ontologies pourraient apparaître telles quelles dans les théories produites, et leur sémantique formelle pourrait être respectée via les règles d'inférences fournies à l'algorithme de PLI. Les règles d'une telle théorie pourraient alors être plus facilement utilisées pour interroger les LOD, et ainsi de vérifier la précision des règles. Cette vérification pourrait permettre de découvrir de nouveaux exemples négatifs qui pourrait être utilisés pour produire une théorie plus spécifique, ou à terme, être proposés comme gènes candidats à des experts des déficiences intellectuelles.

A

Classification des déficiences intellectuelles

Classification en 5 classes des déficiences intellectuelles à partir d'une liste de maladies établie par [Gilissen et al., 2014].

Maladies Métaboliques

OMIM ID	Titre OMIM	GENE	Classe DO
OMIM_203750	ALPHA-METHYLACETOACETIC ACIDURIA	ACAT1	GeneticDisease
OMIM_300387	MENTAL RETARDATION, X-LINKED 63	ACSL4	DiseaseOfMentalHealth
OMIM_609924	AMINOACYLASE 1 DEFICIENCY	ACY1	null
OMIM_103050	ADENYLOSUCCINASE DEFICIENCY	ADSL	null
OMIM_208400	ASPARTYLGLUCOSAMINURIA	AGA	DiseaseOfMetabolism
OMIM_219150	CUTIS LAXA, AUTOSOMAL RECESSIVE, TYPE IIIA	ALDH18A1	null
OMIM_270200	SJOGREN-LARSSON SYNDROME	ALDH3A2	GeneticDisease
OMIM_271980	SUCCINIC SEMIALDEHYDE DEHYDROGENASE DEFICIENCY	ALDH5A1	null
OMIM_266100	EPILEPSY, PYRIDOXINE-DEPENDENT	ALDH7A1	null
OMIM_229600	FRUCTOSE INTOLERANCE, HEREDITARY	ALDOB	null
OMIM_608540	CONGENITAL DISORDER OF GLYCOSYLATION, TYPE Ik	ALG1	DiseaseOfMetabolism
OMIM_607143	CONGENITAL DISORDER OF GLYCOSYLATION, TYPE Ig	ALG12	DiseaseOfMetabolism
OMIM_300884	EPILEPTIC ENCEPHALOPATHY, EARLY INFANTILE, 36	ALG13	null
OMIM_601110	CONGENITAL DISORDER OF GLYCOSYLATION, TYPE Id	ALG3	DiseaseOfMetabolism
OMIM_603147	CONGENITAL DISORDER OF GLYCOSYLATION, TYPE Ic	ALG6	DiseaseOfMetabolism
OMIM_605899	GLYCINE ENCEPHALOPATHY	AMT	DiseaseOfMetabolism
OMIM_250100	METACHROMATIC LEUKODYSTROPHY	ARSA	NervousSystemDisease, DiseaseOfMetabolism
OMIM_228000	FARBER LIPOGRANULOMATOSIS	ASAH1	DiseaseOfMetabolism
OMIM_207900	ARGININOSUCCINIC ACIDURIA	ASL	DiseaseOfMetabolism
OMIM_271900	CANAVAN DISEASE	ASPA	NervousSystemDisease
OMIM_215700	CITRULLINEMIA, CLASSIC	ASS1	DiseaseOfMetabolism
OMIM_248600	MAPLE SYRUP URINE DISEASE	BCKDHB	DiseaseOfMetabolism
OMIM_253260	BIOTINIDASE DEFICIENCY	BTD	DiseaseOfMetabolism
OMIM_613227	CEREBELLAR ATAXIA, MENTAL RETARDATION, AND DYSEQUILIBRIUM SYNDROME 3	CA8	null
OMIM_236200	HOMOCYSTINURIA DUE TO CYSTATHIONINE BETA-SYNTHASE DEFICIENCY	CBS	null
OMIM_237300	CARBAMOYL PHOSPHATE SYNTHETASE I DEFICIENCY, HYPERAMMONEMIA DUE TO	CPS1	DiseaseOfMetabolism
OMIM_213700	CEREBROTENDINOUS XANTHOMATOSIS	CYP27A1	DiseaseOfMetabolism
OMIM_615030	SPASTIC PARAPLEGIA 56, AUTOSOMAL RECESSIVE	CYP2U1	null
OMIM_600721	D-2-HYDROXYGLUTARIC ACIDURIA 1	D2HGDH	DiseaseOfMetabolism
OMIM_615033	SPASTIC PARAPLEGIA 54, AUTOSOMAL RECESSIVE	DDHD2	null
OMIM_270400	SMITH-LEMLI-OPITZ SYNDROME	DHCR7	DiseaseOfMetabolism
OMIM_608093	CONGENITAL DISORDER OF GLYCOSYLATION, TYPE Ij	DPAGT1	DiseaseOfMetabolism
OMIM_608799	CONGENITAL DISORDER OF GLYCOSYLATION, TYPE Ie	DPM1	DiseaseOfMetabolism
OMIM_613068	NEURODEGENERATION DUE TO CEREBRAL FOLATE TRANSPORT DEFICIENCY	FOLR1	null
OMIM_252010	MITOCHONDRIAL COMPLEX I DEFICIENCY	FOXRED1	null
OMIM_229100	GLUTAMATE FORMIMINOTRANSFERASE DEFICIENCY	FTCD	null
OMIM_309549	MENTAL RETARDATION, X-LINKED 9	FTSJ1	null
OMIM_230000	FUCOSIDOSIS	FUCA1	DiseaseOfMetabolism
OMIM_230400	GALACTOSEMIA	GALT	DiseaseOfMetabolism
OMIM_612736	CEREBRAL CREATINE DEFICIENCY SYNDROME 2	GAMT	null
OMIM_612718	CEREBRAL CREATINE DEFICIENCY SYNDROME 3	GATM	DiseaseOfMetabolism
OMIM_231000	GAUCHER DISEASE, TYPE III	GBA	DiseaseOfMetabolism
OMIM_231670	GLUTARIC ACIDEMIA I	GCDH	null
OMIM_233910	HYPERPHENYLALANINEMIA, BH4-DEFICIENT, B	GCH1	null
OMIM_307030	GLYCEROL KINASE DEFICIENCY	GK	null
OMIM_606762	HYPERINSULINEMIC HYPOGLYCEMIA, FAMILIAL, 6	GLUD1	DiseaseOfMetabolism
OMIM_253220	MUCOPOLYSACCHARIDOSIS, TYPE VII	GUSB	DiseaseOfMetabolism
OMIM_309801	LINEAR SKIN DEFECTS WITH MULTIPLE CONGENITAL ANOMALIES 1	HCCS	NervousSystemDisease
OMIM_272800	TAY-SACHS DISEASE	HEXA	NervousSystemDisease, DiseaseOfMetabolism
OMIM_252930	MUCOPOLYSACCHARIDOSIS, TYPE IIIC	HGSNAT	DiseaseOfMetabolism

OMIM_300322	LESCH-NYHAN SYNDROME	HPRT1	DiseaseOfMetabolism
OMIM_300220	MENTAL RETARDATION, X-LINKED, SYNDROMIC 10	HSD17B10	null
OMIM_300438	17-BETA-HYDROXYSTEROID DEHYDROGENASE X DEFICIENCY	HSD17B10	null
OMIM_309900	MUCOPOLYSACCHARIDOSIS, TYPE II	IDS	DiseaseOfMetabolism
OMIM_607014	HURLER SYNDROME	IDUA	DiseaseOfMetabolism
OMIM_614202	MENTAL RETARDATION, AUTOSOMAL RECESSIVE 15	MAN1B1	null
OMIM_248500	MANNOSIDOSIS, ALPHA B, LYSOSOMAL	MAN2B1	DiseaseOfMetabolism
OMIM_248510	MANNOSIDOSIS, BETA A, LYSOSOMAL	MANBA	DiseaseOfMetabolism
OMIM_300615	BRUNNER SYNDROME	MAOA	null
OMIM_277400	METHYLMALONIC ACIDURIA AND HOMOCYSTINURIA, cblC TYPE	MMACHC	DiseaseOfMetabolism
OMIM_236250	HOMOCYSTINURIA DUE TO DEFICIENCY OF N(5,10)-METHYLENETETRAHYDROFOLATE REDUCTASE ACTIVITY	MTHFR	null
OMIM_250940	HOMOCYSTINURIA-MEGALOBLASTIC ANEMIA, cblG COMPLEMENTATION TYPE	MTR	null
OMIM_251000	METHYLMALONIC ACIDURIA DUE TO METHYLMALONYL-CoA MUTASE DEFICIENCY	MUT	null
OMIM_300855	OGDEN SYNDROME	NAA10	null
OMIM_309800	MICROPHTHALMIA, SYNDROMIC 1	NAA10	null
OMIM_252920	MUCOPOLYSACCHARIDOSIS, TYPE IIIB	NAGLU	DiseaseOfMetabolism
OMIM_616116	MENTAL RETARDATION, AUTOSOMAL RECESSIVE 46	NDST1	null
OMIM_300831	CK SYNDROME	NSDHL	null
OMIM_308050	CONGENITAL HEMIDYSPLASIA WITH ICHTHYOSIFORM ERYTHRODERMA AND LIMB DEFECTS	NSDHL	null
OMIM_611091	MENTAL RETARDATION, AUTOSOMAL RECESSIVE 5	NSUN2	null
OMIM_300555	DENT DISEASE 2	OCRL	null
OMIM_309000	LOWE OCULOCEREBRORENAL SYNDROME	OCRL	GeneticDisease
OMIM_311250	ORNITHINE TRANSCARBAMYLASE DEFICIENCY, HYPERAMMONEMIA DUE TO	OTC	DiseaseOfMetabolism
OMIM_261600	PHENYLKETONURIA	PAH	DiseaseOfMetabolism
OMIM_606054	PROPIONIC ACIDEMIA	PCCA	DiseaseOfMetabolism
OMIM_312170	PYRUVATE DEHYDROGENASE E1-ALPHA DEFICIENCY	PDHA1	DiseaseOfMetabolism
OMIM_245349	PYRUVATE DEHYDROGENASE E3-BINDING PROTEIN DEFICIENCY	PDHX	null
OMIM_601815	PHOSPHOGLYCERATE DEHYDROGENASE DEFICIENCY	PHGDH	DiseaseOfMetabolism
OMIM_239300	HYPERPHOSPHATASIA WITH MENTAL RETARDATION SYNDROME 1	PIGV	null
OMIM_212065	CONGENITAL DISORDER OF GLYCOSYLATION, TYPE Ia	PMM2	DiseaseOfMetabolism
OMIM_253280	MUSCULAR DYSTROPHY-DYSTROGLYCANOPATHY (CONGENITAL WITH BRAIN AND EYE ANOMALIES), TYPE A, 3	POMGNT1	MusculoskeletalSystemDisease, NervousSystemDisease
OMIM_613151	MUSCULAR DYSTROPHY-DYSTROGLYCANOPATHY (CONGENITAL WITH MENTAL RETARDATION), TYPE B, 3	POMGNT1	MusculoskeletalSystemDisease, NervousSystemDisease
OMIM_256730	CEROID LIPOFUSCINOSIS, NEURONAL, 1	PPT1	DiseaseOfMetabolism
OMIM_239500	HYPERPROLINEMIA, TYPE I	PRODH	null
OMIM_300661	PHOSPHORIBOSYLPYROPHOSPHATE SYNTHETASE SUPERACTIVITY	PRPS1	null
OMIM_301835	ARTS SYNDROME	PRPS1	GeneticDisease
OMIM_311070	CHARCOT-MARIE-TOOTH DISEASE, X-LINKED RECESSIVE, 5	PRPS1	MusculoskeletalSystemDisease, NervousSystemDisease
OMIM_249900	METACHROMATIC LEUKODYSTROPHY DUE TO SAPOSIN B DEFICIENCY	PSAP	null
OMIM_261640	HYPERPHENYLALANINEMIA, BH4-DEFICIENT, A	PTS	null
OMIM_252900	MUCOPOLYSACCHARIDOSIS, TYPE IIIA	SGSH	DiseaseOfMetabolism
OMIM_238970	HYPERORNITHINEMIA-HYPERAMMONEMIA-HOMOCITRULLINURIA SYNDROME	SLC25A15	DiseaseOfMetabolism
OMIM_309583	MENTAL RETARDATION, X-LINKED, SYNDROMIC, SNYDER-ROBINSON TYPE	SMS	DiseaseOfMentalHealth
OMIM_612379	CONGENITAL DISORDER OF GLYCOSYLATION, TYPE Iq	SRD5A3	DiseaseOfMetabolism
OMIM_612713	KAHRIZI SYNDROME	SRD5A3	null
OMIM_611090	MENTAL RETARDATION, AUTOSOMAL RECESSIVE 12	ST3GAL3	null
OMIM_615006	EPILEPTIC ENCEPHALOPATHY, EARLY INFANTILE, 15	ST3GAL3	null
OMIM_609056	SALT AND PEPPER DEVELOPMENTAL REGRESSION SYNDROME	ST3GAL5	null
OMIM_615476	EPILEPTIC ENCEPHALOPATHY, EARLY INFANTILE, 18	SZT2	null
OMIM_614020	MENTAL RETARDATION, AUTOSOMAL RECESSIVE 14	TECR	null
OMIM_300872	AUTISM, SUSCEPTIBILITY TO, X-LINKED 6	TMLHE	null
OMIM_611093	MENTAL RETARDATION, AUTOSOMAL RECESSIVE 7	TUSC3	null
OMIM_613161	BETA-UREIDOPROPIONASE DEFICIENCY	UPB1	null

OMIM_276880	UROCANASE DEFICIENCY	UROC1	null
OMIM_300577	MENTAL RETARDATION, X-LINKED 91	ZDHHC15	null
OMIM_300799	MENTAL RETARDATION, X-LINKED, SYNDROMIC, RAYMOND TYPE	ZDHHC9	null

Maladies de la Neurogénèse

OMIM ID	Titre OMIM	GENE	Classe DO
OMIM_608097	PERIVENTRICULAR HETEROTOPIA WITH MICROCEPHALY, AUTOSOMAL RECESSIVE	ARFGEF2	PhysicalDisorder
OMIM_608716	MICROCEPHALY 5, PRIMARY, AUTOSOMAL RECESSIVE	ASPM	PhysicalDisorder
OMIM_610042	CORTICAL DYSPLASIA-FOCAL EPILEPSY SYNDROME	CNTNAP2	null
OMIM_300067	LISSENCEPHALY, X-LINKED, 1	DCX	PhysicalDisorder
OMIM_300049	PERIVENTRICULAR NODULAR HETEROTOPIA 1	FLNA	PhysicalDisorder
OMIM_300321	FG SYNDROME 2	FLNA	GeneticDisease
OMIM_304120	OTOPALATODIGITAL SYNDROME, TYPE II	FLNA	null
OMIM_305620	FRONTOMETAPHYSEAL DYSPLASIA 1	FLNA	null
OMIM_311300	OTOPALATODIGITAL SYNDROME, TYPE I	FLNA	null
OMIM_300623	FRAGILE X TREMOR/ATAXIA SYNDROME	FMR1	null
OMIM_300624	FRAGILE X MENTAL RETARDATION SYNDROME	FMR1	GeneticDisease
OMIM_613454	RETT SYNDROME, CONGENITAL VARIANT	FOXG1	DiseaseOfMentalHealth
OMIM_300912	MENTAL RETARDATION, X-LINKED 98	KIAA2022	null
OMIM_303350	MASA SYNDROME	L1CAM	null
OMIM_304100	CORPUS CALLOSUM, PARTIAL AGENESIS OF, X-LINKED	L1CAM	null
OMIM_307000	HYDROCEPHALUS DUE TO CONGENITAL STENOSIS OF AQUEDUCT OF SYLVIUS	L1CAM	null
OMIM_251200	MICROCEPHALY 1, PRIMARY, AUTOSOMAL RECESSIVE	MCPH1	PhysicalDisorder
OMIM_605013	MICROHYDRANENCEPHALY	NDE1	null
OMIM_614019	LISSENCEPHALY 4	NDE1	null
OMIM_300495	AUTISM, SUSCEPTIBILITY TO, X-LINKED 2	NLGN4X	DiseaseOfMentalHealth
OMIM_312080	PELIZAEUS-MERZBACHER DISEASE	PLP1	NervousSystemDisease
OMIM_312920	SPASTIC PARAPLEGIA 2, X-LINKED	PLP1	NervousSystemDisease
OMIM_249500	MENTAL RETARDATION, AUTOSOMAL RECESSIVE 1	PRSS12	null
OMIM_615760	MICROCEPHALY, PROGRESSIVE, WITH SEIZURES AND CEREBRAL AND CEREBELLAR ATROPHY	QARS	null
OMIM_257320	LISSENCEPHALY 2	RELN	PhysicalDisorder
OMIM_269920	INFANTILE SIALIC ACID STORAGE DISEASE	SLC17A5	null
OMIM_604369	SALLA DISEASE	SLC17A5	DiseaseOfMetabolism
OMIM_613671	MENTAL RETARDATION, ANTERIOR MAXILLARY PROTRUSION, AND STRABISMUS	SOBP	null
OMIM_220500	DEAFNESS, ONYCHODYSTROPHY, OSTEODYSTROPHY, MENTAL RETARDATION, AND SEIZURES SYNDROME	TBC1D24	null
OMIM_605021	MYOCLONIC EPILEPSY, FAMILIAL INFANTILE	TBC1D24	null
OMIM_615338	EPILEPTIC ENCEPHALOPATHY, EARLY INFANTILE, 16	TBC1D24	null
OMIM_610031	CORTICAL DYSPLASIA, COMPLEX, WITH OTHER BRAIN MALFORMATIONS 7	TUBB2B	null
OMIM_604317	MICROCEPHALY 2, PRIMARY, AUTOSOMAL RECESSIVE, WITH OR WITHOUT CORTICAL MALFORMATIONS	WDR62	null

Régulation

OMIM ID	Titre OMIM	GENE	Classe DO
OMIM_608629	JOUBERT SYNDROME 3	AHI1	null
OMIM_615493	MENTAL RETARDATION, AUTOSOMAL RECESSIVE 37	ANK3	null
OMIM_304340	PETTIGREW SYNDROME	AP1S2	null
OMIM_614066	SPASTIC PARAPLEGIA 47, AUTOSOMAL RECESSIVE	AP4B1	null
OMIM_613744	SPASTIC PARAPLEGIA 51, AUTOSOMAL RECESSIVE	AP4E1	null
OMIM_612936	SPASTIC PARAPLEGIA 50, AUTOSOMAL RECESSIVE	AP4M1	null
OMIM_614067	SPASTIC PARAPLEGIA 52, AUTOSOMAL RECESSIVE	AP4S1	null
OMIM_300436	MENTAL RETARDATION, X-LINKED 46	ARHGEF6	DiseaseOfMentalHealth
OMIM_300423	MENTAL RETARDATION, X-LINKED, SYNDROMIC, HEDERA TYPE	ATP6AP2	null

OMIM_304150	OCCIPITAL HORN SYNDROME	ATP7A	null
OMIM_309400	MENKES DISEASE	ATP7A	IntegumentarySystemDisease, DiseaseOfMetabolism
OMIM_210600	SECKEL SYNDROME 1	ATR	GeneticDisease
OMIM_300659	MENTAL RETARDATION, X-LINKED 93	BRWD3	null
OMIM_601005	TIMOTHY SYNDROME	CACNA1C	GeneticDisease
OMIM_615474	PRIMARY ALDOSTERONISM, SEIZURES, AND NEUROLOGIC ABNORMALITIES	CACNA1D	null
OMIM_614256	MENTAL RETARDATION, AUTOSOMAL DOMINANT 10	CACNG2	null
OMIM_612580	MENTAL RETARDATION, AUTOSOMAL DOMINANT 3	CDH15	null
OMIM_604804	MICROCEPHALY 3, PRIMARY, AUTOSOMAL RECESSIVE	CDK5RAP2	PhysicalDisorder
OMIM_616080	MICROCEPHALY 12, PRIMARY, AUTOSOMAL RECESSIVE	CDK6	null
OMIM_300672	EPILEPTIC ENCEPHALOPATHY, EARLY INFANTILE, 2	CDKL5	NervousSystemDisease
OMIM_608393	MICROCEPHALY 6, PRIMARY, AUTOSOMAL RECESSIVE	CENPJ	PhysicalDisorder
OMIM_613676	SECKEL SYNDROME 4	CENPJ	GeneticDisease
OMIM_613823	SECKEL SYNDROME 5	CEP152	null
OMIM_614852	MICROCEPHALY 9, PRIMARY, AUTOSOMAL RECESSIVE	CEP152	null
OMIM_615651	LEUKOENCEPHALOPATHY WITH ATAXIA	CLCN2	null
OMIM_300886	MENTAL RETARDATION, X-LINKED, SYNDROMIC 32	CLIC2	null
OMIM_204200	CEROID LIPOFUSCINOSIS, NEURONAL, 3	CLN3	DiseaseOfMetabolism
OMIM_256731	CEROID LIPOFUSCINOSIS, NEURONAL, 5	CLN5	DiseaseOfMetabolism
OMIM_600143	CEROID LIPOFUSCINOSIS, NEURONAL, 8	CLN8	DiseaseOfMetabolism
OMIM_610003	CEROID LIPOFUSCINOSIS, NEURONAL, 8, NORTHERN EPILEPSY VARIANT	CLN8	DiseaseOfMetabolism
OMIM_614418	FEBRILE SEIZURES, FAMILIAL, 11	CPA6	null
OMIM_607417	MENTAL RETARDATION, AUTOSOMAL RECESSIVE 2	CRBN	null
OMIM_615075	MENTAL RETARDATION, AUTOSOMAL DOMINANT 19	CTNNB1	null
OMIM_615362	CEROID LIPOFUSCINOSIS, NEURONAL, 13	CTSF	null
OMIM_300354	MENTAL RETARDATION, X-LINKED, SYNDROMIC, CABEZAS TYPE	CUL4B	null
OMIM_604364	EPILEPSY, FAMILIAL FOCAL, WITH VARIABLE FOCI 1	DEPDC5	null
OMIM_305000	DYSKERATOSIS CONGENITA, X-LINKED	DKC1	IntegumentarySystemDisease
OMIM_614113	MENTAL RETARDATION, AUTOSOMAL DOMINANT 2	DOCK8	null
OMIM_158600	SPINAL MUSCULAR ATROPHY, LOWER EXTREMITY-PREDOMINANT, 1, AUTOSOMAL DOMINANT	DYNC1H1	null
OMIM_614228	CHARCOT-MARIE-TOOTH DISEASE, AXONAL, TYPE 2O	DYNC1H1	null
OMIM_614563	MENTAL RETARDATION, AUTOSOMAL DOMINANT 13	DYNC1H1	null
OMIM_614104	MENTAL RETARDATION, AUTOSOMAL DOMINANT 7	DYRK1A	null
OMIM_614257	MENTAL RETARDATION, AUTOSOMAL DOMINANT 11	EPB41L1	null
OMIM_611225	SPASTIC PARAPLEGIA 18, AUTOSOMAL RECESSIVE	ERLIN2	null
OMIM_305400	AARSKOG-SCOTT SYNDROME	FGD1	GeneticDisease
OMIM_300623	FRAGILE X TREMOR/ATAXIA SYNDROME	FMR1	null
OMIM_300624	FRAGILE X MENTAL RETARDATION SYNDROME	FMR1	GeneticDisease
OMIM_602081	SPEECH-LANGUAGE DISORDER 1	FOXP2	null
OMIM_300849	MENTAL RETARDATION, X-LINKED 41	GDI1	DiseaseOfMentalHealth
OMIM_615473	EPILEPTIC ENCEPHALOPATHY, EARLY INFANTILE, 17	GNAO1	null
OMIM_312870	SIMPSON-GOLABI-BEHMEL SYNDROME, TYPE 1	GPC3	null
OMIM_219000	FRASER SYNDROME	GRIP1	null
OMIM_615871	EPILEPTIC ENCEPHALOPATHY, EARLY INFANTILE, 24	HCN1	null
OMIM_613925	MEGALENCEPHALIC LEUKOENCEPHALOPATHY WITH SUBCORTICAL CYSTS 2A	HEPACAM	null
OMIM_613926	MEGALENCEPHALIC LEUKOENCEPHALOPATHY WITH SUBCORTICAL CYSTS 2B, REMITTING, WITH OR WITHOUT MENTAL RETARDATION	HEPACAM	null
OMIM_218040	COSTELLO SYNDROME	HRAS	GeneticDisease
OMIM_300706	MENTAL RETARDATION, X-LINKED, SYNDROMIC, TURNER TYPE	HUWE1	null
OMIM_300472	CORPUS CALLOSUM, AGENESIS OF, WITH MENTAL RETARDATION, OCULAR COLOBOMA, AND MICROGNATHIA	IGBP1	null
OMIM_608747	INSULIN-LIKE GROWTH FACTOR I DEFICIENCY	IGF1	null
OMIM_308300	INCONTINENTIA PIGMENTI	IKBKG	IntegumentarySystemDisease

OMIM_612780	SEIZURES, SENSORINEURAL DEAFNESS, ATAXIA, MENTAL RETARDATION, AND ELECTROLYTE IMBALANCE	KCNJ10	null
OMIM_612292	BIRK-BAREL MENTAL RETARDATION DYSMORPHISM SYNDROME	KCNK9	GeneticDisease
OMIM_613720	EPILEPTIC ENCEPHALOPATHY, EARLY INFANTILE, 7	KCNQ2	NervousSystemDisease
OMIM_614959	EPILEPTIC ENCEPHALOPATHY, EARLY INFANTILE, 14	KCNT1	null
OMIM_615005	EPILEPSY, NOCTURNAL FRONTAL LOBE, 5	KCNT1	null
OMIM_300257	DANON DISEASE	LAMP2	DiseaseOfMetabolism
OMIM_222448	DONNAI-BARROW SYNDROME	LRP2	null
OMIM_308205	IFAP SYNDROME WITH OR WITHOUT BRESHECK SYNDROME	MBTPS2	null
OMIM_616789	MENTAL RETARDATION AND DISTINCTIVE FACIAL FEATURES WITH OR WITHOUT CARDIAC DEFECTS	MED13L	null
OMIM_610951	CEROID LIPOFUSCINOSIS, NEURONAL, 7	MFSD8	DiseaseOfMetabolism
OMIM_300000	OPITZ GBBB SYNDROME, TYPE I	MID1	null
OMIM_614019	LISSENCEPHALY 4	NDE1	null
OMIM_162200	NEUROFIBROMATOSIS, TYPE I	NF1	GeneticDisease
OMIM_193520	WATSON SYNDROME	NF1	null
OMIM_601321	NEUROFIBROMATOSIS-NOONAN SYNDROME	NF1	null
OMIM_302350	NANCE-HORAN SYNDROME	NHS	null
OMIM_257220	NIEMANN-PICK DISEASE, TYPE C1	NPC1	DiseaseOfMetabolism, ImmuneSystemDisease
OMIM_607625	NIEMANN-PICK DISEASE, TYPE C2	NPC2	DiseaseOfMetabolism, ImmuneSystemDisease
OMIM_163200	SCHIMMELPENNING-FEUERSTEIN-MIMS SYNDROME	NRAS	null
OMIM_614325	PITT-HOPKINS-LIKE SYNDROME 2	NRXN1	null
OMIM_611091	MENTAL RETARDATION, AUTOSOMAL RECESSIVE 5	NSUN2	null
OMIM_300209	SIMPSON-GOLABI-BEHMEL SYNDROME, TYPE 2	OFD1	null
OMIM_300804	JOUBERT SYNDROME 10	OFD1	null
OMIM_311200	OROFACIODIGITAL SYNDROME I	OFD1	GeneticDisease
OMIM_607432	LISSENCEPHALY 1	PAFAH1B1	PhysicalDisorder
OMIM_300558	MENTAL RETARDATION, X-LINKED 30	PAK3	DiseaseOfMentalHealth
OMIM_300088	EPILEPTIC ENCEPHALOPATHY, EARLY INFANTILE, 9	PCDH19	NervousSystemDisease
OMIM_214100	PEROXISOME BIOGENESIS DISORDER 1A (ZELLWEGER)	PEX1	DiseaseOfMetabolism
OMIM_266510	PEROXISOME BIOGENESIS DISORDER 3B	PEX12	DiseaseOfMetabolism
OMIM_614863	PEROXISOME BIOGENESIS DISORDER 4B	PEX6	null
OMIM_215100	RHIZOMELIC CHONDRODYSPLASIA PUNCTATA, TYPE 1	PEX7	GeneticDisease
OMIM_614879	PEROXISOME BIOGENESIS DISORDER 9B	PEX7	null
OMIM_613722	EPILEPTIC ENCEPHALOPATHY, EARLY INFANTILE, 12	PLCB1	NervousSystemDisease
OMIM_613402	MICROCEPHALY, SEIZURES, AND DEVELOPMENTAL DELAY	PNKP	NervousSystemDisease
OMIM_305600	FOCAL DERMAL HYPOPLASIA	PORCN	GeneticDisease
OMIM_153480	BANNAYAN-RILEY-RUVALCABA SYNDROME	PTEN	GeneticDisease
OMIM_158350	COWDEN SYNDROME 1	PTEN	GeneticDisease
OMIM_605309	MACROCEPHALY/AUTISM SYNDROME	PTEN	null
OMIM_151100	LEOPARD SYNDROME 1	PTPN11	GeneticDisease
OMIM_163950	NOONAN SYNDROME 1	PTPN11	GeneticDisease
OMIM_614192	MACROCEPHALY, MACROSOMIA, AND FACIAL DYSMORPHISM SYNDROME	RNF135	null
OMIM_300844	MENTAL RETARDATION, X-LINKED 19	RPS6KA3	DiseaseOfMentalHealth
OMIM_303600	COFFIN-LOWRY SYNDROME	RPS6KA3	GeneticDisease
OMIM_607208	EPILEPTIC ENCEPHALOPATHY, EARLY INFANTILE, 6	SCN1A	NervousSystemDisease
OMIM_613721	EPILEPTIC ENCEPHALOPATHY, EARLY INFANTILE, 11	SCN2A	NervousSystemDisease
OMIM_614306	COGNITIVE IMPAIRMENT WITH OR WITHOUT CEREBELLAR ATAXIA	SCN8A	null
OMIM_614558	EPILEPTIC ENCEPHALOPATHY, EARLY INFANTILE, 13	SCN8A	null
OMIM_214800	CHARGE SYNDROME	SEMA3E	null
OMIM_300434	STOCCO DOS SANTOS X-LINKED MENTAL RETARDATION SYNDROME	SHROOM4	null
OMIM_182212	SHPRINTZEN-GOLDBERG CRANIOSYNOSTOSIS SYNDROME	SKI	MusculoskeletalSystemDisease
OMIM_615905	EPILEPTIC ENCEPHALOPATHY, EARLY INFANTILE, 25	SLC13A5	null
OMIM_300523	ALLAN-HERNDON-DUDLEY SYNDROME	SLC16A2	GeneticDisease

OMIM_269920	INFANTILE SIALIC ACID STORAGE DISEASE	SLC17A5	null
OMIM_604369	SALLA DISEASE	SLC17A5	DiseaseOfMetabolism
OMIM_222730	DICARBOXYLIC AMINOACIDURIA	SLC1A1	null
OMIM_238970	HYPERORNITHINEMIA-HYPERAMMONEMIA-HOMOCITRULLINURIA SYNDROME	SLC25A15	DiseaseOfMetabolism
OMIM_601042	DYSTONIA 9	SLC2A1	null
OMIM_606777	GLUT1 DEFICIENCY SYNDROME 1	SLC2A1	null
OMIM_612126	GLUT1 DEFICIENCY SYNDROME 2	SLC2A1	null
OMIM_300896	CONGENITAL DISORDER OF GLYCOSYLATION, TYPE IIm	SLC35A2	null
OMIM_615553	ARTHROGRYPOSIS, MENTAL RETARDATION, AND SEIZURES	SLC35A3	null
OMIM_266265	CONGENITAL DISORDER OF GLYCOSYLATION, TYPE IIc	SLC35C1	GeneticDisease, DiseaseOfMetabolism
OMIM_229050	FOLATE MALABSORPTION, HEREDITARY	SLC46A1	null
OMIM_300352	CEREBRAL CREATINE DEFICIENCY SYNDROME 1	SLC6A8	null
OMIM_300243	MENTAL RETARDATION, X-LINKED, SYNDROMIC, CHRISTIANSON TYPE	SLC9A6	null
OMIM_300590	CORNELIA DE LANGE SYNDROME 2	SMC1A	null
OMIM_610759	CORNELIA DE LANGE SYNDROME 3	SMC3	null
OMIM_613671	MENTAL RETARDATION, ANTERIOR MAXILLARY PROTRUSION, AND STRABISMUS	SOBP	null
OMIM_613477	EPILEPTIC ENCEPHALOPATHY, EARLY INFANTILE, 5	SPTAN1	NervousSystemDisease
OMIM_612703	MICROCEPHALY 7, PRIMARY, AUTOSOMAL RECESSIVE	STIL	PhysicalDisorder
OMIM_220500	DEAFNESS, ONYCHODYSTROPHY, OSTEODYSTROPHY, MENTAL RETARDATION, AND SEIZURES SYNDROME	TBC1D24	null
OMIM_605021	MYOCLONIC EPILEPSY, FAMILIAL INFANTILE	TBC1D24	null
OMIM_615338	EPILEPTIC ENCEPHALOPATHY, EARLY INFANTILE, 16	TBC1D24	null
OMIM_304700	MOHR-TRANEBJAERG SYNDROME	TIMM8A	null
OMIM_204500	CEROID LIPOFUSCINOSIS, NEURONAL, 2	TPP1	DiseaseOfMetabolism
OMIM_613192	MENTAL RETARDATION, AUTOSOMAL RECESSIVE 13	TRAPPC9	DiseaseOfMentalHealth
OMIM_191100	TUBEROUS SCLEROSIS 1	TSC1	GeneticDisease
OMIM_607341	FOCAL CORTICAL DYSPLASIA OF TAYLOR	TSC1	null
OMIM_613254	TUBEROUS SCLEROSIS 2	TSC2	GeneticDisease
OMIM_300210	MENTAL RETARDATION, X-LINKED 58	TSPAN7	DiseaseOfMentalHealth
OMIM_611603	LISSENCEPHALY 3	TUBA1A	PhysicalDisorder
OMIM_610031	CORTICAL DYSPLASIA, COMPLEX, WITH OTHER BRAIN MALFORMATIONS 7	TUBB2B	null
OMIM_611093	MENTAL RETARDATION, AUTOSOMAL RECESSIVE 7	TUSC3	null
OMIM_300860	MENTAL RETARDATION, X-LINKED, SYNDROMIC, NASCIMENTO TYPE	UBE2A	null
OMIM_244450	KAUFMAN OCULOCEREBROFACIAL SYNDROME	UBE3B	null
OMIM_243800	JOHANSON-BLIZZARD SYNDROME	UBR1	GeneticDisease
OMIM_300676	MENTAL RETARDATION, X-LINKED, SYNDROMIC 14	UPF3B	null
OMIM_224050	CEREBELLAR ATAXIA, MENTAL RETARDATION, AND DYSEQUILIBRIUM SYNDROME 1	VLDLR	null
OMIM_216550	COHEN SYNDROME	VPS13B	null
OMIM_300894	NEURODEGENERATION WITH BRAIN IRON ACCUMULATION 5	WDR45	null
OMIM_614322	SPINOCEREBELLAR ATAXIA, AUTOSOMAL RECESSIVE 12	WWOX	null
OMIM_616211	EPILEPTIC ENCEPHALOPATHY, EARLY INFANTILE, 28	WWOX	null
OMIM_300577	MENTAL RETARDATION, X-LINKED 91	ZDHHC15	null
OMIM_300799	MENTAL RETARDATION, X-LINKED, SYNDROMIC, RAYMOND TYPE	ZDHHC9	null

Régulation de l’expression génétique

OMIM ID	Titre OMIM	GENE	Classe DO
OMIM_615873	HELSMOORTEL-VAN DER AA SYNDROME	ADNP	null
OMIM_309548	MENTAL RETARDATION, X-LINKED, ASSOCIATED WITH FRAGILE SITE FRAXE	AFF2	null
OMIM_148050	KBG SYNDROME	ANKRD11	Syndrome
OMIM_614607	COFFIN-SIRIS SYNDROME 2	ARID1A	null
OMIM_614562	MOVED TO 135900	ARID1B	null
OMIM_300004	CORPUS CALLOSUM, AGENESIS OF, WITH ABNORMAL GENITALIA	ARX	null
OMIM_300215	LISSENCEPHALY, X-LINKED, 2	ARX	PhysicalDisorder

OMIM_300419	MENTAL RETARDATION, X-LINKED, WITH OR WITHOUT SEIZURES, ARX-RELATED	ARX	null
OMIM_308350	EPILEPTIC ENCEPHALOPATHY, EARLY INFANTILE, 1	ARX	NervousSystemDisease
OMIM_309510	PARTINGTON X-LINKED MENTAL RETARDATION SYNDROME	ARX	Syndrome
OMIM_301040	ALPHA-THALASSEMIA/MENTAL RETARDATION SYNDROME, X-LINKED	ATRX	GeneticDisease
OMIM_309580	MENTAL RETARDATION-HYPOTONIC FACIES SYNDROME, X-LINKED, 1	ATRX	null
OMIM_300166	MICROPHTHALMIA, SYNDROMIC 2	BCOR	NervousSystemDisease
OMIM_614756	CEREBELLAR ATAXIA, NONPROGRESSIVE, WITH MENTAL RETARDATION	CAMTA1	null
OMIM_608443	MENTAL RETARDATION, AUTOSOMAL RECESSIVE 3	CC2D1A	null
OMIM_615369	EPILEPTIC ENCEPHALOPATHY, CHILDHOOD-ONSET	CHD2	null
OMIM_180849	RUBINSTEIN-TAYBI SYNDROME 1	CREBBP	GeneticDisease
OMIM_615828	MENTAL RETARDATION, AUTOSOMAL DOMINANT 24	DEAF1	null
OMIM_610253	KLEEFSTRA SYNDROME	EHMT1	null
OMIM_613684	RUBINSTEIN-TAYBI SYNDROME 2	EP300	GeneticDisease
OMIM_613454	RETT SYNDROME, CONGENITAL VARIANT	FOXG1	DiseaseOfMentalHealth
OMIM_600791	DEAFNESS, AUTOSOMAL RECESSIVE 4, WITH ENLARGED VESTIBULAR AQUEDUCT	FOXI1	NervousSystemDisease
OMIM_613670	MENTAL RETARDATION WITH LANGUAGE IMPAIRMENT AND WITH OR WITHOUT AUTISTIC FEATURES	FOXP1	null
OMIM_615074	MENTAL RETARDATION, AUTOSOMAL DOMINANT 18	GATAD2B	null
OMIM_309541	METHYLMALONIC ACIDEMIA AND HOMOCYSTEINEMIA, cblX TYPE	HCFC1	null
OMIM_600430	CHROMOSOME 2q37 DELETION SYNDROME	HDAC4	null
OMIM_300882	CORNELIA DE LANGE SYNDROME 5	HDAC8	null
OMIM_309585	WILSON-TURNER X-LINKED MENTAL RETARDATION SYNDROME	HDAC8	null
OMIM_601536	ATHABASKAN BRAINSTEM DYSGENESIS SYNDROME	HOXA1	GeneticDisease
OMIM_610443	KOOLEN-DE VRIES SYNDROME	KANSL1	null
OMIM_130650	BECKWITH-WIEDEMANN SYNDROME	KCNQ1OT1	GeneticDisease
OMIM_300534	MENTAL RETARDATION, X-LINKED, SYNDROMIC, CLAES-JENSEN TYPE	KDM5C	null
OMIM_300867	KABUKI SYNDROME 2	KDM6A	null
OMIM_156200	MENTAL RETARDATION, AUTOSOMAL DOMINANT 1	MBD5	null
OMIM_105830	ANGELMAN SYNDROME	MECP2	GeneticDisease
OMIM_300055	MENTAL RETARDATION, X-LINKED, SYNDROMIC 13	MECP2	null
OMIM_300260	LUBS X-LINKED MENTAL RETARDATION SYNDROME	MECP2	null
OMIM_300673	ENCEPHALOPATHY, NEONATAL SEVERE, DUE TO MECP2 MUTATIONS	MECP2	null
OMIM_312750	RETT SYNDROME	MECP2	DiseaseOfMentalHealth
OMIM_300895	OHDO SYNDROME, X-LINKED	MED12	null
OMIM_305450	OPITZ-KAVEGGIA SYNDROME	MED12	GeneticDisease
OMIM_309520	LUJAN-FRYNS SYNDROME	MED12	null
OMIM_613668	MICROCEPHALY, POSTNATAL PROGRESSIVE, WITH SEIZURES AND BRAIN ATROPHY	MED17	null
OMIM_614249	MENTAL RETARDATION, AUTOSOMAL RECESSIVE 18	MED23	null
OMIM_613443	MENTAL RETARDATION, AUTOSOMAL DOMINANT 20	MEF2C	null
OMIM_164280	FEINGOLD SYNDROME 1	MYCN	null
OMIM_602535	MARSHALL-SMITH SYNDROME	NFIX	null
OMIM_614753	SOTOS SYNDROME 2	NFIX	null
OMIM_122470	CORNELIA DE LANGE SYNDROME 1	NIPBL	GeneticDisease
OMIM_217095	CONOTRUNCAL HEART MALFORMATIONS	NKX2-6	PhysicalDisorder
OMIM_117550	SOTOS SYNDROME 1	NSD1	GeneticDisease
OMIM_301900	BORJESON-FORSSMAN-LEHMANN SYNDROME	PHF6	GeneticDisease
OMIM_300263	SIDERIUS X-LINKED MENTAL RETARDATION SYNDROME	PHF8	null
OMIM_613038	PITUITARY HORMONE DEFICIENCY, COMBINED, 1	POU1F1	null
OMIM_309500	RENPENNING SYNDROME 1	PQBP1	DiseaseOfMentalHealth
OMIM_612067	DYSTONIA 16	PRKRA	null
OMIM_182290	SMITH-MAGENIS SYNDROME	RAI1	null
OMIM_612313	GLASS SYNDROME	SATB2	null
OMIM_601358	NICOLAIDES-BARAITSER SYNDROME	SMARCA2	null

OMIM_614609	COFFIN-SIRIS SYNDROME 4	SMARCA4	null
OMIM_614608	COFFIN-SIRIS SYNDROME 3	SMARCB1	null
OMIM_300123	MENTAL RETARDATION, X-LINKED, WITH PANHYPOPITUITARISM	SOX3	null
OMIM_188400	DIGEORGE SYNDROME	TBX1	ImmuneSystemDisease
OMIM_192430	VELOCARDIOFACIAL SYNDROME	TBX1	GeneticDisease
OMIM_610954	PITT-HOPKINS SYNDROME	TCF4	null
OMIM_259050	PRIMROSE SYNDROME	ZBTB20	null
OMIM_235730	MOWAT-WILSON SYNDROME	ZEB2	null
OMIM_616083	MENTAL RETARDATION, AUTOSOMAL DOMINANT 30	ZMYND11	null
OMIM_300803	MENTAL RETARDATION, X-LINKED 97	ZNF711	null
OMIM_300498	MENTAL RETARDATION, X-LINKED 45	ZNF81	null

Maladies synaptiques

OMIM ID	Titre OMIM	GENE	Classe DO
OMIM_300607	EPILEPTIC ENCEPHALOPATHY, EARLY INFANTILE, 8	ARHGEF9	NervousSystemDisease
OMIM_300422	FG SYNDROME 4	CASK	GeneticDisease
OMIM_300749	MENTAL RETARDATION AND MICROCEPHALY WITH PONTINE AND CEREBELLAR HYPOPLASIA	CASK	null
OMIM_600513	EPILEPSY, NOCTURNAL FRONTAL LOBE, 1	CHRNA4	null
OMIM_300850	MENTAL RETARDATION, X-LINKED 90	DLG3	null
OMIM_300623	FRAGILE X TREMOR/ATAXIA SYNDROME	FMR1	null
OMIM_300624	FRAGILE X MENTAL RETARDATION SYNDROME	FMR1	GeneticDisease
OMIM_615744	EPILEPTIC ENCEPHALOPATHY, EARLY INFANTILE, 19	GABRA1	null
OMIM_612269	EPILEPSY, CHILDHOOD ABSENCE, SUSCEPTIBILITY TO, 5	GABRB3	null
OMIM_611277	GENERALIZED EPILEPSY WITH FEBRILE SEIZURES PLUS, TYPE 3	GABRG2	null
OMIM_300699	MENTAL RETARDATION, X-LINKED, SYNDROMIC, WU TYPE	GRIA3	null
OMIM_616204	SPINOCEREBELLAR ATAXIA, AUTOSOMAL RECESSIVE 18	GRID2	null
OMIM_611092	MENTAL RETARDATION, AUTOSOMAL RECESSIVE 6	GRIK2	null
OMIM_614254	MENTAL RETARDATION, AUTOSOMAL DOMINANT 8	GRIN1	null
OMIM_245570	EPILEPSY, FOCAL, WITH SPEECH DISORDER AND WITH OR WITHOUT MENTAL RETARDATION	GRIN2A	NervousSystemDisease
OMIM_613970	MENTAL RETARDATION, AUTOSOMAL DOMINANT 6	GRIN2B	null
OMIM_616139	EPILEPTIC ENCEPHALOPATHY, EARLY INFANTILE, 27	GRIN2B	null
OMIM_300143	MENTAL RETARDATION, X-LINKED 21	IL1RAPL1	DiseaseOfMentalHealth
OMIM_614213	NEUROPATHY, HEREDITARY SENSORY, TYPE IIC	KIF1A	null
OMIM_614255	MENTAL RETARDATION, AUTOSOMAL DOMINANT 9	KIF1A	null
OMIM_612581	MENTAL RETARDATION, AUTOSOMAL DOMINANT 4	KIRREL3	null
OMIM_300486	MENTAL RETARDATION, X-LINKED, WITH CEREBELLAR HYPOPLASIA AND DISTINCTIVE FACIAL APPEARANCE	OPHN1	DiseaseOfMentalHealth
OMIM_300271	MENTAL RETARDATION, X-LINKED 72	RAB39B	null
OMIM_311510	WAISMAN SYNDROME	RAB39B	null
OMIM_606232	PHELAN-MCDERMID SYNDROME	SHANK3	null
OMIM_613950	SCHIZOPHRENIA 15	SHANK3	null
OMIM_222730	DICARBOXYLIC AMINOACIDURIA	SLC1A1	null
OMIM_612164	EPILEPTIC ENCEPHALOPATHY, EARLY INFANTILE, 4	STXBP1	NervousSystemDisease
OMIM_300491	EPILEPSY, X-LINKED, WITH VARIABLE LEARNING DISABILITIES AND BEHAVIOR DISORDERS	SYN1	null
OMIM_612621	MENTAL RETARDATION, AUTOSOMAL DOMINANT 5	SYNGAP1	null
OMIM_300802	MENTAL RETARDATION, X-LINKED 96	SYP	null

B

Liste de gènes négatifs pour la caractérisation des gènes responsables de déficiences intellectuelles

NCBI Gene ID	Symbole de gène HGNC
10002	NR2E3
10020	GNE
10049	DNAJB6
10058	ABCB6
10060	ABCC9
10102	TSFM
10210	TOPORS
10461	MERTK
10483	SEC23B
1050	CEBPA
10560	SLC19A2
10594	PRPF8
1073	CFL2
11093	ADAMTS13
11155	LDB3
123016	TTC8
1258	CNGB1
1259	CNGA1
125972	CALR3
1272	CNTN1
1291	COL6A1
1292	COL6A2
1293	COL6A3
133522	PPARGC1B
1410	CRYAB
1411	CRYBA1
1413	CRYBA4
1417	CRYBB3
146059	CDAN1
155	ADRB3
1604	CD55
1605	DAG1
164656	TMPRSS6
1674	DES
1719	DHFR
176	ACAN
1785	DNM2
181	AGRP
1824	DSC2
1829	DSG2
1832	DSP
2010	EMD
203859	ANO5

NCBI Gene ID	Symbole de gène HGNC
2070	EYA4
212	ALAS2
2120	ETV6
2147	F2
2153	F5
2158	F9
2162	F13A1
2178	FANCE
2187	FANCB
2188	FANCF
2189	FANCG
22	ABCB7
2273	FHL1
22954	TRIM32
23092	ARHGAP26
2318	FLNC
2322	FLT3
23345	SYNE1
23365	ARHGEF12
23418	CRB1
23424	TDRD7
23479	ISCU
24	ABCA4
24148	PRPF6
2532	DARC
25821	MTO1
26090	ABHD12
26121	PRPF31
2623	GATA1
2651	GCNT2
2694	GIF
2700	GJA3
2729	GCLC
274	BIN1
282996	RBM20
285489	DOK7
291	SLC25A4
2979	GUCA1B
3026	HABP2
3040	HBA2
3098	HK1
3273	HRG
3299	HSF4

NCBI Gene ID	Symbole de gène HGNC
3339	HSPG2
338557	O3FAR1
3386	ICAM4
3420	IDH3B
3458	IFNG
346007	EYS
35	ACADS
358	AQP1
360	AQP3
3614	IMPDH1
3728	JUP
375790	AGRN
3792	KEL
387700	SLC16A12
388939	C2orf71
390594	KBTBD13
3921	RPSA
4000	LMNA
4026	LPP
4059	BCAM
4088	SMAD3
4117	MAK
4148	MATN3
4159	MC3R
4291	MLF1
43	ACHE
4507	MTAP
4534	MTM1
4607	MYBPC3
4618	MYF6
462	SERPINC1
4624	MYH6
4625	MYH7
4633	MYL2
4634	MYL3
4647	MYO7A
4703	NEB
4869	NPM1
4891	SLC11A2
4976	OPA1
4990	SIX6
50939	IMPG2
50943	FOXP3

NCBI Gene ID	Symbole de gène HGNC
51067	YARS2
51218	GLRX5
5122	PCSK1
51251	NT5C3
51422	PRKAG2
5148	PDE6G
5158	PDE6B
5167	ENPP1
51738	GHRL
5224	PGAM2
5318	PKP2
5350	PLN
5443	POMC
5468	PPARG
54829	ASPN
54968	TMEM70
54977	SLC25A38
55033	FKBP14
55120	FANCL
55214	LEPREL1
55215	FANCI
55750	AGK
55975	KLHL7
5624	PROC
5627	PROS1
5663	PSEN1
5664	PSEN2
5696	PSMB8
570	BAAT
57104	PNPLA2
57158	JPH2
57190	SEPN1
57505	AARS2
57697	FANCM
57798	GATAD1
58	ACTA1
5889	RAD51C
5914	RARA
5972	REN
6010	RHO
6100	RP9
6101	RP1
6102	RP2

NCBI Gene ID	Symbole de gène HGNC
6125	RPL5
613	BCR
6135	RPL11
6204	RPS10
6208	RPS14
6223	RPS19
6229	RPS24
6231	RPS26
6261	RYR1
6262	RYR2
631	BFSP1
6331	SCN5A
64218	SEMA4A
64419	MTMR14
6443	SGCB
6444	SGCD
6492	SIM1
6513	SLC2A1
6563	SLC14A1
675	BRCA2
682	BSG
6901	TAZ
70	ACTC1
7043	TGFB3
7056	THBD
7084	TK2
7112	TMPO
7134	TNNC1
7137	TNNI3
7138	TNNT1
7139	TNNT2
7168	TPM1
7169	TPM2
720	C4A
7273	TTN
7287	TULP1
729920	ISPD
7350	UCP1
7351	UCP2
7352	UCP3
7399	USH2A
7401	CLRN1
7414	VCL
7415	VCP
7439	BEST1
7555	CNBP
768206	PRCD
79188	TMEM43
79443	FYCO1
79728	PALB2
79784	MYH14
79947	DHDDS
80207	OPA3
8021	NUP214
8028	MLLT10
8029	CUBN
80324	PUS1
8048	CSRP3
8106	PABPN1
825	CAPN3
8291	DYSF
83552	MFRP
83700	JAM3
83990	BRIP1
84100	ARL6
84140	FAM161A
8431	NR0B2
84464	SLX4
84466	MEGF10
84668	FAM126A
84701	COX4I2
8482	SEMA7A
85366	MYLK2
8557	TCAP
859	CAV3
861	RUNX1
865	CBFB
867	CBL
8706	B3GALNT1
88	ACTN2
8842	PROM1
9131	AIFM1
9150	CTDP1
91624	NEXN
9197	SLC33A1
9401	RECQL4

NCBI Gene ID	Symbole de gène HGNC
9414	TJP2
9429	ABCG2
9445	ITM2B
9499	MYOT
9531	BAG3
9607	CARTPT
9663	LPIN2
977	CD151
9782	MATR3

C

Théories caractérisant les gènes responsables de déficiences intellectuelles

Theories produced on the whole training set for the five experiments

no-GO theory

Rule_id	Rule_text	positive covered	negative covered
1	is_responsible(A) :- gene_in_reaction(A,'Ubiquinol + Acceptor <=> Ubiquinone + Reduced acceptor').	7	0
2	is_responsible(A) :- gene_in_reaction(A,B), gene_protein(A,C), pp_interaction(C,D), pp_interaction(D,C).	6	0
3	is_responsible(A) :- gene_in_reaction(A,B), gene_protein(A,C), pp_interaction(C,P30480*).	7	0
4	is_responsible(A) :- gene_in_reaction(A,B), gene_ch(A,'1').	14	0
5	is_responsible(A) :- gene_in_reaction(A,B), gene_ch(A,x).	15	2
6	is_responsible(A) :- gene_in_pathway(A,'Alanine and aspartate metabolism').	6	1
7	is_responsible(A) :- gene_in_pathway(A,'Valine, leucine and isoleucine degradation').	11	1
8	is_responsible(A) :- gene_ch_strip(A,'22q13').	6	0
9	is_responsible(A) :- gene_in_pathway(A, 'N-Glycan biosynthesis').	8	0
10	is_responsible(A) :- gene_in_pathway(A, 'Formation of transcription-coupled NER (TC-NER) repair complex').	5	0
11	is_responsible(A) :- gene_in_pathway(A, 'Glycosaminoglycan degradation').	8	0

* : 'MHC class I antigen B*42' is the name of the protein whose UP_id is P30480

G1 theory

Rule_id	Rule_text	positive covered	negative covered
1	is_responsible(A) :- gene_in_reaction(A,B), gene_protein(A,C), protein_bp(C,D), subClassOf(D,'single-organism developmental process').	28	2
2	is_responsible(A) :- gene_in_reaction(A,B), gene_protein(A,C), protein_has_domain(C,D), protein_bp(C,'small molecule metabolic process').	35	2
3	is_responsible(A) :- gene_protein(A,B), protein_bp(B,C), subClassOf(C,'organonitrogen compound metabolic process').	23	1
4	is_responsible(A) :- gene_protein(A,B), protein_bp(B,'cerebral cortex development').	7	1
5	is_responsible(A) :- gene_in_reaction(A,B), gene_protein(A,C), protein_bp(C,D), subClassOf(D,'response to chemical stimulus').	17	0
6	is_responsible(A) :- gene_protein(A,B), protein_bp(B,C), subClassOf(C,'adult locomotory behavior').	8	0
7	is_responsible(A) :- gene_protein(A,B), protein_mf(B,'chromatin binding'), protein_bp(B,C), subClassOf(C,'transcription, dna-dependent').	6	1
8	is_responsible(A) :- gene_in_reaction(A,B), gene_ch(A,'1').	14	0
9	is_responsible(A) :- gene_in_reaction(A,B), gene_protein(A,C), protein_cc(C,D), subClassOf(D,'membrane part').	19	1
10	is_responsible(A) :- gene_ch(A,x), gene_protein(A,B), protein_bp(B,C), subClassOf(C,'system development').	10	0

11	is_responsible(A) :- gene_in_reaction(A,B), gene_protein(A,C), protein_bp(C,'small molecule metabolic process'), pp_interaction(C,D).	23	1
12	is_responsible(A) :- gene_protein(A,B), protein_bp(B,'central nervous system development').	13	1
13	is_responsible(A) :- gene_protein(A,B), protein_mf(B,C), subClassOf(C,'protein binding'), protein_bp(B,D), subClassOf(D,'fatty acid oxidation').	7	1
14	is_responsible(A) :- gene_protein(A,B), protein_bp(B,'peroxisome organization').	9	0
15	is_responsible(A) :- gene_protein(A,B), protein_bp(B,C), subClassOf(C,'system process').	6	1
16	is_responsible(A) :- gene_protein(A,B), protein_mf(B,C), subClassOf(C,'cation binding'), protein_bp(B,D), subClassOf(D,'primary metabolic process').	19	0
17	is_responsible(A) :- gene_protein(A,B), protein_bp(B,C), subClassOf(C,'ion transport').	11	0
18	is_responsible(A) :- gene_protein(A,B), protein_bp(B,C), subClassOf(C,'response to uv'), protein_cc(B,nucleoplasm).	7	1
19	is_responsible(A) :- gene_protein(A,B), protein_bp(B,C), subClassOf(C,'anatomical structure morphogenesis'), protein_bp(B,D), subClassOf(D,'cell surface receptor signaling pathway').	8	0
20	is_responsible(A) :- gene_ch_region(A,xp2), gene_protein(A,B), protein_mf(B,C).	10	0
21	is_responsible(A) :- gene_protein(A,B), pp_interaction(B,C), protein_bp(C,'cell division'), protein_cc(C,nucleus).	8	1
22	is_responsible(A) :- gene_protein(A,B), protein_bp(B,C), subClassOf(C,'alpha-amino acid catabolic process').	19	0

G2 theory

Rule_id	Rule_text	positive covered	negative covered
1	is_responsible(A) :- gene_in_reaction(A,B), gene_protein(A,C), protein_bp(C,D), subClassOf(D,'response to chemical stimulus').	28	2
2	is_responsible(A) :- gene_in_reaction(A,B), gene_protein(A,C), protein_bp(C,'small molecule metabolic process'), protein_cc(C,D), subClassOf(D,'organelle membrane').	38	2
3	is_responsible(A) :- gene_protein(A,B), protein_bp(B,C), subClassOf(C,'neurological system process'), protein_cc(B,nucleus).	9	1
4	is_responsible(A) :- gene_protein(A,B), protein_cc(B,lysosome).	19	2
5	is_responsible(A) :- gene_protein(A,B), protein_mf(B,'chromatin binding'), protein_bp(B,C), subClassOf(C,'cellular macromolecule biosynthetic process').	6	1
6	is_responsible(A) :- gene_in_reaction(A,B), gene_protein(A,C), protein_bp(C,D), subClassOf(D,'system development').	19	1
7	is_responsible(A) :- gene_protein(A,B), protein_has_domain(B,C), protein_bp(B,D), subClassOf(D,'organonitrogen compound metabolic process').	23	0

8	is_responsible(A) :- gene_ch(A,x), gene_protein(A,B), protein_bp(B,C), subClassOf(C,'signal transduction').	11	0
9	is_responsible(A) :- gene_protein(A,B), protein_bp(B,C), subClassOf(C,'cell projection morphogenesis').	11	1
10	is_responsible(A) :- gene_in_reaction(A,B), gene_protein(A,C), protein_bp(C,'small molecule metabolic process'), pp_interaction(C,D).	23	1
11	is_responsible(A) :- gene_protein(A,B), protein_bp(B,'central nervous system development').	13	1
12	is_responsible(A) :- gene_protein(A,B), protein_mf(B,'metal ion binding'), protein_bp(B,C), subClassOf(C,'primary metabolic process'), subClassOf(C,'organic substance metabolic process').	13	0
13	is_responsible(A) :- gene_protein(A,B), protein_bp(B,C), subClassOf(C,'monocarboxylic acid catabolic process').	9	1
14	is_responsible(A) :- gene_protein(A,B), protein_bp(B,C), subClassOf(C,'cytoplasmic transport'), subClassOf(C,'protein import').	12	0
15	is_responsible(A) :- gene_protein(A,B), protein_bp(B,'protein n-linked glycosylation via asparagine').	11	0
16	is_responsible(A) :- gene_protein(A,B), protein_bp(B,C), subClassOf(C,'ion transport'), subClassOf(C,'single-organism transport').	11	1
17	is_responsible(A) :- gene_protein(A,B), protein_bp(B,C), subClassOf(C,'response to uv'), protein_cc(B,nucleoplasm).	7	1
18	is_responsible(A) :- gene_protein(A,B), protein_bp(B,C), subClassOf(C,'alpha-amino acid catabolic process').	19	0
19	is_responsible(A) :- gene_protein(A,B), protein_bp(B,C), subClassOf(C,'carbohydrate derivative metabolic process'), protein_cc(B,D), subClassOf(D,'integral to membrane').	13	0

G3 theory

Rule_id	Rule_text	positive covered	negative covered
1	is_responsible(A) :- gene_in_reaction(A,B), gene_protein(A,C), protein_bp(C,'small molecule metabolic process'), protein_cc(C,'mitochondrial inner membrane').	23	1
2	is_responsible(A) :- gene_protein(A,B), protein_cc(B,lysosome).	19	2
3	is_responsible(A) :- gene_protein(A,B), protein_bp(B,C), subClassOf(C,'regulation of nervous system development'), protein_cc(B,nucleus).	12	1
4	is_responsible(A) :- gene_in_reaction(A,B), gene_protein(A,C), protein_bp(C,D), subClassOf(D,'organonitrogen compound catabolic process').	39	2
5	is_responsible(A) :- gene_protein(A,B), protein_mf(B,'chromatin binding'), protein_bp(B,C), subClassOf(C,'nucleic acid metabolic process').	7	1
6	is_responsible(A) :- gene_protein(A,B), protein_bp(B,C), subClassOf(C,'cell projection morphogenesis').	13	1

7	is_responsible(A) :- gene_protein(A,B), protein_bp(B,C), subClassOf(C,'cellular amino acid metabolic process'), subClassOf(C,'metabolic process').	25	2
8	is_responsible(A) :- gene_protein(A,B), protein_cc(B,C), subClassOf(C,'intracellular membrane-bounded organelle'), subClassOf(C,'intracellular organelle part').	15	0
9	is_responsible(A) :- gene_ch(A,x), gene_protein(A,B), protein_bp(B,C), subClassOf(C,'signal transduction').	13	1
10	is_responsible(A) :- gene_in_reaction(A,B), gene_protein(A,C), protein_bp(C,'small molecule metabolic process'), pp_interaction(C,D).	23	1
11	is_responsible(A) :- gene_ch(A,'2'), gene_protein(A,B), protein_mf(B,C), subClassOf(C,'nucleic acid binding').	6	1
12	is_responsible(A) :- gene_protein(A,B), protein_mf(B,'metal ion binding'), protein_mf(B,C), subClassOf(C,'catalytic activity').	17	1
13	is_responsible(A) :- gene_ch_arm(A,'17p'), gene_protein(A,B), protein_cc(B,C), subClassOf(C,'cell part').	6	0
14	is_responsible(A) :- gene_protein(A,B), protein_bp(B,C), subClassOf(C,'protein glycosylation'), protein_bp(B,D), subClassOf(D,'cellular process').	15	2
15	is_responsible(A) :- gene_protein(A,B), protein_bp(B,'ion transport').	8	0
16	is_responsible(A) :- gene_protein(A,B), protein_bp(B,'carbohydrate metabolic process'), protein_bp(B,C), subClassOf(C,'phosphate-containing compound metabolic process'), protein_cc(B,cytosol).	6	0
17	is_responsible(A) :- gene_protein(A,B), protein_bp(B,C), subClassOf(C,'macromolecule catabolic process'), protein_cc(B,D), subClassOf(D,'intracellular membrane-bounded organelle').	15	0
18	is_responsible(A) :- gene_ch_region(A,xp2), gene_protein(A,B), protein_mf(B,C).	10	0

G4 theory

Rule_id	Rule_text	positive covered	negative covered
1	is_responsible(A) :- gene_in_reaction(A,B), gene_protein(A,C), protein_bp(C,'small molecule metabolic process'), protein_cc(C,'mitochondrial inner membrane').	23	1
2	is_responsible(A) :- gene_protein(A,B), protein_cc(B,lysosome).	19	2
3	is_responsible(A) :- gene_protein(A,B), protein_bp(B,C), subClassOf(C,'regulation of nervous system development'), protein_cc(B,nucleus).	13	2
4	is_responsible(A) :- gene_in_reaction(A,B), gene_protein(A,C), protein_bp(C,D), subClassOf(D,'organonitrogen compound catabolic process').	42	2
5	is_responsible(A) :- gene_protein(A,B), protein_mf(B,'chromatin binding'), protein_bp(B,C), subClassOf(C,'cellular nitrogen compound metabolic process').	7	1
6	is_responsible(A) :- gene_protein(A,B), protein_bp(B,C), subClassOf(C,'cell projection morphogenesis').	13	1

7	is_responsible(A) :- gene_protein(A,B), protein_bp(B,C), subClassOf(C,'cellular amino acid metabolic process'), subClassOf(C,'cellular metabolic process').	32	3
8	is_responsible(A) :- gene_protein(A,B), protein_cc(B,C), subClassOf(C,'membrane-bounded organelle'), subClassOf(C,'organelle part').	15	0
9	is_responsible(A) :- gene_ch(A,x), gene_protein(A,B), protein_bp(B,C), subClassOf(C,'signal transduction').	14	1
10	is_responsible(A) :- gene_protein(A,B), protein_bp(B,C), subClassOf(C,'tube formation').	8	0
11	is_responsible(A) :- gene_protein(A,B), protein_mf(B,'metal ion binding'), protein_bp(B,C), subClassOf(C,'organic substance catabolic process').	16	1
12	is_responsible(A) :- gene_ch_region(A,xp2), gene_protein(A,B), protein_mf(B,C).	10	0
13	is_responsible(A) :- gene_in_reaction(A,B), gene_protein(A,C), protein_cc(C,'endoplasmic reticulum membrane').	15	0
14	is_responsible(A) :- gene_protein(A,B), protein_bp(B,C), subClassOf(C,'generation of precursor metabolites and energy'), protein_bp(B,D), subClassOf(D,'macromolecular complex subunit organization').	5	0
15	is_responsible(A) :- gene_protein(A,B), protein_bp(B,'ion transport').	8	0
16	is_responsible(A) :- gene_protein(A,B), protein_mf(B,C), subClassOf(C,'ion binding'), protein_bp(B,'carbohydrate metabolic process').	19	1

D

Réference des classes d'ontologies

TABLE D.1 – Cette table réference les différents codes ATC utilisés et leur nom de classe complet

Code	Label
A02B	Drugs for peptic ulcer and gastro-oesophageal reflux disease
A02BC	Proton pump inhibitors
A04A	Antiemetics and antinauseants
A06A	Drugs for constipation
A07A	Intestinal antiinfectives
B01A	Antithrombotic agents
B03X	Other antianemic preparations
B05X	I.V. solution additives
C01BB03	Tocainide
C03C	High-ceiling diuretics
C05B	Antivaricose therapy
C07A	Beta blocking agents
C08D	Selective calcium channel blockers with direct cardiac effects
C08DB	Benzothiazepine derivatives
C09A	Ace inhibitors, plain
C10A	Lipid modifying agents, plain
G04BE	Drugs used in erectile dysfunction
G04BE04	Yohimbine
H02A	Corticosteroids for systemic use, plain
H02AA03	Desoxycortone
H02AB	Glucocorticoids
H02AB07	Prednisone
J01XA01	Vancomycin
N02A	Opioids
N02B	Other analgesics and antipyretics
N02BE01	Paracetamol / Acetaminophen
N05B	Anxiolytics
N05C	Hypnotics and sedatives
N06BC	Xanthine derivatives
N06BC01	Caffeine
R05D	Cough suppressants, excl. combinations with expectorants
R06A	Antihistamines for systemic use
R06AA	aminoalkyl ethers
R06AA09	Doxylamine
S01A	Antiinfectives
S01AX	Other antiinfectives in ATC

Table D.2 – Cette table réference les différents codes ICD-9-CM utilisés et leur nom de classe complet

Code ICD-9-CM	Label
280-289	**Diseases of the blood and blood-forming organs**
280	Iron deficiency anemias
285.9	Anemia, unspecified
287.5	Thrombocytopenia, unspecified
390-459	**Diseases of the circulatory system**
390-392	Acute rheumatic fever
420-429	Other forms of heart disease
427	Cardiac dysrhythmias
427.3	Atrial fibrillation and flutter
427.31	Atrial fibrillation
428	Heart failure
428.0	Congestive heart failure, unspecified
428.9	Heart failure, unspecified
580-629	**Diseases of the genitourinary system**
580	Acute glomerulonephritis
586	Renal failure, unspecified
599.8	Other specified disorders of urethra and urinary tract
599.9	Unspecified disorder of urethra and urinary tract
710-739	**Diseases of the musculoskletal system and connective tissue**
710	Diffuse diseases of connective tissue
719.4	Pain in joint

Index

Bibliographie

[Abele et al., 2017] Abele, A., McCrae, J. P., Buitelaar, P., Jentzsch, A., and Cyganiak, R. (2017). Linking Open Data cloud diagram 2017. `http://lod-cloud.net/`.

[Agrawal et al., 1993] Agrawal, R., Imieliński, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. In *Acm sigmod record*, volume 22, pages 207–216. ACM.

[Agrawal et al., 1994] Agrawal, R., Srikant, R., et al. (1994). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499.

[Aït-Kaci et al., 1989] Aït-Kaci, H., Boyer, R., Lincoln, P., and Nasr, R. (1989). Efficient implementation of lattice operations. *ACM Transactions on Programming Languages and Systems (TOPLAS)*, 11(1) :115–146.

[Alam et al., 2015] Alam, M., Buzmakov, A., Codocedo, V., and Napoli, A. (2015). Bridging dbpedia categories and dl-concept definitions using formal concept analysis. In *Proceedings of the 4th International Workshop" What can FCA do for Artificial Intelligence ?", FCA4AI 2015, co-located with the International Joint Conference on Artificial Intelligence (IJCAI 2015), Buenos Aires, Argentina.*, volume 1430.

[AMA, 2004] AMA (2004). *International classification of diseases, 9th revision, clinical modification : physician ICD-9-CM, 2005 : volumes 1 and 2, color-coded, illustrated*, volume 1. American Medical Association.

[Antezana et al., 2009] Antezana, E., Kuiper, M., and Mironov, V. (2009). Biological knowledge management : the emerging role of the Semantic Web technologies. *Briefings in Bioinformatics*, 10(4) :392–407.

[Aronson, 2001] Aronson, A. R. (2001). Effective mapping of biomedical text to the umls metathesaurus : the metamap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association.

[Ashburner et al., 2000] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000). Gene Ontology : tool for the unification of biology. *Nature genetics*, 25(1) :25–29.

[Baader, 2003] Baader, F. (2003). *The description logic handbook : Theory, implementation and applications*. Cambridge university press.

[Baader et al., 2005] Baader, F., Horrocks, I., and Sattler, U. (2005). Description logics as ontology languages for the semantic web. In *Mechanizing Mathematical Reasoning*, pages 228–248. Springer.

[Banda et al., 2016] Banda, J. M., Evans, L., Vanguri, R. S., Tatonetti, N. P., Ryan, P. B., and Shah, N. H. (2016). A curated and standardized adverse drug event resource to accelerate drug safety research. *Scientific data*, 3.

[Barabási et al., 2011] Barabási, A.-L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine : a network-based approach to human disease. *Nature reviews genetics*, 12(1) :56.

[Barrell et al., 2009] Barrell, D., Dimmer, E., Huntley, R. P., Binns, D., O'Donovan, C., and Apweiler, R. (2009). The GOA database in 2009–an integrated Gene Ontology Annotation resource. *Nucleic Acids Research*, 37(1) :396–403.

[Beckett and Berners-Lee, 2011] Beckett, D. and Berners-Lee, T. (2011). Turtle - Terse RDF Triple Language. `https://www.w3.org/TeamSubmission/turtle/`.

[Belleau et al., 2008] Belleau, F., Nolin, M.-A., Tourigny, N., Rigault, P., and Morissette, J. (2008). Bio2RDF : Towards a mashup to build bioinformatics knowledge systems. *Journal of Biomedical Informatics*, 41(5) :706 – 716. Semantic Mashup of Biomedical Data.

[Benabderrahmane et al., 2010] Benabderrahmane, S., Smail-Tabbone, M., Poch, O., Napoli, A., and Devignes, M.-D. (2010). Intelligo : a new vector-based semantic similarity measure including annotation origin. *BMC bioinformatics*, 11(1) :1.

[Bender et al., 2005] Bender, M. A., Farach-Colton, M., Pemmasani, G., Skiena, S., and Sumazin, P. (2005). Lowest common ancestors in trees and directed acyclic graphs. *Journal of Algorithms*, 57(2) :75–94.

[Berners-Lee, 1994] Berners-Lee, T. (1994). Rfc 1630. *Universal Resource Identifiers in WWW : A Unifying Syntax for the Expression of Names and Addresses of Objects on the Network as used in the World-Wide Web*.

[Berners-Lee, 2006] Berners-Lee, T. (2006). Linked data-design issues. *http ://www. w3. org/DesignIssues/LinkedData. html*.

[Berners-Lee et al., 2001] Berners-Lee, T., Hendler, J., Lassila, O., et al. (2001). The semantic web. *Scientific american*, 284(5) :28–37.

[Berthold et al., 2009] Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., Ohl, P., Thiel, K., and Wiswedel, B. (2009). KNIME-the Konstanz information miner : version 2.0 and beyond. *ACM SIGKDD explorations Newsletter*, 11(1) :26–31.

[Berthold et al., 2007] Berthold, M. R., Morik, K., and Siebes, A., editors (2007). *Parallel Universes and Local Patterns*, volume 07181 of *Dagstuhl Seminar Proceedings*.

[Birkhoff, 1949] Birkhoff, G. (1949). Théorie et applications des treillis. In *Annales de l'IHP*, volume 11, pages 227–240.

[Bizer et al., 2009] Bizer, C., Heath, T., and Berners-Lee, T. (2009). Linked Data - The Story So Far. *Int. J. Semantic Web Inf. Syst.*, 5(3) :1–22.

[Bodenreider, 2004] Bodenreider, O. (2004). The unified medical language system (umls) : integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1) :D267–D270.

[Bordat, 1986] Bordat, J.-P. (1986). Calcul pratique du treillis de galois d'une correspondance. *Mathématiques et Sciences humaines*, 96 :31–47.

[Bresso, 2013] Bresso, E. (2013). Organisation et exploitation des connaissances sur les réseaux d'intéractions biomoléculaires pour l'étude de l'étiologie des maladies génétiques et la caractérisation des effets secondaires de principes actifs.

[Callahan et al., 2013] Callahan, A., Cruz-Toledo, J., Ansell, P., and Dumontier, M. (2013). Bio2RDF Release 2 : Improved Coverage, Interoperability and Provenance of Life Science Linked Data. In Cimiano, P., Corcho, O., Presutti, V., Hollink, L., and Rudolph, S., editors, *The Semantic Web : Semantics and Big Data*, volume 7882 of *Lecture Notes in Computer Science*, pages 200–212. Springer Berlin Heidelberg.

[Calvanese et al., 2007] Calvanese, D., De Giacomo, G., Lembo, D., Lenzerini, M., and Rosati, R. (2007). Tractable reasoning and efficient query answering in description logics : The dl-lite family. *Journal of Automated reasoning*, 39(3) :385–429.

[Cimiano et al., 2004] Cimiano, P., Hotho, A., Stumme, G., and Tane, J. (2004). Conceptual knowledge processing with formal concept analysis and ontologies. In *International Conference on Formal Concept Analysis*, pages 189–207. Springer.

[Clocksin and Mellish, 2003] Clocksin, W. F. and Mellish, C. S. (2003). *Programming in PROLOG*. Springer Science & Business Media.

[Copeland et al., 1977] Copeland, K. T., Checkoway, H., McMichael, A. J., and Holbrook, R. H. (1977). Bias due to misclassification in the estimation of relative risk. *American journal of epidemiology*, 105(5) :488–495.

[Coulet et al., 2013] Coulet, A., Domenach, F., Kaytoue, M., and Napoli, A. (2013). Using pattern structures for analyzing ontology-based annotations of biomedical data. In *International Conference on Formal Concept Analysis*, pages 76–91. Springer.

[Couto et al., 2007] Couto, F. M., Silva, M. J., and Coutinho, P. M. (2007). Measuring semantic similarity between gene ontology terms. *Data & knowledge engineering*, 61(1) :137–152.

[Cyganiak et al., 2014] Cyganiak, R., Wood, D., and Lanthaler, M. (2014). RDF 1.1 Concepts and Abstract Syntax. `https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/`.

[De Giacomo et al., 2018] De Giacomo, G., Lembo, D., Lenzerini, M., Poggi, A., and Rosati, R. (2018). Using ontologies for semantic data integration. In *A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years.*, pages 187–202.

[Defays, 1977] Defays, D. (1977). An efficient algorithm for a complete link method. *The Computer Journal*, 20(4) :364–366.

[Dice, 1945] Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26(3) :297–302.

[Donnelly, 2006] Donnelly, K. (2006). Snomed-ct : The advanced terminology and coding system for ehealth. *Studies in health technology and informatics*, 121 :279.

[Dumontier et al., 2014] Dumontier, M., Callahan, A., Cruz-Toledo, J., Ansell, P., Emonet, V., Belleau, F., and Droit, A. (2014). Bio2rdf release 3 : a larger connected network of linked data for the life sciences. In *Proceedings of the 2014 International Conference on Posters & Demonstrations Track*, volume 1272, pages 401–404.

[Dunn, 1973] Dunn, J. C. (1973). A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters.

[Edwards and Aronson, 2000] Edwards, I. R. and Aronson, J. K. (2000). Adverse drug reactions : definitions, diagnosis, and management. *The Lancet*, 356(9237) :1255 – 1259.

[Erling, 2012] Erling, O. (2012). Virtuoso, a hybrid rdbms/graph column store. *IEEE Data Eng. Bull.*, 35(1) :3–8.

[Euzenat et al., 2007] Euzenat, J., Shvaiko, P., et al. (2007). *Ontology matching*, volume 18. Springer.

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

[FDA, 2016] FDA (2016). FDA adverse event reporting system (FAERS). `http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/`. Accessed : 2016-10-01.

[Galárraga et al., 2013] Galárraga, L. A., Teflioudi, C., Hose, K., and Suchanek, F. (2013). Amie : association rule mining under incomplete evidence in ontological knowledge bases. In *Proceedings of the 22nd international conference on World Wide Web*, pages 413–422. International World Wide Web Conferences Steering Committee.

[Gandon and Schreiber, 2014] Gandon, F. and Schreiber, A. T. (2014). RDF 1.1 XML syntax. `https://www.w3.org/TR/rdf-syntax-grammar/`.

[Ganter and Kuznetsov, 2001] Ganter, B. and Kuznetsov, S. O. (2001). Pattern structures and their projections. In *International Conference on Conceptual Structures*, pages 129–142. Springer.

[Ganter and Wille, 1997] Ganter, B. and Wille, R. (1997). *Formal Concept Analysis : Mathematical Foundations*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1st edition.

[Gilissen et al., 2014] Gilissen, C., Hehir-Kwa, J. Y., Thung, D. T., van de Vorst, M., van Bon, B. W., Willemsen, M. H., Kwint, M., Janssen, I. M., Hoischen, A., Schenck, A., et al. (2014). Genome sequencing identifies major causes of severe intellectual disability. *Nature*, 511(7509) :344.

[Goh et al., 2007] Goh, K.-I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., and Barabási, A.-L. (2007). The human disease network. *Proceedings of the National Academy of Sciences*, 104(21) :8685–8690.

[Goldberg et al., 1996] Goldberg, R. M., Mabee, J., Chan, L., and Wong, S. (1996). Drug-drug and drug-disease interactions in the ed : analysis of a high-risk population. *The American journal of emergency medicine*, 14(5) :447–450.

[Gómez-Pérez et al., 2006] Gómez-Pérez, A., Fernández-López, M., and Corcho, O. (2006). *Ontological Engineering : with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web*. Springer Science & Business Media.

[Grahame-Smith DG, 1984] Grahame-Smith DG, A. J. (1984). Adverse drug reactions. pages 132–157.

[Gray et al., 2017] Gray, A. J., Goble, C. A., and Jimenez, R. (2017). Bioschemas : From potato salad to protein annotation. In *International Semantic Web Conference (Posters, Demos & Industry Tracks)*.

[Grimm et al., 2011] Grimm, S., Abecker, A., Völker, J., and Studer, R. (2011). *Ontologies and the Semantic Web*, pages 507–579. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Grisoni et al., 2013] Grisoni, R., Bresso, E., Devignes, M.-D., and Smaïl-Tabbone, M. (2013). Méthodologie et outils pour l'extraction de connaissances par Programmation Logique Inductive (PLI) (Poster). In *13ème Conférence Francophone sur l'Extraction et la Gestion des Connaissances- EGC 2013*, Toulouse, France.

[Guney et al., 2016] Guney, E., Menche, J., Vidal, M., and Barábasi, A.-L. (2016). Network-based in silico drug efficacy screening. *Nature communications*, 7 :10331.

[Hamosh et al., 2005] Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., and McKusick, V. A. (2005). Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. *Nucleic acids research*, 33(DATABASE ISS.) :D514–D517.

[Hartigan-Go and Wong, 2000] Hartigan-Go, K. Y. and Wong, J. Q. (2000). Inclusion of therapeutic failures as adverse drug reactions. *Side Effects of Drugs Annual*, 23 :xxvii–xxxiii.

[Hidalgo et al., 2009] Hidalgo, C. A., Blumm, N., Barabási, A.-L., and Christakis, N. A. (2009). A dynamic network approach for the study of human phenotypes. *PLoS computational biology*, 5(4) :e1000353.

[HJ and G, 1994] HJ, L. and G, B. (1994). Understanding and using the medical subject headings (mesh) vocabulary to perform literature searches. *JAMA*, 271(14) :1103–1108.

[Hoehndorf et al., 2015] Hoehndorf, R., Schofield, P. N., and Gkoutos, G. V. (2015). Analysis of the human diseasome using phenotype similarity between common, genetic, and infectious diseases. *Scientific reports*, 5 :10888.

[Hubert and Arabie, 1985] Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2(1) :193–218.

[Hunter et al., 2008] Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., et al. (2008). InterPro : the integrative protein signature database. *Nucleic acids research*, 37(suppl_1) :D211–D215.

[Inlow and Restifo, 2004] Inlow, J. K. and Restifo, L. L. (2004). Molecular and comparative genetics of mental retardation. *Genetics*, 166(2) :835–881.

[Jaccard, 1901] Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bull Soc Vaudoise Sci Nat*, 37 :547–579.

[Jacomy et al., 2014] Jacomy, M., Venturini, T., Heymann, S., and Bastian, M. (2014). Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. *PloS one*, 9(6) :e98679.

[Jones, 1979] Jones, D. S. (1979). *Elementary information theory*. Clarendon Press.

[Jonquet et al., 2011] Jonquet, C., LePendu, P., Falconer, S. M., Coulet, A., Noy, N. F., Musen, M. A., and Shah, N. H. (2011). NCBO resource index : Ontology-based search and mining of biomedical resources. *J. Web Sem.*, 9(3) :316–324.

[Jupp et al., 2014] Jupp, S., Malone, J., Bolleman, J., Brandizi, M., Davies, M., Garcia, L., Gaulton, A., Gehant, S., Laibe, C., Redaschi, N., et al. (2014). The ebi rdf platform : linked open data for the life sciences. *Bioinformatics*, 30(9) :1338–1339.

[Kakar et al., 2016] Kakar, T., Qin, X., Wunnava, S., and Rundensteiner, E. A. (2016). Towards pharmacovigilance using machine learning to identify unknown adverse reactions triggered by drug-drug interaction.

[Kanehisa and Goto, 2000] Kanehisa, M. and Goto, S. (2000). Kegg : kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1) :27–30.

[Kaufman et al., 2010] Kaufman, L., Ayub, M., and Vincent, J. B. (2010). The genetic basis of non-syndromic intellectual disability : a review. *Journal of neurodevelopmental disorders*, 2(4) :182.

[Kaufman and Rousseeuw, 1987] Kaufman, L. and Rousseeuw, P. (1987). *Clustering by means of medoids*. North-Holland.

[Kaufman and Rousseeuw, 2009] Kaufman, L. and Rousseeuw, P. J. (2009). *Finding groups in data : an introduction to cluster analysis*, volume 344. John Wiley & Sons.

[Kaytoue et al., 2015] Kaytoue, M., Codocedo, V., Buzmakov, A., Baixeries, J., Kuznetsov, S. O., and Napoli, A. (2015). Pattern structures and concept lattices for data mining and knowledge processing. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 227–231. Springer.

[Kinjo et al., 2012] Kinjo, A. R., Suzuki, H., Yamashita, R., Ikegawa, Y., Kudou, T., Igarashi, R., Kengaku, Y., Cho, H., Standley, D. M., Nakagawa, A., and Nakamura, H. (2012). Protein Data Bank Japan (PDBj) : maintaining a structural data archive and resource description framework format. *Nucleic Acids Research*, 40(Database-Issue) :453–460.

[Klyne et al., 2004] Klyne, G., Carroll, J. J., and McBride, B. (2004). RDF/XML Syntax Specification (Revised). `https://www.w3.org/TR/2004/REC-rdf-concepts-20040210/`.

[Knobbe et al., 2008] Knobbe, A., Crémilleux, B., Fürnkranz, J., and Scholz, M. (2008). From Local Patterns to Global Models : The LeGo Approach to Data Mining. In *International Workshop From Local Patterns to Global Models co-located with ECML/PKDD'08*, pages 1–16, Antwerp, Belgium.

[Knublauch et al., 2004] Knublauch, H., Fergerson, R. W., Noy, N. F., and Musen, M. A. (2004). The protégé owl plugin : An open development environment for semantic web applications. In McIlraith, S. A., Plexousakis, D., and van Harmelen, F., editors, *The Semantic Web – ISWC 2004*, pages 229–243, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Köhler et al., 2013] Köhler, S., Doelken, S. C., Mungall, C. J., Bauer, S., Firth, H. V., Bailleul-Forestier, I., Black, G. C., Brown, D. L., Brudno, M., Campbell, J., et al. (2013). The human phenotype ontology project : linking molecular biology and disease through phenotype data. *Nucleic acids research*, 42(D1) :D966–D974.

[Köpcke and Rahm, 2010] Köpcke, H. and Rahm, E. (2010). Frameworks for entity matching : A comparison. *Data & Knowledge Engineering*, 69(2) :197–210.

[Kosub, 2016] Kosub, S. (2016). A note on the triangle inequality for the jaccard distance. *arXiv preprint arXiv :1612.02696*.

[Kourie et al., 2009] Kourie, D. G., Obiedkov, S., Watson, B. W., and van der Merwe, D. (2009). An incremental algorithm to construct a lattice of set intersections. *Science of Computer Programming*, 74(3) :128–142.

[Kuhn et al., 2015] Kuhn, M., Letunic, I., Jensen, L. J., and Bork, P. (2015). The SIDER database of drugs and side effects. *Nucleic acids research*, page gkv1075.

[Lakhal and Stumme, 2005] Lakhal, L. and Stumme, G. (2005). Efficient mining of association rules based on formal concept analysis. In *Formal Concept Analysis, Foundations and Applications*, pages 180–195.

[Langfelder et al., 2007] Langfelder, P., Zhang, B., and Horvath, S. (2007). Defining clusters from a hierarchical cluster tree : the dynamic tree cut package for r. *Bioinformatics*, 24(5) :719–720.

[Larson, 2010] Larson, R. R. (2010). Introduction to information retrieval.

[Lauderdale et al., 1993] Lauderdale, D. S., Furner, S. E., Miles, T. P., and Goldberg, J. (1993). Epidemiologic uses of medicare data. *Epidemiologic Reviews*, 15(2) :319–327.

[LePendu et al., 2013] LePendu, P., Iyer, S. V., Bauer-Mehren, A., Harpaz, R., Mortensen, J. M., Podchiyska, T., Ferris, T. A., and Shah, N. H. (2013). Pharmacovigilance using clinical notes. *Clinical pharmacology & therapeutics*, 93(6) :547–555.

[Lillo-Le Louët et al., 2009] Lillo-Le Louët, A., Toussaint, Y., and Villerd, J. (2009). A qualitative approach to signal mining in pharmacovigilance using formal concept analysis. *Studies in health technology and informatics*, 160(Pt 2) :969–973.

[Lin et al., 1998] Lin, D. et al. (1998). An information-theoretic definition of similarity. In *Icml*, volume 98, pages 296–304.

[Lisi, 2008] Lisi, F. A. (2008). Building rules on top of ontologies for the semantic web with inductive logic programming. *Theory and Practice of Logic Programming*, 8(3) :271–300.

[Lloyd, 1982] Lloyd, S. (1982). Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2) :129–137.

[Lord et al., 2003] Lord, P. W., Stevens, R. D., Brass, A., and Goble, C. A. (2003). Investigating semantic similarity measures across the gene ontology : the relationship between sequence and annotation. *Bioinformatics*, 19(10) :1275–1283.

[Lowe et al., 2009] Lowe, H. J., Ferris, T. A., Hernandez, P. M., Weber, S. C., et al. (2009). STRIDE-an integrated standards-based translational research informatics platform. In *AMIA*.

[Luxenburger, 1991] Luxenburger, M. (1991). Implications partielles dans un contexte. *Mathématiques, informatique et sciences humaines*, 29(113) :35–55.

[MacQueen et al., 1967] MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

[Manda et al., 2012] Manda, P., Ozkan, S., Wang, H., McCarthy, F., and Bridges, S. M. (2012). Cross-ontology multi-level association rule mining in the gene ontology. *PloS one*, 7(10) :e47411.

[Mann and Whitney, 1947] Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60.

[Masse, 2011] Masse, M. (2011). *REST API Design Rulebook : Designing Consistent RESTful Web Service Interfaces*. " O'Reilly Media, Inc.".

[McCray et al., 1994] McCray, A. T., Srinivasan, S., and Browne, A. C. (1994). Lexical methods for managing variation in biomedical terminologies. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 235. American Medical Informatics Association.

[McGuinness, 2002] McGuinness, D. L. (2002). Ontologies come of age. *Spinning the semantic web : bringing the World Wide Web to its full potential*, pages 171–194.

[Meyboom et al., 1997] Meyboom, R. H., Egberts, A. C., Edwards, I. R., Hekster, Y. A., de Koning, F. H., and Gribnau, F. W. (1997). Principles of signal detection in pharmacovigilance. *Drug safety*, 16(6) :355–365.

[Mitchell et al., 1994] Mitchell, J. B., Bubolz, T., Paul, J. E., Pashos, C. L., Escarce, J. J., Muhlbaier, L. H., Wiesman, J. M., Young, W. W., Epstein, R. S., and Javitt, J. C. (1994). Using medicare claims for outcomes research. *Medical care*, pages JS38–JS51.

[Muggleton, 1991] Muggleton, S. (1991). Inductive Logic Programming. *New Generation Computing*, 8(4) :295–318.

[Muggleton and Raedt, 1994] Muggleton, S. and Raedt, L. D. (1994). Inductive logic programming : Theory and methods. *The Journal of Logic Programming*, 19(20) :629–679.

[Müller et al., 2004] Müller, H.-M., Kenny, E. E., and Sternberg, P. W. (2004). Textpresso : an ontology-based information retrieval and extraction system for biological literature. *PLoS biology*, 2(11) :e309.

[Mungall et al., 2016] Mungall, C. J., Koehler, S., Robinson, P., Holmes, I., and Haendel, M. (2016). k-boom : A bayesian approach to ontology structure inference, with applications in disease ontology construction. *bioRxiv*, page 048843.

[Mungall et al., 2017] Mungall, C. J., McMurry, J. A., Köhler, S., Balhoff, J. P., Borromeo, C., Brush, M., Carbon, S., Conlin, T., Dunn, N., Engelstad, M., Foster, E., Gourdine, J., Jacobsen, J. O., Keith, D., Laraway, B., Lewis, S. E., NguyenXuan, J., Shefchek, K., Vasilevsky, N., Yuan, Z., Washington, N., Hochheiser, H., Groza, T., Smedley, D., Robinson, P. N., and Haendel,

M. A. (2017). The monarch initiative : an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Research*, 45(D1) :D712–D722.

[NCBI, 2005] NCBI (2005). *Gene Help : Integrated Access to Genes of Genomes in the Reference Sequence Collection*. National Center for Biotechnology Information.

[Noy et al., 2010] Noy, N., Tudorache, T., Nyulas, C., and Musen, M. (2010). The ontology life cycle : Integrated tools for editing, publishing, peer review, and evolution of ontologies. In *AMIA Annual Symposium Proceedings*, volume 2010, page 552. American Medical Informatics Association.

[Noy et al., 2009] Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D. L., Storey, M.-A., Chute, C. G., and Musen, M. A. (2009). Bioportal : ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research*, 37(2) :W170–W173.

[Pasquier et al., 1999] Pasquier, N., Bastide, Y., Taouil, R., and Lakhal, L. (1999). Efficient mining of association rules using closed itemset lattices. *Information systems*, 24(1) :25–46.

[Pearson, 1895] Pearson, K. (1895). Notes on regression and inheritance in the case of two parents. In *Proceedings of the Royal Society of London*, volume 58, pages 240–242.

[Pesquita et al., 2007] Pesquita, C., Faria, D., Bastos, H., Falcao, A., and Couto, F. (2007). Evaluating go-based semantic similarity measures. In *Proc. 10th Annual Bio-Ontologies Meeting*, volume 37, page 38.

[Poggi et al., 2008] Poggi, A., Lembo, D., Calvanese, D., De Giacomo, G., Lenzerini, M., and Rosati, R. (2008). Linking data to ontologies. In *Journal on data semantics X*, pages 133–173. Springer.

[Prud et al., 2006] Prud, E., Seaborne, A., et al. (2006). Sparql query language for rdf.

[Rand, 1971] Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336) :846–850.

[Rawlins and Thompson, 1977] Rawlins, M. and Thompson, J. (1977). Pathogenesis of adverse drug reactions.

[Rawlins and Thompson, 1981] Rawlins, M. and Thompson, J. (1981). Pathogenesis of adverse drug reactions.

[Razick et al., 2008] Razick, S., Magklaras, G., and Donaldson, I. M. (2008). iRefIndex : a consolidated protein interaction database with provenance. *BMC bioinformatics*, 9(1) :405.

[Resnik, 1995] Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*.

[Resnik et al., 1999] Resnik, P. et al. (1999). Semantic similarity in a taxonomy : An information-based measure and its application to problems of ambiguity in natural language. *J. Artif. Intell. Res.(JAIR)*, 11 :95–130.

[Robinson et al., 2008] Robinson, P. N., Köhler, S., Bauer, S., Seelow, D., Horn, D., and Mundlos, S. (2008). The human phenotype ontology : A tool for annotating and analyzing human hereditary disease. *The American Journal of Human Genetics*, 83(5) :610 – 615.

[Rodríguez-López et al., 2014] Rodríguez-López, R., Reyes-Palomares, A., Sánchez-Jiménez, F., and Medina, M. Á. (2014). Phenuma : a tool for integrating the biomedical relationships among genes and diseases. *BMC bioinformatics*, 15(1) :375.

[Roitmann et al., 2014] Roitmann, E., Eriksson, R., and Brunak, S. (2014). Patient stratification and identification of adverse event correlations in the space of 1190 drug related adverse events. *Frontiers in physiology*, 5.

[Rokach and Maimon, 2005] Rokach, L. and Maimon, O. (2005). Clustering methods. In *Data mining and knowledge discovery handbook*, pages 321–352. Springer.

[Royer, 1997] Royer, R. J. (1997). Mechanism of action of adverse drug reactions : an overview. *Pharmacoepidemiology and drug safety*, 6(S3).

[Rubin et al., 2007] Rubin, D. L., Shah, N. H., and Noy, N. F. (2007). Biomedical ontologies : a functional perspective. *Briefings in bioinformatics*, 9(1) :75–90.

[Sakaeda et al., 2013] Sakaeda, T., Tamon, A., Kadoyama, K., and Okuno, Y. (2013). Data mining of the public version of the fda adverse event reporting system. *Int J Med Sci*, 10(7) :796–803.

[Samwald et al., 2011] Samwald, M., Jentzsch, A., Bouton, C., Kallesøe, C., Willighagen, E. L., Hajagos, J., Marshall, M. S., Prud'hommeaux, E., Hassanzadeh, O., Pichler, E., and Stephens, S. (2011). Linked open drug data for pharmaceutical research and development. *J. Cheminformatics*, 3 :19.

[Schedlbauer et al., 2009] Schedlbauer, A., Prasad, V., Mulvaney, C., Phansalkar, S., Stanton, W., Bates, D. W., and Avery, A. J. (2009). What evidence supports the use of computerized alerts and prompts to improve clinicians' prescribing behavior ? *Journal of the American Medical Informatics Association*, 16(4) :531–538.

[Schlicker et al., 2006] Schlicker, A., Domingues, F. S., Rahnenführer, J., and Lengauer, T. (2006). A new measure for functional similarity of gene products based on gene ontology. *BMC bioinformatics*, 7(1) :302.

[Schmachtenberg et al., 2014] Schmachtenberg, M., Bizer, C., and Paulheim, H. (2014). State of the LOD Cloud 2014. `http://lod-cloud.net/state/state_2014/`.

[Schriml et al., 2011] Schriml, L. M., Arze, C., Nadendla, S., Chang, Y.-W. W., Mazaitis, M., Felix, V., Feng, G., and Kibbe, W. A. (2011). Disease ontology : a backbone for disease semantic integration. *Nucleic acids research*, 40(D1) :D940–D946.

[Sevilla et al., 2005] Sevilla, J. L., Segura, V., Podhorski, A., Guruceaga, E., Mato, J. M., Martinez-Cruz, L. A., Corrales, F. J., and Rubio, A. (2005). Correlation between gene expression and go semantic similarity. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2(4) :330–338.

[Sibson, 1973] Sibson, R. (1973). Slink : an optimally efficient algorithm for the single-link cluster method. *The computer journal*, 16(1) :30–34.

[Singhal et al., 2001] Singhal, A. et al. (2001). Modern information retrieval : A brief overview. *IEEE Data Eng. Bull.*, 24(4) :35–43.

[Spasic et al., 2005] Spasic, I., Ananiadou, S., McNaught, J., and Kumar, A. (2005). Text mining and ontologies in biomedicine : Making sense of raw text. *Briefings in Bioinformatics*, 6(3) :239–251.

[Srinivasan, 2007] Srinivasan, A. (2007). The Aleph Manual. Available at http ://www.comlab.ox.ac.uk/oucl/research/areas/machlearn/Aleph/.

[Stumme and Maedche, 2001] Stumme, G. and Maedche, A. (2001). Fca-merge : Bottom-up merging of ontologies. In *IJCAI*, volume 1, pages 225–230.

[Tversky, 1977] Tversky, A. (1977). Features of similarity. *Psychological review*, 84(4) :327.

[UniProt Consortium, 2016] UniProt Consortium (2016). UniProt : the universal protein knowledgebase. *Nucleic acids research*, 45(D1) :D158–D169.

[US NLM, 2009] US NLM (2009). *UMLS®Reference Manual, Chapter 5. Semantic Network*. US National Library of Medicine.

[US NLM, 2018] US NLM (2018). MEDLINE®/PubMed®Resources Guide. `https://www.nlm.nih.gov/bsd/pmresources.html`.

[Valtchev et al., 2003] Valtchev, P., Grosser, D., Roume, C., and Hacene, M. R. (2003). Galicia : an open platform for lattices. In *Using Conceptual Structures : Contributions to the 11th Intl. Conference on Conceptual Structures (ICCS'03)*, pages 241–254.

[Van Der Merwe et al., 2004] Van Der Merwe, D., Obiedkov, S., and Kourie, D. (2004). Addintent : A new incremental algorithm for constructing concept lattices. In *International Conference on Formal Concept Analysis*, pages 372–385. Springer.

[van Karnebeek and Stockler, 2012] van Karnebeek, C. D. and Stockler, S. (2012). Treatable inborn errors of metabolism causing intellectual disability : a systematic literature review. *Molecular genetics and metabolism*, 105(3) :368–381.

[Vasudevan and Ginzler, 2009] Vasudevan, A. R. and Ginzler, E. M. (2009). Established and novel treatments for lupus : agents in clinical trials are targeting various immunological processes. *The Journal of Musculoskeletal Medicine*, 26(8) :291–291.

[Villerd et al., 2010] Villerd, J., Toussaint, Y., and Lillo-Le Louët, A. (2010). Adverse drug reaction mining in pharmacovigilance data using formal concept analysis. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 386–401. Springer.

[W3C OWL Working Group, 2012] W3C OWL Working Group (2012). OWL 2 Web Ontology Language Document Overview (Second Edition). `https://www.w3.org/TR/owl2-overview/`.

[Warrens, 2008] Warrens, M. J. (2008). On the equivalence of cohen's kappa and the hubert-arabie adjusted rand index. *Journal of Classification*, 25(2) :177–183.

[Wheeler et al., 2007] Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Edgar, R., Federhen, S., Geer, L. Y., Kapustin, Y., Khovayko, O., Landsman, D., Lipman, D. J., Madden, T. L., Maglott, D. R., Ostell, J., Miller, V., Pruitt, K. D., Schuler, G. D., Sequeira, E., Sherry, S. T., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusov, R. L., Tatusova, T. A., Wagner, L., and Yaschenko, E. (2007). Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 35(suppl 1) :D5–D12.

[Whetzel et al., 2011] Whetzel, P. L., Noy, N. F., Shah, N. H., Alexander, P. R., Nyulas, C., Tudorache, T., and Musen, M. A. (2011). BioPortal : enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. *Nucleic acids research*, 39(suppl 2) :W541–W545.

[WHO, 2004] WHO (2004). *International statistical classification of diseases and related health problems*, volume 1. World Health Organization.

[WHO, 2011] WHO (2011). ICD-9-CM Official Guidelines for Coding and Reporting. World Health Organisation. `https://www.cdc.gov/nchs/data/icd/icd9cm_guidelines_2011.pdf`.

[WHOCC, 2013] WHOCC (2013). Guidelines for ATC classification and DDD assignment – 16th edition. World Health Organisation Collaborating Centre for Drug Statistics Methodology. `https://www.whocc.no/filearchive/publications/1_2013guidelines.pdf`.

[Wilcoxon, 1945] Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6) :80–83.

[Wilkinson et al., 2016] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., et al. (2016). The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3.

[Winnenburg et al., 2015] Winnenburg, R., Sorbello, A., and Bodenreider, O. (2015). Exploring adverse drug events at the class level. *Journal of Biomedical Semantics*, 6(1) :18.

[Wu and Palmer, 1994] Wu, Z. and Palmer, M. (1994). Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138. Association for Computational Linguistics.

[Zhou et al., 2014] Zhou, X., Menche, J., Barabási, A.-L., and Sharma, A. (2014). Human symptoms–disease network. *Nature communications*, 5 :4212.

[Zou et al., 2015] Zou, L., Zhang, Z., and Long, J. (2015). A fast incremental algorithm for constructing concept lattices. *Expert Systems with Applications*, 42(9) :4474–4481.

Résumé

Le Web sémantique propose un ensemble de standards et d'outils pour la formalisation et l'interopérabilité de connaissances partagées sur le Web, sous la forme d'ontologies. Les ontologies biomédicales et les données associées constituent de nos jours un ensemble de connaissances complexes, hétérogènes et interconnectées, dont l'analyse est porteuse de grands enjeux en santé, par exemple dans le cadre de la pharmacovigilance. On proposera dans cette thèse des méthodes permettant d'utiliser ces ontologies biomédicales pour étendre les possibilités d'un processus de fouille de données, en particulier, permettant de faire cohabiter et d'exploiter les connaissances de plusieurs ontologies biomédicales.

Les travaux de cette thèse concernent dans un premier temps une méthode fondée sur les structures de patrons, une extension de l'analyse formelle de concepts pour la découverte de co-occurences de événements indésirables médicamenteux dans des données patients. Cette méthode utilise une ontologie de phénotypes et une ontologie de médicaments pour permettre la comparaison de ces événements complexes, et la découverte d'associations à différents niveaux de généralisation, par exemple, au niveau de médicaments ou de classes de médicaments.

Dans un second temps, on utilisera une méthode numérique fondée sur des mesures de similarité sémantique pour la classification de déficiences intellectuelles génétiques. On étudiera deux mesures de similarité utilisant des méthodes de calcul différentes, que l'on utilisera avec différentes combinaisons d'ontologies phénotypiques et géniques. En particulier, on quantifiera l'influence que les différentes connaissances de domaine ont sur la capacité de classification de ces mesures, et comment ces connaissances peuvent coopérer au sein de telles méthodes numériques.

Une troisième étude utilise les données ouvertes liées ou LOD du Web sémantique et les ontologies associées dans le but de caractériser des gènes responsables de déficiences intellectuelles. On utilise ici la programmation logique inductive, qui s'avère adaptée pour fouiller des données relationnelles comme les LOD, en prenant en compte leurs relations avec les ontologies, et en extraire un modèle prédictif et descriptif des gènes responsables de déficiences intellectuelles.

L'ensemble des contributions de cette thèse montre qu'il est possible de faire coopérer avantageusement une ou plusieurs ontologies dans divers processus de fouille de données.

Mots-clés: Bioontologies, Données Ouvertes Liées, Programmation Logique Inductive, Similarité sémantique, Structures de patrons, Web sémantique

Abstract

The semantic Web proposes standards and tools to formalize and share knowledge on the Web, in the form of ontologies. Biomedical ontologies and associated data represents a vast collection of complex, heterogeneous and linked knowledge. The analysis of such knowledge presents great opportunities in healthcare, for instance in pharmacovigilance. This thesis explores several ways to make use of this biomedical knowledge in the data mining step of a knowledge discovery process. In particular, we propose three methods in which several ontologies cooperate to improve data mining results.

A first contribution of this thesis describes a method based on pattern structures, an extension of formal concept analysis, to extract associations between adverse drug events from patient data. In this context, a phenotype ontology and a drug ontology cooperate to allow a semantic comparison of these complex adverse events, and leading to the discovery of associations between such events at varying degrees of generalization, for instance, at the drug or drug class level.

A second contribution uses a numeric method based on semantic similarity measures to classify different types of genetic intellectual disabilities, characterized by both their phenotypes and the functions of their linked genes. We study two different similarity measures, applied with different combinations of phenotypic and gene function ontologies. In particular, we investigate the influence of each domain of knowledge represented in each ontology on the classification process, and how they can cooperate to improve that process.

Finally, a third contribution uses the data component of the semantic Web, the Linked Open Data (LOD), together with linked ontologies, to characterize genes responsible for intellectual deficiencies. We use Inductive Logic Programming, a suitable method to mine relational data such as LOD while exploiting domain knowledge from ontologies by using reasoning mechanisms. Here, ILP allows to extract from LOD and ontologies a descriptive and predictive model of genes responsible for intellectual disabilities.

These contributions illustrates the possibility of having several ontologies cooperate to improve various data mining processes.

Keywords: Bioontologies, Inductive Logic Programming, Linked Open Data, Pattern structures, Semantic similarity, Semantic Web