UNIVERSITÉ DE LORRAINE

École doctorale IAEM Lorraine

Mises en correspondances de données, textes et connaissances pour la découverte de connaissances biomédicales

MÉMOIRE

présenté et soutenu publiquement le 16 décembre 2019

pour l'obtention d'une

habilitation à diriger des recherches de l'Université de Lorraine

(mention informatique)

par

Adrien Coulet

Composition du jury

Rapporteurs :	Nathalie Aussenac-Gilles	Directrice de Recherche, CNRS
	Sarah Cohen-Boulakia	Professeure, Université Paris-Sud
	Olivier Curé	Maître de conférences HDR, Université Paris-Est Marne-la-Vallée
Examinateurs :	Olivier Dameron	Maître de conférences HDR, Université de Rennes 1
	Marie-Dominique Devignes	Chargée de recherche, CNRS
	Anne Gégout-Petit	Professeure, Université de Lorraine
Parrain scientifique :	Amedeo Napoli	Directeur de Recherche, CNRS

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

Mis en page avec la classe thesul.

Abstract

In computer science, knowledge can take many forms: it can be formalized in text, a format difficult to handle for machines; it can be represented in knowledge bases, then interpretable to some extend by machines; or being latent in a set of data awaiting to be analysed. These variously formalized representations of knowledge coexist in many fields, and in particular in biomedicine, our domain of interested. However they are most of the time not compared with each other, eventhough this comparison has many potentials. Comparison would allow to provide a unique access to available knowledge; to detect agreements between various sources; to evaluate knowledge discovery; and to reuse knowledge to guide knowledge discovery. Without achieving all of them, we present in this manuscript several contributions that use mappings between data, texts and ontologies with these goals in mind. First, we consider data and text annotations, i.e., mappings between data, text and ontologies. Then, we consider annotations of Electronic Health Records and how these can help analyse clinical data. Finally, we explore the mining of knowledge graphs, a particular setting where data and knowledge are already connected. In each case, we will explain how knowledge may guide knowledge discovery.

Keywords: knowledge discovery, knowledge comparison, biomedical informatics, data mining, text mining, machine learning, precision medicine

Résumé

Les connaissances peuvent revêtir des formes multiples en informatique : elles peuvent être écrites en langage naturel dans des textes, un format difficile à manipuler pour les machines ; elles peuvent être formalisées dans des bases de connaissances alors appréhendables dans une certaine mesure grâce aux outils du Web Sémantique ; ou encore être présentes, de façon sous-jacente, dans un ensemble de données en attente d'analyse. Ces trois types de représentations de connaissances plus ou moins formalisées, plus ou moins explicites coexistent dans de nombreux domaines, et c'est notamment le cas dans le domaine biomédical auquel nous nous intéressons particulièrement. Malgré cette coexistence, ces types de connaissances ne sont généralement pas confrontés les uns aux autres. Il paraît pourtant très utile d'être capable de les comparer et cela pour les quatre raisons suivantes : cela permet d'intégrer un panorama de connaissances disponibles et de proposer un accès unique à celui-ci ; de détecter des accords et désaccords entre les sources ; d'évaluer des extractions de connaissances à partir de données ; et enfin de faciliter l'extraction de connaissances en s'appuyant sur des connaissances pré-existantes. Sans forcement les atteindre tous les quatre, nous présentons dans ce mémoire différentes contributions qui mettent en correspondance données, textes et ontologies avec ces objectifs en vue. Dans un premier temps nous abordons les mises en correspondances entre ontologies et données ou textes. Dans un second temps nous considérons l'utilisation de ces correspondances pour l'analyse de données et en particulier de données cliniques. Enfin, nous présentons des approches de fouille de données appliquées à des graphes de connaissances, un contexte particulier où données et connaissances sont déjà liées. Nous nous attacherons dans tous les cas à expliquer comment les connaissances peuvent guider la découverte de connaissances.

Mots-clés: extraction de connaissances, comparaison de connaissances, informatique biomédicale, fouille de données, fouille de texte, apprentissage, médecine de précision

Remerciements

Je remercie tout d'abord Nathalie Aussenac-Gilles, Sarah Cohen-Boulakia et Olivier Curé pour avoir accepté de relire ce mémoire et pour l'avoir fait avec beaucoup de bienveillance. Merci également à Olivier Dameron et Anne Gégout-Petit pour avoir accepté d'examiner ce travail. Comme tout ceci est aussi un peu de ta faute Marie-Dominique, je te remercie particulièrement pour avoir accepté d'examiner mon travail, une nouvelle fois. Et merci à toi Amedeo pour le parrainage et au delà pour avoir trouvé le bon équilibre entre direction et liberté pour que je m'épanouisse comme Orpailleur.

"It takes a village to raise a child" Le travail décrit dans ce mémoire n'est pas le fruit de mon seul travail, mais celui de nombreux échanges et collaborations. Je remercie Gabin, Joël, Kevin, Mohsen et Pierre pour les nombreuses réalisations de vos postdocs, thèses de science, de pharmacie, soutenues ou à venir. C'est avant tout votre travail et je vous remercie de me laisser le raconter dans ce document. Merci à tous les Orpailleurs passés et présents. Merci Malika, Chedy, Jean, Nicolas et Yannick pour les conseils réguliers et les bons moments. Merci Miguel de reprendre la suite, avec ton propre style. Longue vie à l'équipe Pépite ! Merci à tous les collègues : Florent Domenach ; Mehwish Alam, Olfa Makkaoui, Yassine Marzougui ; Bastien Rance, Cédric Bousquet, Clément Jonquet et les PractiKPharma-cien(ne)s ; Céline Bonnet, Ndeye Coumba Ndiaye, Nadine Petitpain, Philippe Jonveaux, Céline Bonnet et Nicolas Girerd du CHRU. Merci à tous les collègues de Telecom Nancy et aux collègues administratifs et notamment à Antoinette, Anne, Anne-Marie, Carole, Delphine, Emmanuelle, Marine, Sylvie pour me faciliter énormément la vie.

Many thanks to Stanford friends! Mark, Michel, Nigam, Russ thanks for always putting research in new perspective and for keeping alive the French connection of the BMIR. Nigam, thanks for opening the door of your amazing lab for my last 2 years. Yael, Paea, NCBO and PharmGKB people, thanks for all the exchanges.

"Love lives here" Thanks to Gwyn Dukes and everyone from the Bechtel I-Center community. Everyday I'm amazed to have friends all around the word. Merci Castaña & Jorge, Clément & Isa, Domitille, Erwan, Gilles, Helsa & Scott, Julie, Louise, Mariki & Shaun, Michaël, Nicolas, Pierre-Frédéric, Seb & Manue, J&E, C&C, Mayumi, et les autres... Merci à la famille Martin pour son soutien sans faille. Merci à mes parents et à mon frère pour l'inconditionnalité.

Camille, Auguste, Célestin *"may you stay forever young"*.

Aude, merci de continuer de bousculer ma vie. *"Banana, you rock !"*

Sommaire

Résumé **i**

Remerciements **iii**

Table des figures **vii**

Introduction **1**

1 Distinctions entre données, textes et connaissances 1
2 Les relations réciproques entre données et connaissances 2
3 La découverte de connaissances guidée par les connaissances 6
4 La médecine de précision comme domaine d'application 6
5 Le contexte : projets de recherche et encadrements 8
6 Contenu et organisation du mémoire 9

Chapitre 1 Annotation de données pour l'extraction de connaissances **11**

1.1 Introduction 11
1.2 Annotation de bases de données biologiques avec le Bioportal et le Resource Index 18
1.3 Analyse d'annotations pour la recherche d'associations entre financement de recherche et publication 21
1.4 Les structures de patrons pour considérer les connaissances associées aux annotations 25
1.5 Quelques mots sur les textes 36
1.6 Discussion générale 38

Chapitre 2 Des annotations pour l'analyse de données cliniques **41**

2.1 Introduction 41
2.2 La recherche de cooccurrence d'effets indésirables médicamenteux 42
2.3 Prédiction de la nécessité de réduire la dose d'un médicament, avant sa première prescription 48
2.4 Discussion générale 60

Chapitre 3 Découverte de connaissances à partir de graphes de connaissances 63
3.1 Introduction 63
3.2 Prédiction de liens dans un graphe de connaissances 65
3.3 Recherche de régularités dans un graphe de connaissances 82
3.4 Discussion générale 91

Perspectives de recherche 93

Annexes 97

Annexe A Annexe du Chapitre 1 97

Annexe B Annexes du Chapitre 2 99

Annexe C Annexes du Chapitre 3 103
C.1 Définitions des règles de réconciliation définies sur PGxLOD 103
C.2 Exemples d'application des règles 106

Bibliographie 107

Table des figures

1 Représentation du processus d'Extraction de Connaissances à partir des Bases de Données (ECBD, aussi appelé *knowledge discovery*), d'après [Fayyad *et al.*, 1996]. 4

2 Exemple de connaissance pharmacogénomique illustrant l'impact des variants du gène CYP2D6 sur la réponse à la codéine. 7

1.1 Processus de peuplement du Resource Index illustré par l'exemple de l'annotation automatique de la bases de données GEO (Gene Expression Omnibus). Figure issue de [Jonquet *et al.*, 2010]. 19

1.2 Génération et utilisation des annotations du Resource Index pour l'analyse de l'activité de recherche biomédicale par maladie, suivant deux dimensions : les financements et les publications par maladie. La partie A schématise la Disease Ontology ; la partie B représente le processus d'annotation ; la partie C présente les attributs utilisés pour joindre ou agréger les données. 22

1.3 Représentation à deux niveaux d'agrégation différents (A et B) des financements et impacts de publications par maladie (ou classes de maladies). 23

1.4 Détail de l'ontologie *NCI Thesaurus* et des types sémantiques de l'UMLS associés aux concepts. Les lignes pointillées associent à chaque concept de l'ontologie son type sémantique comme définie par le *Semantic Network* de l'UMLS. 25

1.5 Exemple d'annotation de la base de données DrugBank avec l'ontologie NCI Thesaurus. (a) La partie gauche représente un morceau de l'ontologie *NCI Thesaurus* ; (b) la partie droite un extrait du document DB01082 de DrugBank à propos de la streptomycine. Les flèches représentent les annotations. 28

1.6 Illustration figurative de la structure de patron $(G, (\mathcal{D}, \sqcap), \delta)$ proposée pour l'analyse d'annotations de documents. G est l'ensemble des documents annotés. $\delta(g_i)$ sont les annotations du document g_i. $(\mathcal{D}, \sqcap)$ est l'ensemble des descriptions organisé dans un semi-treillis 29

1.7 Le demi-treillis des enveloppes convexes associées au contexte présenté Tableau 1.3 et au NCI Thesaurus 32

1.8 Représentations de la couverture des enveloppes convexes des trois dimensions non-vides des annotations de l'ensemble des deux documents "Drug1" et "DB01082" 33

1.9 Représentation du treillis obtenu par la structure de patrons exemple définie par le contexte présenté Tableau 1.3. 34

1.10 Annotations syntaxiques et sémantiques d'une phrase (Figure issue de [Garten, 2010]). 37

2.1 Vue générale de notre approche de prédiction de changement de dose. (1) Les annotations par des concepts ontologiques des dossiers patients sont extraites. (2) Les évènements de changement de doses sont identifiés. (3) Les annotations les plus caractéristiques des changements de dose sont sélectionnées pour construire des profils phénotypiques. (4) Ces profils sont utilisés pour construire une matrice où les patients sont décrits avec les annotations les plus caractéristiques. (5) La matrice résultante est utilisée pour entraîner puis (6) évaluer deux modèles de forêts d'arbres aléatoires : l'un pour prédire les réductions de dose, l'autre pour prédire les augmentations. 49

2.2 Définition des intervalles de changement ou de continuation de la dose des prescriptions. La partie haute **(a)** présente les trois types d'intervalles, chacun délimité par deux prescriptions d_1 et d_2 faites à t_1 et t_2. Aucun autre médicament avec la même molécule n'est prescrit durant un intervalle. Les deux panneaux du bas positionnent sur la ligne de temps les attributs utilisés par les modèles prédictifs, avant la première prescription. **(b)** Les trois types d'attributs sont les concepts ICD9 utilisés pour le diagnostic (*diag*), les conditions annotées dans les notes cliniques (*cond*) et les commandes d'examens biologiques (*lab*). Puisque les diagnostics et conditions sont des annotations faites avec des concepts d'ontologies, ils sont généralisés selon la hiérarchie de concepts pour enrichir les annotations (*i.e.*, de (b) on obtient (c)). Par exemple comme $diag_3$ est plus général que $diag_1$ il est également associé à l'historique du patient. 50

2.3 Ce tableau donne les nombres d'attributs phénotypiques de chaque type qui composent les profils phénotypiques après chaque étape de filtrage. Les étapes de filtrages RR et IC utilisent le risque relatif (RR) et le contenu en information (IC) associés aux attributs. La méthode *elim* permet de filtrer les attributs générés lors de l'expansion des annotations avec les ontologies [Alexa *et al.*, 2006]. *elim* ne peut pas être appliqué aux examens biologiques car ceux-ci ne sont pas encodés avec une ontologie et n'ont donc pas été étendus. 54

2.4 Résultats de l'évaluation de la prédiction des réductions de doses à partir des profils phénotypiques. Deux modes d'évaluation ont été testés : une validation croisée à 10 feuillets et une validation où seules les données de la dernière année de données ont été retirées pour constituer l'ensemble de test (*hold last year out*). Ici seuls les 100 premiers codes diagnostics, conditions mentionnées dans les notes et examens biologiques sont conservés dans les profils phénotypiques. Seuls les médicaments et groupes de médicaments avec une mesure F (F-m) 0,7 avec le second mode d'évaluation sont rapportés ici, plus les résultats de l'ensemble des médicaments P450 et de la classe ATC L. L fait référence à la classes *Antineoplastic and immunomodulating agents* et H est *Systemic hormonal preparations, excluding sex hormones and insulins*. Les résultats de classe L sont les meilleurs obtenus en validation croisée, cependant ils ne sont pas calculables avec le mode *hold last year out* à cause de profils phénotypiques vides pour la dernière année (2014). |instances| correspond au nombre d'instances dans l'ensemble d'entraînement. Les précisions (P) et rappels (R) sont données entre parenthèse après la mesure F. Les résultats complets sont disponibles dans les suppléments de [Coulet *et al.*, 2018]. 56

2.5 Performances de nos modèles prédictifs par types d'attributs phénotypiques. Les 100 premiers attributs de chaque type sont obtenus suivant les valeurs p associées aux attributs. Les performances sont celles de la validation croisée à 10 feuillets. Les 300 attributs sont la combinaison des trois 100 premiers attributs de chaque type. 57

3.1 Les concepts et relations de PGxO, autour du concept central `Pharmacogenomic Relationship`. 66

3.2 Une relation pharmacogénomique extraite de PharmGKB le 11 août 2018 et son instanciation des concepts de PGxO. Pour faciliter la lecture nous avons parfois utilisé les labels à la place des URI. Un seul médicament et un seul variant sont représentés, alors que la relation implique en réalité plus de composants. L'annotation clinique d'origine est disponible à `https://www.pharmgkb.org/gene/PA356/clinicalAnnotation/1184648909`. pharmgkb2triples est le nom de notre script d'extraction. v2018-03-05 est une référence à la version de PharmGKB utilisée. . 69

3.3 Exemple de phrase (PMID=18370849) manuellement annotée avec quatre entités et une relation pharmacogénomique. 70

3.4 Deux exemples de graphes RDF sur lesquels la première règle de réconciliation identifie deux relations pharmacogénomiques qui sont identiques. Le lien `owl:sameAs` résulte de l'application de cette règle sur le graphe. 77

3.5 Trois exemples de graphe RDF sur lesquels la règle de réconciliation 2 identifie des liens de type `skos:broaderMatch` entre des relations pharmacogénomiques. . 78

3.6 Exemple jouet de concepts d'ontologie instanciés par des entités RDF. Les instanciations sont représentées par des flêches pointillées et la subsomption par une flèche continue. Par exemple, $\mathtt{r}_1$ instancie $\mathtt{k}_1$ et $\mathtt{k}_2$, et $\mathtt{k}_2$ est subsumé par $\mathtt{k}_3$. . . . 87

3.7 A gauche, le treillis de concept *annoté* construit à partir du contexte du Tableau 3.12 et annoté avec les concepts de l'ontologie de la Figure 3.6. A droite, l'ordre induit des annotations. Le treillis, ses annotations et leur ordre sont représentés selon la notation réduite. Les sujets (*i.e.*, extensions) sont en noir, les prédicats (*i.e.*, intentions) sont en gris et les annotations sont entre crochés avec la notation $\{\cdot\}_A$. Les concepts formels sont arbitrairement annotés de 1 à 6. 88

A.1 Correspondances entre la *Disease Ontology* et les sources des taux de mortalité de l'OMS et du CDC. 98

B.1 25 groupes de médicaments associés aux enzymes P450, organisés selon 3 critères : l'enzyme P450 avec laquelle ils interagissent, le type de relation avec cet enzyme P450 et leur classe ATC (1er niveau seulement). Un médicament peut appartenir à plusieurs groupes. La taille de chaque noeud est proportionnelle à la taille du groupe qui est donné en parenthèse. 100

B.2 Exemple de profil phénotypique. L'exemple est réduit aux 10 premiers diagnostics (A), conditions (B), et examens biologiques (C) qui sont observés avant une prescription de *tacrolimus* chez les patients qui nécessitent ensuite une réduction de dose. Chaque phénotype est associé à une valeur p significative (test hypergéométrique, p<0.05, correction de Bonferroni) et est trié selon la valeur des valeurs absolues du logarithme du risque relatif (RR) (représentées dans la première colonne sur une échelle de 0 à 2). Pour aider l'interprétation de ces profils, le nombre d'articles de PubMed qui mentionnent à la fois le médicament (tacrolimus) et le phénotype sont fournis (seconde colonne, sur une échelle de 0 à 180). Par exemple 45 articles mentionnent à la fois le tacrolimus et la candidose (*candidiasis*). (A) Les diagnostics sont les codes ICD-9-CM associés avec les visites de patients ; (B) les conditions sont les phénotypes mentionnés dans le texte de notes cliniques ; (C) les examens de laboratoires commandés. Certains noms d'examens sont préfixés avec "*NO*" pour indiquer une relation négative (RR<1) entre l'examen de laboratoire est la réduction de dose de tacrolimus. 101
B.3 Trois exemples de profils phénotypiques associés avec les réductions de doses des *labetalol* (a), *sildenafil* (b) et *warfarin* (c). Voir la Figure B.2 pour plus d'indications sur la lecture des profils phénotypiques. 102

C.1 Exemple d'application de la règle 3. 106
C.2 Exemple d'application de la règle 4. 106
C.3 Exemple d'application de la règle 5. 106

Introduction

1 Distinctions entre données, textes et connaissances

En informatique, *données, information* et *connaissances* se distinguent classiquement par le niveau d'interprétation qui leur est associé [Schreiber and Akkermans, 2000, Wille, 2002]. Les *données* sont des signes, possiblement organisés selon une syntaxe, mais non associés à une interprétation. Il s'agit d'un signal brut. L'*information* est un ensemble de données auquel est associé un message, mais la façon d'interpréter ce message est variable et généralement informelle. L'interprétation du message peut être faite par un utilisateur humain lors par exemple de la lecture de données textuelles ou d'une structure, comme un modèle de données qui distingue et lie des sous-parties des données permettant de leur associer implicitement un sens. Les *connaissances*, sont quant-à elles constituées de données, d'une syntaxe et d'un langage, par exemple une logique, qui permet d'associer à leur représentation une interprétation non ambiguë. Les connaissances ainsi formalisées peuvent être manipulées et mises en action aussi bien par un humain que par une machine. La représentation de connaissances est pour cette raison un objectif clé de l'intelligence artificielle [Kayser, 1997] et celui-ci est rendu possible entre autres grâce aux outils et standards du Web Sémantique [Berners-Lee *et al.*, 2001].

Ces définitions clarifient bien les notions de données et connaissances, cependant la notion d'information demeure relativement vague, et pour cette raison, nous éviterons de l'utiliser dans ce document. Une précision supplémentaire pour bien distinguer données et connaissances est de clarifier l'utilisation du terme "base de connaissances" qui en informatique fait référence à une représentation de connaissances instanciée par des objets, mais qui dans d'autres contextes désigne une base de données experte. Dans le domaine biomédical c'est par exemple le cas pour la base de protéines UniProtKB [Consortium, 2018] ou la base de relations gène-médicament PharmGKB (the Pharmacogenomics Knowledge Base) [Whirl-Carrillo *et al.*, 2012]. Dans ce cas, les connaissances ne sont pas formelles, et d'un point de vue purement informatique, il s'agit de données et de relations structurées par le schéma de la base. Cependant les données de ces bases sont interprétées lors de leur consultation par les experts et peuvent alors dans une certaine mesure être considérées comme des connaissances de domaine. Ces bases de données permettent de répertorier des relations de façon plus synthétique que les articles scientifiques. Cela facilite leur réutilisation notamment en analyse de données. D'ailleurs, le terme *background knowledge* est utilisé sans véritable distinction dans la littérature pour faire référence à l'utilisation soit de bases de données expertes soit d'ontologies.

La frontière entre données et connaissances est relativement perméable et peut varier selon le point de vue ou l'application visée. La considération des textes en est un bon exemple, puisqu'ils sont soit des données, soit des connaissances, selon que ce soit une machine ou un humain qui les considère. En effet pour un agent informatique classique, le texte reste une séquence de caractères ou de mots, arrangés en phrases suivant une certaine structure grammaticale, *i.e.*, des données et une syntaxe. Pour l'humain, la lecture du texte ne peut pas être dissociée de son

interprétation. Celle-ci se fait par la mise en correspondance du texte avec les connaissances du lecteur et des éléments de contexte, pas forcement disponibles dans le texte lui-même. Dans ce cas, la même source est données et connaissances, seulement la machine n'est pas équipée du système de raisonnement et des connaissances contextuelles permettant d'interpréter le contenu du texte.

En représentation des connaissances la frontière entre données et connaissances est également perméable et dépend du problème que l'on cherche à résoudre. Par exemple en Logique de Description (LD), cette frontière est en théorie matérialisée par la distinction entre les parties terminologique (la TBox) et assertionelle (la ABox) d'une base de connaissances [Baader *et al.*, 2010]. Le problème est que le positionnement de la frontière TBox/ABox peut varier selon l'application. Il peut être intéressant dans certains cas de représenter une même entité soit comme un objet, soit comme une classe d'objets qui pourra alors être instanciée. Si l'on cherche par exemple à représenter des connaissances pharmacologiques, un médicament comme le paracetamol peut-être considéré comme un objet, c'est-à-dire un individu au sens des LD, qui instancie la classe des anti-douleurs et le fait que le patient n°42 soit traité avec du paracetamol peut alors être représenté par l'instanciation d'un prédicat binaire *estTraitéPar* par les deux individus concernés, *i.e.*, le patient n°42 et le paracetamol. Mais le paracetamol peut également être vu comme une classe qui sera instanciée par des individus qui font référence aux matérialisations de ce médicament, *i.e.*, des pilules, gélules ou autres formulations du paracétamol, ce qui permet de représenter plus finement et de façon traçable les objets administrés au patient n°42.

2 Les relations réciproques entre données et connaissances

L'intérêt de disposer de connaissances formalisées dans un langage de représentation de connaissances est que celles-ci sont en partie interprétables par une machine, et qu'elles peuvent ainsi être considérées au sein de mécanismes de raisonnement automatiques qui aident à mieux tirer parti des données disponibles. Ceci peut présenter un avantage dans des tâches comme l'intégration de données, la recherche d'information ou la fouille de données. En intégration de données notamment, représentations de connaissances et raisonnement peuvent faciliter l'alignement des schémas des différentes sources à intégrer [Saïs, 2007, Euzenat and Shvaiko, 2013]. Ceci est particulièrement intéressant quand les vocabulaires des schémas sont différents à cause de granularités différentes ou de langues différentes par exemple. En recherche d'information, connaissances et raisonnement permettent la réécriture de requêtes et la définition de distances sémantiques qui améliorent la pertinence des documents retrouvés [Andreasen *et al.*, 2011]. En fouille de données, ils peuvent permettre de réduire l'espace de recherche en définissant des contraintes sur celui-ci ou inversement de produire des descripteurs additionnels des données offrant ainsi plus d'opportunités pour généraliser. La Section 4 du Chaptire 2 de ma thèse de doctorat [Coulet, 2008] propose des exemples, pour chaque étape de l'ECBD où des connaissances peuvent jouer un rôle. Une pratique nouvelle et qui pour cette raison est absente de ma thèse est la constitution d'*embeddings* d'éléments de connaissances. Les embeddings sont des représentations sous forme de vecteurs de nombres réels des représentations initiales des instances considérées dans une tâche d'apprentissage. La constitution des embeddings consiste alors à réduire l'ensemble de descripteurs complexes et associés à des domaines de valeurs variés, à des vecteurs numériques représentés dans un espace continu et plus réduit que l'espace d'origine [Bengio *et al.*, 2003]. Les embeddings embarquent classiquement différents descripteurs des instances mais ceux ci peuvent également être complétés avec des éléments de connaissances associés aux instances [Bordes *et al.*, 2011, Monnin *et al.*, 2019b].

Dans tous les cas, utiliser connaissances et raisonnement pour mieux manipuler des données, demande de disposer de relations entre ces données et connaissances. J'ai d'ailleurs choisi d'articuler ce mémoire autour de ces relations que j'appellerai *annotations* ou *correspondances* pour souligner leur côté réciproque. En effet une unité de connaissance décrit un ensemble de données, dans une représentation formelle ; mais réciproquement un ensemble de données groupées sur la base d'une régularité définit une connaissance. Ces correspondances peuvent avoir des formes diverses et cette section en présente quelques exemples.

L'instanciation Suivant le paradigme classique des langages par objets, l'instanciation d'une classe par un objet constitue une relation entre données et connaissances dans le mesure où la classe appartient à une représentation de connaissances telle qu'une ontologie. La sémantique associée à l'instanciation est que la classe décrit un ensemble d'objets et si un objet l'instancie il satisfait les propriétés qui définissent cette classe [Masini *et al.*, 1989]. Cela correspond en LD à l'instanciation de concept, par distinction avec un second type d'instanciation possible dans ce formalisme : l'instanciation de rôle. Dans ce second cas, le rôle est un prédicat binaire qui est instancié par deux objets (ou individus en LD) appelés sujet et objet, lesquels sont reliés par le prédicat. Instanciations de concepts et de rôles sont représentés en LD par des axiomes de la forme $C(a)$ et $r(a, b)$ où l'individu a instancie le concept C et la paire d'individus a, b le rôle r [Baader *et al.*, 2010]. L'ensemble de ces axiomes constitue la partie assertionnelle, ou ABox, des bases de connaissances de LD.

L'annotation Une autre façon de mettre en correspondance données et connaissances est l'annotation. La notion d'annotation est définie de façon assez flexible et correspond à une paire (*unité de données, concept d'ontologie*) qui signifie que l'unité de données (un n-uplet, un document, un mot dans un document par exemple) est liée dans une certaine mesure au concept. L'interprétation qui y est associée est variable, pouvant se rapprocher dans certains cas de l'instanciation et dans d'autres de la définition d'une propriété. De façon générale, il est possible de considérer que toute meta-donnée (*i.e.*, une donnée qui décrit des données) représentée à l'aide d'éléments de connaissance constitue une annotation. Par exemple, les articles scientifiques référencés dans la base PubMed, sont annotés avec des concepts de l'ontologie MeSH pour faciliter leur recherche [U.S. National Library of Medicine, 2018]. La génération de ces annotation est semi-automatique et leur signification est que le document annoté est en rapport avec le concept MeSH utilisé pour annoter. Ces annotations sont utilisées pour la recherche de documents dans PubMed. Elles permettent de retrouver des articles à partir de mots clés et cela même si le mot clé n'est pas mentionné dans le texte de l'article. Si l'article a été annoté avec le concept correspondant, ou un des enfants du concept dans la hiérarchie de MeSH il sera retrouvé. Un autre exemple d'annotation est l'annotation des protéines avec les termes de Gene Ontology (GO) [Ashburner *et al.*, 2000]. GO est un ensemble de trois hiérarchies de termes qui sert de vocabulaire de référence pour définir les fonctions, processus biologiques et compartiments cellulaires associés aux protéines. Dans ce cas l'interprétation des annotations correspond plus à la spécification de propriétés que l'on pourrait appeler respectivement *aPourFonction, estImpliquéDansLeProcessus, estTrouvéDansLeCompartiment*. Un autre exemple d'annotation est celui de portions de textes dans les documents que constituent les ressources linguistiques comme les corpora annotés. Par exemple l'annotation des entités nommées avec des concepts ontologiques, comme les noms de médicaments dans les résumés de la base PubMed annotés avec les concepts de la classification des médicaments ATC [WHO Collaborating Centre for Drug Statistics Methodology, 2018]. Ce type d'an-

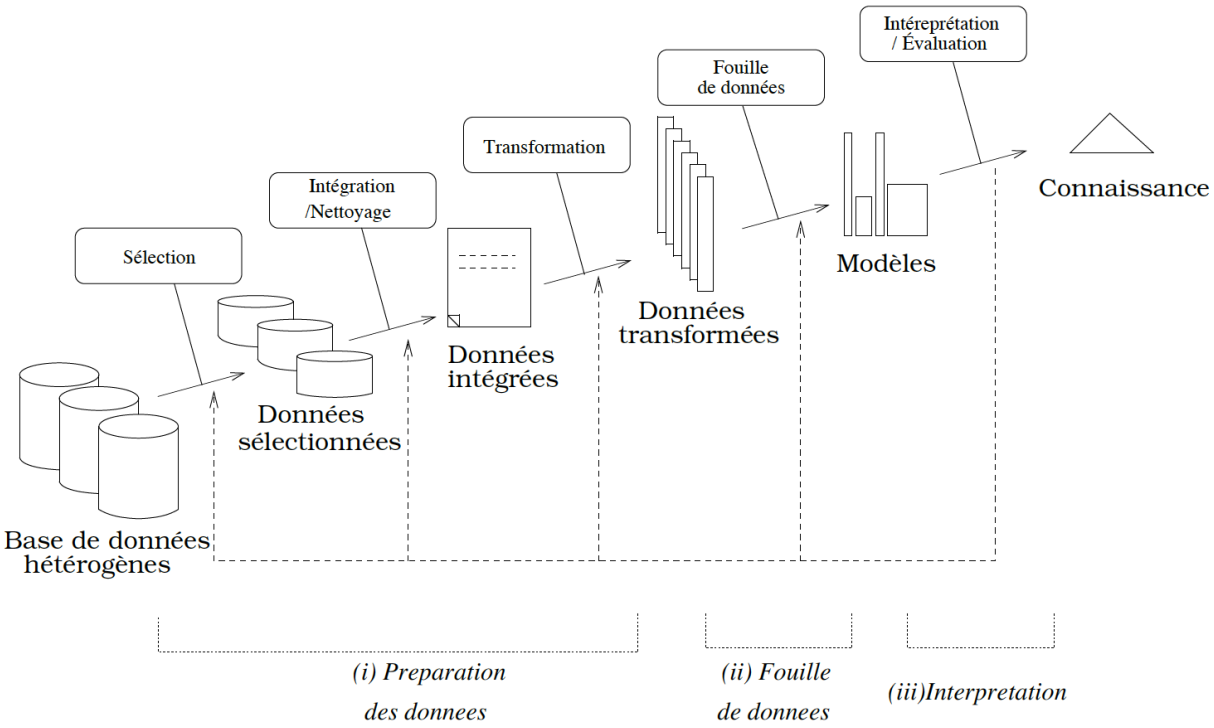


FIGURE 1 – Représentation du processus d'Extraction de Connaissances à partir des Bases de Données (ECBD, aussi appelé *knowledge discovery*), d'après [Fayyad *et al.*, 1996].

notations est plus complexe lorsqu'il s'agit d'annoter des relations entre entités ou des anaphores. Quand ces annotations sont de bonne qualité, elles peuvent servir d'ensemble d'apprentissage pour entraîner des modèles supervisés ou de référence pour évaluer ces modèles. Si nous suivons notre exemple d'annotation de médicaments, un modèle peut être entraîné à reconnaître automatiquement les mentions de médicaments et les performances de cette tâche sont évaluées en estimant comment le modèle réussit à reproduire les annotations de référence.

La découverte de connaissances Un type moins direct de correspondance entre données et connaissances que j'aimerais développer ici est la correspondance qui existe entre un sous-ensemble de données partageant une régularité et l'unité de connaissances que cette régularité permet d'identifier au sein d'un ensemble plus large de données. Ici, la correspondance se fait entre d'une part des régularités (motifs, clusters, classes, concepts, etc. produits par la fouille de données) et d'autre part des éléments de connaissances qui caractérisent bien cet ensemble de données.

Cette mise en correspondance peut-être également décrite en s'appuyant sur le processus classique d'extraction de connaissances à partir de données (ECBD ou *KD* pour *Knowledge Discovery* en anglais) présenté Figure 1 [Fayyad *et al.*, 1996]. Ce processus débute par la considération de données hétérogènes, qui sont préparées (*i.e.*, intégrées, structurées), fouillées, pour générer des régularités qui sont ensuite interprétées pour donner naissance à des connaissances. Dans la description classique du processus d'ECBD, les connaissances finales ne sont pas formelles, ces connaissances sont plutôt le résultat de l'interprétation de l'expert. Il s'agit dans le cas qui nous intéresse de faire que cette étape d'interprétation conduise non seulement à une connaissance informelle, mais que la régularité observée soit associée à une représentation de

connaissances formelle à l'aide d'une ontologie. Il résulte dans ce cas un lien entre un élément de connaissance et un sous-ensemble de données regroupé par la fouille.

Pour illustrer cette correspondance, considérons trois exemples avec trois types de régularités générées à partir de données : un *cluster* généré par un algorithme de clustering non-supervisé comme les K-moyen [Steinhaus, 1956] ; une *classe* définie à partir des instances d'entraînement d'un modèle supervisé tel qu'une régression logistique ; et un *concept formel* construit lors du calcul d'un treillis en Analyse Formelle de Concepts (AFC) [Ganter and Wille, 1999]. Dans ces trois cas, le cluster, la classe et le concept formel ne sont pas des connaissances à proprement parler, mais des parties de constructions mathématiques qui sont porteuses d'un certain niveau de connaissance et qu'il est très tentant d'associer à une sémantique formelle pour en faciliter l'interprétation. Ce type d'association a été proposé dans différents travaux, qui associent les attributs sur lesquels sont construits les clusters, classes ou concepts à des concepts d'ontologies afin de caractériser ces constructions.

Par exemple les données transcriptomiques qui permettent d'étudier la façon avec laquelle les gènes s'expriment de façon différente selon les tissus ou les conditions, sont classiquement analysées par clustering. Les gènes qui sont sur- ou sous- exprimés dans les mêmes conditions sont alors groupés en clusters. Il est cependant très délicat pour les experts biomédicaux de donner un sens à ces clusters et de comprendre les fonctions ou processus particuliers qui pourraient être communs à ces gènes et alors être associés à la condition étudiée. Pour faciliter cette interprétation, les annotations des gènes avec l'ontologie Gene Ontology (GO) sont analysées par une analyse par enrichissement [Draghici *et al.*, 2003, Subramanian *et al.*, 2005]. Cette analyse tire partie du vocabulaire contrôlé et des hiérarchies offerts par GO pour mettre en avant les concepts de GO les plus caractéristiques de chaque cluster de gènes. Au final chaque cluster est associé à des concepts GO qui décrivent des fonctions moléculaires, processus biologiques ou localisations cellulaires plus particulièrement utilisés pour annoter les gènes du cluster. Ces types de concepts sont plus parlant pour les experts que la liste initiale de noms de gènes, facilitant ainsi l'interprétation de l'analyse. Pour ce qui nous intéresse le résultat de l'enrichissement permet d'associer un ensemble de concepts ontologiques à un cluster.

Plusieurs méthodes basées sur l'AFC proposent de mettre en correspondance les concepts et relations définis par un treillis avec des éléments ontologiques. Ciminao *et al.* ont décrit différentes façons de faire de telles correspondances entre concepts formels et concepts ontologiques et comment elles peuvent servir soit à guider l'AFC, soit à aider la construction d'ontologie [Cimiano *et al.*, 2004]. Par exemple, Shi *et al.* utilise les construits produits par l'AFC pour structurer et compléter le contenu d'un wiki sémantique [Shi *et al.*, 2011]. Concepts formels et concepts ontologiques sont particulièrement proches et le fait que les premiers sont définis à partir de régularités au sein des données, permet de confronter une vue apprise à partir des données aux concepts ontologiques potentiellement définis en intention. Nous avons utilisé ce genre de correspondances dans les travaux décrits dans les Chapitres 1 (1.4), 2 (2.2) et 3 (3.3.2).

Dans le cas plus complexe d'un modèle supervisé de classification binaire, la description des deux classes est purement définie par la matrice creuse composée des embeddings des instances. Dans ce cas on peut considérer que la classe est définie en extension, *i.e.*, seulement par des exemples (et leurs propriétés). La mise en correspondance d'une telle description (comme une matrice creuse) avec une ontologie n'est pas facile. Mais une idée est d'enrichir la composition des embeddings décrivant les instances avec des éléments de connaissances définis au préalable dans l'ontologie de sorte que la description des classes inclue ce genre d'éléments et que la considération d'une nouvelle instance par le modèle puisse également considérer ces éléments.

Après avoir appris des éléments de connaissances à partir de données, la complétude de l'ontologie associée et son adéquation avec les données analysées sont les critères qui entrent en

discussion. En effet, il faut effectivement que les éléments de connaissance émergeant des données puissent trouver une correspondance dans l'ontologie utilisée. Quand ce n'est pas le cas, il peut être intéressant de considérer le résultat de la fouille pour compléter ou corriger l'ontologie.

3 La découverte de connaissances guidée par les connaissances

Un axe de recherche cher à l'équipe Orpailleur du Loria-Inria Nancy est de guider le processus d'ECBD, non plus seulement par les connaissances de l'analyste, mais également par des connaissances exprimées dans un langage de représentation des connaissances, interprétable par une machine. Puisque ce processus est itératif, ces connaissances peuvent pré-exister ou résulter d'une itération précédente du processus d'ECBD. Dans la version classique de l'ECBD, les connaissances résultantes sont bien réutilisées à l'itération suivante du processus, mais celles-ci sont les connaissances de l'analyste et ne sont pas formalisées. L'idée est alors de les formaliser de sorte que leur utilisation soit plus automatique et intégrée au processus. Nous appelons le processus résultant l'ECCD pour Extraction de Connaissance guidée par les Connaissances du Domaine (ou KDDK en anglais pour *Knowledge Discovery guided by Domain Knowledge*). Il a été décrit de façon générale dans [Lieber *et al.*, 2006] et nous l'avons instancié dans des travaux présentés dans ce mémoire, notamment des travaux d'AFC [Coulet *et al.*, 2013], d'ingénierie des connaissances [Monnin *et al.*, 2017b] et de fouille de texte [Coulet *et al.*, 2010]. Dans tous les cas de figure, l'utilisation des connaissances pour guider le processus s'appuie sur une mise en correspondance entre données et connaissances. Comment utiliser les annotations et les connaissances associées pour guider l'une ou l'autre des étapes de l'extraction de connaissances est une des problématiques transversale de ce mémoire et un fil de lecture.

4 La médecine de précision comme domaine d'application

L'extraction de connaissances est une activité motivée par les applications et dans le cas des travaux présentés dans ce mémoire, les applications sont toutes motivées par le développement d'une médecine de précision et plus particulièrement de la pharmacogénomique (noté dans ce document `PGx`, abréviation classique de l'anglais *pharmacogenomics*)

La **médecine de précision** est une médecine qui prend au maximum en considération les données disponibles sur le patient, c'est-à-dire son histoire, son alimentation, ses profils génétiques et moléculaires, afin d'adapter la prise en charge à l'individu [Goldman and Goldman, 2019]. Une notion proche de celle-ci est la médecine personnalisée, mais parce qu'il n'est pas particulièrement nouveau de considérer l'individu en médecine, nous lui préférons le terme précision, qui met l'accent sur le fait que la personnalisation se fait avec plus de précision, en considérant davantage de données. La disponibilité des données de l'individu et leur considération vis-à-vis de connaissances biomédicales générales est cruciale et donne un rôle particulier à l'informatique médicale pour la mise en œuvre d'une telle médecine [Frey *et al.*, 2016]. Les contributions de ce manuscrit s'intéressent d'un point de vue informatique à étudier comment la considération des données du patient d'une part et des textes et connaissances plus générales d'autre part participent au développement d'une médecine plus précise.

Un domaine clé pour la mise en oeuvre d'une médecine de précision est la **pharmacogénomique** (`PGx`). Celle-ci étudie comment les variations génétiques impactent la réponse individuelle aux médicaments [Weinshilboum and Wang, 2017]. La `PGx` est d'une importance particulière car les connaissances bien validées de `PGx` contribuent déjà à la médecine de précision par la production de guides de bonnes pratiques cliniques (*clinical guidelines* en anglais) comme ceux du

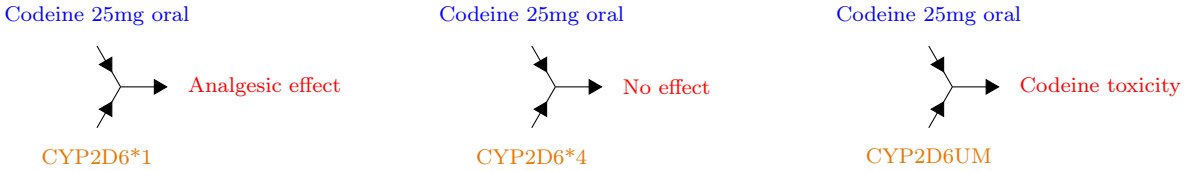


FIGURE 2 – Exemple de connaissance pharmacogénomique illustrant l'impact des variants du gène CYP2D6 sur la réponse à la codéine.

CPIC (Clinical Pharmacogenetics Implementation Consortium) qui décrivent comment mieux choisir et doser les médicaments en prenant en considération le profil génétique des patients, afin de réduire le nombre d'effets indésirables [Relling and Klein, 2011, Lin and Chung, 2019]. Les unités de connaissances en `PGx` ont typiquement la forme de relations ternaires (variant génétique–médicament–effet indésirable), signifiant que l'effet indésirable risque d'être observé chez les patients porteurs du variant et exposés au médicament. Par exemple, une relation pharmacogénomique bien étudiée est la suivante : G6PD:202A–chloroquine–anémie qui signifie que les patients porteurs de la version 202A du gène G6PD et traités avec la chloroquine (un antipaludique) ont des risques de présenter une anémie (un niveau anormalement bas de globules rouges dans le sang). La Figure 2 représente un second exemple de connaissance pharmacogénomique impliquant la codéine et trois versions du gène CYP2D6.

Plusieurs particularités font que les connaissances `PGx` sont délicates à manipuler en informatique. Tout d'abord, les connaissances de l'état de l'art en `PGx` sont réparties entre les bases de données biomédicales spécialisées comme PharmGKB [Whirl-Carrillo *et al.*, 2012] et la littérature biomédicale [Garten, 2010]. Ensuite, il se trouve qu'une grande partie (~80%) de ces connaissances n'ont été observées que de façon sporadique, et pour cette raison ne sont pas encore utilisées en pratique clinique [PharmGKB web page, 2019]. En effet, ces connaissances résultent souvent d'études difficiles à reproduire et qui ne remplissent pas les standards statistiques classiques à cause de la petite taille des populations de patients considérées. Ceci est causé notamment par la rareté des variants génétiques étudiés et l'effet modéré de certains variants génétiques, masqués ou modulés par d'autres facteurs génétiques ou non [Ioannidis, 2013, Zineh *et al.*, 2013]. Par exemple, les patients avec le variant 202A du gène G6PD, connu pour causer des anémies chez les patients traités par chloroquine, ont aussi été identifiés, à partir d'observations plus sporadiques, comme présentant des anémies quand ils sont traités avec le glyburine, un autre médicament utilisé dans le traitement du diabète de type 2. La relation `PGx` G6PD:202A–glyburide–anemia est présente dans la littérature et peut en être extraite automatiquement puisqu'elle est décrite dans deux articles de PubMed (PMID 15126005 et 21147013). Cependant ces articles sont des publications de type "cas clinique", rapportant des observations faites sur seulement un patient. Cette relation nécessite d'être soumise à investigation sur des populations plus grandes avant de pouvoir être validée et éventuellement utilisée en clinique. Si cette relation est confirmée, il est possible d'imaginer que de façon routinière les cliniciens souhaitant prescrire du glyburide à un patient demandent un génotypage du variant G6PD:202A et choisissent un médicament alternatif si le patient est à risque pour une anémie causée par le glyburide.

L'exemple précédent est intéressant car il illustre combien il peut être bénéfique d'être capable de comparer le contenu de bases de données (où sont les connaissances bien validées), de la littérature (où sont les connaissances en besoin de validation) et de bases de données cliniques (où sont les données qui pourraient être analysées pour confirmer ou modérer les connaissances

de la littérature). Pour être automatisé, ce scénario nécessite des approches d'extraction de connaissances à partir de textes, de comparaison de connaissances et de fouille de données. C'est l'objectif du project ANR PractiKPharma (2016-2020) que de développer des approches qui permettent ces extractions et comparaisons de connaissances `PGx` [?]. Les résultats de ce projet sont présentés dans ce mémoire.

Au delà de la `PGx`, la médecine personnalisée soulève de nombreux autres défis auxquels la gestion des connaissances et la fouille de données peuvent apporter des solutions. En particulier, certains travaux présentés ici concernent l'extraction et la normalisation de phénotypes complexes à partir de textes [Hassan, 2017] ; et la caractérisation génotypique et phénotypique des maladies complexes comme les déficiences intellectuelles (DI) [Personeni, 2018].

5 Le contexte : projets de recherche et encadrements

Les travaux décrits dans ce mémoire ont été menés dans le cadre de projets de recherches et d'encadrement de thèses dans lesquels j'ai été impliqué après ma thèse de doctorat, c'est-à-dire de octobre 2008 à décembre 2019. Cette section en présente les grandes lignes.

Une description plus détaillée de mon profil est faite sur mon curriculum vitæ disponible en ligne à l'adresse `https://tinyurl.com/y67kdf7h`. J'ai volontairement séparé les deux documents pour que ce mémoire ne concerne que l'aspect scientifique de l'habilitation.

Le NCBO et PharmGKB J'ai eu la chance de participer conjointement aux projets du National Center for Biomedical Ontologies (NCBO) et de PharmGKB (the Pharmacogenomics Knowledge Base) durant mon postdoctorat à l'Université Stanford (2008-10). Mes contributions à ces projets ont consisté en la participation au développement d'un index global du contenu de bases de données biologiques à l'aide d'ontologies (*i.e.*, le NCBO Resource Index) [Jonquet *et al.*, 2011] et au développement de méthodes d'extraction de connaissances `PGx` normalisées à partir de texte [Coulet *et al.*, 2010].

Le PEPS EXPLOD-BioMed et la thèse de Gabin Personeni Ce Projet Exploratoire Premier Soutien d'un an (2013) a permis de démarrer une collaboration avec l'équipe de génétique humaine du CHRU de Nancy. Le projet portait sur la fouille de données ouvertes et liées pour la recherche de gènes responsables des déficiences intellectuelles. Le PEPS a en particulier permis de définir le sujet de la thèse de G. Personeni et de soutenir son démarrage. Cette thèse nous a ensuite offert l'opportunité de poursuivre sur ce thème [Personeni, 2018].

Le projet ANR Hybride et la thèse de Mohsen Hassan Le projet Hybride (2013-16) porté par Yannick Toussaint du Loria s'est intéressé au développement d'approches de fouille de données hybrides numériques – symboliques pour la caractérisation des maladies rares. Au sein de ce projet, j'ai co-encadré la thèse de Mohsen Hassan qui portait sur une tâche d'extraction et de normalisation de relations maladie–symptôme à partir de textes [Hassan, 2017].

La thèse de pharmacie de Kevin Dalleau J'ai encadré Kevin Dalleau, un étudiant avec une double compétence forte, puisqu'il combinait à l'époque études de pharmacie et d'informatique (à TELECOM Nancy). Son intérêt pour les deux domaines nous ont mené à des travaux sur la prédiction de liens dans des jeux de données liées biomédicaux qui ont été publiés dans une conférence [Dalleau *et al.*, 2015] et un journal [Dalleau *et al.*, 2017] internationaux. Sa thèse

d'exercice de pharmacie a reçu en 2018 le prix de la thèse d'exercice à caractère expérimental de l'Université de Lorraine.

Le projet ANR PractiKPharma et la thèse de Pierre Monnin Ce projet de 4 ans (2016-20) dont je suis porteur a pour objectif l'extraction et la comparaison de connaissances pharmacogénomiques à partir de sources diverses : littérature, bases de données expertes, entrepôt de dossiers patients électroniques [**?**]. La thèse de Pierre Monnin est financée dans le cadre de ce projet et est axée sur la comparaison de connaissances, notamment à l'aide de l'AFC.

La collaboration avec le Stanford Center for Biomedical Informatics (BMIR) Cette collaboration a été soutenue par une équipe-associée Inria (2014-19) commune aux Orpailleurs du Loria et aux équipes de Nigam H. Shah et de Michel Dumontier à Stanford. L'équipe associée a facilité nos collaborations autour de la découverte de connaissances à partir de données de patients. Elle a notamment permis à G. Personeni de réaliser un séjour de 4 mois dans l'équipe de M. Dumontier et à moi-même de passer 2 ans de délégation (2017-19) dans l'équipe de N. Shah. J'ai pu y développer des travaux sur la prédiction de réponses aux médicaments et y étudier les thèmes développés dans la section intitulée Perspectives de recherche.

6 Contenu et organisation du mémoire

Ce manuscrit n'a pas pour ambition de lister exhaustivement les travaux de recherches que j'ai pu mener depuis mon doctorat, et ne le fait pas. Il ambitionne plutôt de présenter les travaux qui illustrent le mieux les utilisations que j'ai pu faire de correspondances entre données et connaissances dans des travaux d'analyses de données ou d'extraction de connaissances.

Les trois chapitres décrivent des méthodes où interviennent des connaissances, et chaque chapitre se distingue par le type de données auquel sont associées les connaissances : des bases de données biomédicales pour le Chapitre 1 ; des entrepôts de données cliniques pour le Chapitre 2 ; des graphes de connaissances pour le Chapitre 3.

En particulier, le **Chapitre 1** traite d'un processus d'annotation de bases de données à l'aide d'une large collection d'ontologies et de la constitution résultante d'un index sur ces bases, appelé le *Resource Index*. Le chapitre illustre ensuite l'intérêt d'un tel ensemble d'annotations avec la description de deux travaux d'analyse qui ont été menés sur le Resource Index. Le premier qui est très appliqué s'est intéressé à mettre en évidence des associations, par sujet de recherche (*i.e.*, par maladie), entre financement de recherche et publication. Le second travail est plus théorique et propose d'adapter une extension d'AFC appelée les *structures de patrons* pour analyser les annotations et permettre de considérer les connaissances dans l'analyse. Enfin nous évoquons brièvement à la fin du chapitre nos travaux sur l'annotation automatique ou manuelle de textes pour l'extraction de relations à partir de la littérature.

Le **Chapitre 2** reprend le thème de l'utilisation d'annotations et des connaissances associées mais pour l'analyse d'un type de données particulier : les entrepôts de données cliniques. Le chapitre rapporte deux expérimentations menées principalement sur les données cliniques de l'hôpital de l'Université Stanford. La première reprend le cadre de l'AFC et des structures de patrons pour la mise en évidence de cooccurrences fréquentes d'effets indésirables médicamenteux dans un groupe de patients. Le problème est que la façon selon laquelle se manifeste un effet indésirable causé par un même médicament est diverse chez les individus et qu'en plus, ces manifestations sont rapportées dans les dossiers patients de façon hétérogène. L'intérêt d'expérimenter la combinaison structures de patrons – ontologies est de permettre une comparaison flexible de

ces événements complexes pour mettre en évidence des effets indésirables souvent associés. Le second travail rapporté dans ce chapitre est assez différent puisqu'il s'agit d'expériences autour d'un modèle prédictif entraîné sur l'historique des patients. Nous avons évalué la capacité de données phénotypiques classiquement disponibles dans les dossiers patients à prédire pour un nouveau patient à qui l'on souhaite prescrire un nouveau médicament si ce patient risque de bénéficier d'une prescription à dose réduite ou à dose forte.

Le **Chapitre 3** s'intéresse à la fouille de graphes de connaissances, c'est-à-dire à un espace où les données de différentes origines et connaissances associées sont connectées et représentées dans un formalisme homogène. Dans ce cas, les annotations sont des éléments clés puisqu'elles définissent dans le graphe les liens entre données et connaissances. Elles sont alors indispensables si l'on souhaite considérer les connaissances dans le processus de fouille. Le chapitre s'articule autour de deux tâches classiques dont l'objectif est la complétion des graphes de connaissances : la prédiction de liens, et l'enrichissement des ontologies associées. Concernant la prédiction de liens nous décrivons deux contributions, la première à propos de la prédiction de liens gène-médicament dans un graphe de connaissances pharmacogénomiques ; la seconde à propos de la prédiction de liens d'identité (*i.e.*, qui spécifient que deux entités font référence au même objet) dans le même graphe. Concernant la complétion des connaissances nous décrivons d'abord une expérience utilisant la programmation logique inductive pour apprendre, à partir de données ouvertes et liées, des descriptions concernant des classes de gènes. Nous finissons ce chapitre par une utilisation originale de l'AFC où nous proposons d'annoter les concepts formels d'un treillis pour mettre en évidence des nouveaux axiomes qui pourront venir enrichir une ontologie.

Mes perspectives de recherches concluent ce mémoire.

Chapitre 1

Annotation de données pour l'extraction de connaissances

Sommaire

1.1 Introduction 11
1.1.1 Quatre types d'annotations, selon leur origine 12
1.1.2 La sémantique des annotations 14
1.2 Annotation de bases de données biologiques avec le Bioportal et le Resource Index 18
1.3 Analyse d'annotations pour la recherche d'associations entre financement de recherche et publication 21
1.3.1 Données et ontologie 21
1.3.2 Processus d'annotation 22
1.3.3 Analyses 23
1.3.4 Discussion et conclusion 23
1.4 Les structures de patrons pour considérer les connaissances associées aux annotations 25
1.4.1 L'analyse formelle de concepts 25
1.4.2 Les structures de patrons 26
1.4.3 Adaptation des structures de patrons aux annotations 27
1.4.4 Exemple 31
1.4.5 Discussion et conclusion 33
1.5 Quelques mots sur les textes 36
1.6 Discussion générale 38

1.1 Introduction

En analyse de données, il est souvent fait distinction entre *données structurées*, *semi-structurées* et *non-structurées*. Les données structurées et semi-structurées sont associées à un schéma de données plus ou mois strict qui facilite leur interprétation et documente leur composition. Dans le cas de données structurées, les objets (ou nuplets si l'on se place dans le paradigme relationnel plutôt qu'objet) peuvent être décrits par la composition d'un ensemble d'attributs nommés, ordonnés, typés, et associés à une valeur atomique. Les données semi-structurées sont associées à un schéma plus lâche, qui permet d'associer à un attribut une portion de texte ou un ensemble

TABLE 1.1 – Exemples d'annotations de données structurées et semi-structurées à partir des bases de données UniProt et DrugBank. `INS` est le symbole de l'insuline, `GO, ATC, DBCOND, ICD9CM, tax` sont des abréviations pour les ontologies Gene Ontology, ATC, DrugBank Conditions, ICD9-CM

Type d'annotation	*Source de données*	*Relation attribut-valeur dans la source*	*Annotation*
Préexistante	UniProt [1]	`Function(INS, GO:0006006)`	`Function(INS, GO:0006006)`
	DrugBank [2]	`ATC(warfarin, ATC:B01AA)`	`ATC(warfarin, ATC:B01AA)`
Préexistante, traduite	DrugBank [3]	`Condition(warfarin, DBCOND:0085133)`	`Condition(warfarin, ICD9CM:431)`
A partir d'un champ texte contrôlé	DrugBank [4]	`TargetOrganism(abacavir,` *"Human immunodeficiency virus 1 "*)	`TargetOrganism(abacavir,tax:11676)`
A partir d'un champ texte libre	DrugBank[3]	`Indication(warfarin,`*"Secondary prevention of stroke and transient ischemic attacks..."*)	`Indication(warfarin, ICD9CM:431)` `Indication(warfarin, ICD9CM:435)`

de valeurs. A contratio, les données non-structurées ne sont associées à aucun schéma particulier et les objets y sont décrits sans cadre prédéfini, comme c'est le cas d'un texte en langage naturel.

En guise d'introduction de ce chapitre, nous présentons comment les attributs des données structurées ou semi-structurées peuvent être annotés dans le cas simple d'annotations à l'aide de concepts ontologiques ; puis quelle sémantique peut être associée à ces annotations. Ensuite nous présentons nos contributions avec, en Section 1.2, la description d'un processus d'annotation automatique de bases de données biologiques, puis son utilisation pour la constitution d'une base d'annotations [Jonquet *et al.*, 2010, Jonquet *et al.*, 2011]. La suite illustre l'intérêt d'un tel ensemble d'annotations avec la description de deux travaux d'analyse qui ont été menés sur le Resource Index. Le premier qui est très appliqué s'est intéressé à mettre en évidence des associations, par sujet de recherche (*i.e.*, par maladie), entre financement de recherche et publication [Liu *et al.*, 2012]. Le second travail est plus théorique et propose d'adapter une extension d'AFC appelée les *structures de patrons* pour analyser les annotations et permettre de considérer les connaissances dans l'analyse [Coulet *et al.*, 2013]. Enfin nous évoquons brièvement à la fin du chapitre nos travaux sur l'annotation automatique ou manuelle de textes pour l'extraction de relations à partir de la littérature.

1.1.1 Quatre types d'annotations, selon leur origine

Pour clarifier la suite de ce document, nous définissons quatre types d'annotations selon leur origine. Ces types sont listés et illustrés dans le Tableau 1.1 puis détaillés dans ce qui suit. Ces types d'annotations ne constituent pas des ensembles disjoints, il est par exemple tout à fait envisageable qu'une annotation soit préexistante, traduite et associée à du texte libre.

Les annotations préexistantes Parfois, les valeurs associées à un attribut sont déjà des concepts ontologiques, et ne nécessitent alors pas d'être annotées. Il peut être considéré dans

1. Les fonctions de l'insuline humaine dans UniProt : `https://www.uniprot.org/uniprot/A6XGL2#function`
2. Les codes ATC de la warfarine dans DrugBank : `https://www.drugbank.ca/drugs/DB00682#references`
3. Les indications de la warfarine dans DrugBank : `https://www.drugbank.ca/drugs/DB00682#pharmacology`
4. Les cibles de l'abacavir dans DrugBank : `https://www.drugbank.ca/drugs/DB01048#targets`

ce cas que l'annotation a déjà été réalisée. C'est par exemple le cas des valeurs associées à l'attribut *Function* des protéines répertoriées dans la base UniProt, qui associe à une protéine un ensemble de fonctions moléculaires et/ou de processus biologiques à l'aide des concepts de la Gene Ontology (GO). Comme illustré dans le Tableau 1.1, l'insuline humaine (`https://www.uniprot.org/uniprot/P01308`) est déjà associée dans UniProt au processus biologique `Glucose metabolic process` avec le concept `GO:0006006`. Ces annotations préexistantes dans une source de données peuvent avoir diverses origines : elle peuvent avoir été faites manuellement par des humains experts du domaine, appelés *annotateurs* ; elles peuvent avoir été suggérées automatiquement à des experts qui les valident avant qu'elles ne viennent enrichir les données ; elles peuvent être générées complètement automatiquement, sans validation humaine. Les annotations produites selon les deux premières modalités sont généralement considérées de bonne qualité, alors qu'avec la dernière, leur qualité peut être légitimement questionnée.

Les annotations préexistantes, traduites Il arrive que l'ontologie utilisée dans une source de données ne soit pas standard ou ne soit pas celle souhaitée pour une application particulière. Il est alors nécessaire de traduire les annotations dans les termes d'une ontologie plus standard ou plus adaptée au cas d'utilisation. Par exemple, la base de données de médicament DrugBank associe à chaque médicament un ensemble de conditions (sans qu'il soit précisé s'il s'agit d'indications ou d'effets secondaires), et celles-ci sont désignées avec les identifiants et labels d'une ontologie locale, appelée *DrugBank Conditions*. On peut facilement imaginer que pour une étude épidémiologique, un chercheur ait besoin que les conditions associées aux médicaments soient décrites par les codes standards de la classification internationale des maladie ICD9-CM. En effet, ceux-ci sont classiquement utilisés pour encoder les diagnostics des patients dans les entrepôts de données cliniques. Lier des connaissances à ce genre de données cliniques est alors facilité si l'on utilise cette ontologie standard. Il s'agit alors de traduire les annotations disponibles dans DrugBank dans les termes de la nouvelle ontologie. Ce processus peut se faire par l'utilisation d'alignements (parfois appelés *mappings*) préexistants entre ontologies ou par le calcul d'un nouvel alignement avec une méthode de mise en correspondance d'ontologies (voir [Euzenat and Shvaiko, 2013] pour un panorama de ces méthodes). Par exemple, le Tableau 1.1 illustre comment la warfarine, un anticoagulant, qui est associée dans DrugBank à la condition *Strokes* (identifiant : `DBCOND0085133`), encodée avec l'ontologie locale des Conditions de DrugBank, peut être à son tour associée au concept *Intracerebral hemorrhage* (identifiant : `431`) d'ICD9-CM à l'aide d'une hypothétique méthode d'alignement.

Cependant, dans le cas général, les valeurs des attributs des données structurées ne sont pas déjà associées à des concepts ontologiques. Il est alors utile de les annoter soi-même en considérant les valeurs observées pour un attribut afin de les associer avec les concepts adéquats d'une ontologie cible, (s'ils existent !).

Les annotations de texte contrôlé L'annotation avec des concepts ontologiques est un problème relativement simple lorsque les valeurs prises par l'attribut se limitent à un terme ou un mot et que le domaine de cet attribut est limité, alors des méthodes de mise en correspondance de chaînes de caractères entre les valeurs possibles et les labels associés aux concepts d'une ontologie cible permettent de proposer des annotations automatiques. Nous pouvons ainsi imaginer annoter l'attribut `organism of target` de DrugBank qui décrit l'espèce de l'organisme chez lequel on trouve la molécule cible du médicament avec la taxonomie des espèces de la *NCBI Taxonomy Database*. Dans l'exemple du Tableau 1.1, c'est l'organisme *HIV1* qui porte la molécule cible de l'antiretroviral *abacavir*. Dans cet exemple, le domaine des valeurs possibles de l'attribut est

assez restreint et est retrouvé de façon non ambiguë dans les labels des concepts de la taxonomie des espèces.

Cependant l'annotation est souvent rendue plus difficile par le fait que les valeurs en question ne se limitent pas à un mot mais sont des expressions complexes, en langage naturel. Dans ce cas on ne parle plus de données structurées mais de données semi-structurées.

Les annotations de texte libre Dans le cas de valeurs dont le domaine est le langage naturel, il est nécessaire non seulement de trouver les concepts adéquats dans l'ontologie cible, mais également d'isoler les termes à considérer au sein du texte. Cette distinction peut être guidée par l'utilisation d'outils de traitement de la langue (de *NLP* pour l'anglais *Natural Language Processing*) . Par exemple, le Tableau 1.1 présente un extrait de la base DrugBank avec l'attribut *Indications* de la warfarine dont la valeur est "*Secondary prevention of stroke and transient ischemic attacks in patients with rheumatic mitral valve disease but without atrial fibrillation*". L'annotation de cet attribut avec des concepts ontologiques est complexe car il faut dans un premier temps localiser correctement les phénotypes puis distinguer ceux qui sont des indications (*stroke*, *transient ischemic attacks*) de ceux qui sont des éléments de contexte (*rheumatic mitral valve disease* et *atrial fibrillation*). Ensuite, il faut associer aux portions de texte décrivant des indications des concepts ontologiques. Nous pouvons identifier ici trois tâches de NLP : la reconnaissance d'entité nommées, la normalisation, la contextualisation.

1.1.2 La sémantique des annotations

Le terme "annotation" ne fait pas référence à une notion unique, mais à plusieurs qui ont en commun d'associer une portion de données à un élément de connaissance. D'ailleurs le terme peut aussi bien faire référence au processus d'annotation, c'est-à-dire la création d'une association donnée–connaissance, qu'au résultat de ce processus : par exemple une paire portion de texte–concept ontologique. Une annotation peut être plus complexe que cette simple paire. Si l'on considère par exemple la portion de texte *"'warfarin-induced bleeding"*, celle-ci peut être annotée dans sa globalité avec un concept `Drug Response`, mais également en partie avec l'annotation de *warfarin* par `Drug` et *bleeding* par `Phenotype`. Il est également possible d'annoter une relation de causalité entre *warfarin* et *bleeding*. Un des objectifs de ce mémoire est de lever l'ambiguïté autour de la notion d'annotation en clarifiant ce que nous entendons exactement par là dans chacun des travaux abordés. Une façon de clarifier cela est de proposer une formalisation des annotations et de leur associer une sémantique, c'est-à-dire un façon non-ambiguë de les interpréter. Ceci n'est ni évident ni classique, car de nombreuses ontologies (du domaine biomédical mais pas seulement) sont des vocabulaires contrôlés organisés selon une hiérarchie et n'ont pas été développés suivant les standards du Web sémantique.

Nous proposons dans cette section de présenter différentes interprétations qui peuvent être associées aux annotations, et cela en traduisant les exemples d'annotations du Tableau 1.1 en axiomes de Logique de Description (LD) auxquels il est associé une sémantique précise en termes ensemblistes comme décrit dans le Tableau A 1.1 de l'Annexe 1 de [Baader *et al.*, 2010].

Instanciation ou subsomption Une annotation peut être vue comme le typage d'une donnée avec un concept ontologique. L'annotation peut alors être représentée soit par une assertion de concept, soit par une subsomption, selon que l'on choisisse de représenter l'unité de données comme un individu (le pendant des objets en LD) ou comme un concept.

TABLE 1.2 – Exemples d'interprétations possibles d'annotations sous forme d'axiomes de Logique de Description (LD). Les exemples d'annotations reprennent ceux proposés dans le Tableau 1.1

Annotation	*Axiomes de LD*
`catégorie_wikipédia(paris, VilleOrganisatrice)`	`VilleOrganisatrice(paris)`
`ATC(warfarin, ATC:B01AA)`	ATC :B01AA(warfarin) *ou* warfarin ⊑ ATC :B01AA
`Function(INS, GO:0006006)`	Axiomes 1.1, 1.2 ou 1.3

`Condition(warfarin, ICD9CM:431)`	Suivant l'exemple des Axiomes 1.2
	`ICD9CM:431(intracerebral_hemorrhage)`
	`condition(warfarin,intracerebral_hemorrhage)`

`TargetOrganism(abacavir,tax:11676)`	`Target(abacavir,reverse_transcriptase)`
	`tax:11676(hiv1)`
	`Organism(reverse_transcriptase,hiv1)`

Un exemple simple est l'annotation de la page Wikipedia de la ville de Paris [5] avec le concept "Ville organisatrice des jeux olympiques d'été", noté ci-après `VilleOrganisatrice` et dont l'interprétation est assez naturellement une instanciation et peut alors être représentée par l'axiome d'assertion de concept en LD

`VilleOrganisatrice(paris)`

qui est associé à l'interprétation suivante : `paris` est un individu qui est une `VilleOrganisatrice` et par là remplit toutes les conditions nécessaires à l'appartenance à ce concept, comme par exemple avoir un maire et une date d'ouverture des jeux.

Si le choix de représentation semble assez naturel pour l'exemple précédent, les choses le sont moins quand l'unité de données à annoter peut être considérée comme un individu ou un concept. C'est par exemple le cas pour les médicaments qui peuvent être vus comme un individu (la molécule) ou un concept (l'ensemble des matérialisations du médicament, *i.e.*, comprimés, gélules, etc. avec cette molécule). Chaque représentation peut se justifier selon l'application visée. Ainsi si l'on considère l'annotation de la warfarine dans la base DrugBank [6] avec la classe de médicament ATC `ATC:B01AA - Vitamin K antagonists`, celle-ci peut-être interprétée comme une subsomption et alors représentée en LD par l'axiome suivant :

$$\texttt{Warfarin} \sqsubseteq \texttt{ATC:B01AA}$$

(ou $\texttt{Warfarin} \equiv \texttt{ATC:B01AA03}$ si l'on considère une annotation avec le niveau inférieur). Dans ce cas il est possible de représenter les matérialisations de la warfarine par des individus instances du concept `Warfarin`. Par exemple le comprimé de warfarine dont l'identifiant est THX1138 que le patient 42 a pris ce matin peut être représenté par l'axiome

`Warfarin(warfarin_tablet_THX1138).`

Cependant représenter la warfarine comme un individu permet de représenter l'annotation avec l'axiome suivant `B01AA(warfarin)`. Ceci peut être adéquat notamment car cela permet de lier directement la molécule aux patients ayant été traités avec le seul axiome `treatedWith(patient42, warfarin)` ce qui constitue un raccourci pratique pour l'écriture de requêtes quand la traçabilité ou l'inventaire des comprimés de médicaments n'est pas en question.

5. Page de la ville de Paris sur wikipédia.fr : `https://fr.wikipedia.org/wiki/Paris`
6. Les codes ATC de la warfarine dans DrugBank : `https://www.drugbank.ca/drugs/DB00682#references`

Il s'agit ici d'un choix de modélisation qui doit être motivé par l'application visée : est-ce que l'on veut tracer les comprimés administrés dans un hôpital ou rechercher les patients ayant été exposés à une même molécule ? En d'autres termes ce choix matérialise la limite dans une base de connaissances entre TBox et ABox. C'est-à-dire entre concepts et individus. Dans les deux cas, il est sain de se demander ce que représente le concept ou l'individu `warfarin` dans le domaine en question et ce que cela implique en terme d'utilisabilité.

Une alternative à citer est la possibilité de représenter la warfarine à la fois comme un concept et comme un individu. Cela est possible et peut être pratique du point de vue de la modélisation de connaissances, mais la contrepartie est que la plupart des mécanismes de raisonnement perdent leur décidabilité quand ils ne peuvent distinguer concepts et individus.

Condition nécessaire ou instanciation de rôle S'il ne s'agit pas d'un typage, une annotation peut être vue comme une propriété de l'objet annoté. Dans ce cas, nous distinguons trois façons de représenter l'annotation, et nous illustrons dans ce qui suit ces représentations avec le même exemple, tiré de la base UniProt, de l'annotation de l'insuline humaine (`INS`) avec le concept `GO:0006006` (métabolisme du glucose) de la Gene Ontology.

La première représentation proposée fait appel à une expressivité relativement élevée puisqu'elle implique la définition d'un nouveau concept représentant l'ensemble des individus ayant la propriété définie par l'annotation et cela avec un quantificateur existentiel, noté $\exists$. Par exemple :

$$\begin{aligned} \texttt{ProtMetaboGlucose} &\sqsubseteq \texttt{Protein} \sqcap \exists\ \texttt{hasFunction.GO:0006006} \\ \texttt{ProtMetaboGlucose(INS)} & \end{aligned} \tag{1.1}$$

où le premier axiome introduit un concept `ProtMetaboGlucose` et le décrit en lui associant une condition nécessaire qui implique que l'ensemble des individus qui l'instancie sont des protéines et ont au moins une fonction, celle d'être impliquée dans le métabolisme du glucose. Le second axiome précise que l'insuline humaine instancie ce concept et satisfait donc la condition. C'est une façon élégante de représenter l'annotation car elle définit implicitement que l'individu `INS` est lié à un individu anonyme qui instancie la classe `GO:0006006`. L'expressivité de cette représentation est celle de la logique $\mathcal{EL}$ qui inclut le quantificateur existentiel et la conjonction.

Une seconde représentation de cette annotation revient à créer de façon explicite un individu qui vient instancier le concept utilisé pour l'annotation. Suivant notre exemple, nous créons un individu `my_fct` qui intervient dans deux assertions, une de concept, et une de rôle :

$$\begin{aligned} &\texttt{GO:0006006(my_fct)} \\ &\texttt{hasFunction(INS, my_fct)} \end{aligned} \tag{1.2}$$

Les avantages de cette représentation sont d'être explicite et de ne nécessiter que l'expressivité du RDF associé à la ABox, car elle ne nécessite pas de quantificateur ou d'opérateur logique absent du RDF.

Une troisième représentation consiste à utiliser directement la classe de l'annotation dans l'assertion de rôle, ce qui donne avec notre exemple :

$$\texttt{hasFunction(INS, GO:0006006)} \tag{1.3}$$

Cette représentation a la particularité d'utiliser `GO:006006` à la fois comme un concept puisqu'il est défini en tant que tel dans la hiérarchie des concepts de GO et comme un individu puisque il participe à une assertion de rôle. Cette représentation est incompatible avec le langage OWL-DL et les raisonneurs associés. Elle peut être envisagée si l'ontologie de référence est encodée

en OWL Full ou dans un langage distinct comme SKOS [Jupp *et al.*, 2008]. Les mécanismes de raisonnement classiques sont indécidables dans le cas de OWL Full et dans celui de SKOS, il n'existe pas de raisonneur standard associé, mais un raisonneur peut être développé et implanter un ensemble de règles qui définiront la sémantique souhaitée. C'est par exemple ce que fait le raisonneur associé au gestionnaire de base de connaissance KiWi et à la librairie Apache Marmotta [Schaffert *et al.*, 2009].

En conclusion, le choix de représentation pour une propriété sera guidé par le niveau d'expressivité souhaité et le besoin en terme de raisonnement. Les travaux présentés dans ce mémoire utilise alternativement les représentations 1.1 ou 1.2, selon les besoins en expressivité et passage à l'échelle.

Multiplicité Une simple annotation entre un concept ontologique et une unité de données, peut en réalité représenter une relation plus complexe, par exemple indirecte, et nécessiter une multiplicité d'axiomes pour être représentée de façon non ambiguë. Par exemple, l'annotation présentée dans la cinquième ligne du Tableau 1.2 représente une relation indirecte entre un médicament, sa molécule cible et l'espèce porteuse de cette cible. Une représentation rigoureuse nécessite de lister explicitement les intermédiaires implicites de cette relation. Cela fait appel à une connaissance experte et semble difficilement automatisable. Dans l'exemple en question, il est nécessaire de savoir que le médicament abacavir cible un enzyme appelée *Reverse transcriptase* et que c'est la Reverse transcriptase du virus HIV 1 qui est ciblée en particulier par le médicament. Il est en revanche possible de définir un noeud blanc à la place de l'individu `Reverse_transcriptase` ou encore d'autoriser que les objets du prédicat `Target` qui sont normalement des protéines puissent de façon plus générale être également des organismes.

Ce dernier exemple illustre bien que formaliser une annotation n'est pas toujours trivial et que cela nécessite de faire des choix qui pourront être motivés par l'application visée. Dans tous les cas, cette étape de formalisation nous semble indispensable au développement d'applications et à des processus d'ECBD qui tirent parti non seulement des données mais également des connaissances associées.

La suite de ce chapitre présente le *Resource Index*, une ressource qui regroupe des annotations faites sur plusieurs bases de données biologiques, puis présente deux travaux d'analyse d'un tel ensemble d'annotations.

1.2 Annotation de bases de données biologiques avec le Bioportal et le Resource Index

Le Bioportal [Noy *et al.*, 2009] est un ensemble de services dont la vocation est de faciliter l'utilisation des ontologies biomédicales. Il a été développé dans le cadre d'un projet étasunien appelé le NCBO, pour National Center for Biomedical Ontologies et financé par le NIH (National Institute of Health). Ces services sont disponibles en ligne à `https://bioportal.bioontology.org/`, et parmi ses services, le Bioportal héberge une grande collection d'ontologies biomédicales (815 ontologies le 9 octobre 2019) pour faciliter leur partage et leur évolution. Le Bioportal permet également de naviguer dans la hiérarchie de ces ontologies, de les télécharger, d'explorer les correspondances (ou *mappings*) définies entre ontologies et de rechercher des concepts de façon unifiée sur l'ensemble de sa collection d'ontologies.

Le service du Bioportal qui nous intéresse ici plus particulièrement est le Resource Index [Jonquet *et al.*, 2010, Jonquet *et al.*, 2011]. Il s'agit d'une base d'annotations de la forme (*portion de texte*, *concept ontologique*) obtenues par une annotation automatique relativement simple du contenu semi-structuré d'un ensemble de bases de données biologiques. L'intérêt principal de ces annotations est qu'elles sont faites à la fois avec plusieurs ontologies (toutes celles du Bioportal, c'est-à-dire 245 en 2011, 815 en 2019) et sur plusieurs bases de données biologiques (22 en 2011, 48 depuis 2015). Plus précisément, chaque annotation est décrite : pour la portion de texte annotée par *(1)* la base de données où l'annotation est faite, *(2)* l'identifiant de l'objet annoté, *(3)* l'attribut annoté puis *(4)* la position des caractères annotés dans le texte associé à l'attribut ; et pour le concept ontologique par *(a)* l'ontologie source et *(b)* l'identifiant local du concept.

L'annotation automatique des portions de textes des bases de données est réalisée avec un outil appelé l'Annotator [Jonquet *et al.*, 2009] qui lui même utilise l'outil Mgrep [Dai *et al.*, 2008] qui compare le texte à un dictionnaire d'entités nommées constitué à partir des ontologies du Bioportal. Le processus général de peuplement du Resource Index incluant notamment la génération d'annotations avec l'Annotator est décrit Figure 1.1. Ce processus peut considérer deux types d'annotations : des annotations préexistantes et de nouvelles annotations qu'il génère lui même à partir de portions de texte (ces annotations sont respectivement appelées *reported vs. mgrep* dans la 3^{me}étape du processus représenté Figure 1.1), il peut générer des annotations dites indirectes à l'aide des parents dans l'ontologie du concept d'origine ou à l'aide des mappings définis dans le Bioportal.

En 2010, nous avons utilisé l'ensemble des ontologies disponibles sur le Biorpotal et l'Annotator pour générer des annotations sur 20 bases de données biologiques telles que DrugBank, UniProt ou GEO (voir [Jonquet *et al.*, 2009] pour la liste exhaustive). L'ensemble des annotations résultantes constitue le Resource Index, une source originale qui a deux intérêts : permettre de retrouver toutes les annotations faites sur une entité décrite dans une base de données ; et inversement et de façon plus intéressante, permettre de retrouver parmi plusieurs bases toutes les entités annotées par un même concept. La dernière version peuplée en 2015 du Resource Index est interrogeable en ligne à `https://bioportal.bioontology.org/resource_index`). En 2010, il contenait environ 3 milliards d'annotations directes (*i.e.*, sans considérer les annotations calculées avec la fermeture transitive ou les alignements d'ontologies) sur 3,2 millions d'entrées. Depuis sa dernière mise à jour en 2015, il contient 95 milliards annotations sur 40 millions d'entrées.

Au delà des chiffre et des capacités d'interrogation par objet annoté ou par concept ontologique, le Resource Index offre des possibilités inédites en termes d'analyse de données car il met à disposition de façon concrète un vaste ensemble de correspondances entre données et connais-

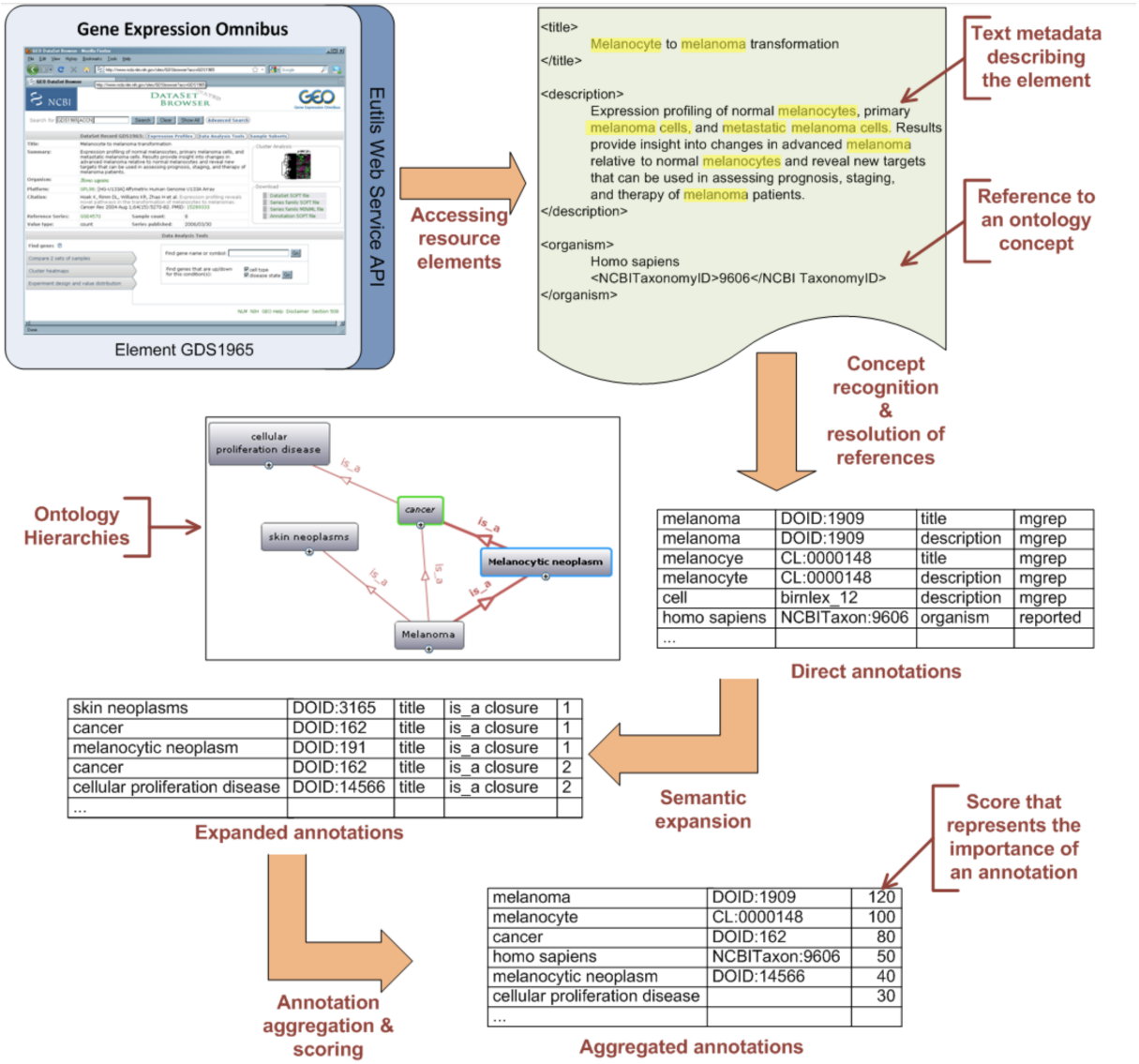


FIGURE 1.1 – Processus de peuplement du Resource Index illustré par l'exemple de l'annotation automatique de la bases de données GEO (Gene Expression Omnibus). Figure issue de [Jonquet *et al.*, 2010].

sances et car celles-ci permettent de connecter des sources de données initialement indépendantes. Les deux sections suivantes présentent un travail original d'analyse de données rendu possible par le Resource Index (Section 1.3), puis une méthode qui utilise une extension de l'AFC, *i.e.*, les structures de patrons, pour tirer parti dans l'analyse de données des ontologies associées (Section 1.4).

1.3 Analyse d'annotations pour la recherche d'associations entre financement de recherche et publication

Ce travail est une preuve de concept de la faisabilité d'utiliser des annotations générées automatiquement à partir d'une ontologie pré-existante, pour analyser conjointement des sources de données initialement disjointes. Pour cela, nous nous sommes intéressés à combler un manque : l'absence d'analyse qui relie les données concernant le financement de la recherche et les données sur les publications scientifiques [Liu *et al.*, 2012]. Il existe des sources pour l'un et pour l'autre, mais leur utilisation conjointe n'est pas évidente car elles ne partagent pas une terminologie commune pour indexer les bourses de financement d'une part et les publications de l'autre. Par exemple les descriptions de bourses du NIH disponibles dans la base de données RePORT sont indexées par type de recherche, conditions et catégorie de maladies alors que les articles biomédicaux de la base PubMed sont indexés avec les termes MeSH.

Nous avons donc proposé dans ce travail de nous intéresser aux thèmes de recherches que constituent les maladies et de considérer pour chaque maladie l'association entre financement et publication. Pour cela nous avons utilisé le processus d'annotation associé au Resource Index pour annoter avec une ontologie unique trois sources de données considérées dans l'analyse : deux concernant les financements de recherche, *Research Crossroads (RC)* et *the Scholary Database(SRC)* et une concernant les publications, PubMed.

1.3.1 Données et ontologie

Research Crossroads est une base de données propriétaire qui intègre les données sur les bourses attribuées par 33 organismes internationaux de financement de la recherche comme le NIH, la NSF, la Commission Européenne, etc. Celle-ci est complétée par la base ouvert *the Scholary Database* pour constituer au final une base unique qui contient l'organisme de financement, le montant, l'année et le récipiendaire de bourses de recherche entre 1997 et 2007, soit 81 858 bourses pour un total de 327 milliards de dollars.

Concernant les données de publications scientifiques, nous avons utilisé la base de données de publications biomédicales PubMed, que nous avons restreinte aux articles de journaux (excluant les éditoriaux, les lettres, les reviews par exemple), publiés entre 1997 et 2007 (pour être cohérent avec les données de financements) et aux auteurs affiliés avec une institution de recherche américaine (pour réduire le domaine de notre analyse et parce que les données de financement des organismes étasuniens étaient plus complètes). Le résultat de ces restrictions est un ensemble de 2.4 millions d'articles.

Pour structurer notre analyse, nous voulions considérer les données par thème de recherche et pour ces thèmes nous avons choisi les maladies et avons choisi comme maladies celles définies dans la Disease Ontology (DO) [Osborne *et al.*, 2009]. La DO est une ontologie construite manuellement par des experts à partir des concepts de l'UMLS qui décrivent des maladies. Pour cette raison, ses concepts ont tous une correspondance avec les concepts de l'UMLS qui présentent l'intérêt d'être associés à de nombreux labels alternatifs, incluant notamment les pluriels et des acronymes. Ces labels sont inclus dans la constitution du dictionnaire utilisé par l'Annotator et ainsi considérés pour l'annotation.

En plus de ces sources, nous avons voulu avoir une estimation de l'importance des différentes maladies et avons pour cela récupéré les chiffres des taux de mortalité des maladies à partir des rapports émis en 2005 par l'OMS (organisation mondiale pour la santé) pour 17 maladies [Mathers *et al.*, 2009] et celui émis en 2007 par le CDC (Center of Disease Control) étasunien [Xu *et al.*, 2007] pour 12 maladies. Nous avons mis en correspondance manuellement les causes

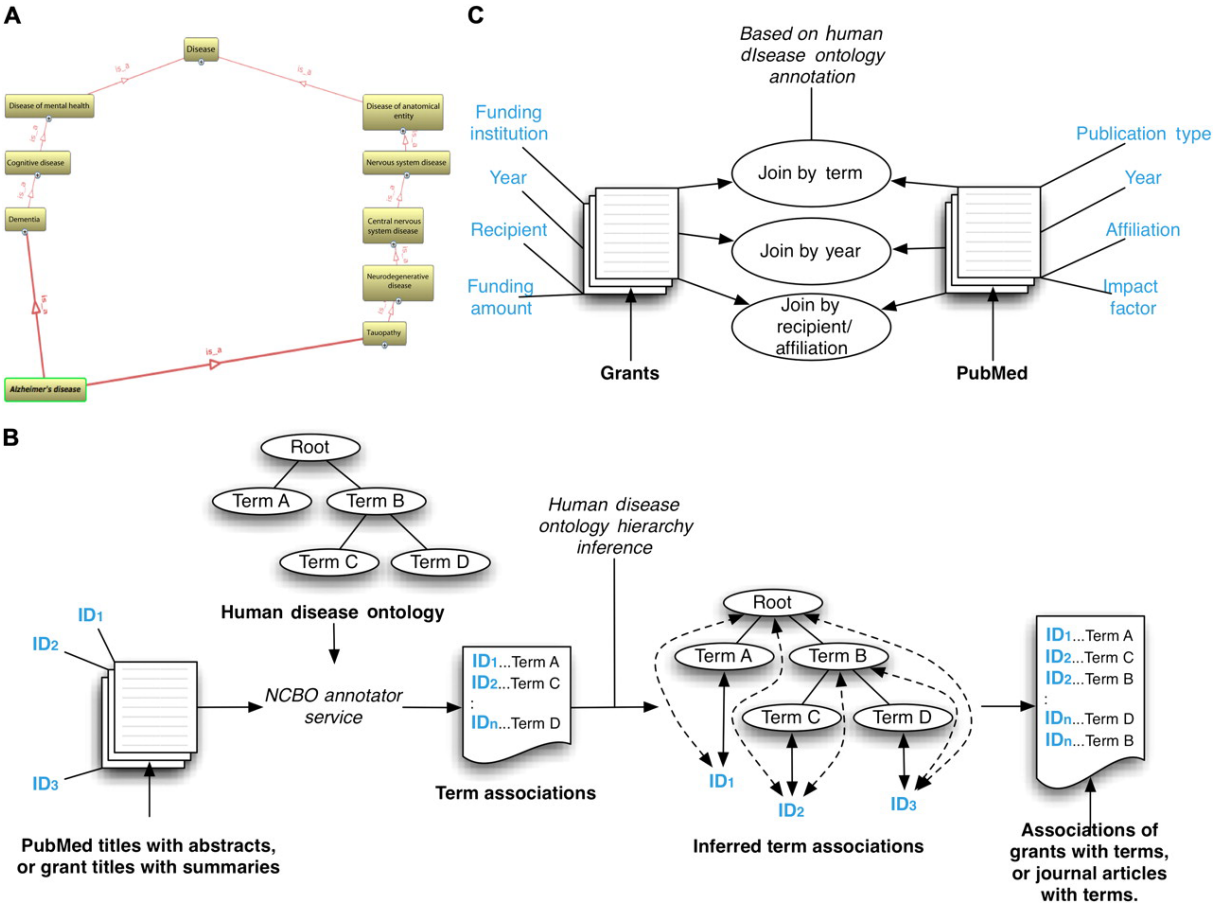


FIGURE 1.2 – Génération et utilisation des annotations du Resource Index pour l'analyse de l'activité de recherche biomédicale par maladie, suivant deux dimensions : les financements et les publications par maladie. La partie A schématise la Disease Ontology ; la partie B représente le processus d'annotation ; la partie C présente les attributs utilisés pour joindre ou agréger les données.

de mortalités mentionnés dans ces rapports avec les termes DO. Ces correspondances sont en Annexe A.1.

1.3.2 Processus d'annotation

Nous avons annotés les résumés descriptifs des bourses et les résumés des publications avec les termes définis dans la DO. La Figure 1.2 (partie B) présente notre instanciation du processus de peuplement du Resource Index pour l'annotation de RC, SCD et PubMed.

En sortie de l'annotation nous produisons une table des financement par maladie (*i.e.*, le thème de recherche), organisme de financement, institution de recherche lauréate et année. L'annotation des publications produit une table similaire où les articles sont associés à des maladies (*i.e.*, le thème de recherche), une institution de recherche auteure de l'article et à l'année. L'utilisation de la DO, ainsi que la normalisation des noms des principales institutions de recherches étasuniennes fait que nous pouvons joindre les deux tables sur trois attributs : le thème de recherche, l'institution lauréate/auteure et l'année offrant la possibilité d'analyser conjointement financements et publications (Figure 1.2, partie C).

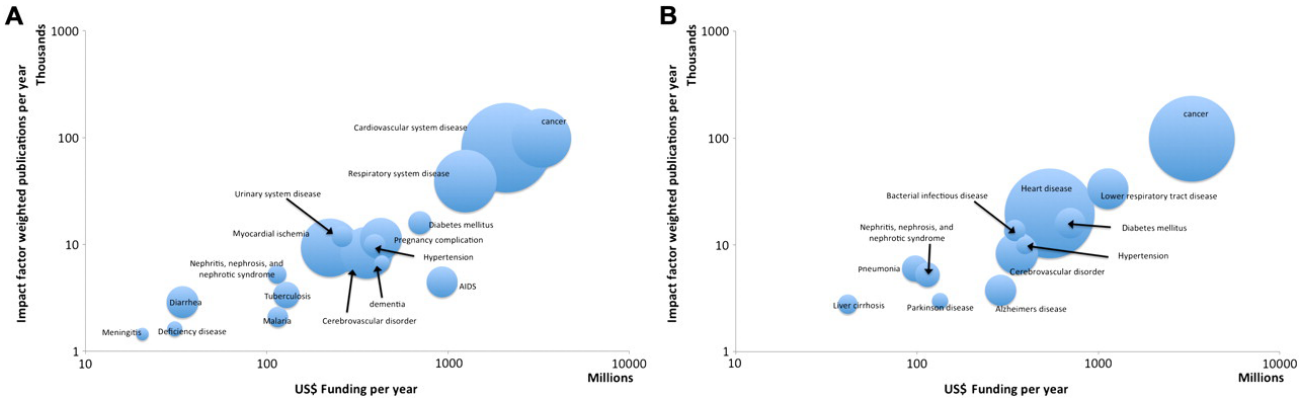


FIGURE 1.3 – Représentation à deux niveaux d'agrégation différents (A et B) des financements et impacts de publications par maladie (ou classes de maladies).

1.3.3 Analyses

Nous avons produit plusieurs analyses dans [Liu *et al.*, 2012] que nous ne détaillerons pas ici, notamment des clusterings des thèmes financés par les mêmes organismes ou des organismes finançant les mêmes thèmes, et des évolutions des financements dans le temps pour ces clusters. Nous proposons de discuter l'analyse que nous trouvons la plus originale et valorisante pour l'utilisation des annotations : la comparaison par maladie des dotations, publications et taux de mortalité présenté par la Figure 1.3.

L'axe des abscisses de la Figure représente le montant moyen (en dollars) alloué annuellement aux thèmes de recherche entre 1997 et 2007. Ce nombre reflète mieux selon nous l'activité de recherche que le nombre absolu de bourses sur cette même période. L'axe des ordonnées représente le nombre de publications moyen par thème sur la même période. Ce nombre de publication est pondéré par le facteur d'impact du journal où l'article a été publié. Cela revient à additionner les facteurs d'impact de chaque article. Cette pondération donne plus d'importance aux articles publiés dans des journaux à facteur d'impact élevé.

1.3.4 Discussion et conclusion

L'utilisation d'ontologies pour la réalisation de cette analyse est intéressante car la richesse de la DO permet de créer des annotations nombreuses dans les différentes sources considérées, même si leur champ lexical diffère, et l'organisation des concepts en hiérarchie permet de généraliser et grouper les maladies de façon intéressante car cela permet de comparer les taux de mortalité comptabilisés aux États-Unis par le CDC et ceux compatbilisés par l'OMS au niveau mondial. En résumé l'on peut dire que la richesse des termes associés aux concepts de l'ontologie permet d'annoter des sources qui utilisent des vocabulaires différents et que la hiérarchie de ces concepts permet de comparer les trois sources en trouvant quand c'est nécessaire un niveau de granularité supérieur mais consensus. En pratique la hiérarchie de l'ontologie est utilisée pendant la phase d'annotation durant laquelle les documents (descriptions de bourses et résumés de publications) sont annotés avec le concept trouvé dans le document et tous ses parents dans la hiérarchie de l'ontologie. Ceci a l'inconvénient de générer beaucoup d'annotations, mais l'avantage d'éviter le calcul de la fermeture transitive des annotations au moment de leur comparaison.

Subjectivement, nous observons sur la Figure 1.3 une certaine linéarité dans le lien entre financement des thèmes et nombre de publications. Il ressort de cette linéarité générale que

quelques thèmes sont décalés, illustrant une relation différente entre investissement et publication comme pour le SIDA (*AIDS* en anglais sur la Figure 1.3) et la malaria sur le volet A de la Figure ou les maladies de Parkinson et Alzheimer dans le volet B. Pour ces quatre exemples, nous observons un nombre de publications plus faible que les thèmes associés à des niveaux d'investissement similaires. Nous n'observons jamais l'inverse, c'est-à-dire un thème de recherche qui nécessiterait moins d'investissement pour des résultats de même ou de plus grande ampleur.

Une autre observation intéressante est la distinction entre les résultats des deux volets de la Figure 1.3 selon des niveaux d'agrégation différents. Ce niveau d'agrégation est dicté par les maladies pour lesquelles nous avons pu récolter les taux de mortalité. Ceux-ci regroupent sur chaque volet des ensembles de concepts DO différents et donc des montants de financement et des niveaux de publication différents. En d'autres termes, agréger deux sphères qui peuvent être en milieu de graphique, peut créer une sphère unique positionnée plus haut et plus à droite. C'est par exemple ce que l'on observe pour les sphères *Cardiovascular disorder*, *Hypertension*, *Heart disease* dans le panel B dont l'agrégation dans le panel A compose la sphère *cardiovascular system disease*. Notons que, hormis ce niveau d'agrégation différent, la différence entre les panels A et B de la Figure ne concerne que la taille des sphères, *i.e.*, le taux de mortalité fourni soit par l'OMS, soit par le CDC. La différence d'agrégation a un rôle particulier et impacte une interprétation naïve si l'on établit par exemple le classement des deux thèmes les plus financés. Alors la liste diffère selon le volet que l'on regarde : 1. *Cancer*, 2. *Cardiovascular system disease* pour l'analyse A avec les agrégations de l'OMS, et 1. *Cancer*, 2. *Lower respiratory tract disease* pour l'analyse B avec les agrégations du CDC. Cela illustre bien l'importance du choix du niveau de représentation des résultats et son impact potentiel sur la visualisation et l'interprétation.

D'un point de vue plus général ce travail d'analyse illustre l'intérêt des annotations et des liens qu'elles procurent vers une ontologie. Ici l'ontologie sert à deux choses : elle fournit un référentiel commun qui permet d'analyser conjointement des sources même si le vocabulaire varie entre sources, ou au cours du temps (*i.e.*, c'est par exemple ce que l'on observe avec le vocabulaire utilisé dans les demandes de financement qui varient avec le temps) ; l'ontologie fournit également une hiérarchie de concepts qui permet de comparer des données décrites à des niveaux de granularité différents. C'est le cas avec les maladies utilisées par l'OMS et le CDC, plus générales que celles que l'on trouve dans les demandes de financement et les publications.

Cette analyse fait un usage relativement simple des annotations et connaissances associées. Une seule ontologie (DO) est utilisée et les connaissances produites ne concernent pas de combinaison de concepts, alors que les documents sont associés à plusieurs annotations. On pourrait imaginer poser la question du niveau de financement et de publication du thème "cancer chez les malades de Parkinson", mais nous ne sommes pas allés jusque là dans ce travail. Aussi, un seul niveau d'agrégation des résultats est utilisé, celui imposé par les concepts mis en correspondance avec les sources OMS et CDC. Nous avons observé que ces niveaux d'agrégation impactent l'interprétation. Pour mieux tirer parti des connaissances associées aux données par les annotations, il nous paraît intéressant de proposer des méthodes qui permettent d'analyser des données décrites par plusieurs annotations (faites avec des ontologies différentes ou une même ontologie) et qui considèrent tous les niveaux hiérarchiques des ontologies associées. C'est justement l'objectif de la contribution décrite dans la section suivante.

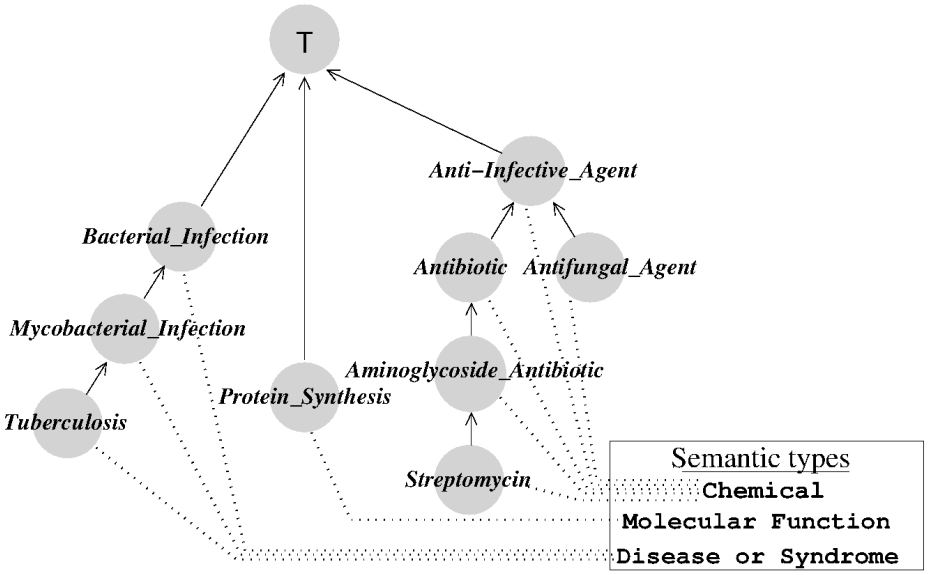


FIGURE 1.4 – Détail de l'ontologie *NCI Thesaurus* et des types sémantiques de l'UMLS associés aux concepts. Les lignes pointillées associent à chaque concept de l'ontologie son type sémantique comme définie par le *Semantic Network* de l'UMLS.

1.4 Les structures de patrons pour considérer les connaissances associées aux annotations

Ce travail était motivé par la volonté de trouver des associations entre des concepts de catégories sémantiques différentes et cela à partir d'annotations faites sur un ensemble de documents possiblement disparates. Les catégories sémantiques peuvent être définies indifféremment au sein d'une même ontologie ou d'ontologies différentes. Pour illustrer ce travail nous adoptons l'exemple de la recherche d'associations entre des concepts de *types sémantiques* différents au sein d'une même ontologie. Les types sémantiques sont une partie importante de l'UMLS [Bodenreider, 2004] où ils participent au *Semantic Network* . Il s'agit de catégories générales comme *Organism, Tissue, Chemical, Disease or Syndrome*, etc. qui sont utilisées comme une classification de haut niveau pour les ontologies de l'UMLS [McCray, 2003]. Ainsi, chaque concept des ontologies de l'UMLS est associé à un ou plusieurs types sémantiques. La liste complète des types sémantiques est visible en ligne à `https://www.nlm.nih.gov/research/umls/META3_current_semantic_types.html` (visité le 15 octobre 2019). La Figure 1.4 illustre cette association entre des concepts de l'ontologie *NCI Thesaurus* et leurs types sémantiques.

Nous souhaitons donc trouver des associations entre concepts de types sémantiques différents, à partir du fait que ces concepts sont utilisés pour annoter de mêmes documents biomédicaux. Par exemple, si un expert cherche des pistes sur les mécanismes moléculaires utilisés par un médicament, il peut chercher une association entre les concepts *Antibiotic* et *Inflammation* et alors explorer si d'autres concepts du type *Molecular function* sont associés aux deux premiers.

1.4.1 L'analyse formelle de concepts

L'analyse formelle de concept (AFC ou *FCA* en anglais pour *Formal Concept Analysis*) est un ensemble d'outils mathématiques pour l'analyse de données et la découverte de connaissances. L'AFC est notamment fondée sur la notion de treillis de concepts (ou treillis de Galois), une

structure organisant suivant deux ordres partiels un ensemble de concepts associant à des objets une description commune. Ce treillis peut être vu comme la construction à partir de données, d'ensembles d'objets ayant des propriétés communes et l'organisation de ces ensembles selon deux hiérarchies. Ce type de structure est particulièrement intéressant en ingénierie des connaissances. En effet, elle propose une construction apprise à partir des données qui peut être comparée à une hiérarchie de concepts classiquement présente dans les ontologies de domaines [Alam, 2015]. Pour cette raison principale nous avons souhaité étudier comment l'AFC pourrait servir à l'analyse d'annotations.

Pour une introduction simple à l'AFC, nous dirigeons le lecteur vers la section 1.3.1 de ma thèse de doctorat [Coulet, 2008]. Pour plus de détails nous dirigeons le lecteur vers [Ganter and Wille, 1999].

Nous rappelons ici seulement les notations et définitions utiles à la présentation de notre travail.

Un *contexte formel* (G, M, I) est défini par un ensemble G d'objets, un ensemble M d'attributs et une relation binaire $I \subseteq G \times M$. $(g, m) \in I$ signifie que "l'objet g est relié à l'attribut m par la relation I". Deux opérateurs de dérivation peuvent être définis entre ensembles d'objets et ensembles d'attributs comme suit $\forall A \subseteq G, B \subseteq M$:

$$A' = \{m \in M : \forall g \in A, (g, m) \in I\}$$

$$B' = \{g \in G : \forall m \in B, (g, m) \in I\}$$

Les deux opérateurs $(\cdot)'$ définissent une connexion de Galois entre l'ensemble des parties des objets $\wp(G)$ et l'ensemble des parties des attributs $\wp(M)$. Une paire $(A, B), A \subseteq G, B \subseteq M$, est un *concept formel* ssi $A' = B$ et $B' = A$. A est appelé l'*extension* et B l'*intention* du concept. L'ensemble de tous les concepts formels, ordonnés suivant l'inclusion des extensions (ou de façon duale par inclusion des intentions), *i.e.*, $(A_1, B_1) \leq (A_2, B_2)$ ssi $A_1 \subseteq A_2$ (ou de façon duale $B_2 \subseteq B_1$), constitue un treillis complet [Barbut and Monjardet, 1970], appelé *treillis de concepts*.

Malgré la bonne adéquation de l'AFC pour des travaux d'ingénierie des connaissances, son utilisation nécessite des adaptations pour traiter des objets complexes comme les annotations. En particulier, si l'on considère l'annotation de documents avec une ou plusieurs ontologies, les annotations peuvent être représentées comme une paire $(document, ensembledeconcepts)$ difficilement représentable sous la forme d'une relation binaire alors que les outils classiques d'AFC manipulent uniquement des relations binaires, entre un ensemble d'objets et un ensemble d'attributs. De plus nous souhaitons considérer les relations entre concepts définis dans l'ontologie pour analyser et comparer les annotations. Nous considérons pour ces raisons les méthodes qui permettent d'utiliser l'AFC avec des données complexes.

Une approche classique pour cela est le *scaling* qui consiste à transformer des données non binaires en données binaires [Ganter and Wille, 1999]. Mais le scaling s'accompagne de plusieurs limites comme la transformation arbitraire de données, la perte d'information et la génération d'un nombre d'attributs très important qui rend difficile l'interprétation et la visualisation des résultats. Kaytoue *et al.* donnent des exemples du processus de scaling et discutent ses limites dans [Kaytoue *et al.*, 2011b].

1.4.2 Les structures de patrons

Une alternative au scaling est les *structures de patrons* qui permettent d'analyser directement (*i.e.*, sans transformation) des données complexes dans la mesure où les descriptions associées aux objets peuvent être organisées dans un demi-treillis, ou en d'autres termes dans la mesure où un ordre partiel existe entre les descriptions des objets [Ganter and Kuznetsov, 2001]. Ce

que nous appelons description est l'ensemble des attributs associés à un objet ou un ensemble d'objets. Il s'agit ainsi en AFC classique d'un ensemble d'attributs binaires, mais le type de ces attributs peut être plus complexe avec les structures de patrons : des intervalles numériques [Kaytoue *et al.*, 2011a], des ensembles d'attributs [Ganter and Kuznetsov, 2001] ou des graphes [Kuznetsov and Samokhin, 2005]. Dans le cadre des structures de patrons, un objet a une description dans un demi-treillis où l'opérateur de similarité $\sqcap$ est défini. Cet opérateur permet de caractériser la similarité de deux descriptions, *i.e.*, ce que deux descriptions ont en commun. Le demi-treillis matérialise l'ordre partiel entre les descriptions. L'*opérateur de similarité* $\sqcap$ est associé à une opération de subsomption ($\sqsubseteq$) qu'il faut distinguer de l'opération de subsomption en LD, définie en extension car elle porte sur les instances des concepts. Il s'agit ici d'une opération "inverse" puisqu'elle est définie en intention, *i.e.*, sur les descriptions. Nous avons proposé dans ce travail d'adapter les structures de patrons à l'analyse d'annotations et cela notamment en définissant un opérateur de similarité qui prend en considération l'ordre des concepts défini dans une (ou plusieurs) ontologie(s).

Formellement, dénotons G un ensemble d'objets et $(\mathcal{D}, \sqcap)$ un demi-treillis des descriptions des objets associé à l'opération de similarité $\sqcap$ et $\delta : G \longrightarrow \mathcal{D}$ est une fonction associant chaque objet à sa description. $(G, (\mathcal{D}, \sqcap), \delta)$ est une structure de patrons. Les éléments de $\mathcal{D}$ sont les descriptions ou *patterns* et sont ordonnés par une relation de subsomption $\sqsubseteq$ telle que $\forall c, d \in \mathcal{D}$, $c \sqsubseteq d \Longleftrightarrow c \sqcap d = c$. Une structure de patrons $(G, (\mathcal{D}, \sqcap), \delta)$ fait apparaître deux opérateurs de dérivation notés tous les deux par $(\cdot)^{\square}$:

$$A^{\square} = \prod_{g \in A} \delta(g) \qquad pour\ A \subseteq G$$
$$d^{\square} = \{g \in G | d \sqsubseteq \delta(g)\} \qquad pour\ d \in (\mathcal{D}, \sqcap).$$

Ces opérateurs constituent une connexion de Galois entre l'ensemble des parties des objets $\wp(G)$ et $(\mathcal{D}, \sqcap)$. Les concepts de patrons de $(G, (\mathcal{D}, \sqcap), \delta)$ sont des paires de la forme (A, d), $A \subseteq G$, $d \in (\mathcal{D}, \sqcap)$, telles que $A^{\square} = d$ et $A = d^{\square}$. Pour un concept de patrons (A, d), d est l'intention du patron et la description commune à tous les objets de A, l'extension du patron. Suivant l'ordre partiel $(A_1, d_1) \leq (A_2, d_2) \Leftrightarrow A_1 \subseteq A_2 \ \ (\Leftrightarrow d_2 \sqsubseteq d_1)$, l'ensemble de tous les concepts de patrons forme un treillis complet appelé treillis de concepts de patrons (ou treillis de patrons). L'opérateur $(\cdot)^{\square\square}$ est une fermeture et les intentions des patrons sont fermées.

1.4.3 Adaptation des structures de patrons aux annotations

Tout d'abord, considérons un ensemble de catégories sémantiques ou dimensions

$$\mathcal{ST} = \{\mathtt{ST}_1, \mathtt{ST}_2, ..., \mathtt{ST}_k\}$$

où chaque $\mathtt{ST}_i$ est une catégorie sémantique. Considérons également un ensemble de documents G annotés avec des concepts de l'ontologie de référence $\mathcal{O}$ qui appartiennent à de types sémantiques de $\mathcal{ST}$. Alors si g est un document, l'annotation de g avec $\mathcal{O}$ et les types $\mathcal{ST}$ est notée

$$(g, \langle \mathtt{ST}_1, \mathtt{ST}_2, ..., \mathtt{ST}_k \rangle)$$

où $\mathtt{ST}_i$ est l'ensemble des concepts annotant g pour la dimension i de $\mathcal{ST}$.

Par exemple la Figure 1.5 montre l'annotation du document DB01082 (la page DrugBank de la streptomycine) avec trois concepts de l'ontologie *NCI Thesaurus*. Si l'on considère les dimensions de $\mathcal{ST}$ ={"$Disease or Syndrome''$, "$Bacterium''$, "$Molecular Function''$, "$Chemical''$}. Alors l'annotation de DB01082 peut être notée :

$$(\text{DB01082}, \langle \{Tuberculosis\},\ \{\},\ \{Protein_Synthesis\},\ \{Streptomycin\} \rangle)$$

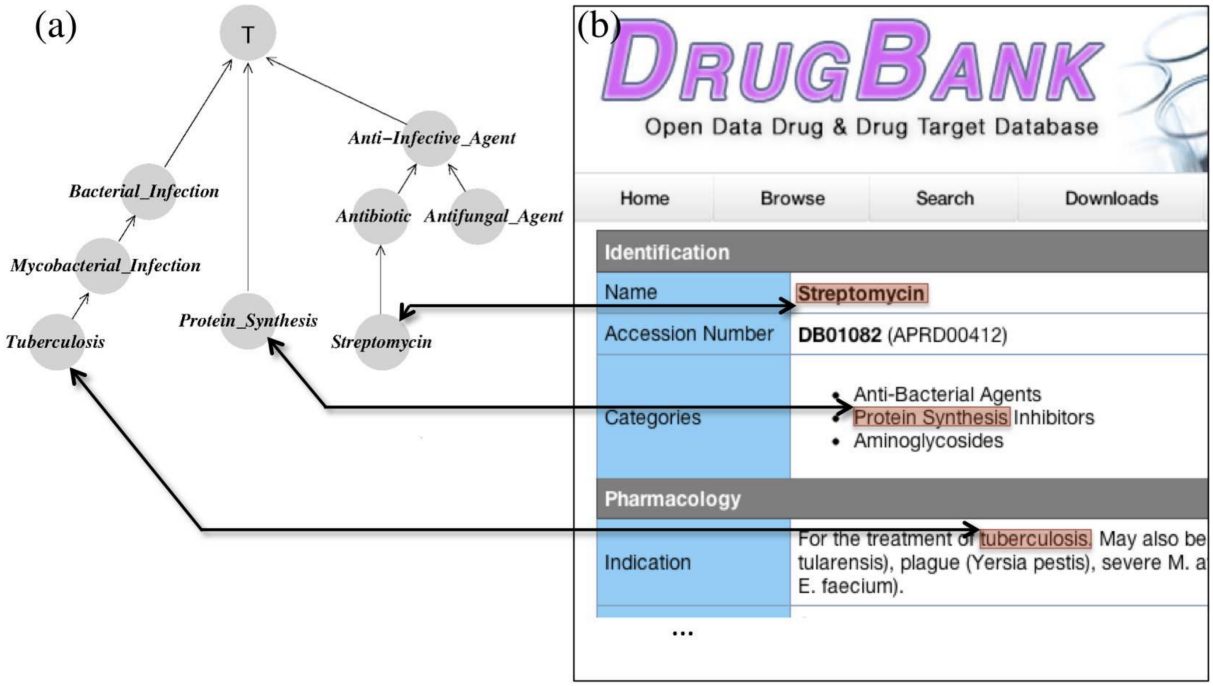


FIGURE 1.5 – Exemple d'annotation de la base de données DrugBank avec l'ontologie NCI Thesaurus. (a) La partie gauche représente un morceau de l'ontologie *NCI Thesaurus* ; (b) la partie droite un extrait du document DB01082 de DrugBank à propos de la streptomycine. Les flèches représentent les annotations.

Nous définissons la structure de patrons $(G, (\mathcal{D}, \sqcap), \delta)$ dédiée à l'analyse d'annotations de documents, composée de :

- $G = \{g_1, g_2, ..., g_n\}$ un ensemble de documents annotés ;
- $\mathcal{O}$ une ontologie de référence et $C(\mathcal{O})$ l'ensemble des concepts de $\mathcal{O}$;
- $\mathcal{ST} = \{\mathtt{ST}_1, \mathtt{ST}_2, ..., \mathtt{ST}_k\}$ l'ensemble des dimensions ou catégories sémantiques qui définissent la taille de $\mathcal{ST}$ et les dimensions du vecteur d'annotations ;
- $\mathcal{D} = \mathcal{P}(\mathtt{ST}_1) \times \mathcal{P}(\mathtt{ST}_2) \times ... \times \mathcal{P}(\mathtt{ST}_k))$ où $\mathcal{P}(\mathtt{ST}_i))$ est l'ensemble des parties de l'ensemble des concepts de la dimension $\mathtt{ST}_i$. $\mathcal{D}$ est un treillis complet (et donc un demi-treillis). Les éléments de $\mathcal{D}$ seront nommés *patrons ontologiques* ;
- $(\mathcal{D}, \sqcap)$ est l'ensemble des descriptions organisé dans un semi-treillis. L'opérateur de similarité $\sqcap$ est défini plus loin.
- $\delta : G \to \mathcal{D}$ est la fonction qui associe un document $g_i \in G$ à ses annotations, *i.e.*, à sa description dans $\mathcal{D}$ ou plus précisément à un vecteur de $\mathcal{D}$,

$$\delta(g_i) = (g_i, \langle \mathtt{ST}_1(g_i), \mathtt{ST}_2(g_i), ..., \mathtt{ST}_k(g_i) \rangle)$$

où $\mathtt{ST}_j(g_i)$ est l'ensemble des concepts de la dimension $\mathtt{ST}_j$ annotant g_i.

La Figure 1.6 illustre les différents composants de la structure de patron proposée.

Pour terminer la définition de notre structure de patron, il reste à définir l'*opération de similarité* $\sqcap$ entre deux descriptions $\delta(g_1)$ et $\delta(g_2)$:

$$\delta(g_1) = (g_1, \langle \mathtt{ST}_1(g_1), \mathtt{ST}_2(g_1), ..., \mathtt{ST}_k(g_1) \rangle)$$

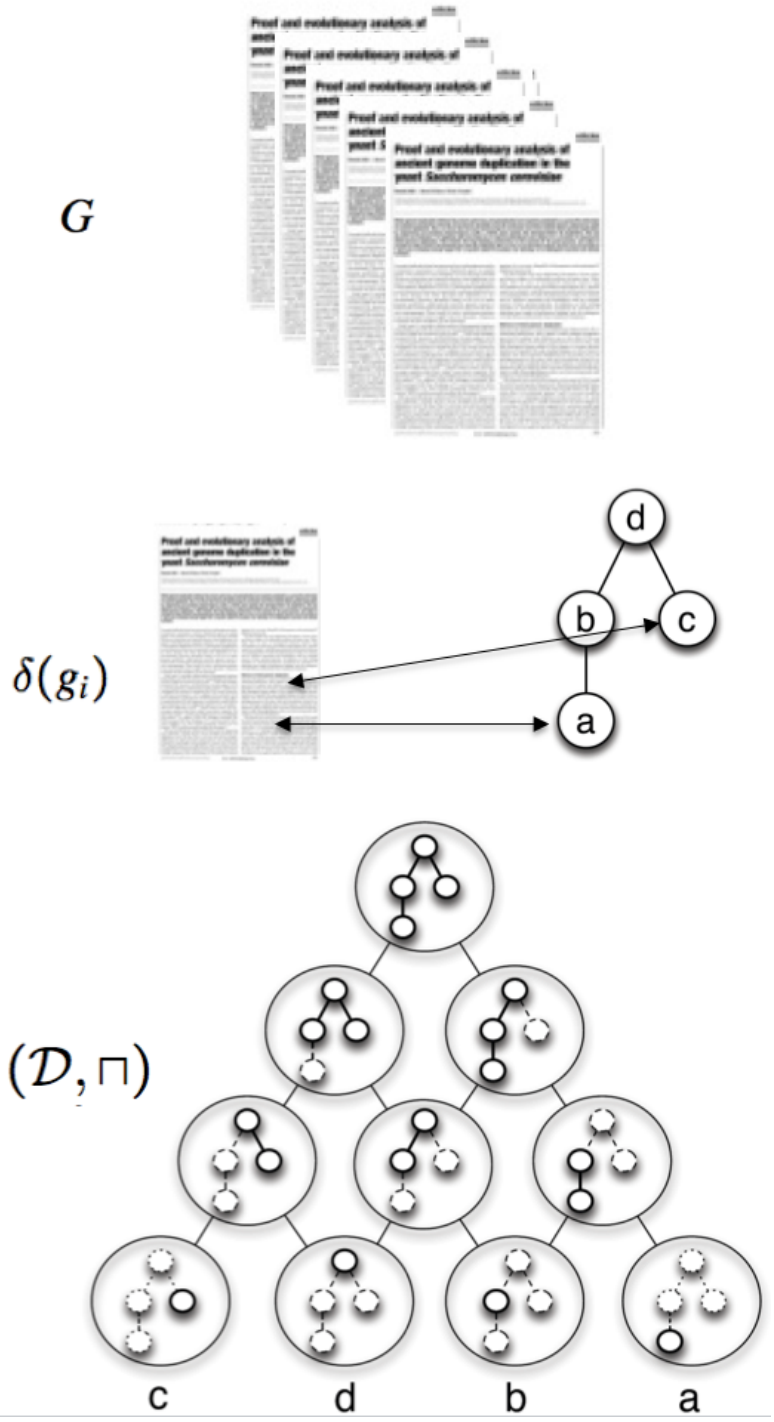


FIGURE 1.6 – Illustration figurative de la structure de patron $(G, (\mathcal{D}, \sqcap), \delta)$ proposée pour l'analyse d'annotations de documents. G est l'ensemble des documents annotés. $\delta(g_i)$ sont les annotations du document g_i. $(\mathcal{D}, \sqcap)$ est l'ensemble des descriptions organisé dans un semi-treillis

$$\delta(g_2) = (g_2, \langle \mathtt{ST}_1(g_2), \mathtt{ST}_2(g_2), ..., \mathtt{ST}_k(g_2) \rangle)$$

$$\delta(g_1) \sqcap \delta(g_2) = \langle \mathtt{ST}_1(g_1) \sqcap \mathtt{ST}_1(g_2), \mathtt{ST}_2(g_1) \sqcap \mathtt{ST}_2(g_2), ..., \mathtt{ST}_k(g_1) \sqcap \mathtt{ST}_k(g_2) \rangle$$

où $\mathtt{ST}_1(g_1) \sqcap \mathtt{ST}_1(g_2)$ est l'*enveloppe convexe* dans $\mathcal{O}$ des concepts de $\mathtt{ST}_1(g_1)$ et $\mathtt{ST}_2(g_2)$. La définition de l'enveloppe convexe est donnée dans ce qui suit. Cependant le semi-treillis figuratif représenté en bas de la Figure 1.6 donne l'intuition du fonctionnement de l'opérateur de similarité sur les descriptions. Si l'on considère deux annotations l'une faite avec le concept b l'autre avec le concept c de l'ontologie exemple donnée sur cette figure, leur description peut être trouvées dans le demi-treillis respectivement au dessus du b et du c. La description résultante de l'opération de similarité peut être trouvée si l'on parcourt le demi-treillis de bas en haut jusqu'à trouver le parent commun à b et c.

Pour définir l'enveloppe convexe, nous utilisons une première opération notée `lcs` qui retourne le subsumant commun le plus spécifique. Étant donnée une ontologie $\mathcal{O}$ et deux concepts c_1 et c_2, le subsumant commun le plus spécifique, noté $\mathtt{lcs}(c_1, c_2)$, est le concept de $\mathcal{O}$ le plus spécifique qui subsume à la fois c_1 et c_2 selon $\mathcal{O}$. Pour simplifier, $\mathcal{O}$ est ici une ontologie en $\mathcal{EL}$ sans cycle dans les définitions de concepts. Avec cette restriction le `lcs` de deux concepts existe toujours [Baader *et al.*, 1999]. Par exemple, dans la Figure 1.5 le `lcs` de *Streptomycin* et $Antifungal_Agent$ est $Anti-Infective_Agent$.

Plus généralement, le `lcs` peut être défini (récursivement) pour un ensemble de concepts de taille quelconque $C_n = \{c_1, c_2, ..., c_n\}$:

$$\forall n \in N,\ \mathtt{lcs}(C_n) = \mathtt{lcs}(\mathtt{lcs}(C_{n-1}), c_n)$$

Nous définissons l'*enveloppe convexe* $\mathtt{CVX}(c_1, c_2)$ des deux concepts c_1 and c_2 comme l'ensemble des concepts $\{x_1, x_2, ..., x_n\}$ qui vérifient :

— $x_i \leqslant \mathtt{lcs}(c_1, c_2)$, et

— soit $\begin{cases} x_i \geqslant c_1 \text{ et } x_i \wedge c_1 \equiv c_1 & \text{ou} \\ x_i \geqslant c_2 \text{ et } x_i \wedge c_2 \equiv c_2 \end{cases}$

— $x_i \not\equiv \top$

Comme pour le `lcs`, l'enveloppe convexe peut être généralisée (récursivement) à un ensemble de concepts $C_p = \{c_1, c_2, ..., c_p\}$:

$$\forall p \in N,\ \mathtt{CVX}(C_p) = \mathtt{CVX}(\mathtt{CVX}(C_{p-1}), c_p)$$

Nous avons appelé cette opération "enveloppe convexe" par analogie avec l'enveloppe convexe définie en géométrie euclidienne. Dans ce contexte, l'enveloppe convexe d'un ensemble de points est le sous ensemble convexe minimal de points qui contient l'ensemble. Dans notre cas, l'enveloppe convexe d'un ensemble de concepts est l'ensemble minimal de concepts qui inclut ces concepts et est borné par ces concepts.

L'opérateur de similarité sur les descriptions s'applique alors sur deux vecteurs de même dimension et retourne un vecteur où les composants sont les enveloppes convexes de l'union des deux ensembles de concepts initiaux. Formellement :

$$\delta(g_1) = (g_1, \langle \mathtt{ST}_1(g_1), \mathtt{ST}_2(g_1), ..., \mathtt{ST}_k(g_1) \rangle)$$

$$\delta(g_2) = (g_2, \langle \mathtt{ST}_1(g_2), \mathtt{ST}_2(g_2), ..., \mathtt{ST}_k(g_2) \rangle)$$

$$\delta(g_1) \sqcap \delta(g_2) = \langle \mathtt{ST}_1(g_1) \sqcap \mathtt{ST}_1(g_2), \mathtt{ST}_2(g_1) \sqcap \mathtt{ST}_2(g_2), ..., \mathtt{ST}_k(g_1) \sqcap \mathtt{ST}_k(g_2) \rangle$$

où

$$\mathtt{ST}_i(g_1) \sqcap \mathtt{ST}_i(g_2) = \mathtt{CVX}(\mathtt{ST}_i(g_1) \cup \mathtt{ST}_i(g_2)).$$

TABLE 1.3 – Un contexte où les objets sont des documents (ou pages) de DrugBank et les attributs des types sémantiques. Chaque document est annoté avec un ensemble de concepts du NCI Thesaurus (notre ontologie de référence) et associé à différents types sémantiques. Le document DB01082 de DrugBank (sur la quatrième ligne) est annoté avec trois concepts, notamment le concept *Tuberculosis* de type sémantique "Disease or Syndrome".

G $\mathcal{ST}$	**Disease or Syndrome**	**Bacterium**	**Molecular Function**	**Chemical**
Drug1	{Tuberculosis, Bacterial_Infection}	{}	{Protein_Synthesis}	{Antibiotic, Antifungal_Agent}
Drug2	{Bacterial_Infection}	{}	{Protein_Synthesis}	{}
Drug3	{Tuberculosis, Bacterial_Infection}	{}	{}	{Anti-Infective_Agent}
DB01082	{Tuberculosis}	{}	{Protein_Synthesis}	{Streptomycin}
Drug5	{Tuberculosis, Bacterial_Infection}	{}	{}	{Antibiotic, Antifungal_Agent}

Nous remarquons que la définition de cet opérateur de similarité peut être comparée à la définition du même opérateur sur les intervalles numériques, comme l'enveloppe convexe de deux intervalles (voir [Kaytoue *et al.*, 2011a] pour des exemples). Comme avec les intervalles, nous avons la propriété suivante :

$$\delta(g_1) \sqcap \delta(g_2) = \delta(g_1) \text{ iff } \delta(g_1) \sqsubseteq \delta(g_2)$$

De façon duale il est possible de définir une opération *inf* ou $\sqcup$ sur les descriptions, faisant de $(\mathcal{D}, \sqcap, \sqcup)$ un treillis complet. Cette opération n'est pas nécesssaire pour la définition des structures de patrons, mais dans notre cas, elle existe de par la propriété de l'espace des descriptions $\mathcal{D}$. L'opérateur inf sur deux descriptions $\delta(g_1)$ et $\delta(g_2)$ est :

$$\delta(g_1) \sqcup \delta(g_2) = \langle \mathtt{ST}_1(g_1) \sqcup \mathtt{ST}_1(g_2), \mathtt{ST}_2(g_1) \sqcup \mathtt{ST}_2(g_2), ..., \mathtt{ST}_k(g_1) \sqcup \mathtt{ST}_k(g_2) \rangle$$

où

$$\mathtt{ST}_i(g_1) \sqcup \mathtt{ST}_i(g_2) = \mathtt{CVX}(\mathtt{ST}_i(g_1)) \cap \mathtt{CVX}(\mathtt{ST}_i(g_2)).$$

En pratique, le résultat de l'opétation inf est l'ensemble des concepts communs entre les deux enveloppes convexes de $\mathtt{ST}_i(g_1)$ et $\mathtt{ST}_i(g_1)$.

1.4.4 Exemple

Le Tableau 1.3 donne une exemple de contexte adapté à notre structure de patrons. Notamment la quatrième ligne présente l'exemple courant des annotations du document DB01082 de DrugBank (à propos de la streptomycine) et introduit Figure 1.5. Chaque colonne correspond à une dimension, ici un type sémantique de l'UMLS. Les cases sont remplies avec les labels de l'ensemble de concepts du type sémantique (défini par la colonne) qui annote le document (défini par la ligne).

Par exemple, $\mathtt{CVX}(Streptomycin, Antifungal_Agent) = \{Anti\text{–}Infective_Agent, Antibiotic, Antifungal_\ Agent, Streptomycin\}$.

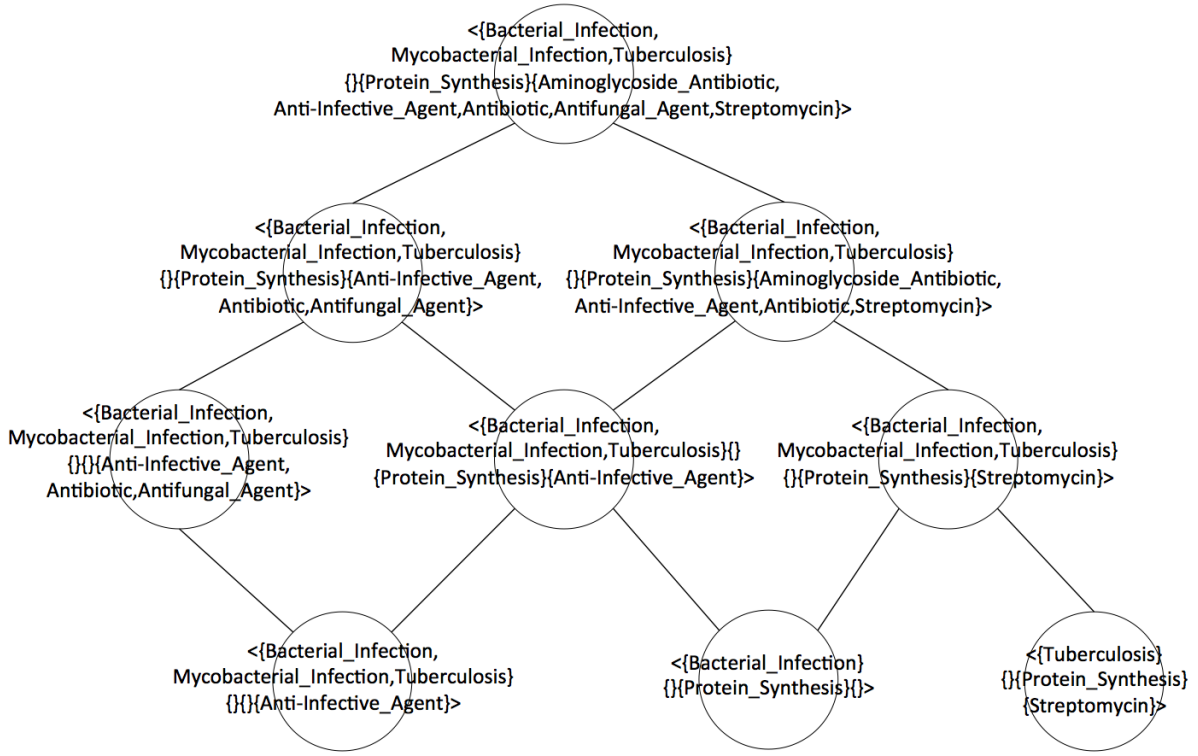


FIGURE 1.7 – Le demi-treillis des enveloppes convexes associées au contexte présenté Tableau 1.3 et au NCI Thesaurus

Le demi-treillis $(\mathcal{D}, \sqcap)$ de patrons (*i.e.*, des enveloppes convexes) défini par l'opération de similarité à partir du contexte précédent (et du NCI Thesaurus) est donnée en Figure 1.7.

Pour illustrer l'opérateur de similarité, voyons comment il agit sur les descriptions de "Drug1" et "DB01082" (δ(Drug1) et δ(DB01082)) qui sont fournies dans le Tableau 1.3 :

$$\begin{aligned}\delta(\text{Drug1}) \sqcap \delta(\text{DB01082}) = \\ \langle \{&Bacterial_Infection, Mycobacterial_Infection, Tuberculosis\}, \\ &\{\}, \\ &\{Protein_Synthesis\}, \\ &\{Anti- \\ Infective_Agent, &Antibiotic, Antifungal_Agent, Streptomycin\}\rangle.\end{aligned}$$

Pour un rendu plus visuel, la Figure 1.8 permet de visualiser la couverture des trois enveloppes convexes sur les trois dimensions non vides des description associées aux annotations des deux documents.

De façon duale, l'opérateur inf entre "Drug1" et "DB01082" donne :

$$\delta(\text{Drug1}) \sqcup \delta(\text{DB01082}) = \langle \{Tuberculosis\}, \{\}, \{Protein_Synthesis\}, \{\} \rangle.$$

Nous notons que l'intersection des enveloppes convexes peut être vide comme dans la seconde et quatrième dimension de l'exemple précédent. Cependant, il peut aussi être noté que même si $\delta(g_1)$ and $\delta(g_2)$ n'ont pas d'élément en commun, ils peuvent toujours avoir un inf, comme illustré dans l'exemple suivant. Ne considérons qu'une seule dimension et supposons que l'on

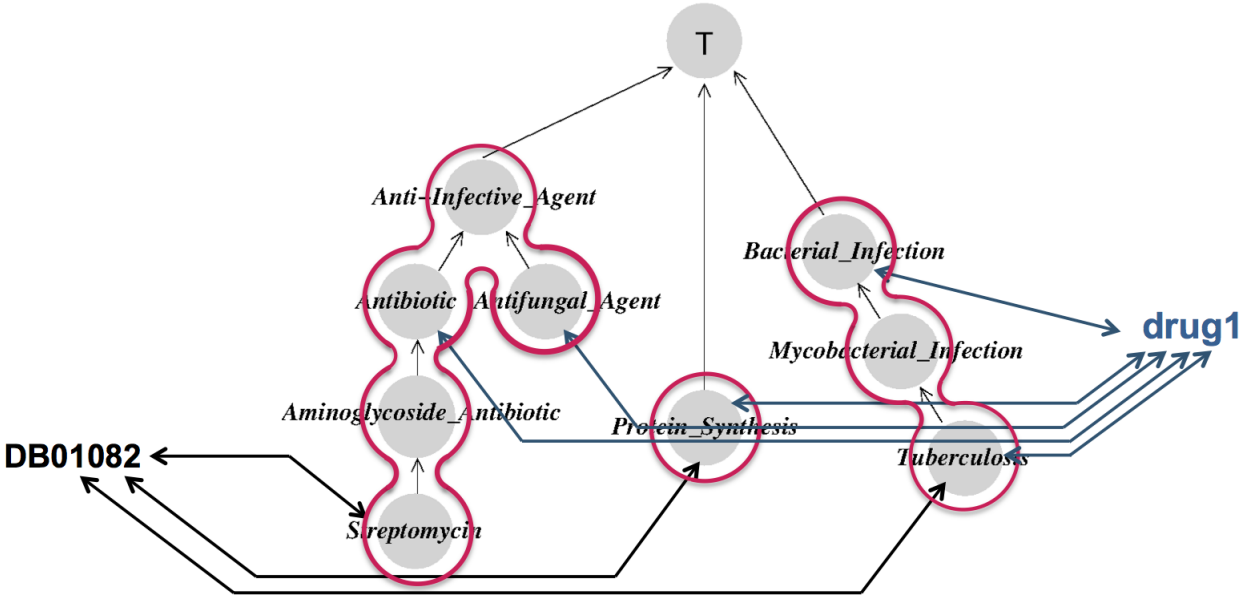


FIGURE 1.8 – Représentations de la couverture des enveloppes convexes des trois dimensions non-vides des annotations de l'ensemble des deux documents "Drug1" et "DB01082"
.

utilise l'ontologie représentée Figure 1.5 :

$$\delta(g_1) = \langle\{Bacterial_Infection, Tuberculosis\}\rangle$$

$$\delta(g_2) = \langle\{Mycobacterial_Infection\}\rangle.$$

Les résultats de l'opérateur de similarité et inf. sur ces deux descriptions sont :

$$\delta(g_1) \sqcap \delta(g_2) = \langle\{Bacterial_Infection, Mycobacterial_Infection, Tuberculosis\}\rangle$$

et

$$\delta(g_1) \sqcup \delta(g_2) = \langle\{Mycobacterial_Infection\}\rangle.$$

De plus nous remarquons que nous n'avons pas $\delta(g_1) \sqcap \delta(g_2) = \delta(g_1)$ car $\delta(g_1)$ n'est pas une enveloppe convexe et donc nous n'avons pas non plus $\delta(g_1) \sqsubseteq \delta(g_2)$.

Le treillis complet correspondant à la structure de patrons donnée Tableau 1.3 et à l'ontologie NCI Thesaurus est représenté Figure 1.9. Le concept du sommet a l'intention avec la plus grande description et par conséquent son intention inclut tous les documents. Si l'on traverse le treillis vers le bas, les concepts présentent des extensions plus spécialisées et des intentions plus générales.

1.4.5 Discussion et conclusion

La navigation dans ce treillis et l'interprétation de ses concepts illustre les résultats que peut fournir la méthode originale présentée. Les structures de patrons permettent une classification des documents sur la base de leurs annotations et cela en respect des connaissances représentées dans une ontologie. Cela démontre la capacité de l'AFC à considérer des données complexes et que l'opérateur de similarité sur lequel s'appuie les structures de patrons peut être défini en considération d'une ontologie. C'est en effet ce que font les opérateurs `lcs` et `CVX` qui nous ont permit de définir $\sqcap$.

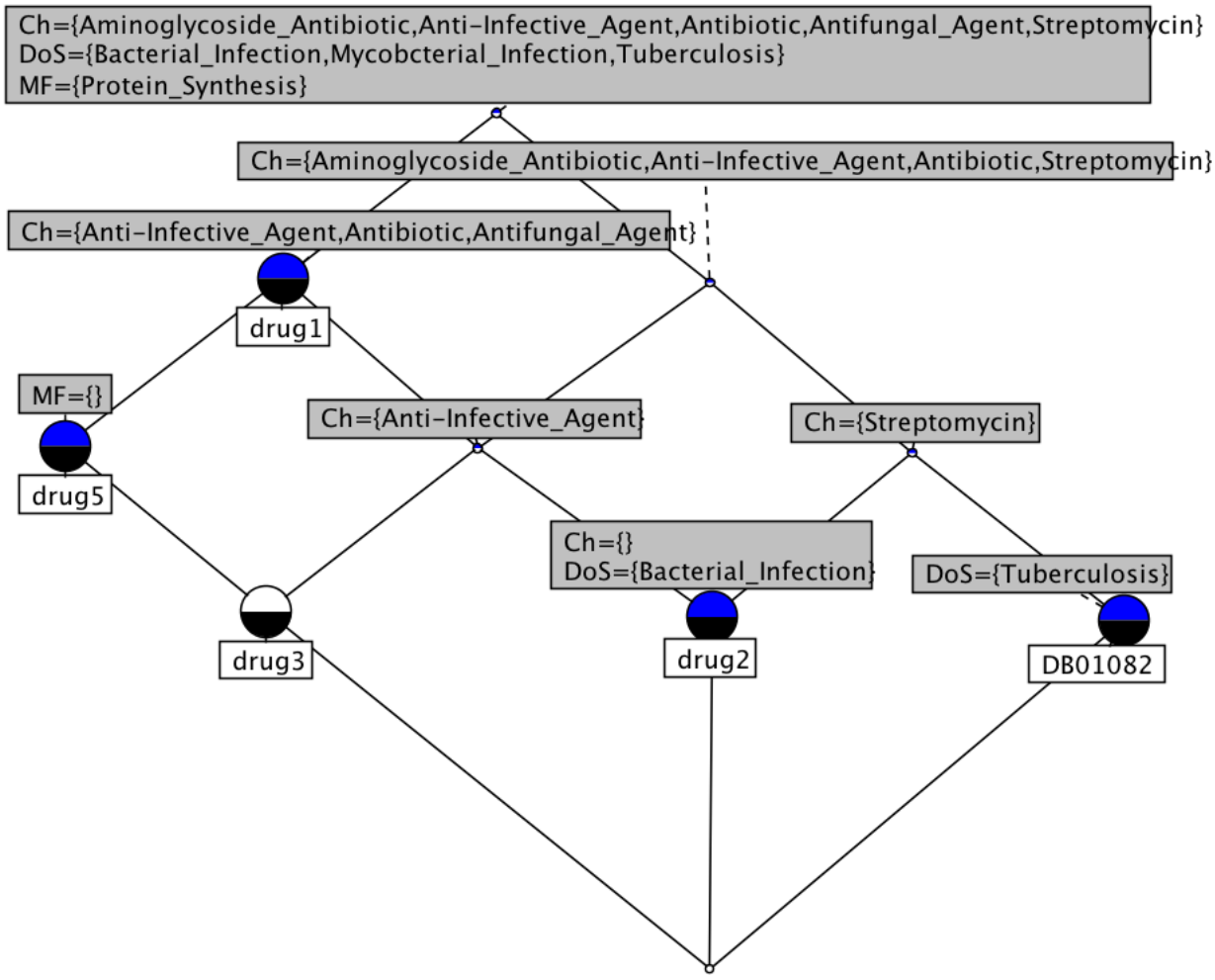


FIGURE 1.9 – Représentation du treillis obtenu par la structure de patrons exemple définie par le contexte présenté Tableau 1.3.

.

Dans [Coulet *et al.*, 2013] nous avons ainsi proposé pour la première fois une adaptation des structures de patrons qui prend en considération des connaissances de domaine. Cette approche a été réutilisée dans d'autres travaux comme par exemple celui de Mehwish Alam *et al.* sur l'utilisation de l'AFC pour apprendre des définitions de concepts à partir de graphes de connaissances [Alam *et al.*, 2015] ou le travail présenté dans le chapitre suivant qui fait usage des structures de patrons et de plusieurs ontologies pour comparer des objets complexes : les effets indésirables de médicaments. Un défi mis en avant par le travail présenté ici est la gestion de la quantité de concepts de patrons construits. Pour mesurer cette quantité nous avons construit un contexte proche de l'exemple précédent avec les 4 mêmes types sémantiques, mais 25 antibiotiques au lieu de 5 et l'ensemble de leurs annotations dans le *Resource Index* (présenté Section 1.2). Avec notre méthode, ce contexte produit 204 801 concepts fermés ce qui fait un treillis de concepts large et complexe à analyser. Le défi est alors d'identifier des concepts plus intéressants au regard d'une métrique. La définition de telles métriques a motivé plusieurs travaux de notre collaborateur Florent Domenach sur la similarité des concepts formels [Domenach, 2013, Domenach and Portides, 2014, Domenach, 2017].

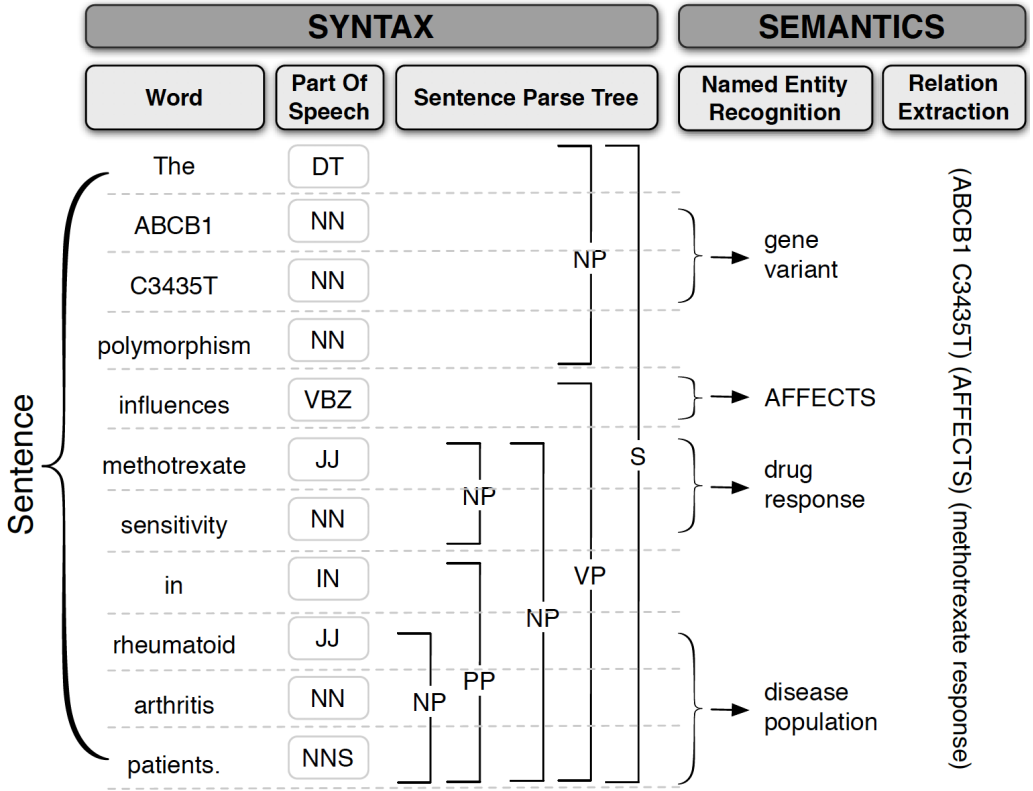
1.5 Quelques mots sur les textes

J'ai fait le choix de ne pas consacrer un chapitre entier à l'annotation et l'extraction de connaissances à partir de textes en langage naturel, bien que cela ait été une partie importante de mon activité de recherche d'après thèse. Ce choix n'est pas motivé par le fait que ces travaux sont en décalage avec ce qui est présenté ici. Ils sont très en phase avec la ligne directrice de ce mémoire qui est la création et l'utilisation d'annotations ontologiques pour la découverte de connaissances. Ce choix correspond plutôt à la volonté de ne pas alourdir outre mesure ce mémoire. Pour ne pas complètement les occulter, je propose de les décrire très brièvement dans cette section, en tâchant de mettre en relief la façon dont les connaissances de domaines sont intervenues pour guider la fouille de textes.

De façon générale nous nous sommes intéressés à l'extraction de relations à partir de texte et en particulier l'extraction de relations entre des entités complexes. Notre première contribution dans ce domaine a été le développement d'une méthode d'extraction de relations pharmacogénomiques à partir de la littérature [Coulet *et al.*, 2010] dont la principale particularité était de normaliser les entités complexes (*e.g.*, les réponses à des médicaments, les variants génétiques) qui entrent en jeux dans les relations. Cela nous permet de reconnaître que dans certains cas, des phrases qui n'ont pas un mot en commun, peuvent en réalité rapporter une relation identique. Nous avons pour cela construit un ensemble de règles d'extraction (qui utilise les dépendances grammaticales entre les mots d'une phrase) et utilisé une ontologie qui représente les dépendances grammaticales entre termes qui peuvent être utilisées pour décrire une même entité. La partie inférieure de la Figure 1.10 donne deux phrases très différentes et montre qu'après normalisation celles-ci sont comparables à plusieurs niveaux. La partie supérieure de la Figure montre les différents constituants syntaxiques et sémantiques d'une des phrases. Un des intérêts de cette normalisation est de pouvoir synthétiser l'état des connaissances de ce domaine sous la forme d'un graphe qui peut ensuite être fouillé en vue de la découverte de connaissances. Pour cette raison nous avons appliqué notre méthode d'extraction et normalisation à l'ensemble des résumés de PubMed publiés avant 2008 et avons mis en réseau les relations extraites [Coulet *et al.*, 2011]. L'analyse du graphe de connaissances résultant a ensuite été la source de plusieurs travaux de l'équipe du Prof. Russ Altman à Stanford [Percha *et al.*, 2012, Percha and Altman, 2013].

Une des limites de cette approche est que les règles d'extraction sont écrites manuellement à partir de l'observation d'exemples, ce qui la rend peut généralisable. Pour cette raison nous avons appris automatiquement de telles règles dont nous avons évalué la qualité dans un scénario d'extraction de relations maladie–symptôme [Hassan *et al.*, 2015].

Dans le cadre du projet ANR PractiKPharma où l'on souhaite comparer les connaissances pharmacogénomiques de l'état de l'art, nous nous sommes intéressés à nouveau à l'extraction de relations pharmacogénomiques, mais cette fois avec des méthodes d'apprentissage profond. Le constat était que c'est ce type d'approche qui obtient les meilleures performances de l'état de l'art mais qu'il n'existe pas de corpus annoté de taille suffisante pour entraîner un modèle d'extraction de relations pharmacogénomiques avec un modèle d'apprentissage profond. Pour cela, nous avons étudié d'une part la possibilité d'entraîner un modèle pour une tâche d'extraction cible à partir d'un corpus développé pour une tâche source différente de la tâche cible. D'autre part nous avons lancé la création d'un corpus annoté manuellement pour ce domaine. Dans les deux cas nous avons obtenu des résultats concluants. Nous avons mis en évidence des modèles et des descripteurs du texte qui permettent de faire de l'apprentissage par transfert pour la tâche d'extraction de relations [Legrand *et al.*, 2018]. En parallèle, nous avons construit PGxCorpus, un corpus de 945 phrases où 2 875 relations sont annotées manuellement et où chaque phrase a été vue au moins 4 fois et par 4 annotateurs différents [Legrand *et al.*, 2019].



sentence	gene variant	relationship	phenotype
The ABCB1 C3435T polymorphism influences methotrexate sensitivity in rheumatoid arthritis patients.	ABCB1 C3435T	AFFECTS	methotrexate response
A variant 2677A allele of the MDR1 gene affects fexofenadine disposition.	ABCB1 2677A	AFFECTS	fexofenadine response

FIGURE 1.10 – Annotations syntaxiques et sémantiques d'une phrase (Figure issue de [Garten, 2010]).

1.6 Discussion générale

Des annotations simples Les annotations dont nous avons parlé dans cette section sont des annotations très simples où une portion de texte ou un attribut est associé à un concept ontologique. Nos travaux en fouille de texte nous ont montré que les sources de données et les objets biomédicaux qu'elles représentent sont souvent plus complexes et nécessitent des annotations capables de refléter leur subtilité. Un exemple en est les annotations composites comme avec *carbamazepine hypersensitivity* où l'ensemble peut être annoté avec un concept `DrugResponse` ou `Phenotype` mais la sous-partie *carbamazepine* est un nom de médicament et peut donc être annotée avec le concept `Drug`. Nous avons alors deux annotations, l'une imbriquée dans l'autre et il est en plus possible d'associer une sémantique à cette imbrication. Dans ce cas l'hypersensibilité est une réaction qui est causée par la carbamazepine. Cette sémantique est en partie associée au fait que carbamazepine a un rôle de modulateur dans le nom composé. Il qualifie en quelque sort la sensibilité et le fait que nous sachions que c'est un médicament nous dit également que ce phénotype est qualifié par un nom de médicament. Il serait à mon avis très intéressant de considérer ce genre de complexité dans les tâches d'annotation et d'extraction de connaissances.

Un autre exemple d'annotation plus complexe est l'annotation de relations par exemple binaires, qui décrivent la mention dans un texte d'un lien entre deux entités. Le type du lien en question peut être défini dans une ontologie qui offre un ordre pratique pour comparer des relations, à la condition que la relation soient faites entre des entités qui elles-même sont comparables. En effet les entités associées peuvent être annotées avec des ontologies distinctes, ce qui augmente le nombre de référentiels à considérer et mettre en correspondance pour comparer et distinguer une relation vis à vis d'une autre. Par ailleurs, le niveau d'hétérogénéité de ces relations fait que les besoins en termes de normalisation et d'agrégation sont importants pour des tâches d'analyse de données.

L'annotation par dictionnaire Les annotations générées par l'*Annotator* du BioPortal et regroupées dans le *Resource index* sont obtenues par une approche par dictionnaire. Les labels associés aux concepts des ontologies du BioPortal sont utilisés pour constituer un grand dictionnaire, dont le contenu est comparé à un nouveau texte traité pour retrouver la mention de concepts définis dans le dictionnaire. Lors de la comparaison entre le texte et les entrées du dictionnaire, il est possible d'ajouter des modules linguistiques pour prendre en compte les pluriels ou les lemmes. Malgré ces améliorations possibles, l'approche par dictionnaire a certaines limites. Par exemple il existe des ambiguïtés dans le dictionnaire, c'est-à-dire qu'une chaîne de caractère fait référence à plusieurs entrées et donc à plusieurs concepts, ce qui est source de faux positifs dans l'annotation automatique. Les approches de *désambiguïsations* classiques utilisent le contexte des mots pour distinguer le sens auquel ils sont associés dans un document particulier. Avec l'idée de prendre en considération le contexte dans lequel est fait une annotation des améliorations de l'*Annotateur* ont été proposées [Tchechmedjiev *et al.*, 2018]. Il permet ainsi de détecter si un concept est nié ou s'il est mentionné dans un contexte hypothétique à l'aide de listes des termes particuliers qui sont recherchés en amont et en aval de la mention d'un concept. Une approche alternative pour désambiguïser est de considérer la *sémantique distributionnelle* associée à un concept. Dans ce cas la sémantique d'un concept est définie par la distribution des mots observés dans son voisinage au sein d'un corpus d'apprentissage [Collobert *et al.*, 2011]. Les mots voisins et leur distribution respective sont encodés dans des vecteurs numériques appelés *embeddings*. Ces représentations bénéficient de bonnes performances notamment pour la reconnaissance d'entités nommées et résolvent dans une certaine mesure les problèmes d'ambiguïté. Cependant, elles nécessitent pour cela d'être entraînées sur des corpus annotés de grande taille

qui ne sont pas toujours disponibles, lorsque l'on s'intéresse à des entités rares ou complexes.

Apprentissage faiblement supervisé L'apprentissage faiblement supervisé a l'attrait de pouvoir tirer parti d'annotations automatiques faites sans supervision, de qualité moyenne mais obtenues en grand nombre, pour palier le manque d'un grand ensemble de données manuellement annotées. Ce genre d'approche se satisfait d'un ensemble de données annotées de petite taille et des erreurs d'annotation que peut faire une annotation automatique [Ratner *et al.*, 2017, Dao *et al.*, 2018]. L'idée est alors de tirer parti du bruit de l'annotation automatique connu sur le jeu d'entraînement et de généraliser à propos de ce bruit pour considérer les annotations faites automatiquement sur le reste des données et alors l'objectif est de pouvoir considérer un ensemble d'entraînement plus grand, dont on a une estimation du bruit. Il serait intéressant je pense d'annoter manuellement un petit ensemble des données du Resource Index et d'utiliser le restant dans un processus faiblement supervisé pour entraîner un nouveau modèle capable d'annoter de nouveaux documents.

Chapitre 2

Des annotations pour l'analyse de données cliniques

Sommaire

2.1 Introduction 41
2.2 La recherche de cooccurrence d'effets indésirables médicamenteux 42
2.2.1 Motivation 42
2.2.2 Données et ontologies 42
2.2.3 Opérateur de similarité 44
2.2.4 Comparaison entre STRIDE et FAERS 45
2.2.5 Analyse statistique des associations entre EIM 45
2.2.6 Discussion et conclusion 46
2.3 Prédiction de la nécessité de réduire la dose d'un médicament, avant sa première prescription 48
2.3.1 Les données 48
2.3.2 Les attributs considérés 51
2.3.3 La sélection d'attributs 52
2.3.4 La tâche d'apprentissage 53
2.3.5 Résultats 55
2.3.6 Discussion et conclusion 55
2.4 Discussion générale 60

2.1 Introduction

Dernièrement, nous nous sommes intéressés à l'analyse d'une source de données particulièrement riche et prometteuse : les dossiers patients électroniques (DPE ou *EHR* en anglais pour *Electronic Health Records*). La considération de ces données dans la prise en charge des patients peut avoir de grands impacts sociétaux, à condition de lever certains verrous associés à leur utilisation [Jensen *et al.*, 2012]. Généralement les DPE sont constitués au fur et à mesure de l'activité clinique hospitalière puis sont extraits à intervalles réguliers dans des entrepôts de données cliniques (ou EDC) dédiés notamment à la recherche [Canuel *et al.*, 2015]. Ces données sont typiquement semi-structurées puisqu'elles combinent des attributs structurés et certains champs textes libres comme les notes cliniques. Mais les DPE sont également composés d'annotations,

notamment des annotations pré-existantes, comme les diagnostics encodés avec l'ontologie ICD (*International Classification of Diseases* en angais ou CIM pour Classification internationale des maladies) et des annotations créées à partir du traitement des notes cliniques.

Ce chapitre présente deux contributions qui s'intéressent à ce type de données en mettant l'accent tant que possible sur l'utilisation des annotations et donc des ontologies lors de leur analyse. La première contribution fait appel à la combinaison structures de patrons - ontologies comme décrit dans le chapitre précédent pour identifier des effets indésirables médicamenteux (EIM) cooccurrents fréquemment [Personeni *et al.*, 2017]. La seconde contribution concerne le développement de modèles prédictifs qui estiment à partir de DPE le risque pour un nouveau patient de nécessiter une réduction de dose médicamenteuse, et cela avant que le médicament ne lui soit prescrit [Coulet *et al.*, 2018].

2.2 La recherche de cooccurrence d'effets indésirables médicamenteux

2.2.1 Motivation

Les effets indésirables médicamenteux (EIM) apparaissent inégalement dans différents sous-groupes de patients. Leurs causes sont multiples : génétiques, métaboliques, interactions de médicaments. Des études précédentes ont montré que les EIM pouvaient être détectés et étudiés en analysant des DPE [LePendu *et al.*, 2013]. Nous souhaitions dans ce travail analyser les DPE pour explorer si un groupe de patients sensible aux effets secondaires d'un médicament est également sensible aux effets secondaires d'un autre. Pour cela, nous proposons une méthode pour identifier des EIM fréquemment associés dans des sous-groupes de patients. Les manifestations des EIM et leur description dans les DPE étant variables et complexes nous avons proposé d'utiliser les structures de patrons [Ganter and Kuznetsov, 2001] associées à des ontologies pour autoriser la généralisation lors de la comparaison des EIM entre-eux. Nous avons utilisé un jeu de DPE de patients diagnostiqués avec le Lupus Erythémateux Disséminé (LED), une maladie auto-immune. Ces patients sont souvent sujets aux EIM de par les traitements du LED et ceux des maladies opportunistes qui l'accompagnent [Vasudevan and Ginzler, 2009].

2.2.2 Données et ontologies

Données et représentation des EIM Nous avons pour ce travail utilisé deux sources de données cliniques distinctes : STRIDE qui est l'EDC de l'hôpital universitaire de Stanford [Lowe *et al.*, 2009] et FAERS [U.S. Food & Drug Administration, 2018 visitée le 17102019] qui est une base de données non pas de DPE mais de rapports d'EIM renseignée par des patients, professionnels de la santé et laboratoires pharmaceutiques aux Etats-Unis et gérée par la FDA (Food and Drug Administration).

Quelle que soit la source de données, un EIM est un évènement complexe qui peut impliquer plusieurs médicaments, et se manifester sous la forme de plusieurs phénotypes. Un EIM peut ainsi être caractérisé par un ensemble de médicaments et un ensemble de phénotypes. On représentera alors 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. Afin de faciliter la comparaison entre des EIM, on considérera des ensembles d'ingrédients actifs de médicaments, plutôt que des ensembles de formes commerciales de médicaments. Ainsi, on utilisera ici le terme "médicament" pour désigner un ingrédient (ou principe) actif.

Afin d'identifier des EIM fréquemment cooccurrents **à partir des DPE de STRIDE**, nous avons travaillé sur un sous-ensemble de 6 869 DPE anonymisés, extraits de STRIDE, de patients diagnostiqués avec le LED. Ce sous-ensemble documente 451 000 visites à l'hôpital, avec leurs dates relatives, diagnostics encodés avec ICD9-CM (International Classification of Diseases, Clinical Modification) et prescriptions sous la forme de listes d'ingrédients, représentés par leurs identifiants RxNorm. Nous avons identifié de façon relativement naïve des EIM candidats, puis sélectionné les patients qui en présentent au moins deux. Nous établissons alors pour chaque patient, une liste d'EIM candidats. Pour deux visites consécutives dans le DPE, nous extrayions l'ensemble des médicaments prescrits (une prescription) dans la première visite D_i, et les phénotypes P_i diagnostiqués durant la seconde. L'intervalle entre deux visites doit être d'au plus 14 jours. Nous pensons qu'il est raisonnable de supposer qu'un effet secondaire se manifeste dans un court délai après la prescription, de plus, nous avons pu observer qu'augmenter la borne haute de cet intervalle (*i.e.*, 14 jours) n'augmente pas le nombre de patients retenus dans notre corpus (voir [Personeni *et al.*, 2017] pour plus de détails). Un EIM candidat C_i est alors un couple d'ensembles $C_i = (D_i, P_i)$. Nous ne conservons dans P_i que les phénotypes listés dans SIDER 4.1 comme effets secondaires d'un médicament de D_i. SIDER est une base de données d'indications et d'effets secondaires de médicaments (Kuhn, et al., 2016). Nous excluons ensuite les candidats où P_i est vide. Nous excluons aussi un EIM candidats (D_1, P_1) s'il existe un autre EIM candidat (D_2, P_2) pour le même patient tel que $D_1 \subseteq D_2$: en effet, si une prescription est répétée pour un patient, cela indique qu'elle n'a pas été jugée dangereuse pour lui. Par ce processus de sélection, nous obtenons à partir de STRIDE un corpus de 3 286 EIM provenant de 548 patients présentant au moins 2 EIM.

Pour constituer un ensemble similaire d'EIM **à partir des rapports de FAERS**, nous avons utilisé la ressource AEOLUS [Banda *et al.*, 2016] qui propose des outils pour mettre en correspondance les médicaments avec les vocabulaires RxNorm et les phénotypes avec le vocabulaire SNOMED CT. Nous avons utilisé ces outils pour reconstruire une base de rapports FAERS sur la période 2012 quatrième trimestre à 2016 second trimestre inclus. Chaque rapport d'EIM 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 également 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. Comme pour STRIDE, nous avons exclu les EIM où l'ensemble des médicaments est déjà inclus dans un autre EIM. Comme pour STRIDE, nous ne conservons que les patients (*i.e.*, les cas) avec au moins 2 EIM, puisque l'on s'intéresse à des associations entre EIM. Avec ces contraintes, nous avons pu extraire de FAERS 570 patients pour un total de 1 148 EIM.

Ontologies Nous utilisions trois ontologies biomédicales : ICD9-CM qui décrit des classes de phénotypes et qui est utilisé dans STRIDE pour encoder les diagnostics ; SNOMED CT qui est une ontologie médicale que nous utilisons pour décrire les phénotypes issus de FAERS à l'aide des correspondances offertes par AEOLUS ; et l'Anatomical Therapeutic Chemical Classification System (ATC) qui décrit des classes de médicaments et est utilisée dans les deux sources. Nous considérons seulement la hiérarchie de classes de ces ontologies afin de généraliser la description des phénotypes et des prescriptions extraits des sources. De plus, pour simplifier le problème nous utilisons uniquement les trois niveaux les plus spécifiques d'ATC : sous-groupes pharmacologiques, sous-groupes chimiques, substances chimiques.

TABLE 2.1 – Exemple de représentation d'EIM pour trois patients. Les trois n'ont que deux EIM, *i.e.*, deux vecteurs $\langle D_i, P_i \rangle$. Chaque EIM n'est composé que d'un seul médicament et un seul phénotype sauf le premier EIM du patient P3 qui associe deux médicaments à un phénotype. Les codes ATC et ICD9-CM du tableau sont associés aux labels suivants : $H02AB07$=Prednisone, $N02BE01$=Paracétamol, $ICD599.8$=Other specified disorders of urethra and urinary tract, $ICD599.9$=Unspecified disorder of urethra and urinary tract, $ICD719.4$=Pain in joint.

Patient	**Description**
P1	$\{\langle\{H02AB07\}, \{ICD599.8\}\rangle, \langle\{N02BE01\}, \{ICD599.9\}\rangle\}$
P2	$\{\langle\{H02AB07\}, \{ICD599.9\}\rangle, \langle\{H02AB07\}, \{ICD719.4\}\rangle\}$
P3	$\{\langle\{H02AB07, N02BE01\}, \{ICD599.9\}\rangle, \langle\{N02BE01\}, \{ICD719.4\}\rangle\}$

2.2.3 Opérateur de similarité

On définit ici une structure de patrons $(G, (D, \sqcap), \delta)$ où : G est un ensemble de patients, D est un ensemble de descriptions représentant chacune un ensemble d'EIM, et δ est la fonction associant à un patient la description de ses EIM. On définira cette structure de patrons de manière à permettre l'utilisation de l'ontologie ATC pour décrire les médicaments ainsi qu'une ontologie de phénotypes (ICD9-CM ou SNOMED CT) pour décrire les effets secondaires. Dans notre cas, afin d'éviter une sur-généralisation des descriptions d'EIM, on exclut les trois niveaux les plus généraux de l'ontologie de phénotypes. La Tableau 2.1 présente un exemple de la représentation de données utilisées avec la structure de patrons décrite ici, exprimée avec les ontologies ATC et ICD9-CM.

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.

On définit d'abord un opérateur $\sqcap_{\mathcal{O}}$ permettant la comparaison de deux ensembles de classes d'une ontologie $\mathcal{O}$ tel que, pour x et y deux ensembles de classes d'une ontologie :

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

où $max(\leq_i, S)$ est l'unique ensemble des élements maximaux appartenant à un ensemble S pour un ordre partiel $\leq_i$, tel que $max(\leq_i, S) = s | \nexists x.(s \leq_i x), x, s \in S$.

On définit ensuite un opérateur permettant la comparaison deux à deux d'EIM $\sqcap_{EIM}$ tel que :

$$\langle D_x, P_x \rangle \sqcap_{EIM} \langle D_y, P_y \rangle = \begin{cases} \langle D_x \sqcap_{\mathcal{O}} D_y, P_x \sqcap_{\mathcal{O}} P_y \rangle \text{ si les deux dimensions sont} \neq \varnothing \\ \langle \varnothing, \varnothing \rangle \text{sinon} \end{cases}$$

L'opérateur $\sqcap_{EIM}$ applique l'opérateur de similarité $\sqcap_{\mathcal{O}}$ sur les deux dimensions du vecteur représentant l'EIM, en utilisant pour chaque dimension l'ontologie associée 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 s'ils n'ont aucun médicament ou phénotype en commun.

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

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

TABLE 2.2 – Statistiques sur le processus d'AFC et d'extraction de règles d'associations

Données	**STRIDE (DPE)**	**FAERS (rapports d'EIM)**
#patients	548	570
#EIM	3 286	1 148
#concepts dans le treillis	≈2.5 millions	≈22 700
#règles extraites	≈9 millions	≈18 500
#règles conservées après filtrage	913	493
#règles avec un support > 8	15	151
Support maximal	10	27

Cet opérateur peut s'appliquer aux données extraites des DPE ou de FAERS, 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.

2.2.4 Comparaison entre STRIDE et FAERS

La structure de patrons présentée dans la section précédente permet de construire un treillis de concepts duquel on extrait des règles d'associations, avec un support et une confiance d'au moins 5 et 0,75, respectivement. La Table 2.2 présente quelques statistiques sur ce processus.

On observe d'abord que l'expérience sur FAERS produit un nombre de règles beaucoup moins important qu'avec les DPE de STRIDE. Cependant, comme dans l'expérience sur les DPE, beaucoup de ces règles contiennent des associations triviales, notamment 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 $\langle\{d_1\},\{p_1\}\rangle \rightarrow \langle\{d_1,d_2\},\{p_1\}\rangle$ semble triviale dans le cas de l'étude des EIM). 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 ne soient comparable par $\leq_{\mathcal{O}}$.

On observe également que l'expérience utilisant les données de FAERS génère un treillis de concepts beaucoup plus petit que celle utilisant les données de DPE (100 fois moins de concepts), et ce malgré un nombre de patients comparable. Cependant après filtrage des règles, leur nombre est dans le même ordre de grandeur (seulement 2 fois moins). Cette différence peut s'expliquer par les différences entre les deux jeux de données : le jeu de données de DPE 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 DPE 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/`.

2.2.5 Analyse statistique des associations entre EIM

Pour chaque paire de classes de médicaments ATC (l, r), nous avons cherché l'ensemble des règles extraites des DPE et de FAERS qui ont 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. Nous avons ensuite calculé 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. Ce ratio exprime la fréquence à laquelle les règles associent un EIM

impliquant un médicament de la classe r à un EIM impliquant un médicament de la classe l. Nous avons utilisé un Z-test pour évaluer la significativité statistique de l'écart obtenu à l'écart attendu si les associations entre EIM étaient extraites aléatoirement (voir les Figure 2 et 3 de [Personeni *et al.*, 2017] pour l'ensemble des résultats). On observe notamment quelques associations d'intérêt entre classes ATC (avec p $<$ 0 :001). Par exemple, on constate parmi les règles extraites des DPE, 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. 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.

Parce que les DPE extraits de STRIDE ne représentent qu'une petite partie de cet entrepôt de 2 millions de DPE, pour les 15 règles ayant le plus grand support (entre 8 et 10) de celles extraites de ces DPE, nous avons calculé leur support sur la totalité de l'entrepôt STRIDE, cela pour évaluer la généralité de ces règles. On observe que pour ces 15 règles, le nombre de DPE les vérifiant varie entre 33 à 326 pour une moyenne de 86,2. Cela illustre que les règles d'association extraites à partir de DPE de patients atteints de LED peuvent être pertinentes en dehors du jeu de données initial.

2.2.6 Discussion et conclusion

Ce travail montre que les structures de patrons permettent de fouiller des objets réels complexes (comme les EIM) en considérant les connaissances représentées dans plusieurs ontologies. L'utilisation des ontologies biomédicales permet une comparaison flexible des EIM, notamment en comparant des EIM *(1)* causés par des médicaments d'une même classe et *(2)* ayant des manifestations potentiellement distinctes mais proches dans l'ontologie des phénotypes. Il est cependant nécessaire de considérer la complexité des algorithmes d'AFC combinés aux opérateurs de similarité proposés. Le nombre de concepts générés est grand, notamment car l'on augmente les possibilités de trouver une description générale commune à de nombreux patients. Cette augmentation nous a imposé un post-traitement important des résultats d'AFC pour pouvoir interpréter les résultats.

Une limite des expérimentations avec les données réelles est le manque d'intersection entre les résultats obtenus avec les deux sources de données. Cela est notamment dû à la faible qualité des correspondances existantes entre les ontologies de phénotype ICD9-CM et SNOMED CT utilisées respectivement dans STRIDE et FAERS. Cela peut également s'expliquer par la différence de nature des deux jeux de données : FAERS est constitué d'EIM rapportés librement par des patients (entre autres), tandis que à partir des DPE les EIM sont automatiquement extraits des dossiers, sans d'ailleurs que la qualité de cette extraction n'ait été évaluée. De plus, 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.

Une limite de la représentation des EIM utilisée dans cette étude est la considération des seuls EIM apparus dans une fenêtre de 14 jours après la prescription. Ainsi, seuls des EIM à court terme ont été considérés. La représentation des EIM pourrait être enrichie avec le délai entre la prescription et l'apparition du phénotype indésirable. 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 limite de l'extraction de règles d'association est que les règles extraites n'expriment pas une relation causale entre les EIM associés. En effet, 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.

A notre connaissance, cette approche est unique tout d'abord dans sa capacité à comparer des représentations détaillées d'EIM en considérant de multiples ontologies mais également dans son objectif d'extraire des associations entre EIM fréquemment associés. Nous avons démontré la flexibilité de cette approche en l'appliquant à deux sources différentes de données et des ontologies 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 second EIM impliquant le médicament d.

2.3 Prédiction de la nécessité de réduire la dose d'un médicament, avant sa première prescription

Dans ce travail nous avons exploré la faisabilité d'utiliser trois types de descripteurs phénotypiques présents dans les DPE : les codes diagnostics ICD9-CM (*i.e.*, des annotations préexistantes) ; les conditions mentionnées dans le texte des notes cliniques annotés par un concept SNOMED CT (*i.e.*, des annotations obtenues à partir de texte libre) ; les examens de laboratoire commandés par les patients (*i.e.*, des attributs structurés non associés à une ontologie) pour prédire la nécessité de réduire la dose initiale d'un médicament, avant que celui-ci ne soit prescrit à un patient [Coulet *et al.*, 2018].

Une vue générale de la méthode que nous avons adoptée pour cet expérience est présentée Figure 2.1. Les étapes principales y sont identifiées et sont décrites brièvement dans la légende de la Figure.

2.3.1 Les données

Les DPE de STRIDE Nous avons ici encore utilisé les DPE de STRIDE, l'EDC de l'hôpital de l'Université Stanford [Lowe *et al.*, 2009]. La version de STRIDE utilisée contient des données de 1 250 825 patients venus à l'hôpital entre 2008 et 2014, constituant 49 086 060 visites, 27 049 309 notes cliniques, 19 435 069 commandes de médicaments (dont 2 891 470 pour les seuls médicaments considérés dans cette étude) et 165 141 675 commandes d'examens biologiques.

Les changements de dose Nous avons défini des évènements particuliers au sein de STRIDE : les *intervalles de changements de dose*. Ceux-ci sont définis comme suit : il s'agit d'intervalles de temps de moins de 20 jours pendant lesquels une molécule active de médicament a été prescrite deux fois (et deux fois seulement) à un même patient, suivant la même route d'administration (*e.g.*, voie orale, intra-veineuses) et rapportés dans les DPE avec la même unité. Nous distinguons ainsi trois types de changements de dose :

- les réductions de dose, pour lesquelles la seconde prescription est faite avec une dose plus faible que la première ;
- les augmentations de dose, pour lesquelles la seconde prescription est faite avec une dose plus élevée que la première ;
- les continuations de la dose, qui ne sont pas des changements de dose à proprement parler, mais que nous considérons également et pour lesquelles la dose prescrite est inchangée entre les deux prescriptions.

Une représentation schématique de ces intervalles est présentée Figure 2.2. La durée de 20 jours est un choix arbitraire, qui est soutenu par le fait que dans notre étude la longueur moyenne des intervalles est de 3,64 jours avec un écart type de 4,41.

Les réductions de doses sont soit une réduction dans la quantité de molécule prescrite (*e.g.*, 2g toutes les 2 heures → 1g toutes les 2 heures) ou une réduction dans la fréquence d'administration (*e.g.*, 2g toutes les 2 heures → 2g toutes les 4 heures). Respectivement les augmentations de doses sont des augmentations soit de la quantité soit de la fréquence de la prescription. Nous avons éliminé les valeurs extrêmes en excluant les intervalles les plus courts, c'est-à-dire les 10% les plus courts (ce qui correspond aux intervalles de moins de 6 heures) ; et en excluant également les intervalles les plus longs (les 10% les plus long également). Une particularité supplémentaire de la préparation de données est que nous ne considérons les continuations de dose que pour les patients qui n'ont jamais connu d'augmentation ou de réduction de dose. En revanche, les augmentations et réductions peuvent précéder ou suivre des intervalles de continuation de la dose.

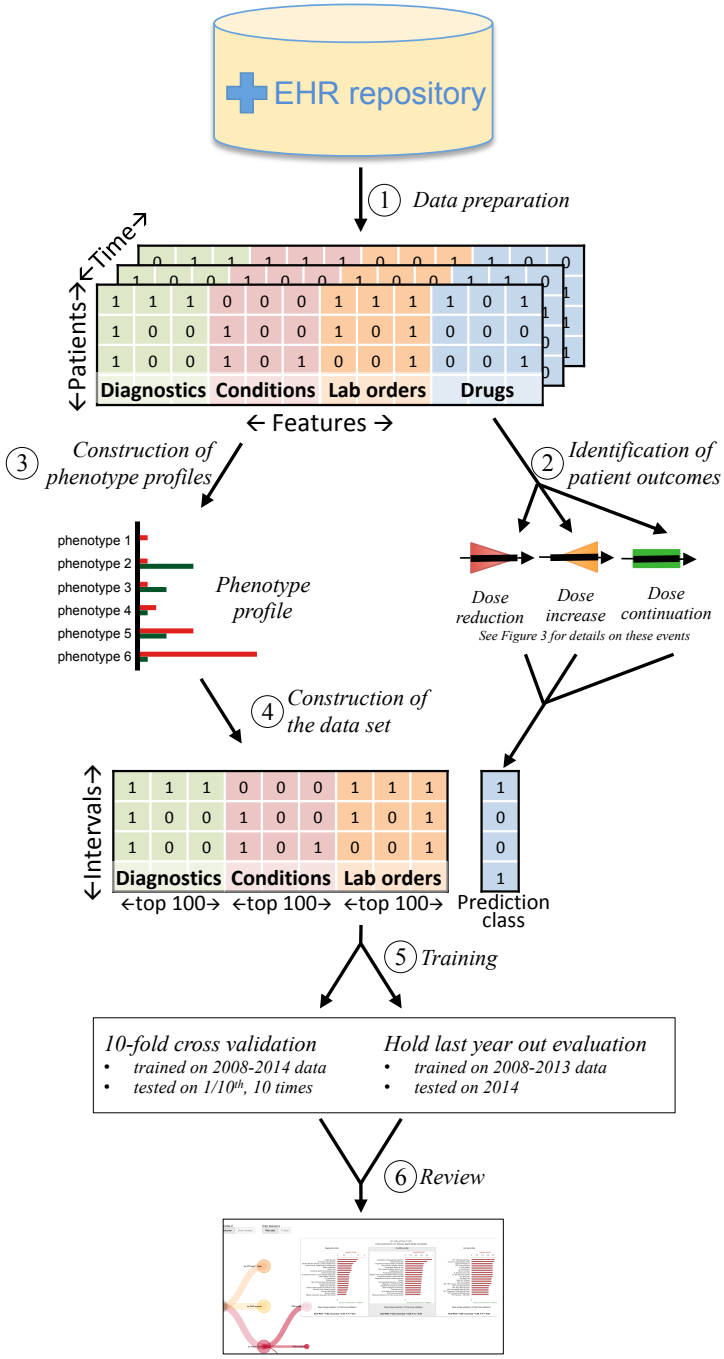


FIGURE 2.1 – Vue générale de notre approche de prédiction de changement de dose. (1) Les annotations par des concepts ontologiques des dossiers patients sont extraites. (2) Les évènements de changement de doses sont identifiés. (3) Les annotations les plus caractéristiques des changements de dose sont sélectionnées pour construire des profils phénotypiques. (4) Ces profils sont utilisés pour construire une matrice où les patients sont décrits avec les annotations les plus caractéristiques. (5) La matrice résultante est utilisée pour entraîner puis (6) évaluer deux modèles de forêts d'arbres aléatoires : l'un pour prédire les réductions de dose, l'autre pour prédire les augmentations.

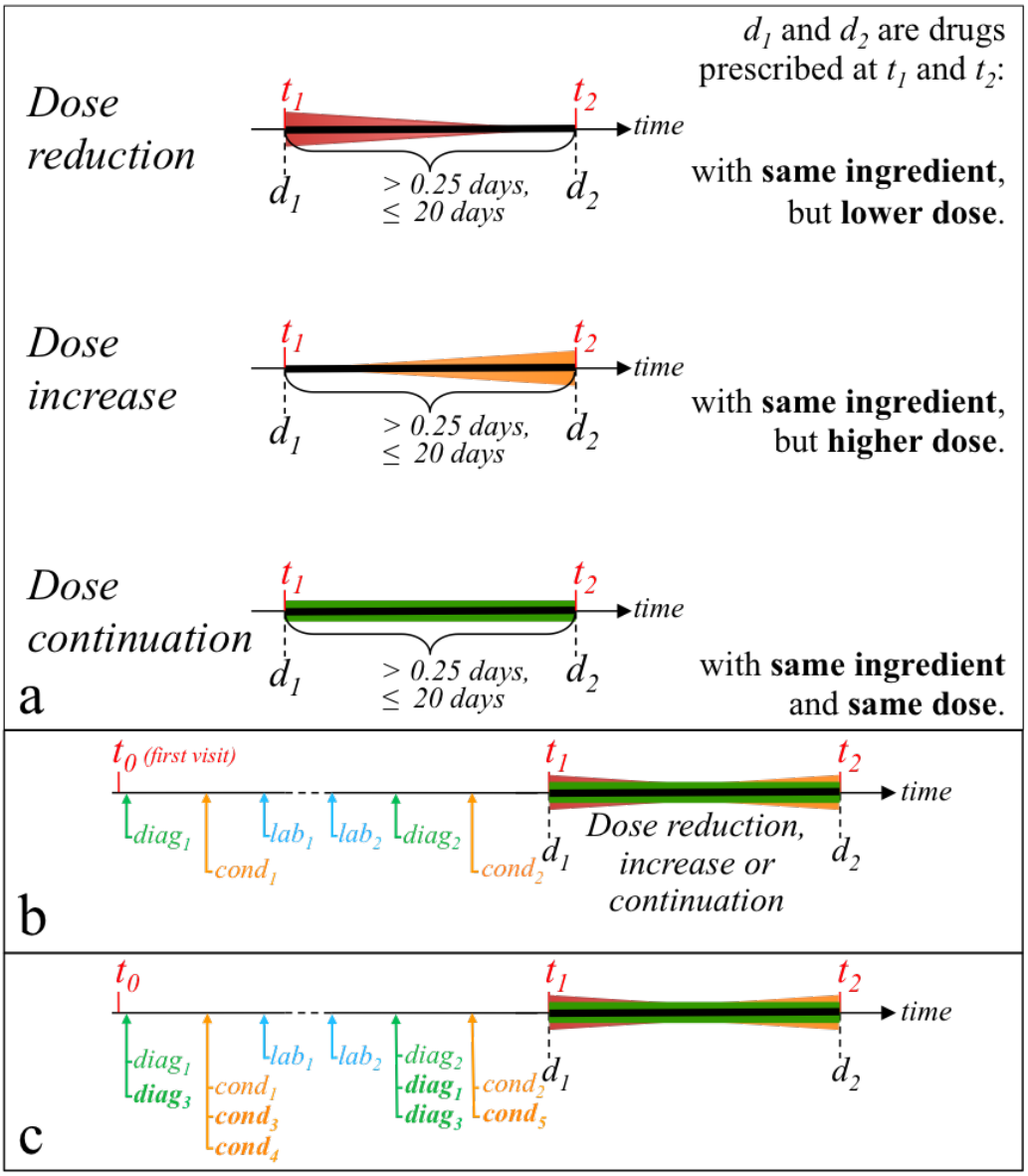


FIGURE 2.2 – Définition des intervalles de changement ou de continuation de la dose des prescriptions. La partie haute **(a)** présente les trois types d'intervalles, chacun délimité par deux prescriptions d_1 et d_2 faites à t_1 et t_2. Aucun autre médicament avec la même molécule n'est prescrit durant un intervalle. Les deux panneaux du bas positionnent sur la ligne de temps les attributs utilisés par les modèles prédictifs, avant la première prescription. **(b)** Les trois types d'attributs sont les concepts ICD9 utilisés pour le diagnostic (*diag*), les conditions annotées dans les notes cliniques (*cond*) et les commandes d'examens biologiques (*lab*). Puisque les diagnostics et conditions sont des annotations faites avec des concepts d'ontologies, ils sont généralisés selon la hiérarchie de concepts pour enrichir les annotations (*i.e.*, de (b) on obtient (c)). Par exemple comme $diag_3$ est plus général que $diag_1$ il est également associé à l'historique du patient.

Au final, nous avons identifié 50 704 réductions, 60 719 augmentations et 176 140 continuations de dose dans les prescriptions des médicaments considérés dans cette étude.

Les médicaments considérés Nous n'avons considéré dans cette étude que les médicaments dont le métabolisme est impacté par les enzymes de la famille des *cytochromes P450*. En effet les membres de cette famille sont connus pour interagir avec de nombreux médicaments et xénobiotiques. Nous appellerons ici ces médicaments les médicaments P450. Flockhart propose une liste de ces médicaments [Flockart, 2007] qui est établie manuellement par des pharmacologues experts des enzymes P450 et qui associe à chaque médicament les références bibliographiques qui soutiennent l'association. En suivant cette liste, nous avons identifié 205 médicaments P450 que nous avons groupés en 25 ensembles non disjoints, groupés selon 3 critères distincts : le premier niveau de la classification ATC, l'enzyme P450 associé et l'impact sur l'activité de l'enzyme (substrat, inhibiteur ou inducteur). Cette classification de nos médicaments est visible en annexe B.1

Pour réduire le biais des médicaments très fréquemment prescrits, nous avons exclu ceux prescrits plus de 55 000 fois dans STRIDE. Des 205 médicaments groupés en 25 ensembles, nous avons exclu ceux associés avec trop peu d'intervalles de changement de dose en imposant un seuil minimum arbitraire de 300 intervalles (≥150 réductions/augmentations et ≥150 continuations). Seuls 34 médicaments et 23 groupes en possédaient suffisamment.

2.3.2 Les attributs considérés

Comme évoqué en préambule de cette section, nous considérons trois types d'attributs, tous observés avant la première prescription des intervalles de changement de dose. Le positionnement des attributs considérés par rapport aux changements de dose est représenté Figure 2.2.

Les **codes diagnostics** sont encodés avec l'ontologie ICD9-CM. Ces codes sont associés à chaque visite du patient pour documenter les raisons principales de son admission et les principaux évènements qui ont pu se produire durant son séjour.

Les **conditions** sont des maladies ou symptômes mentionnés dans le texte des notes cliniques. Celles-ci sont générées selon l'annotation automatique décrite dans [LePendu *et al.*, 2013]. Ce processus est une adaptation légère du processus utilisé pour le peuplement du Resource Index décrit Section 1.2. Ce processus ignore les concepts qui sont niés ou qui sont mentionnés dans l'historique familial du patient. Ses performances rapportées dans [LePendu *et al.*, 2013] sont 74% de sensitivité et 96% de specificité. Les conditions considérées dans ce travail sont uniquement les concepts de l'ontologie SNOMED-CT qui sont associés aux types sémantiques [Bodenreider, 2004] suivant : *'Disease or Syndrome'*, *'Mental or Behavioral Dysfunction'*, *'Cell or Molecular Dysfunction'*, *'Event'*, *'Sign or Symptom'*, *'Anatomical Abnormality'*, *'Neoplastic Process'*.

Les **commandes d'examens biologiques** (ou de tests de laboratoire) sont des données structurées qui indiquent si un examen biologique, par exemple la saturation en oxygène a été commandé pour le patient. Malheureusement, au moment où nous avons réalisé ce travail notre vue de STRIDE ne contenait pas les valeurs associées à ces examens. Aussi seulement 2 examens sur 10 étaient associés à l'époque à une ontologie (LOINC). Pour cette raison nous avons utilisé les codes locaux d'examens sans considérer leur correspondance éventuelle dans une ontologie de référence. Même si les commandes d'examens biologiques ne sont pas à proprement parler des attributs phénotypiques, nous les considérons comme tels ici car ils peuvent être vus comme de bons "proxies" des observations faites sur le phénotypes. En effet, les examens discriminants

ne seront pas ceux qui sont commandés pour chaque patient, mais ceux qui sont faits suite à l'observation d'un phénotype particulier.

Nous avons utilisé les hiérarchies des ontologies pour ajouter des annotations avec les concepts parents du concept initial. Cela génère un ensemble d'attributs étendu comme illustré par la Figure 2.2. Par exemple, si une réduction de dose est associée avec le concept *Stomatitis* de SNOMED CT, il sera automatiquement associé avec ses parents *Inflammatory disorder of digestive tract* et *Disorder of digestive tract*. Cette généralisation a pour objectif de capturer plus de descripteurs des patients avec l'espoir que ceux-ci offrent plus d'opportunités au modèle de prédiction de trouver des similarités entre patients et ainsi de généraliser plus facilement.

2.3.3 La sélection d'attributs

Pour sélectionner un sous-ensemble d'attributs, nous avons utilisé l'approche de Lependu *et al.* qui propose d'utiliser l'*analyse par enrichissement* pour mettre en avant des maladies ou des phénotypes sur-représentés dans un ensemble d'annotations [LePendu *et al.*, 2011]. Ce type d'analyse est classiquement utilisé en transcriptomique pour générer des profils d'expression de gènes, ce qui permet d'étudier comment l'expression des gènes varie selon les conditions telles qu'une maladie ou un traitement [Subramanian *et al.*, 2005]. Un "profil" dans le cas de la transcriptomique est un ensemble de gènes sur- ou sous- exprimés de façon statistiquement significative dans une condition plutôt qu'une autre, comme par exemple en présence d'une maladie ou en son absence. Suivant [LePendu *et al.*, 2011] la mesure de l'expression des gènes peut être remplacée par des attributs phénotypiques comme les codes diagnostics (ou les conditions mentionnées dans les notes cliniques ou les examens biologiques commandés) présents dans les DPE, alors l'analyse par enrichissement met en avant les codes ICD9 (ou les conditions ou les examens biologiques) qui sont sur-représentés dans un groupe de patients par rapport à un autre.

Dans ce travail, nous avons construit des profils phénotypiques des patients à qui il a été prescrit un médicament P450 en comparant les patients qui ont subi une réduction de dose pour ce médicament avec ceux qui n'ont eu que des continuations de dose ; et en comparant les patients qui ont subi une augmentation de dose avec ceux qui n'ont eu que des continuations. La sur-représentation d'un attribut (un gène dans le cas classique, un phénotype pour nous) est quantifiée par sa *valeur p* à l'aide (en général) du test hypergéométrique. La valeur p quantifie la probabilité que l'attribut soit associé à une réduction (ou une augmentation) de dose par hasard. En plus de la valeur p, nous avons également calculé pour chaque attribut son *risque relatif* (RR) et son *contenu en information* (*IC* pour *information content* en anglais). Le RR quantifie l'importance de l'association entre l'attribut et la réduction (ou augmentation) de dose. L'IC est utilisé pour filtrer les attributs qui peuvent être soit trop communs, soit trop rares dans notre entrepôt clinique. Les lignes qui suivent détaillent la façon dont sont calculées ces trois métriques qui nous permettent de sélectionner des attributs.

Calcul de la valeur p avec le test hypergéométrique Dans notre cas, la valeur p quantifie le fait que l'association entre un attribut phénotypique et une réduction (ou augmentation) de dose d'un médicament P450 ou d'un groupe de médicament P450, soit due au hasard. Plus petite est la valeur p, plus grande est la significativité statistique entre l'attribut et la variable réponse (réduction ou continuation, augmentation ou continuation). Dans le cas des réductions de dose, le calcul de la valeur p pour un attribut a et un médicament d nécessite 4 paramètres :

— M, la taille de la population, c'est ici le cardinal de l'ensemble des réductions dr_d et continuation dc_d de dose pour le médicament d, soit $M = |dr_d \cup dc_d|$.
— m, le cardinal de l'ensemble des réductions de dose pour d soit $|dr_d|$.

— N, la taille de l'échantillon, ici le cardinal du nombre de réductions ou continuation associées avec l'attribut a soit $N = |dr^a \cup dc^a|$.
— n, le cardinal de l'ensemble des réductions de dose pour d associées à a, $|dr^a_d|$.

Alors la valeur p hypergéométrique est calculée comme la probabilité de tirer aléatoirement n ou plus intervalles dans la population M en m tirages.

Calcul du risque relatif Pour un médicament d et un attribut a, le risque relatif (RR) se calcule avec les même paramètres que la valeur p suivant la formule suivante :

$$RR(d,a) = \frac{\frac{n}{m}}{\frac{N-n}{M-m}}$$

Calcul du contenu en information Pour un attribut a, le contenu en information (IC) est

$$IC(a) = -log(\frac{k}{K})$$

où classiquement K est le nombre total de documents et k le nombre de documents associés au mot clé k. Dans notre cas, les définitions de K et k dépendent des types d'attributs phénotypiques considérés. Pour les codes diagnostics, les conditions des notes et les commandes d'examens biologiques K sont respectivement le nombre total de visites, notes ou commandes ; et k est le nombre de visites, de notes ou de commandes associées à a.

Composition des profils phénotypiques Les profils phénotypiques sont initialement composés des attributs dont la valeur p est $< 0,05$ avec le test hypergéométrique. Lors d'un premier filtre nous éliminons les attributs avec $0.5 < RR(d,a) < 2$. Puis un second filtre élimine les attributs dont l'IC est soit dans le premier, soit dans le dernier quartile. Par exemple pour les conditions mentionnées dans les notes cliniques, nous ne conservons que les conditions qui satisfont $4.25 < IC(a) < 12.75$. Le choix de ces valeurs est fait arbitrairement à partir de l'observation de la distribution des valeurs de l'IC. Un troisième filtre réalise une modification du seuil de significativité ($p < 0.05$) avec la correction de Bonferroni pour éviter les biais dus aux comparaisons multiples [Holm, 1979]. Enfin, nous appliquons la méthode *elim* pour filtrer les attributs ajoutés lors de l'expansion des annotations par fermeture transitive [Alexa *et al.*, 2006]. En effet, un souci avec l'expansion des annotations est que les concepts très généraux de l'ontologie tendent à dominer les résultats finaux. La méthode *elim* limite cet effet en éliminant les concepts lorsque l'on trouve pour une concept un concept plus spécifique associé à la variable réponse (réduction, augmentation ou continuation de dose) avec une valeur p égale ou plus basse.

Le Tableau 2.3 donne le nombre d'attributs obtenu après chaque étape de filtrage. L'ordre dans lequel sont appliqués les filtres a été choisi pour des raisons pratiques de calcul, mais ils peuvent être exécutés dans n'importe quel ordre. Les profils phénotypiques produits peuvent être explorés à l'aide d'un outil de visualisation à l'adresse `https://snowball.loria.fr/p450/`. Quatre exemples de profils sont montrés en Annexes B.2 et B.3.

2.3.4 La tâche d'apprentissage

Après ces filtrages, les profils phénotypiques sont utilisés pour constituer les ensembles d'entraînement et de test qui servent à construire et évaluer nos modèles prédictifs. En particulier, nous avons entraîné deux types de modèles de classification binaire avec l'algorithme des forêts

		Dose reduction	Dose increase	Dose continuation	Total
Intervals		50,704	60,719	176,140	287,563
Patients		22,571	25,381	56,902	69,308
Diagnostic codes	*before expansion*	1,434,606	1,623,309	4,088,965	7,146,880
	after expansion	3,687,022	4,193,204	10,811,286	18,691,512
Conditions from clinical note	*before expansion*	441,746	505,235	1,173,403	2,120,384
	after expansion	4,110,928	4,728,006	11,487,221	20,326,155
Lab test orders		6,773,097	7,906,040	17,896,716	32,575,853

FIGURE 2.3 – Ce tableau donne les nombres d'attributs phénotypiques de chaque type qui composent les profils phénotypiques après chaque étape de filtrage. Les étapes de filtrages RR et IC utilisent le risque relatif (RR) et le contenu en information (IC) associés aux attributs. La méthode *elim* permet de filtrer les attributs générés lors de l'expansion des annotations avec les ontologies [Alexa *et al.*, 2006]. *elim* ne peut pas être appliqué aux examens biologiques car ceux-ci ne sont pas encodés avec une ontologie et n'ont donc pas été étendus.

d'arbres aléatoires [Breiman, 2001] : l'un pour prédire la survenue d'une réduction de la dose plutôt qu'une continuation et l'autre type pour prédire l'augmentation de la dose plutôt que la continuation. Pour chaque type de prédiction (réduction ou augmentation) nous avons entraîné deux modèles pour chaque médicament P450 et pour chaque groupe de médicaments dont la distinction est l'ensemble d'attributs considérés. Le premier est entraîné avec des profils de 300 attributs maximum, composés des 100 premiers attributs de chaque type (diagnostic, condition, examen). Le second est entraîné avec tous les attributs statistiquement significatifs. Tous les ensembles d'entraînement sont équilibrés en sous-échantillonnant à la taille de la classe (changement ou continuation) la plus petite. Les attributs présents dans l'historique des patients sont encodés avec des 1, les attributs absents ou manquants sont encodés avec des 0. Les médicaments et groupes de médicaments avec moins de 300 instances ou avec des profils phénotypiques vides (*i.e.*, aucun attribut n'a à la fois une valeur p, un RR et un IC suffisants) ne sont pas considérés. Cette élimination réduit à 34 le nombre de médicaments P450 et à 23 le nombre de groupes de médicaments pour lesquels nous avons entraîné des modèles.

Chaque modèle est évalué suivant deux modes : une validation croisée à 10 feuillets et une validation où nous avons seulement exclu les intervalles mesurés lors de la dernière année (suivant le modèle *leave one out*) que nous gardons pour le test. Dans chaque mode d'évaluation la sélection des attributs a été réalisée sur l'ensemble d'entraînement seulement. Cela veut dire que des profils phénotypiques distincts ont été calculés pour chaque feuillet de la validation croisée. Dans le second mode de validation, les données de l'année 2014 sont exclues de l'entraînement. Celles-ci représentent respectivement 17,5%, 18,0% et 18,9% de l'ensemble des réductions, augmentations et continuations de dose (enregistrées entre 2008 et 2014). Ce second mode d'évaluation a pour vocation à mesurer la capacité de nos modèles à faire des prédictions sur des données complètement nouvelles. C'est en théorie une évaluation plus contraignante si l'on considère la non-stabilité des processus d'apprentissage [Jung and Shah, 2015]. Les détails complets des paramètres choisis pour l'entraînement des modèles et leur évaluation sont présentés dans [Coulet *et al.*, 2018].

2.3.5 Résultats

En résumé nous avons considéré 34 médicaments P450 pour lesquels nous avions au moins 300 intervalles de changement ou de continuation de dose. Nous avons construit des profils phénotypiques en comparant les attributs des patients qui avaient subi un changement de dose avec ceux des patients qui n'en avaient pas subi. Et nous avons enfin évalué la capacité de ces attributs à prédire la survenue pour des nouveaux patients d'une réduction ou d'une augmentation de dose. Avec cette méthode nous avons montré que nos modèles étaient efficaces pour prédire une réduction de dose (AUC-ROC $>$ 0,7) pour 23 médicaments sur 34 et pour 22 groupes de médicaments sur 23. En revanche, ils ne le sont pas pour prédire les augmentations de dose. Les Tableaux 2.4 et 2.5 illustrent ces résultats avec les performances obtenues lors de l'évaluation de nos modèles. En particulier, le Tableau 2.4 montre que de façon générale avec le mode d'évaluation le plus contraignant (*Hold last year out*) les performances sont légèrement inférieures mais restent décentes. Dans l'objectif de développer des modèles prédictifs pour la médecine, considérer un ensemble de données de test nouveau, comme c'est le cas avec cette évaluation *Hold last year out* est plus approprié car elle reflète les conditions réelles d'utilisation d'un tel modèle. Nous notons ici que dans cette évaluation prospective, nous perdons un certain nombre de candidats qui dans ce contexte n'ont plus suffisamment d'intervalles associés ($>$300) pour construire ensembles d'entraînement et de test ou alors n'ont plus d'attributs significativement associés au changement de dose. Malgré ce contexte plus contraignant, nous sommes capables de prédire (AUC-ROC $>$ 0,7) les réductions de dose pour 10 médicaments sur les 29 initiaux. Le Tableau 2.5 montre les performances obtenues avec les 100 meilleurs attributs de chaque catégorie indépendemment puis combinés. Les commandes d'examens biologiques sont nettement les meilleurs prédicteurs et suffisent à prédire des réductions de dose alors que les autres types d'attributs (codes diagnostics ICD et conditions mentionnées dans les notes) ne suffisent pas à prédire un changement de dose. Ces deux Tableaux montrent également les performances obtenues pour le groupe de tous les médicaments P450 retenus. Celles-ci sont assez médiocres (mesure F=0,64) illustrant que la capacité de prédiction est quelque chose d'observé pour certains médicaments, certains groupes de médicaments, mais n'est pas généralisable à tous.

Pour l'interprétation de ces résultats, nous avons développé une interface web qui permet de visualiser les profils phénotypiques de chaque médicament et chaque groupe de médicaments et leurs performances prédictives. Cet outil, disponible à `https://snowball.loria.fr/p450/`, a été utilisé par trois médecins pour proposer des éléments d'interprétation clinique de la capacité des attributs à prédire le besoin d'une réduction de dose. Des exemples d'interprétation du rôle d'attributs précis dans les résultats de prédiction sont disponibles dans la Section *Interpretation* de l'article [Coulet *et al.*, 2018].

2.3.6 Discussion et conclusion

Nous pensons que les résultats positifs obtenus dans ce travail sont riches d'enseignements. Ils démontrent qu'il est possible de prédire à partir de données phénotypiques recueillies dans les DPE avant prescription qu'une réduction de dose va avoir lieu. Et qu'ainsi il pourrait être envisagé de prescrire une dose réduite en première intention. Un aspect intéressant est que la prédiction se fait individuellement au regard de l'historique du patient. Ces résultats sont encourageants mais les résultats lèvent également plusieurs questions que nous discutons dans ce qui suit, notamment pour diriger nos prochains efforts de recherche.

Drug or drug set	10-fold cross-validation			Hold last year out		
	|instances|	AUC-ROC	F-m (P; R)	|instances|	AUC-ROC	F-m (P; R)
Labetalol	353	0.85	0.77 (0.78; 0.77)	314	0.95	**0.86** (0.87; 0.86)
Tacrolimus	709	0.94	**0.89** (0.89; 0.89)	598	0.94	**0.86** (0.86; 0.86)
Itraconazole	292	0.88	0.80 (0.80; 0.80)	460	0.86	0.85 (0.88; 0.86)
Sildenafil	419	0.88	0.80 (0.81; 0.80)	338	0.90	0.80 (0.80; 0.80)
Methadone	941	0.90	0.81 (0.82; 0.81)	866	0.82	0.74 (0.78; 0.75)
Warfarin	2851	0.82	0.74 (0.74; 0.74)	2630	0.79	0.72 (0.72; 0.72)
Hydrocortisone	4853	0.90	0.83 (0.83; 0.83)	4288	0.80	0.71 (0.72; 0.71)
H	17302	0.86	0.78 (0.78; 0.78)	15366	0.76	0.70 (0.70; 0.70)
2C9	21607	0.79	0.71 (0.71; 0.71)	19212	0.73	0.70 (0.70; 0.70)
L	12119	0.93	0.88 (0.88; 0.88)	10354	—	—
All P450-drugs	91267	0.70	0.64 (0.64; 0.64)	80614	—	—

FIGURE 2.4 – Résultats de l'évaluation de la prédiction des réductions de doses à partir des profils phénotypiques. Deux modes d'évaluation ont été testés : une validation croisée à 10 feuillets et une validation où seules les données de la dernière année de données ont été retirées pour constituer l'ensemble de test (*hold last year out*). Ici seuls les 100 premiers codes diagnostics, conditions mentionnées dans les notes et examens biologiques sont conservés dans les profils phénotypiques. Seuls les médicaments et groupes de médicaments avec une mesure F (F-m) 0,7 avec le second mode d'évaluation sont rapportés ici, plus les résultats de l'ensemble des médicaments P450 et de la classe ATC L. L fait référence à la classes *Antineoplastic and immunomodulating agents* et H est *Systemic hormonal preparations, excluding sex hormones and insulins*. Les résultats de classe L sont les meilleurs obtenus en validation croisée, cependant ils ne sont pas calculables avec le mode *hold last year out* à cause de profils phénotypiques vides pour la dernière année (2014). |instances| correspond au nombre d'instances dans l'ensemble d'entraînement. Les précisions (P) et rappels (R) sont données entre parenthèse après la mesure F. Les résultats complets sont disponibles dans les suppléments de [Coulet *et al.*, 2018].

Drug or drug set	F-measure			
	Top 100 diagnostics	Top 100 conditions	Top 100 labs	Top 300 features
Labetalol	0.37	0.41	0.81	0.77
Tacrolimus	0.41	0.43	0.90	0.89
Itraconazole	0.36	0.49	0.79	0.80
Sildenafil	0.39	0.42	0.78	0.80
Methadone	0.38	0.43	0.82	0.81
Warfarin	0.38	0.45	0.76	0.74
Hydrocortisone	0.40	0.49	0.82	0.83
H	0.37	0.48	0.78	0.78
2C9	0.36	0.42	0.71	0.71
L	0.41	0.33	0.87	0.88
All P450-drugs	0.36	0.45	0.64	0.64

FIGURE 2.5 – Performances de nos modèles prédictifs par types d'attributs phénotypiques. Les 100 premiers attributs de chaque type sont obtenus suivant les valeurs p associées aux attributs. Les performances sont celles de la validation croisée à 10 feuillets. Les 300 attributs sont la combinaison des trois 100 premiers attributs de chaque type.

Association entre réduction de la dose et EIM L'hypothèse de départ de ce travail est qu'un changement de dose est un marqueur d'une réponse inappropriée au traitement comme une EIM ou une absence de réponse. Cependant, les réductions de dose font partie intégrante du protocole de prescription de certains médicaments. Dans ce cas, prédire la réduction de dose n'est pas intéressant pour la pratique clinique. Dans notre étude, les glycocorticoïdes comme l'hydrocortisone et le dexamethasone ont des protocoles qui recommandent une réduction de dose indépendemment de la réponse au traitement. Nos résultats pour ces médicaments ne sont pas directement pertinents. Pour évaluer l'importance du nombre de médicaments pour lesquels il est normal de réduire la dose, nous avons passé en revue manuellement les protocoles de prescriptions des notices de médicaments associées aux 23 médicaments pour lesquels la prédiction fonctionne. Pour 14 médicaments sur les 23 une réduction de dose est recommandée uniquement en cas de réponse indésirable au traitement. Les protocoles de certains médicaments comme le tacrolimus ou la warfarine mentionnent le besoin d'ajuster la dose (dans les deux sens possibles) selon la réaction du patient. Dans ce cas nos modèles peuvent aider à construire un système d'aide à la prescription qui peut proposer au regard de l'historique du patient une dose réduite dès l'entame du traitement. Ceci pourrait permettre de réduire le temps pour atteindre une dose stable de traitement et ainsi réduire le nombre d'EIM.

Limites Une limite de notre approche et que nous ne considérons pas le cas où la survenue d'un EIM est pris en charge non pas par une réduction de la dose, mais par un arrêt du médicament et éventuellement un changement de molécule. Considérer ces alternatives permettrait de développer des modèles plus complets et plus performants. Une limite difficile à expliquer est l'échec de l'approche à prédire les augmentations de dose. En premier abord nous pourrions penser que cela peut s'expliquer par le fait que l'"absence de réponse" n'est pas toujours renseignée dans les DPE, notamment car c'est une absence plutôt qu'une observation en tant que telle. Capturer cela n'est certainement pas possible avec les attributs phénotypiques simples que nous avons considérés. Cependant, il faut rappeler que la prédiction se fait sur les attributs observés avant la prescription, et donc avant la réponse ou l'absence de réponse. Nous pensons que prédire les augmentations de doses serait possible, mais que cela nécessiterait une modélisation différente et plus précise des attributs considérés. Une amélioration relativement simple pourrait être la considération de plus d'attributs : non seulement les commandes d'examens biologiques, mais également leur résultat, les médicaments prescrits, etc. Aussi les méthodes d'apprentissage profond comme les réseaux de neurones récurrents seraient d'ailleurs assez adaptées à la prise en considération d'un grand nombre d'attributs et devraient améliorer les performances de nos modèles.

Conclusion En conclusion nous avons démontré dans ce travail que l'historique phénotypique des patients disponible dans les DPE peut être utilisé pour identifier quels patients risquent de subir une réduction de dose. Un perspective intéressante serait de développer un système d'aide à la prescription individualisée combinant tests génétiques et historique phénotypique des patients. La présence de données génétiques permettrait par exemple de valider et évaluer les performances de modèles basés sur le phénotype. En effet, plusieurs projets comme PREDICT à l'Université Vanderbuilt ou RIGHT à la Mayo Clinic [Roden *et al.*, 2018, Volpi *et al.*, 2018, Denny *et al.*, 2017] ont démontré avec succès que les marqueurs génétiques de la variabilité de la réponse aux médicaments sont utiles pour personnaliser les traitement et réduire les EIM. Bien qu'utiles ces données génétiques sont le plus souvent absentes ou difficiles à obtenir en raison du coût, du délai ou de l'aspect non routinier de la procédure [Ioannidis, 2013]. Dans ce cas particulier, les modèles prédictifs basés seulement sur le phénotype pourraient être une

alternative viable aux tests pharmacogénomiques.

2.4 Discussion générale

Le besoin d'historique, de continuité et de portabilité Une limite à l'utilisabilité des modèles prédictifs comme celui que nous avons présenté ici est la nécessité de disposer de l'historique du patient si l'on souhaite faire une prédiction. C'est une limite dans le sens où cela exclut la considération de nouveaux patients, qui arriveraient pour la première fois dans une structure de soins et pour qui le DPE serait vierge. Cela milite pour le développement de DPE transversaux, qui suivent les patients tout au long de leur vie, même si cela est à mettre en perspective avec les questions de propriété, d'hébergement, de confidentialité des données que cela pose. Les données que nous utilisons sont collectées uniquement lors de séjours hospitaliers, des fenêtres de temps riches en données, mais entre lesquelles nous n'avons aucune donnée. Dans cette mesure, les données issues des systèmes d'assurance santé sont moins détaillées mais moins discontinues. Pour revenir au besoin d'historique, dans le cas où un individu change d'hôpital, d'assurance ou de pays, nous pourrions imaginer que le modèle prédictif puisse considérer un historique constitué ailleurs à la condition qu'il existe un standard d'échange des données de santé. C'est le cas et nous pouvons imaginer que des initiatives comme les modèles FHIR (Fast Healthcare Interoperability Resources) [Mandel *et al.*, 2016] et OMOP CDM (Observational Medical Outcomes Partnership Common Data Model) [Voss *et al.*, 2015] permettent de tels échanges. Un avantage déjà établi de ces standards d'échange est leur utilisation pour démontrer la portabilité des systèmes prédictifs en santé [Hripcsak *et al.*, 2015]. En effet les modèles entraînés sur les données d'un établissement ou d'une compagnie d'assurance marchent rarement aussi bien dans un contexte différent, mais ces standards permettent à des établissements partenaires d'entraîner indépendemment le même modèle chacun sur leurs données pour réaliser des méta-analyses.

Utilisabilité : Explicabilité et équité Les bonnes performances que permettent l'augmentation du volume de données de santé récoltées et le développement des méthodes d'apprentissage profond laissent parfois dire qu'avec les données adaptées (en terme de variables, de qualité et de volume) il serait possible de prédire n'importe quelle autre variable. Cette hypothèse n'empêche pas qu'il puisse être utile de développer et évaluer des modèles, mais présente l'intérêt de placer le curseur de la faisabilité de la prédiction sur le problème suivant : l'utilité des modèles [Shah *et al.*, 2019]. Si une variable peut-être prédite, est-il véritablement utile de le faire ? Est ce que le système associé sera adopté par les utilisateurs [Bates *et al.*, 2003] ? Ce sont des questions particulièrement intéressantes dans le domaine médical où les médecins exercent une activité experte dans un environnement contraint. Pour certaines tâches comme le diagnostic de maladies communes, ils n'ont pas besoin d'assistance, pour d'autres ils pourraient en bénéficier mais ne l'adopteront pas forcement si le système est trop contraignant. Nous pensons que pour que des systèmes prédictifs puissent être adoptés en santé ils doivent au moins satisfaire deux conditions : être *explicables*, *i.e.*, être capables de fournir des éléments de connaissances interprétables qui expliquent la décision qu'ils proposent [Ribeiro *et al.*, 2016] ; être équitables, *i.e.*, assurer que les performances du modèle sont constantes, ou au moins au dessus d'un seuil minimal pour tous les individus et en particulier pour ceux appartenant à des sous-groupes potentiellement sous-représentés dans les données ou considérés comme fragiles [Barocas *et al.*, 2019].

Avantages et limites des études cliniques pour la découverte de connaissances L'établissement de connaissances biomédicales, et particulièrement celles qui concernent le médicament, sont classiquement établies par des études cliniques randomisées. Les contraintes imposées par ce genre d'études font qu'il est possible de définir un cadre statistique particulier à partir duquel des connaissances comme l'*effet moyen d'un traitement* sont inférées [Rubin, 2005].

Parmi ces contraintes citons les critères d'inclusion des patients et la sélection aléatoire des cas et contrôles qui garantissent respectivement une certaine similarité entre patients et une indépendance entre la probabilité de recevoir le traitement et celle d'avoir la réponse attendue au traitement. La contrepartie de ces contraintes est qu'il est coûteux de recruter un ensemble d'individus de taille suffisante et de les suivre pour réaliser une étude clinique randomisée. Le coût élevé fait que les études sont réalisées une fois et n'évaluent que l'effet moyen recherché du médicament. Les effets "collatéraux" (indésirables comme bénéfiques [Jung *et al.*, 2014]) d'un médicament et les effets spécifiques à des groupes minoritaires ne sont généralement pas recherchés. La disponibilité des données observationnelles, c'est-à-dire des données recueillies lors de l'activité de soins hors d'un objectif d'étude clinique comme celles que nous avons utilisées, constitue une opportunité intéressante au regard des pans de connaissances délaissés par les études cliniques randomisées [Gombar *et al.*, 2019]. Un défi que nous abordons dans nos perspectives de recherche est l'étude des méthodes qui permettent d'adapter l'analyse des données observationnelles pour les rapprocher le plus possible des études cliniques randomisées [Rubin, 2005]. Pour illustrer cette adaptation, les deux critères mentionnés précédemment imposent des adaptations. En effet, dans le cadre d'études observationnelles les patients ne sont pas initialement sélectionnés sur la base de critères d'inclusion et le traitement ne leur est pas attribué au hasard. Il s'agit donc de définir une cohorte et de choisir une méthode d'analyse adaptée à ce nouveau contexte.

Chapitre 3

Découverte de connaissances à partir de graphes de connaissances

Sommaire

3.1 Introduction 63
3.2 Prédiction de liens dans un graphe de connaissances 65
3.2.1 PGxLOD : un graphe de connaissances pharmacogénomiques 65
3.2.2 Prédiction de liens Gène-Médicament 70
3.2.3 Prédiction de l'identité, ou à défaut d'une similarité 75
3.3 Recherche de régularités dans un graphe de connaissances 82
3.3.1 Apprentissage de descriptions des gènes responsables des DI 82
3.3.2 Complétion d'ontologies à partir des données du LOD 86
3.3.3 Discussion et conclusion 90
3.4 Discussion générale 91

3.1 Introduction

La notion de *graphe de connaissances* est une appellation générique pour une base de connaissances représentée sous la forme d'un graphe, sans référence à un standard ou un formalisme particulier [Ehrlinger and Wöß, 2016, McCusker *et al.*, 2018]. Le terme "connaissances" implique cependant que les nœuds et les arcs de ce type de graphes sont associés à une représentation de connaissances, sans imposer de formalisme. Dans ce mémoire nous nous intéressons au type particulier de graphes de connaissances qui implémentent les principes du Web sémantique et que nous appellerons avec les anglicismes *Linked Data* (LD) ou *Linked Open Data* (LOD), quand leur contenu est libre [Bizer *et al.*, 2009]. Ces graphes de connaissances nous intéressent particulièrement car ils proposent un cadre où des données de sources diverses et les connaissances associées sont regroupées au sein d'une même structure. Ils constituent ainsi une opportunité unique pour étudier dans quelle mesure les connaissances peuvent aider le processus de découverte de connaissances. Deux autres avantages particuliers aux LD sont d'être représentées suivant un format standard qui inclut l'utilisation d'URI (*Unified Resources Identifier*) pour référer de façon unique aux éléments du graphe et l'utilisation du langage RDF (*Resource Description Framework*) pour encoder le graphe sous la forme de triplets de la forme *(sujet, prédicat, objet)*. Dans le cas des LD, les connaissances associées sont des ontologies RDF ou OWL qui sont connectées aux données et font souvent partie du graphe, mais pas systématiquement. En effet, pour les ontologies

les plus expressives les axiomes de logiques de description ne sont en général pas représentés en triplets. En revanche, les annotations dont nous avons discuté la représentation formelle au début du Chapitre 1 participent souvent à la composition de ce type de graphes et cela que l'on choisisse de les représenter comme des instanciations de concepts ou comme des propriétés. Une dimension qui est systématiquement associée aux LD est l'idée que ce type de graphes de connaissances est constitué de sources différentes, reliées entre elles (d'où l'utilisation du terme *Linked*). La publication de façon ouverte de ces données et leur inter-connexion dans ce que nous appelons alors le LOD fait partie d'un effort communautaire pour la construction du Web sémantique [Bizer *et al.*, 2009], où les ressources du Web sont interprétables non seulement par les humains mais également par les machines [Berners-Lee *et al.*, 2001]. Dans cette optique les LOD connectant des données entre-elles constituent une opportunité unique pour l'intégration de données et la découverte de connaissances.

Cela est particulièrement le cas dans le domaine des sciences de la vie où de nombreuses données sont ouvertes mais publiées hors standard et sans concertation [Antezana *et al.*, 2009]. Par conséquence dès lors que quelqu'un souhaite utiliser des données de sources multiples pour un projet, il lui faut faire face à des questions d'intégration de données. Plusieurs initiatives comme Bio2RDF [Callahan *et al.*, 2013], EBI RDF platform [Jupp *et al.*, 2014], Linked Open Drug Data (LODD) [Samwald *et al.*, 2011] ou PDBj [Kinjo *et al.*, 2012] se sont donné l'objectif de regrouper des sources biomédicales diverses suivant les standards du LOD afin de faciliter leur utilisation de façon intégrée. Il résulte de ces initiatives une grande collection de données biomédicales disponibles dans un format standard et même si chacune a des défauts, elles offrent un terrain d'expérimentation nouveau à la fouille de données.

Les contributions que nous décrivons dans ce chapitre sont des travaux de fouille appliqués aux LOD où données et parfois connaissances sont utilisées conjointement. En particulier nous décrivons des travaux de prédiction de liens dans les LOD (Section 3.2) qui visent à compléter les liens entre données liées ; et des travaux de recherche de régularités au sein du LOD (Section 3.3) qui visent à obtenir des descriptions de concepts qui peuvent être utilisées pour compléter les connaissances associées aux données.

3.2 Prédiction de liens dans un graphe de connaissances

La prédiction de liens est une tâche classique dans le domaine de la fouille de graphe mais aussi du Web sémantique. Il s'agit, étant donné un graphe, de compléter celui-ci avec de nouveaux arcs valides. Les deux contributions décrites dans cette section concernent la prédiction de liens dans un graphe de connaissances et ont été appliquées au même graphe de connaissances. Il s'agit d'un graphe pour le domaine de la pharmacogénomique que nous avons construit et appelé PGxLOD. Nous commencerons par décrire PGxLOD et sa construction dans la section suivante. Puis, nous décrirons la première contribution dont l'objectif est la mise en évidence par apprentissage supervisé des liens entre gènes et médicaments dans ce graphe de connaissances [Dalleau *et al.*, 2017]. La contribution suivante concerne la prédiction d'un type de liens particulier : la recherche de correspondances, ou à défaut de similarité, entre des objets complexes qui représentent des relations n-aires [Monnin *et al.*, 2019a]. Les premiers auteurs de ces deux contributions, Kevin Dalleau et Pierre Monnin, sont respectivement des doctorants en pharmacie et en informatique que j'ai encadré.

3.2.1 PGxLOD : un graphe de connaissances pharmacogénomiques

PGxLOD est un graphe de connaissances pharmacogénomiques assemblé à partir de données d'origines diverses. PGxLOD a été assemblé dans le cadre du projet ANR PractiKPharma dont l'objectif principal est la comparaison de connaissances dans ce domaine. PGxLOD se veut être un espace où l'on peut représenter et requêter les connaissances de l'état de l'art en pharmacogénomique, afin d'autoriser d'éventuelles comparaisons.

PGxO : une ontologie très simple pour la pharmacogénomique

Pour définir quelques éléments de structure sur les connaissances que nous souhaitons manipuler dans PGxLOD, c'est-à-dire les *relations pharmacogénomiques*, nous avons défini une ontologie appelée PGXO [Monnin *et al.*, 2017a]. Celle-ci est constituée des concepts et rôles nécessaires à la représentation de ces relations que nous considérons comme l'unité de connaissance en pharmacogénomique. PGxO s'organise autour d'un concept central appelé `PharmacogenomicRelationship` qui est en relation avec dix autres concepts comme illustré Figure 3.1. L'objectif de cette ontologie très simple n'est pas de contraindre les données pour assurer une complétude ou une cohérence, mais d'offrir quelques éléments de structure qui permettent la définition de trois composants autour des relations pharmacogénomiques, et par là facilite leur comparaison.

Nous considérons qu'une relation pharmacogénomique est composée d'au moins deux des trois composants suivants : un médicament (ou plus), un facteur génétique (ou plus), un phénotype (ou plus). Pour permettre plus de flexibilité le composant médicament peut être un médicament ou bien un phénotype qui dépend d'un médicament comme *carbamazepine hypersensitivity*, et de même le composant génétique peut être un facteur génétique (gène, variant ou haplotype) ou un phénotype qui dépend d'un facteur génétique comme par exemple *VKORC1 gene expression*.

Suivant les bonnes pratiques classiques [Fernandez-Lopez and Corcho, 2010] : PGxO a été mis en correspondance avec trois ontologies de domaines à large spectre (MeSH, NCiT and SNOMED CT) et avec quatre ontologies existantes pour le domaine de la pharmacogénomique (SO-Pharm, PO, PHARE et Genomic CDS) ; PGxO est partagée librement sur le Bioportal où elle peut être visualisée et téléchargée à l'adresse `https://bioportal.bioontology.org/ontologies/PGXO/` ; PGxO est mise à jour de façon cyclique et est au 28 octobre 2019 à sa cinquième version.

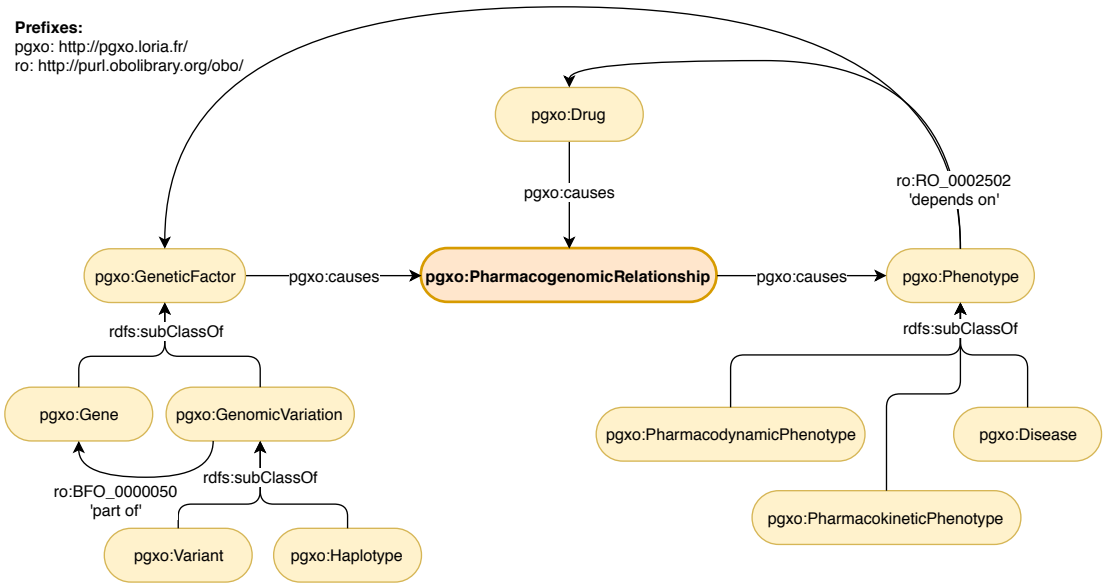


FIGURE 3.1 – Les concepts et relations de PGxO, autour du concept central `Pharmacogenomic Relationship`.

Les entités de PGxLOD

PGxLOD est peuplé d'entités issues de sources de données diverses originellement en RDF ou pas. Nous avons choisi ces sources de façon arbitraires, mais avons orienté nos choix vers des sources qui proposent des relations entre deux des 3 entités principales de la PGx, *i.e.*, les gènes (et leurs variants), les médicaments et les réponses aux médicaments. De plus nous avons préféré des sources dont les relations sont typées. Nous nous sommes restreints à deux sources par types de relations, ainsi nous avons sélectionné :

- ClinVar [Landrum *et al.*, 2014] et DisGeNET [Piñero *et al.*, 2015] pour les relations gène–phénotype ;
- SIDER [Kuhn *et al.*, 2016] et MediSpan (une base de données propriétaire sur le médicament) pour les relations phénotype–médicament ;
- DrugBank [Wishart *et al.*, 2008] et UniProt [Consortium, 2018] pour les relations gène–médicament [7].

Le contenu de ces bases a été transformé en RDF lorsqu'il ne l'était pas puis vient instancier les concepts de PGxO associés. Le Tableau 3.1 montre le nombre d'entités issues de chaque source et le concept qu'elles instancient dans PGxLOD.

A ce stade PGxLOD ne contient aucune relation PGx à proprement parler, mais les entités qui les composent, ainsi que des correspondances entre les entités retrouvées dans différentes sources de données. Pour définir ces correspondances nous avons utilisé les identifiants standards tels que le NCBI Gene ID ou les CUI de l'UMLS retrouvés dans plusieurs sources et des références croisées proposées par des services tiers comme `biodb.jp` [Imanishi and Nakaoka, 2009] et RxNav [Zeng *et al.*, 2007].

Nous appelons dans la suite du mémoire le LOD résultant de l'agrégation de ces entités pharmacogénomiques PGxLOD $v1$. Ceci nous permet de le distinguer de sa version enrichie de

7. Uniprot ne contient pas de relations gène-médicament mais les références croisées nécessaires à mettre en correspondance les gènes de DrugBank avec d'autres ressources.

TABLE 3.1 – Les entités de PGxLOD $v1$: leur type, leur nombre et leur source d'origine.

Source	Genes	Variants	Drugs	Diseases	Phenotypes
ClinVar	21 487	103 219	0	0	6 837
DisGeNET	85 893	49 279	0	38 727	6 092
DrugBank	4 300	0	7 740	0	0
MediSpan	0	0	5 820	2 481	0
SIDER	0	0	25 479	6 291	0
UniProt	25 456	0	0	0	0
Total	137 136	152 498	39 039	47 499	12 929

TABLE 3.2 – Main statistics of PGxLOD v2

PGxO Concept	Number of instances
`Drug`	51 459
`GeneticFactor`	386 802
`Gene`	172 881
`GenomicVariation`	213 911
`Haplotype`	33
`Variant`	204 875
`Phenotype`	88 247
`Disease`	47 573
`PharmacodynamicPhenotype`	63
`PharmacokineticPhenotype`	44
`PharmacogenomicRelationship`	68 431
from PharmGKB	2 701
from the literature	65 720
from EHR studies	10

relations PGx de provenance diverses, que nous appelons PGxLOD $v2$. La distinction est importante car la première contribution décrite plus loin utilise PGxLOD $v1$ et la seconde PGxLOD $v2$.

Les relations pharmacogénomiques de PGxLOD

Dans un deuxième temps nous avons instancié PGxLOD et plus particulièrement son concept `PharmacogenomicRelationship` avec des relations pharmacogénomiques de trois origines différentes : la base de données de référence PharmGKB, la littérature scientifique, des Dossiers Patients Électroniques (DPE). L'essentiel de ces relations se font entre les entités de PGxLOD $v1$, mais certaines nécessitent l'ajout de nouvelles entités. Le nombre d'entités et de types d'entités a pour cette raison augmenté dans PGxLOD $v2$ comme l'illustre le Tableau 3.2. Il montre également le nombre et l'origine des relations pharmacogénomiques ajoutées. Les trois paragraphes suivants présentent comment ces relations ont été extraites de leur source d'origine pour instancier PGxLOD.

Depuis PharmGKB Une des trois sources de relations pharmacogénomiques est PharmGKB [Whirl-Carrillo *et al.*, 2012], la base de données de référence pour la pharmacogénomique. Celle-

ci est construite autour d'*annotations cliniques* qui décrivent des relations pharmacogénomiques entre des gènes (ou leurs variants), des médicaments et les réponses aux médicaments. Ces annotations sont produites par les *curateurs* de PharmGKB après considération de la littérature biomédicale et de recommandations des agences de santé comme la U.S. Food and Drug Administration (FDA). Nous avons écrit des scripts d'extraction de données spécialement pour PharmGKB. La Figure 3.2 donne un exemple de relation extraite de PharmGKB et ajoutée à PGxLOD. Elle représente une relation entre l'haplotype TPMT*3C et le médicament *azathioprine*. Cette relation est associée à des meta-données de provenance à l'aide des concepts de l'ontologie PROV-O [Lebo *et al.*, 2013]. Ces meta-données sont par exemple la version de PharmGKB et la version du script d'extraction utilisées. Cela permet notamment la coexistence dans le LOD de relations extraites de diverses versions de PharmGKB ou avec différents scripts.

Depuis la littérature biomédicale Des relations pharmacogénomiques sont issues d'une extraction automatique de relations (et de leur provenance) à partir de la littérature scientifique. Pour cela nous avons constitué un corpus de 307 phrases issues de résumés de PubMed, chacune manuellement annotée par 3 annotateurs différents, avec les entités et les relations pharmacogénomiques. Un exemple de phrase annotée est représenté Figure 3.3. À partir de ce corpus nous avons entraîné deux modèles d'apprentissage supervisé, le premier pour reconnaître les entités nommées dans le texte et le second pour extraire les relations entre ces entités. Le détail de ces modèles est donné dans [Monnin *et al.*, 2019a]. Le corpus et les modèles sont des versions naïves de ceux construits pour PGxCorpus [Legrand *et al.*, 2019]. Les modèles ont été appliqués sur un corpus non annoté de 176 397 phrases de résumés PubMed pour en extraire des relations pharmacogénomiques, auxquelles nous avons ajouté les relations annotées manuellement dans le corpus de départ. Cela a permis d'extraire 65 720 relations. Le Tableau 3.3 donne les nombres d'entités reconnues que nous avons pu mettre en correspondance avec des ontologies ou bases de données de référence. Les nombres de l'avant dernière ligne (PGxLOD) correspondent aux nombre d'entités qui n'ont pas pu être mises en correspondance avec d'autres ressources. Alors un identifiant local est créé.

Table 3.3 – Nombre d'entités uniques reconnues dans notre corpus de test et mises en correspondance avec les bases de données et ontologies de référence.

Base de données / Ontologie	Drug	Gene	GenomicVariation	Phenotype
MeSH	1 600	n/a	n/a	1 625
ChEBI	285	n/a	n/a	n/a
ATC	78	n/a	n/a	n/a
NCBI Gene	n/a	4 907	n/a	n/a
dbSNP	n/a	n/a	803	n/a
MEDDRA	n/a	n/a	n/a	0
PGxLOD	6 449	5 905	7 937	22 335
Total	8 412	10 812	8 740	23 960

A partir d'études sur des dossiers patients électroniques Dans une optique exploratoire, nous avons également extrait manuellement 10 relations PGx à partir de la lecture de dix études faites sur des DPE et des banques d'échantillons biologiques associées. Par exemple

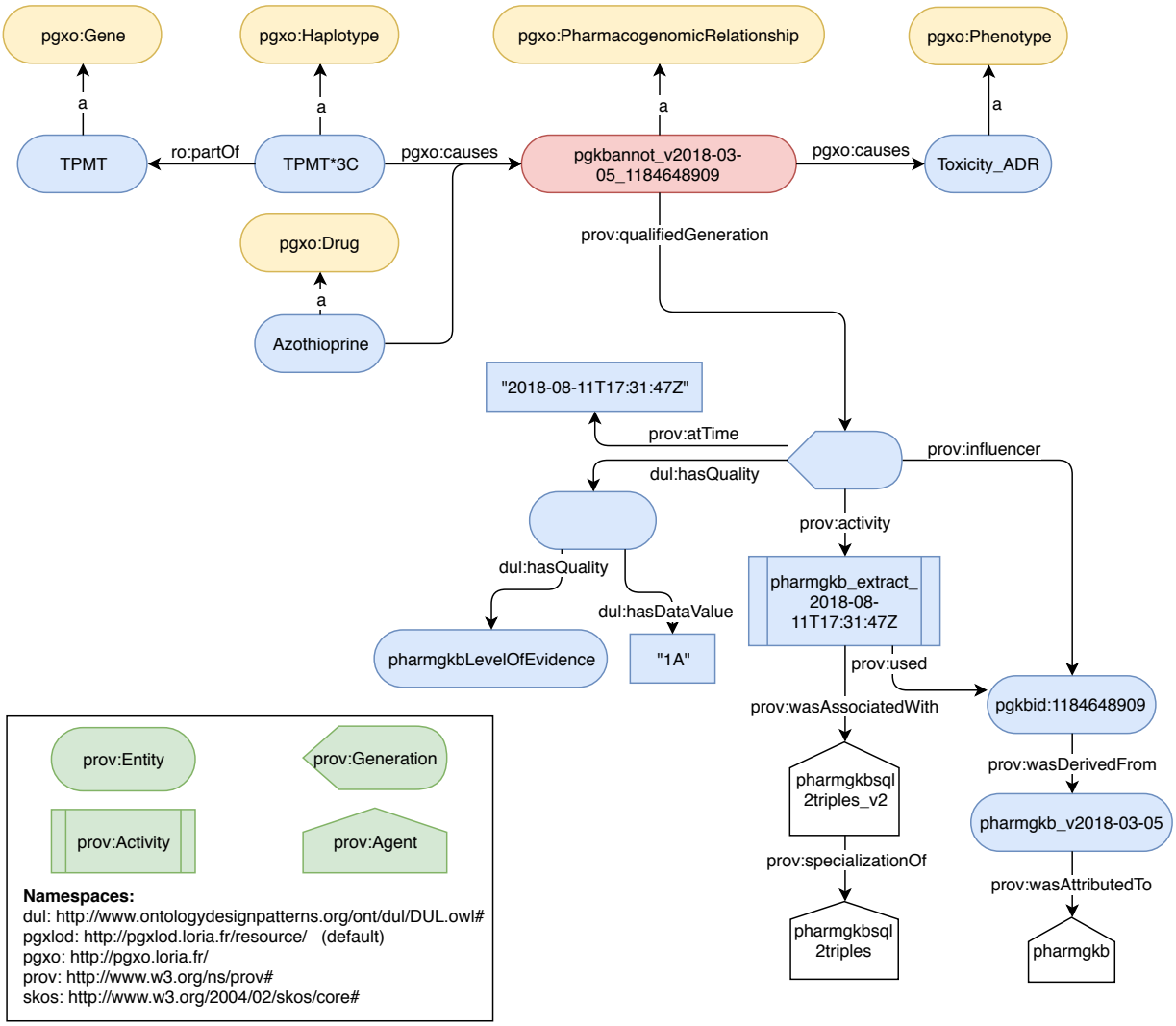


FIGURE 3.2 – Une relation pharmacogénomique extraite de PharmGKB le 11 août 2018 et son instanciation des concepts de PGxO. Pour faciliter la lecture nous avons parfois utilisé les labels à la place des URI. Un seul médicament et un seul variant sont représentés, alors que la relation implique en réalité plus de composants. L'annotation clinique d'origine est disponible à `https://www.pharmgkb.org/gene/PA356/clinicalAnnotation/1184648909`. pharmgkb2triples est le nom de notre script d'extraction. v2018-03-05 est une référence à la version de PharmGKB utilisée.

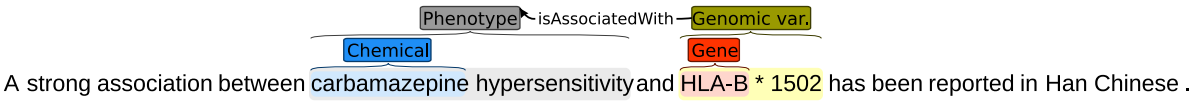


FIGURE 3.3 – Exemple de phrase (PMID=18370849) manuellement annotée avec quatre entités et une relation pharmacogénomique.

[Kawai *et al.*, 2014] rapporte une association (Odd ratio = 2,05, 95%) entre l'haplotype CYP2C9 *3 et des saignements sévères chez les patients traités avec la warfarine. Cette association a été mise en évidence à partir des DPE et de la banque d'échantillons biologiques de l'hôpital de l'Université Vanderbuilt. Il s'agit donc ici de formaliser manuellement les connaissances résultant de leur étude et de les intégrer à PGxLOD. Chaque étude considérée a permis d'isoler une relation pharmacogénomique et de l'intégrer à PGxLOD.

Il est important de noter que l'objectif des extractions à partir de la littérature et des DPE n'est pas la qualité ou la quantité des relations extraites, mais la faisabilité de la mise en correspondance des connaissances d'origine diverses. Dans le cas des relations extraites de la littérature, leur qualité est moyenne (précision=0,67) et dans le cas des relations extraites des DPE elles sont peu nombreuses à ce stade, à cause de la difficulté à automatiser cette extraction.

3.2.2 Prédiction de liens Gène-Médicament

La contribution [Dalleau *et al.*, 2017] présentée dans cette section s'intéresse à évaluer la capacité des données assemblées dans PGxLOD à identifier des pharmacogènes, c'est-à-dire des gènes impliqués dans les réponses aux médicaments. Ceci revient à une tâche de prédiction de liens gène-médicament où le lien décrit le fait qu'il existe un impact du gène sur la réponse au médicament. Nous avons étudié cette hypothèse sur PGxLOD $v1$, décrit dans la section précédente. La particularité de cette version est qu'elle ne contient pas de relations pharmacogénomiques à proprement parler, uniquement des entités impliquées dans ces types de relations et les propriétés de ces entités.

La tâche d'apprentissage et les ensembles d'entraînement et de test

Pour évaluer la capacité des données à prédire des liens nous définissons une tâche supervisée de classification des nœuds d'un graphe. Les nœuds considérés pour la classification représentent chacun une *paire candidate* composée d'un gène et d'un médicaments (deux autres nœuds). Chaque paire candidate est à classer de façon binaire comme étant une paire pharmacologiquement associée ou non.

Pour constituer notre ensemble d'entraînement nous définissons deux classes de nœuds paires particuliers : les *positifs* et les *négatifs*. Nos exemples positifs sont les paires gène–médicament associées dans la base de référence PharmGKB avec un *niveau d'évidence* 1 ou 2. En effet, les relations de PharmGKB sont associées à un niveau d'évidence (de 1 à 4) qui distingue les relations bien validées de celles qui ne le sont pas suffisamment [PharmGKB web page, 2019]. Les relations de niveaux 1 et 2 ont été observées sur de grands échantillons ou reproduites alors que celles de niveaux 3 et 4 ont été observées de façon plus sporadique. Suivant cette définition nous avons obtenu un ensemble de 91 paires positives au regard de la version du 1/10/2015 de PharmGKB. Pour constituer un ensemble d'exemples négatifs, nous tirons au hasard une paire

gène-médicament à partir de l'ensemble des paires possibles, mais nous nous assurons que cette paire est absente de DGIdb (the *Drug Gene Interaction database*), une base qui regroupe les relations gène–médicament de plusieurs sources [Wagner *et al.*, 2016], notamment PharmGKB. Le fait que DGIdb inclut de nombreuses relations gène-médicament inférées automatiquement nous laisse penser qu'une relation a plus de chance d'être négative si elle n'est pas référencée dans DGIdb. Suivant cette approche nous avons généré deux ensembles de négatifs : l'un de 91 paires, d'une taille équilibrée vis-à-vis des positifs et l'autre de 182.

Nous considérons pour notre ensemble de test les 1760 paires gène–médicament insuffisamment validées au regard de PharmGKB, *i.e.*, les paires associées aux niveaux d'évidence 3 ou 4. Ceci veut dire que notre approche s'intéresse plus particulièrement à identifier parmi les relations supposées (mais non validées) les paires qui mériteraient d'être considérée en priorité pour de plus amples analyses.

Nous avons expérimenté cette tâche d'apprentissage avec deux méthodes d'apprentissage supervisé différentes qui nécessitent deux préparations de données différentes : les forêts d'arbres aléatoires (*RF* pour *Random Forest* en anglais) [Breiman, 2001] et les *Graph Kernels* (GK, nous garderons l'anglicisme).

Préparation des données pour les forêts d'arbres

Les LOD sont représentées sous la forme de graphes alors que la plupart des algorithmes de classification comme les RF prennent en entrée une matrice (instance × attribut). Il est donc nécessaire de formater les données de PGxLOD sous la forme d'une matrice où chaque ligne représente une instance et chaque colonne représente un attribut décrivant les instances. Pour décrire une paire gène-médicament, nous proposons d'encoder sous la forme d'attributs les chemins qui partent soit du gène, soit du médicament dans PGxLOD. Pour réduire la quantité d'attributs alors possibles nous avons simplifié les chemins possibles entre les entités principales (gènes, médicaments, phénotypes) en des chemins de longueur 1. En plus, nous avons sélectionné manuellement quelques propriétés qui qualifient les médicaments, gènes et phénotypes dans le graphe et considéré uniquement celles-ci pour faire des attributs. Malgré cette simplification du graphe, il demeure que dans la matrice une seule paire est en général décrite par plusieurs instances qui chacune décrit une combinaison possible de chemins et de propriétés. Cependant nous souhaitons une classification par paire gène-médicament alors qu'à ce stade les RF nous donneraient une classification par instance et donc plusieurs classifications pour une même paire. Pour éviter cela, nous faisons un traitement supplémentaire sur les groupes d'instances (appelés *sac d'instances*) qui qualifient une même paire [Amores, 2013]. Nous calculons la moyenne pondérée des probabilités associées aux instances qui représentent la confiance du modèle dans sa décision. Si p_i est la probabilité estimée pour le sac i, n_i est la taille du sac i, p_{ij} la probabilité associée à l'instance j du sac i et Class_{ij} la décision de classification proposée par le modèle pour l'instance j du sac B_i. La probabilité du sac i est alors

$$p_i = \frac{\sum_{j=1}^{n_i} a \times p_{ij}}{n_i}, \text{ avec}$$

$$a = \begin{cases} -1 & si \quad \text{Class}_{ij} = 0 \\ 1 & si \quad \text{Class}_{ij} = 1. \end{cases}$$

Pour chaque sac B_i, $p_i \in [-1, 1]$. Si chaque instance d'un sac est associée avec une confiance forte (une probabilité proche de 1) pour être classée comme positive alors p_i sera proche de 1. Si le sac d'instances est associé à une forte confiance pour une classification négative, p_i sera proche

de -1. p_i proche de 0 signifie que l'on ne peut pas décider entre classe positive ou négative avec une forte confiance.

Préparation des données pour les graph kernels

Les *graph kernels* (GK) présentent l'avantage de prendre en entrée directement des données sous forme de graphe. De plus, nous avons utilisé la librairie Mustard library (`https://github.com/Data2Semantics/mustard`) qui prend en entrée des graphes RDF et réduit ainsi les besoins de formatage. [de Vries and de Rooij, 2015] proposent un cadre général, implémenté dans Mustard avec un ensemble de fonctions kernels pour générer les attributs décrivant les nœuds à classer. Pour cela ils proposent le fonctionnement et le cadre suivant. Dans un premier temps, le voisinage à une certaine profondeur du nœud est extrait. Dans un second temps, des sous-structures prédéfinies sont comptées au sein de ce voisinage. Les attributs de chaque nœud sont alors le nombre de chacune des sous-structures trouvées dans le voisinage.

[de Vries and de Rooij, 2015] proposent de considérer au choix 3 types de sous-structures :

- *Sac de labels* : c'est un ensemble de labels de nœuds et arcs dans le voisinage du nœud.
- *Chemin*, associé à une longueur maximale : un chemin est un ensemble ordonné de labels de nœuds et d'arcs.
- *Sous-arbre*, associé à une profondeur maximale : un sous-arbre est un graphe acyclique enraciné au nœud de départ ou non.

En modifiant la taille du voisinage et le type de sous-structure, les vecteurs décrivant les nœuds et utilisés pour calculer leur similarité sont différents. Cette similarité est calculée par la combinaison des fonctions kernels (*i.e.*, les fonctions à noyaux en français) et d'un algorithme d'apprentissage : ici nous avons choisi une machine à vecteur support (SVM pour *Support Vector Machine* en anglais).

Résultats

Résultats des forêts d'arbres Nous avons comparé deux modèles entraînés respectivement avec un ensemble d'exemples équilibrés (91 positifs et 91 négatifs) et déséquilibrés (91 positifs et 182 négatifs). Les performances des deux modèles sont évaluées par validation croisée à 10 feuillets. Nous avons entraîné et évalué notre modèle avec la bibliothèque Weka.

Nous observons que le modèle avec les classes équilibrées sur-apprend clairement avec une mesure F= 0.996. Cela est probablement dû au fait que lors de la préparation des données les paires négatives pour lesquelles il y a beaucoup moins de données dans PGxLOD sont décrites avec beaucoup moins d'instances ce qui entraîne un déséquilibre important entre le nombre d'instances positives (108,038) et négatives (3,197) dans la matrice. Un ensemble plus grand d'exemple négatifs produit une matrice plus équilibrée et réduit le sur-apprentissage. Avec ce déséquilibre entre classes nous obtenons une mesure F= 0.729 (voir le Tableau 3.4) Avec ce dernier modèle nous avons classé les 1760 paires (représentées par 984460 instances) de notre ensemble de test.

Résultats des graph kernels Nous avons réalisé plusieurs expérimentations pour chercher les paramètres (sous-structure, taille du voisinage, etc.) les plus adaptés à nos kernels. Ces expériences sont des validations croisées à 10 feuillets répétées 10 fois avec une initialisation aléatoire différente des paramètres à chaque itération. Chaque feuillet est lui-même le lieu d'une autre validation croisée à 10 feuillets pour optimiser les paramètres du modèle d'apprentissage, *i.e.*,

TABLE 3.4 – Résultats de l'évaluation de performance de notre second modèle de forêt d'arbres (RF) ou les classes de relations gène-médicament positives et négatives sont déséquilibrées (91 pos., 182 nég.). L'évaluation est une validation croisée à 10 feuillets.

Classe	*Précision*	*Rappel*	*Mesure F*
1 *(positive)*	0,804	0,998	0,891
0 *(négative)*	0,994	0,547	0,706
Moyenne pondérée	0,728	0,735	0,729

TABLE 3.5 – Résultats de la validation croisée à 10 feuillets de notre approche GK+SVM. Les modèles sont entraînés avec 91 paires positives et 91 négatives pour le jeu *équilibré*, 91 et 182 pour le jeu *déséquilibré*. Les métriques rapportées sont la mesure F, sa moyenne pondérée selon la taille des classes et l'aire sous la courbe ROC (AUC-ROC).

	Mesure F	*Mesure F Moy*	*AUC-ROC*
Équilibré	0,770	0,761	0,840
Déséquilibré	0,746	**0,807**	0,905

du SVM. Nous avons entraîné et évalué les graph kernels avec la bibliothèque Mustard et utilisé l'implémentation C-SVM de la librairie LibSVM pour les SVM.

En résumé, nous avons observé que le type de représentation de voisinage et de sous-structures recherchées dans ce voisinage impactait seulement légèrement les performances avec un avantage pour le voisinage représenté sous forme d'arbre (vs. graphe) et des sous-structures sous forme de sous-arbres (vs. sacs de labels et chemins). Nous avons aussi observé que la profondeur du voisinage n'impactait plus les performances au delà de 4, que la contrainte racinaire impacte très négativement les performances et qu'une contrainte sur la fréquence minimum des nœuds et arcs n'impacte pas réellement les performances en terme de mesure F mais permet de réduire le temps de calcul. Le détail de ces expériences est présenté dans [Dalleau *et al.*, 2017]. Ces expériences nous ont permis de sélectionner les paramètres les plus adaptés à notre tâche d'apprentissage et à nos données.

Avec les paramètres choisis (sous-arbres, non enracinés, de profondeur maximum 4 notamment) nous avons entraîné deux modèles entraînés avec un jeu d'entraînement équilibré pour l'un et déséquilibré pour l'autre. Ici encore, le modèle est évalué avec une validation croisée à 10 feuillets répétée 10 fois avec des initialisations différentes. Pour chaque feuillet nous avons optimisé les paramètres du SVM avec une nouvelle validation croisée à 10 feuillets. Faire cette validation croisée supplémentaire, plutôt que de sélectionner simplement les paramètres du SVM de la meilleure évaluation de la phase expérimentale précédente permet d'éviter un biais de sélection du modèle qui pourrait amener à une évaluation trop optimiste des performances [Cawley and Talbot, 2010]. Le Tableau 3.5 présente les résultats de l'évaluation des deux modèles entraînés équilibré et déséquilibré. La meilleure mesure F moyenne est 0,807 pour le modèle déséquilibré. Avec ce modèle nous avons classé les 1760 instances de notre ensemble de test.

Combinaison des résultats et interprétation Nous avons récupéré les deux listes de 20 premières paires candidates classées avec la meilleure confiance par notre RF d'une part et notre

GK de l'autre, puis réalisé l'intersection de ces deux listes. La liste finale a été soumise pour interprétation à une biologiste experte en pharmacogénomique (Dr Ndeye-Coumba Ndiaye de l'Université de Lorraine) qui a pu, à l'aide de ses connaissances propres et de la littérature estimé l'intérêt de ces paires pour de plus amples investigations. La liste résultante des paires prédites par les deux approches combine à la fois des paires inédites et des paires intensivement étudiées. Quasiment toutes sont liées au traitement du cancer. La présence de paires très étudiées peut être vue comme le signe positif que notre approche réussit à retrouver ce genre d'exemple. Le fait qu'elles soient très étudiées mais tout de même dans notre ensemble de test (niveau d'évidence 3–4 selon PharmGKB) illustre le fait que leur association est très probablement encore soumise à discussion. Des éléments d'interprétation supplémentaires sur les paires candidates les plus intéressantes sont discutés dans [Dalleau *et al.*, 2017].

Discussion et conclusion

Nous avons considéré en première instance l'algorithme des RF [Dalleau *et al.*, 2015] car il avait été appliqué avec succès à la prédiction de liens médicament–médicament (*i.e.*, d'interactions médicamenteuses) à partir d'un graphe RDF simplifié [Percha *et al.*, 2012]. Ce travail nous a montré que les RF étaient difficiles à adapter à la fouille de graphe et que la préparation des données oblige à faire des choix de simplification et à privilégier des chemins dans le graphe. De plus, dans notre cas les GK atteignent de meilleures performances que les RF, ce qui est en grande partie dû au fait que la librairie Mustard permet de considérer directement des graphes RDF et que par conséquent nous avons moins simplifié les données. De plus l'expérimentation avec cette librairie nous donne des éléments d'interprétation intéressants sur les attributs (notamment avec les sous-structures les plus pertinentes et leur distance dans le graphe du nœud de départ) qui pourraient être les plus pertinents à considérer dans la fouille de nos LOD. Cependant cela est à modérer par le fait qu'il est compliqué de savoir quelles parties du graphe exactement ont la plus grande contribution et cela car les sous-structures sont encodées dans un vecteur numérique difficile à décrypter. Ceci pourrait motiver l'exploration d'autres méthodes de fouille de graphes comme la recherche de sous-graphe fréquents qui pourraient être plus informatifs sur l'importance relative des sous-structures dans la classification. En particulier des approches comme gBoost [Saigo *et al.*, 2009] qui construit progressivement des patterns, ou gSpan [Yan and Han, 2002] qui énumère les sous-graphes fréquents pourraient permettre de construire et contrôler les attributs utilisés lors de la classification.

D'un point de vue applicatif, notre choix initial de ressources peut être justement questionné. En réalité, au delà du choix des ressources qui est toujours délicat, notre choix d'utiliser le cadre théorique du Web sémantique et des standards et outils associés aux LOD n'est pas anodin. Ce choix fort est clairement motivé par le fait qu'il est aisé dans ce cadre d'ajouter ou retirer des sources dans l'idée que notre approche puisse au final tirer parti des meilleurs attributs initialement répartis entre sources. Nous pourrions d'ailleurs imaginer un travail qui évalue l'impact de l'ajout et du retrait de sources de données, donnant ainsi l'opportunité d'évaluer la complémentarité des sources.

Néanmoins, notre objectif de départ qui était d'évaluer la capacité de PGxLOD $v1$ à prédire des liens gène-médicament à partir d'un premier agrégat de données liées a bien été exploré et démontré.

3.2.3 Prédiction de l'identité, ou à défaut d'une similarité

Un type de lien particulièrement utile à prédire dans les graphes de connaissances qui lient des sources distinctes est celui qui spécifie que deux entités sont deux références à un même objet. Nous appelons ce type de lien *l'identité*. Dans les LOD, cette tâche revient à la prédiction des liens `owl:sameAs` entre les entités dupliquées. Nous nous sommes particulièrement intéressés à ce problème [Monnin *et al.*, 2019a] au sein de PGxLOD *v*2, la version qui intègre des relations pharmacogénomiques de trois origines distinctes : la base experte PharmGKB, la littérature et des DPE. C'est d'ailleurs un des objectifs de PGxLOD que d'offrir un cadre pour comparer les connaissances de ce domaine. En effet, les connaissances pharmacogénomiques de l'état de l'art ont des niveaux de validation très différents [Ioannidis, 2013], et c'est ce que les niveaux d'évidence spécifiés dans PharmGKB soulignent. Il existe en effet des connaissances très bien validées et connues car l'impact du facteur génétique est fort. C'est par exemple le cas des variations génétiques des enzymes P450 qui métabolisent de nombreux médicaments. Mais pour d'autres facteurs génétiques qui ont un impact moindre, combiné et difficile à isoler les uns des autres il est plus délicat de valider les connaissances les concernant. Il nous semblait pour cette raison important d'être capable de combiner et comparer d'une part les connaissances de l'état de l'art et d'autre part d'être capable d'instancier ces connaissances avec des cas observés puis enregistrés dans les dossiers patients [Coulet and Smaïl-Tabbone, 2016]. Ces points sont des objectifs du projet ANR PractiKPharma.

Nous présentons dans cette section une façon de comparer les connaissances pharmacogénomiques représentées dans PGxLOD. Dans ce graphe, une connaissance a la forme d'un nœud qui représente une relation n-aire entre un médicament, un facteur génétique et une réponse à un médicament (ou plusieurs de ces éléments). Nous cherchons des méthodes qui permettent de comparer ces relations n-aires pour préciser si elles font référence à la même connaissance, ou à défaut si elle font référence à des connaissances comparables. En effet il peut s'agir d'une connaissances plus générale ou d'une connaissance dans une certaine mesure similaire à la première. Un fait particulier est que ces unités de connaissances ne sont pas décrites par un label, une meta-donnée ou une propriété particulière, mais elles sont définies par l'ensemble des entités qu'elles mettent en relation. Nous avons proposé une méthode de comparaison de ces connaissances assez simple et symbolique puisqu'elle s'appuie sur des règles définies manuellement appelées *règles de réconciliation*. Nous évoquerons brièvement en perspective une seconde méthode que nous avons proposée plus récemment pour la même tâche qui, elle, est numérique et apprend du voisinage des nœuds pour les comparer avec une méthode d'apprentissage profond [Monnin *et al.*, 2019b].

Les règles de réconciliation

Nous avons défini cinq règles simples pour permettre une comparaison pair à pair des relations pharmacogénomiques représentées dans PGxLOD. Ces règles sont capables d'identifier que deux relations de provenances distinctes sont en fait des références à : une même unité de connaissance (pour la règle 1) ; une unité de connaissance plus spécifique (pour les règles 2, 3, 4) ; une unité dans une certaine mesure similaire (pour la règle 5).

Les cinq règles et leur utilisation sont décrites en Annexe C.1. Nous décrivons ici la première, les quatre suivantes sont plus complexes, mais suivent le même principe. Les règles comparent un sous-ensemble d'entités deux à deux. Dans notre cas ces entités sont les instances du concept `PharmacogenomicRelationship`. Pour cette comparaison les règles considèrent trois ensembles liés à la relation (*i.e.*, ses composants) : l'ensemble des médicaments, des fac-

teurs génétiques et des phénotypes liés. Plus formellement, en considérant `r`, une instance de `PharmacogenomicRelationship` dans la base de connaissances $\mathcal{KB}$, nous définissons les trois ensembles suivants :

Notation 1 *Soit D, l'ensemble des instances de* `Drug` *qui causent* `r`*, défini comme*

$$D = \{\mathtt{d} \mid \mathcal{KB} \vDash \mathtt{Drug(d)} \text{ and } \mathcal{KB} \vDash \mathtt{causes(d,r)}\}$$

Notation 2 *Soit G, l'ensemble des instances de* `GeneticFactor` *qui causent* `r`*, défini comme*

$$G = \{\mathtt{g} \mid \mathcal{KB} \vDash \mathtt{GeneticFactor(g)} \text{ and } \mathcal{KB} \vDash \mathtt{causes(g,r)}\}$$

Notation 3 *Soit P, l'ensemble des instances de* `Phenotype` *causés par* `r`*, défini comme*

$$P = \{\mathtt{p} \mid \mathcal{KB} \vDash \mathtt{Phenotype(p)} \text{ and } \mathcal{KB} \vDash \mathtt{causes(r,p)}\}$$

En utilisant ces notations, la comparaison de deux relations n-aires notées $\mathtt{r}_1$ et $\mathtt{r}_2$, se fait selon la comparaison de leurs composants respectifs D_1, G_1, P_1 et D_2, G_2, P_2. Ainsi la première règle de réconciliation identifie si deux relations font référence à la même connaissance et est formalisée simplement comme suit

Règle 1 $D_1 = D_2$ AND $G_1 = G_2$ AND $P_1 = P_2 \quad \Rightarrow \quad \mathtt{owl:sameAs(r_1, r_2)}$

Si deux relations associent les mêmes ensembles de médicaments, facteurs génétiques et réponses, alors il s'agit de la même relation. Alors le lien $\mathtt{owl:sameAs(r_1, r_2)}$ peut être ajouté à la base de connaissances. Par exemple si l'on considère le graphe RDF représenté dans la partie supérieure de la Figure 3.4. Nous avons :

— $D_1 = D_2 = \{\mathtt{warfarin}\}$
— $G_1 = G_2 = \{\mathtt{CYP2C9}\}$
— $P_1 = P_2 = \{\mathtt{cardiovascular_diseases_inst1}\}$

La partie gauche de la règle est vraie et le lien $\mathtt{owl:sameAs(r_1, r_2)}$ peut être ajouté à la base de connaissances. La partie inférieure de la Figure 3.4 illustre le cas où les ensembles d'instances d'un composant peuvent être considérés égaux après avoir été complétés par un mécanisme de raisonnement qui interprète le prédicat `owl:sameAs`.

Sans détailler leur définition ici, les règles 2, 3 et 4 concluent sur le fait qu'une relation est plus spécifique qu'une autre en ajoutant à la base de connaissances le lien $\mathtt{skos:broadMatch(r_1, r_2)}$. La Figure 3.5 illustre l'application de la règle 2 à des graphes RDF légèrement différents. La règle 5 conclut quant à elle en proposant un lien $\mathtt{skos:relatedMatch\ (r_1, r_2)}$ qui décrit une certaine similarité. Les définitions des autres règles de réconciliation sont en Annexe C.1.2 avec des exemples d'applications en C.2.

Application des règles de réconciliation sur PGxLOD

Nous avons exécuté nos règles de réconciliation sur PGxLOD $v2$. Comme chaque nœud relation est comparé à tous les autres, 4682733330(*i.e.*, 68430×68431) comparaisons ont été effectuées. Ces comparaisons génèrent dans PGxLOD des liens `owl:sameAs`, `skos:broadMatch` et `skos:relatedMatch` dénombrés respectivement dans les Tableau 3.6, 3.7) et 3.8. Entre autres, nous observons que 66 relations issues de PharmGKB sont identiques et génèrent 132 liens `owl:sameAs` (comme ce prédicat est symétrique nous créons explicitement un lien dans chaque

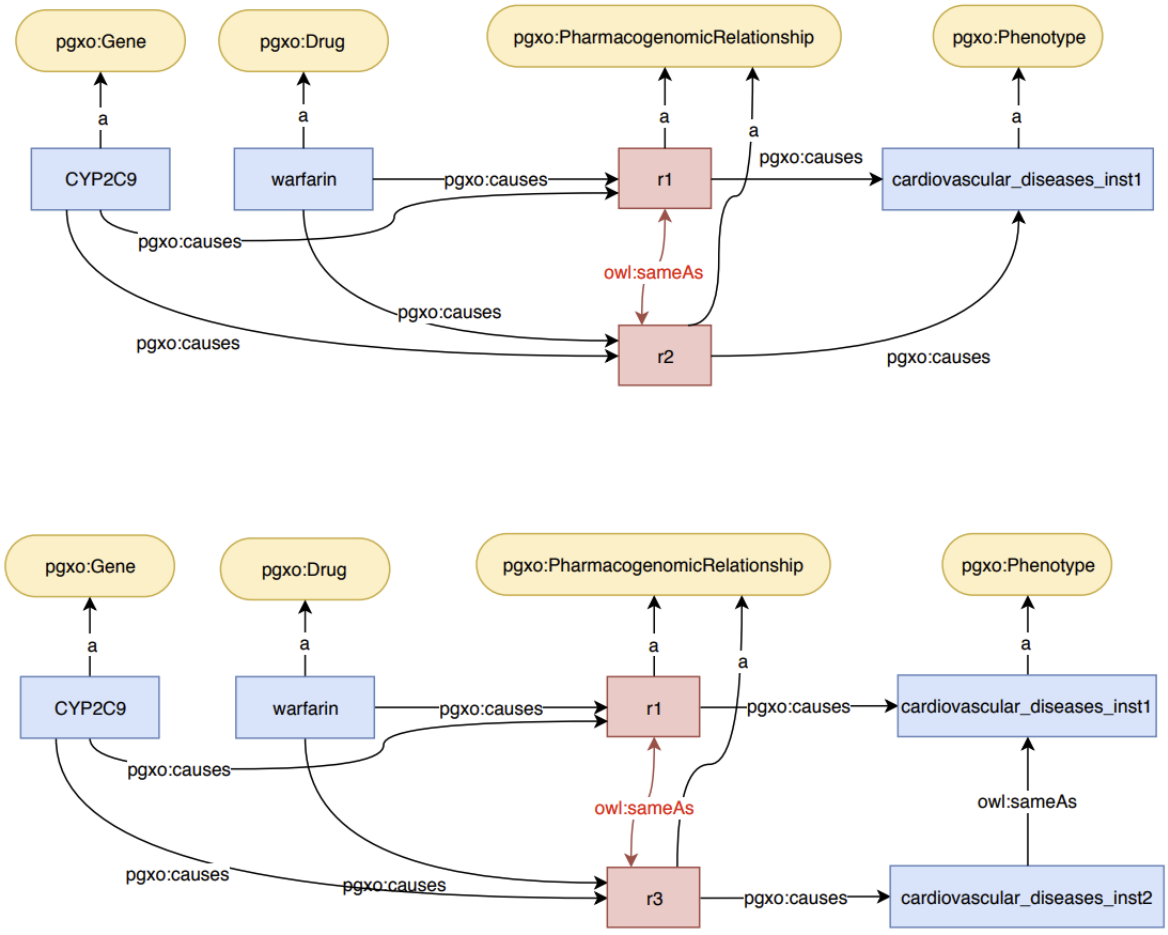


FIGURE 3.4 – Deux exemples de graphes RDF sur lesquels la première règle de réconciliation identifie deux relations pharmacogénomiques qui sont identiques. Le lien `owl:sameAs` résulte de l'application de cette règle sur le graphe.

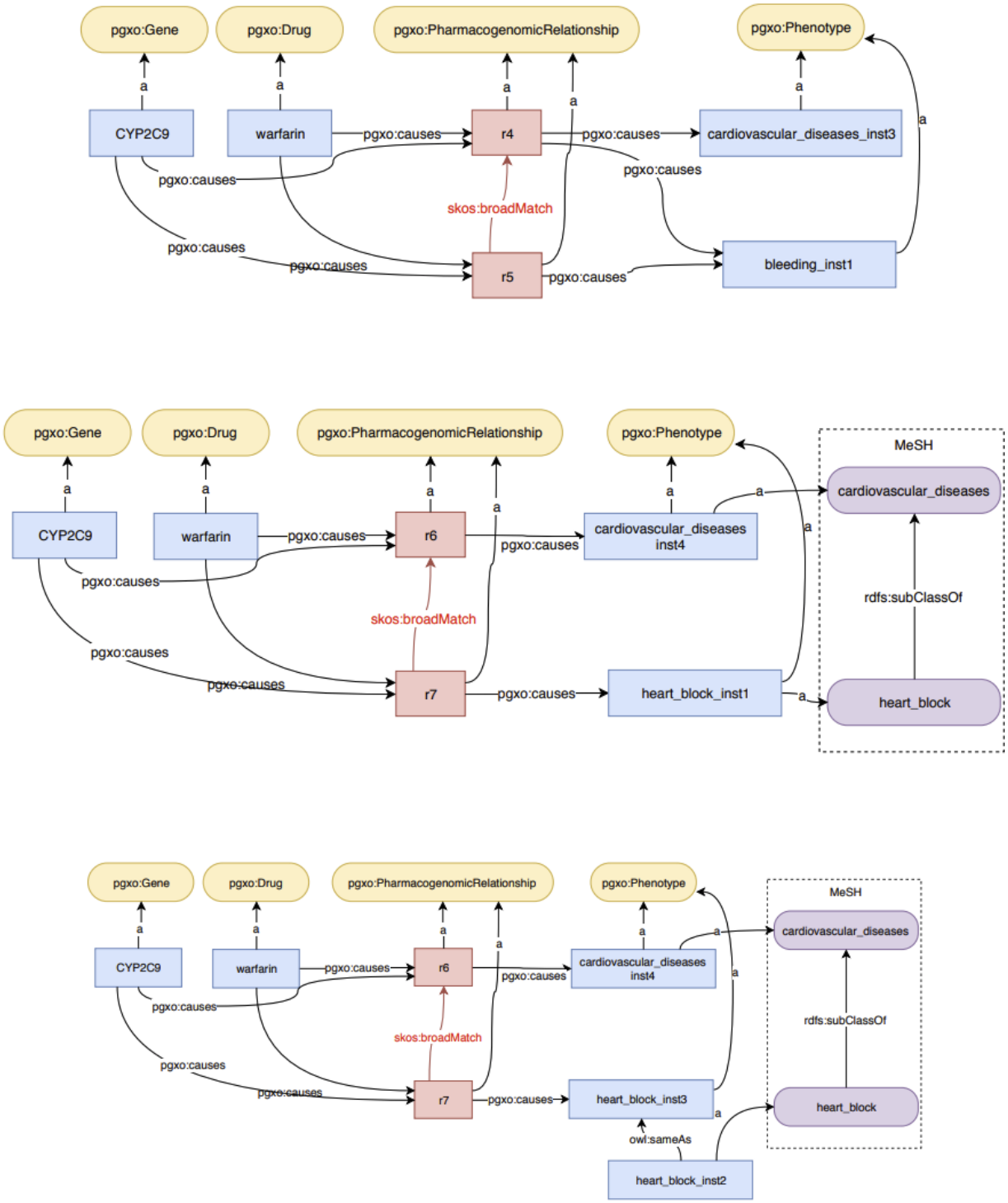


FIGURE 3.5 – Trois exemples de graphe RDF sur lesquels la règle de réconciliation 2 identifie des liens de type `skos:broaderMatch` entre des relations pharmacogénomiques.

direction, donc deux par paire de relations identiques). Aussi, 14 relations extraites de la littératures sont identifiées comme plus génériques que ce qui est rapporté dans les études sur les DPE que nous avons considérés. Nous observons également que les liens `skos:broadMatch` sont trouvés parfois entre des sources distinctes alors que les liens `owl:sameAs` et `skos:relatedMatch` sont trouvés uniquement au sein d'une même source dans cette expérimentation. Une explication possible est qu'il existe peu de correspondances entre les vocabulaires utilisés dans les différentes sources. C'est le cas des phénotypes extraits de PharmGKB qui sont représentés très gros grain et sans vocabulaire contrôlé (*e.g.*, *Toxicity/ADR*, *Efficacy*) et par conséquent sont difficiles à comparer à ceux extraits des autres sources.

TABLE 3.6 – Nombre de liens `owl:sameAs` entre les relations pharmacogénomiques d'origine distincte.

	DPE	Littérature	PharmGKB
DPE	0	0	0
Littérature	0	109 078	0
PharmGKB	0	0	132

TABLE 3.7 – Nombre de liens `skos:broadMatch` entre les relations pharmacogénomiques d'origine distincte. L'origine du domaine des liens est en ligne et celle du co-domaine en colonne.

	DPE	Littérature	PharmGKB
DPE	0	14	0
Littérature	0	133 762	0
PharmGKB	0	974	894

TABLE 3.8 – Nombre de liens `skos:relatedMatch` entre les relations pharmacogénomiques d'origine distincte. L'origine du domaine des liens est en ligne et celle du co-domaine en colonne.

	DPE	Littérature	PharmGKB
DPE	0	0	0
Littérature	0	37 864	0
PharmGKB	0	0	0

Discussion et conclusion

Un élément important à considérer est que les connaissances de PGxLOD sont dans sa version actuelle d'une qualité moyenne et ne sont pas toutes adaptées à une comparaison. C'est notamment le cas des relations issues de PharmGKB dont les phénotypes sont très génériques et ne sont pas représentés dans un vocabulaire standard, ce qui rend difficile leur comparaison. Une perspective évidente serait de compléter PGxLOD avec les correspondances définies dans l'UMLS [Bodenreider, 2004] ou le BioPortal du NCBO [Ghazvinian *et al.*, 2009] ce qui enrichirait

l'ensemble des labels associés aux concepts et offrirait plus de voix possibles pour des comparaisons entre sources dont les vocabulaires sont distincts. Cela serait particulièrement utile si nous extrayions systématiquement des connaissances des DPE. En effet, dans ce type de sources, à la fois les vocabulaires utilisés dans les textes cliniques et les ontologies utilisées pour encoder l'activité clinique (comme ICD pour les codes diagnostics) sont différents de ceux utilisés dans les bases de données biologiques et la littérature. Une autre limite associée à la représentation des phénotypes est que les phénotypes qui décrivent des réponses aux médicaments sont des expressions complexes qu'il est difficile de capturer à partir d'un dictionnaire ou d'un modèle entraîné sur un petit corpus. Il est nécessaire dans le cas de ces expressions complexes et rares de développer des méthodes ad-hoc.

Les relations extraites de la littérature sont identifiées par un modèle relativement naïf, entraîné sur un petit corpus (307 phrases). Cela fait que la qualité de l'extraction est limitée et que des faux positifs viennent ajouter du bruit au contenu de PGxLOD (la précision du modèle est 0,67). Pour améliorer cela nous avons assemblé un corpus plus grand, de plus de 900 phrases avec l'objectif de constituer une ressources ouverte et utile pour le domaine [Legrand *et al.*, 2019]. Ce corpus s'appelle PGxCorpus et est utilisé depuis peu (le 4 novembre 2019) pour entrainer des modèles moins naïfs qui pourront alimenter la prochaine version de PGxLOD ($v3$?) avec des relations de meilleure qualité. Nous pouvons espérer que cette amélioration, comme celle de la considération des phénotypes complexes de PharmGKB permettra des comparaisons plus précises et plus exploitables au niveau applicatif.

L'aspect manuel de l'alimentation des relations issues des DPE est également une limite. Une de nos perspectives est de chercher des patients dans les entrepôts de données cliniques qui instancient les unités de connaissances pharmacogénomiques. Cela permettrait de discuter le niveau d'évidence associé à ces connaissances. En effet, une connaissance observée sporadiquement pourrait ainsi être illustrée par de nouveaux cas et être alors renforcée, ou modérée selon le profil de ces nouveaux cas. Cette instanciation n'est cependant pas si évidente à réaliser. Une première raison est le manque de données génétiques associées aux DPE. Les études exemples dont les résultats ont été ajoutés à PGxLOD sont des études assez exceptionnelles, où les DPE sont associés à des banques d'échantillons biologiques, échantillons qui ont permis un séquençage de variants d'intérêt et peuvent alors venir instancier le composant génétique de nos relations. Nous pouvons espérer qu'avec des projets comme le *100,000 genomes project* [8] ou *France–Génomique* [9], ce genre d'études devienne dans le futur plus commun. Une solution alternative qui a quelques avantages est de se contenter des DPE sans donnée génétique, et d'utiliser des marqueurs phénotypiques présents dans les DPE comme *proxy* des variants génétiques. Le travail de [Neuraz *et al.*, 2013] est un bon exemple de cette approche. L'avantage d'une telle approche est que les marqueurs phénotypiques peuvent intégrer, en plus de l'effet génétique, des effets environnementaux influençant la réponse au médicament. C'est également cette idée qui a motivé notre travail de prédiction de réponse aux médicaments à partir de phénotypes (Section 2.3).

Malgré ses limites, l'application de nos règles de réconciliation sur PGxLOD a mis en évidence des premiers liens entre les connaissances des différentes sources qui sont pourtant très hétérogènes (en terme de granularité, de type de langage utilisé clinique ou scientifique, etc.). En cela la mise en oeuvre de cette comparaison et la découvertes de premières correspondances sont déjà des résultats prometteurs que nous cherchons à améliorer.

D'un point de vue plus théorique, la comparaison de connaissances dans le LOD est un sujet de

8. `https://www.genomicsengland.co.uk/about-genomics-england/the-100000-genomes-project/`
9. `https://www.france-genomique.org/`

recherche partagé au sein de la communauté du Web sémantique [Cheatham *et al.*, 2017]. Notons à ce propos que notre approche à base de règles est généralisable et est originale dans la mesure où nous réconcilions des relations n-aires au regard des entités qu'elles relient [Monnin *et al.*, 2018b]. Dans cette mesure il est intéressant de positionner notre approche par rapport à celle des LinkKeys [Atencia *et al.*, 2019] qui propose d'apprendre à partir de régularités dans les données les propriétés qui définissent des clés au même sens que celles définies par les dépendances fonctionnelles des bases de données relationnelles. Ce genre de propriétés permet, comme dans le relationnel, de garantir l'unicité des entités et alors soit de fusionner les entités dupliquées soit d'établir un lien `owl:sameAs` entre elles. Il pourrait être intéressant de généraliser et reformuler notre approche dans les termes des LinkKeys.

Un élément émergeant de ce travail est qu'au dela du besoin d'identifier l'identité, comparer des sources très hétérogènes nécessite aussi d'identifier et de quantifier la similarité. C'est quelque chose que nous avons cherché à capturer avec le lien `skos:broadMatch` dans le cas de la spécialisation, et de façon plus générale avec le lien `skos:relatedMatch`. Il demeure que nos règles de réconciliation ne sont ni très générales, ni très flexibles. Pour cette raison nous nous sommes intéressés à des approches capables d'identifier et quantifier une similarité. En particulier, nous avons récemment proposé deux travaux qui ne sont pas détaillés dans ce mémoire mais vers lesquels j'aimerais orienter le lecteur intéressé par le sujet. Le premier peut être vu comme une première étape vers un clustering des relations n-aires définies dans des LOD [Monnin *et al.*, 2019b]. Nous avons défini une représentation vectorielle (un *embedding*) pour ces relations à partir de leur voisinage dans le graphe, mais au lieu d'utiliser les graph kernels comme dans [Dalleau *et al.*, 2017], nous avons utilisé un réseau de neurones convolutif adapté aux graphes. Nos premiers résultats montrent que nous arrivons à isoler des "profils de similarité" qui correspondent aux trois niveaux de similarité que nous avons implicitement définis avec nos règles (`owl:sameAs`, `skos:broadMatch` et `skos:relatedMatch`). Le second travail qui répond à ce même besoin de quantification de la similarité a été appliqué à une recherche de similarité entre des maladies décrites par des annotations faites avec plusieurs ontologies [Personeni *et al.*, 2018]. Dans ce cas nous sommes hors du LOD, mais le fait que nos objets, des maladies génétiques, soient décrits par des annotations GO (pour leurs gènes responsables) et HPO (pour les phénotypes qui les caractérisent), les lie avec des connaissances comme c'est le cas dans le LOD. Nous avons montré que dans ce cas nous pouvions définir une distance sémantique qui prend en considération des descriptions complexes exprimées avec plusieurs ontologies, puis utiliser celle-ci pour une tâche de classification. Cette distance pourrait être utilisée pour faire un clustering qui groupe des éléments proches au regard d'ensembles d'annotations complexes.

Cette question ramène également à la question de la prise en considération des connaissances associées aux LOD dans les deux travaux décrits dans cette section. Dans le travail de prédiction des liens gène-médicament (3.2.2), des annotations sont utilisées à deux niveaux : les annotations GO des gènes sont utilisées comme attributs pour la prédiction, et les annotations des phénotypes avec les CUI de l'UMLS sont utilisées comme référence croisée pour établir des correspondances entre les phénotypes de différentes sources. En somme, les annotations sont utilisées comme des propriétés mais au dela de ça la sémantique associée n'est pas utilisée dans ce travail. Dans le travail sur la prédiction de lien d'identité (3.2.3), les connaissances associées aux données du LOD sont cette fois utilisées. Elles le sont par les règles de réconciliation qui considèrent à la fois la hiérarchie des ontologies et les liens `owl:sameAs` pour prédire de nouveaux liens. Ceci est illustré dans les Figures 3.4 et 3.5.

Au delà de la prédiction de liens, il est intéressant de fouiller les LOD pour mettre en évidence des connaissances plus générales qui peuvent enrichir les connaissances de domaine. La section suivante présente deux travaux qui illustrent cette idée.

3.3 Recherche de régularités dans un graphe de connaissances

Nous nous intéressons dans cette section à la fouille de LOD pour compléter des représentations de connaissances. Nous décrivons d'abord une expérience avec la programmation logique inductive pour apprendre, à partir de données ouvertes et liées, des descriptions concernant des classes de gènes. Nous présentons ensuite une utilisation originale de l'analyse formelle de concept où nous proposons d'annoter les concepts formels d'un treillis pour mettre en évidence de nouveaux axiomes qui pourront venir enrichir une ontologie.

3.3.1 Apprentissage de descriptions des gènes responsables des DI

Nous avons proposé d'utiliser les LOD et notamment des annotations que l'on peut y trouver avec une ontologie (la Gene Ontology, notée ci-après GO), pour caractériser et classer des gènes responsables de déficiences intellectuelles (DI). Pour cela nous avons utilisé la Programmation Logique Inductive (PLI) [Muggleton, 1991] pour produire un modèle formel de ces objets. En effet, la PLI permet d'apprendre un modèle caractérisant un ensemble d'objets qui lui est donné en entrée. En PLI, ce type de modèle est appelé une *théorie* et consiste en un ensemble de règles écrites en logique du premier ordre. La capacité de ces règles à bien décrire les objets et notamment leur appartenance à leur classe vis à vis d'objets n'y appartenant pas découle de la prise en compte de leurs propriétés et peut être évaluée notamment en testant l'efficacité des règles à reproduire une classification experte. L'avantage de composer une théorie avec de telles règles est qu'elle peut être évaluée objectivement. De plus, elle peut être interprétée, plus subjectivement par un expert du domaine qui peut lire ces règles et évaluer leur pertinence au regard des ses connaissances expertes. Dans ce travail, nous avons construit plusieurs théories capables de décrire et classer les gènes responsables de DI. Nous avons utilisé pour cela des propriétés décrivant ces objets trouvés dans le LOD et notamment des annotations ontologiques. Nous avons porté une attention particulière à évaluer l'intérêt d'utiliser ces annotations et les connaissances qui leurs sont associées en comparant les performances en terme de classification des théories qui considèrent un peu, beaucoup ou pas du tout de connaissances.

Notre travail a commencé par des étapes de sélection et d'intégration de données du LOD en lien avec les DI (et donc indépendemment de PGxLOD), et de transformation des triplets RDF en un format compatible avec la PLI. Nous ne détaillerons pas ces étapes ici, pour nous concentrer sur la dernière étape où annotations et connaissances associées sont utilisées dans un processus de fouille. En pratique nous avons constitué un ensemble d'exemples positifs (des gènes responsables de DI) et un ensemble d'exemples négatifs (des gènes probablement pas responsables de DI) pour apprendre avec la PLI des théories caractérisant les exemples positifs vis à vis des négatifs. L'objectif de chaque théorie est de proposer une caractérisation la plus couvrante possible des exemples positifs qui exclut les exemples négatifs. Nous avons évalué la caractérisation proposée par chaque théorie en mesurant sa capacité à reproduire une classification experte.

Sélection des exemples positifs et négatifs

Nos exemples positifs sont définis par une liste de gènes responsables de DI selon l'étude de l'état de l'art de [Inlow and Restifo, 2004]. Les exemples négatifs quant à eux sont des gènes qui à notre connaissance ne causent pas ce type de maladie car ces gènes sont associés à des phénotypes bien distincts de ceux des DI, selon nos collaborateurs experts (le Prof. Philippe Jonveaux et le Dr Céline Bonnet du CHRU de Nancy). La liste de positifs extraits de [Inlow and Restifo, 2004] est composée de 282 exemples positifs. La liste de négatifs a été constituée par la sélection manuelle

avec les experts d'une liste de phénotypes définis dans OMIM clairement non-associés selon eux aux DI. De plus, nous nous sommes assurés qu'une déficience intellectuelle n'était pas un des symptômes associés à ces maladies (dans la section *clinical synopsis* d'OMIM). A partir de cette liste de phénotypes nous avons extrait d'OMIM la liste des gènes responsables de ces phénotypes.

La liste de phénotypes dont une DI n'est pas un symptôme est la suivante : *deafness*, *retinitis pigmentosa*, *obesity*, *cataract*, *muscular dystrophy*, *myopathy*, *hemolytic anemia*, *anemia*, *complement component deficiency*, *osteoarthritis*, *ectodermal dyspolasia*, *thrombophilia*. À partir de cette liste, nous avons extrait un sous-ensemble de ces gènes, précisément 267 exemples négatifs, sélectionnés de façon à ce que la proportion de gènes responsables de chaque phénotype soit la même dans notre échantillon que dans la liste complète des gènes responsables.

Nous avons ensuite requêté le LOD de façon contrôlée pour en extraire des descripteurs des gènes positifs et négatifs. Les propriétés sélectionnées sont des associations entre les gènes et des pathways et réactions extraits de KEGG et Reactome, des protéines extraites de iRefIndex, des domaines protéiques extraits de InterPro, une position chromosomique extraite de NCBI Gene, des annotations GO extraites de GOA et les parents des concepts utilisés dans les annotations extraits du BioPortal. Ces données ont été extraites de trois sources de LOD : le projet Bio2RDF [Callahan *et al.*, 2013], la plateforme RDF de l'EBI [Jupp *et al.*, 2014] et le SPARQL *endpoint* du Bioportal [Noy *et al.*, 2009]. Le détail des sources et des chemins utilisés pour extraire ces propriétés sont disponibles dans [Personeni *et al.*, 2014].

Programmation Logique Inductive avec des ontologies

Nous avons réalisé 5 expériences de fouille à partir du jeu de données obtenu. Chaque expérience contraint différemment le niveau de la hiérarchie des termes GO utilisable pour généraliser les concepts associés aux exemples par les annotations. La théorie résultante de chaque expérience de fouille peut être utilisée dans un but descriptif et prédictif. La description des gènes responsables de DI a ainsi été évaluée qualitativement en collaboration avec nos collaborateurs experts des DI. Nous n'aborderons pas cette évaluation ici. Nous orientons le lecteur qui voudrait en savoir plus à ce propos vers [Personeni, 2018]. Nous nous limiterons ici à l'évaluation du pouvoir prédictif des règles en mesurant par une validation croisée dans quelle mesure elles reproduisent la classification experte en gènes positifs vs. négatifs.

Nous avons choisi la PLI car elle permet de prendre en considération en plus des données sur les exemples, les connaissances de domaine qui leur sont associées. Pour cela, il est possible de représenter ces connaissances sous la forme de règles d'inférence exprimées en logique du premier ordre. Ces règles sont considérées lors de l'étape d'apprentissage et enrichissent les descriptions de chaque exemple.

Afin d'évaluer la contribution des connaissances contenues dans les ontologies de domaine, nous contraignons la transitivité de la relation `subClassOf` à s'arrêter à différents niveaux de généralisation. Pour autoriser n étapes de généralisation sur l'ontologie considérée, nous définissons $2n$ règles d'inférence qui seront considérées lors de la construction des théories. Ces règles de la forme `tête :- corps` se lisent de la façon suivante : la tête de la règle est vraie si le corps est vrai. La virgule représente la conjonction.
Ainsi nous créons une règle pour chaque $i \in [2, n]$ exprimant la transitivité du prédicat `subClassOf` (*i.e.*, `rdfs:subClassOf`) au i-ième degré :

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

Une règle pour chaque $i \in [1, n]$:

$$\texttt{subClassOf(X,Y) :- subClassOf}_\texttt{i}\texttt{(X,Y).}$$

TABLE 3.9 – Mesures sur les règles des théories produites par nos cinq 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

Une règle représentant la réflexivité de `subClassOf` :

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

Nous avons appelé la première expérience $G1$. Celle-ci utilise les propriétés des gènes en termes de protéines, réactions, pathways, etc. et inclut également les annotations GO, et les concepts parents au premier degré de ces annotations, dénotés explicitement (*i.e.*, non inférés) par la propriété `rdfs:subClassOf`. De cette façon, nous proposons quatre expériences pour n variant de 1 à 4, appelées $G1$, $G2$, $G3$ et $G4$ utilisant chacune respectivement jusqu'à 1, 2, 3 ou 4 étapes de généralisation. Nous ajoutons à ces quatre expériences une dernière expérience, appelée $no-GO$ dans laquelle les annotations GO sont exclues. Cette dernière expérience a pour objectif d'analyser l'apport des prédicats qui ne sont pas des annotations GO.

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

- *rulesize* = 6, le nombre maximal de termes composant 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.

Notons que le paramètre *noise* permet aux règles de la théorie de tolérer quelques exceptions, c'est-à-dire de couvrir quelques exemples négatifs. Ceci est important dans un contexte comme le nôtre (*i.e.*, les LOD) où les données sont potentiellement bruitées.

Résultats

Chaque expérience produit une théorie décrivant les gènes responsables de DI. La Table 3.9 présente quelques métriques sur les différentes théories permettant de constater les effets d'un niveau de généralisation variable sur les règles obtenues.

Lorsque l'on retire les annotations GO ($no-GO$), le nombre de règles dans la théorie est divisé par deux et le nombre moyen d'exemples couverts par les règles diminue de 14 à 8,4. Cela indique que les termes GO aident à construire des règles qui décrivent la classe des gènes responsables de DI. Avec GO, on constate que plus le niveau de généralisation possible est haut, plus le nombre de règles est réduit (de 22 à 16 règles quand on augmente de 1 à 4 le niveau de généralisation). De plus, les règles obtenues couvrent davantage d'exemples en moyenne (de 14 à 16,2). Ces résultats confirment qu'avec l'addition de niveaux de généralisation, les théories

TABLE 3.10 – Résultats de la classification pour les cinq expériences, par validation croisée *leave-one-out*. R : Rappel, Spec. : Spécificité, P : Précision.

Expérience	*R*	*Spec.*	*P*
$no-GO$	26,6	**94,4**	59,6
$G1$	47,9	81,3	64,1
$G2$	55,7	80,5	67,8
$G3$	55,7	81,7	68,3
$G4$	**57,1**	83,1	**69,8**

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 ensembles de règles différents permettent de prédire si un gène donné est responsable de DI.

Pour cela nous avons évalué avec une validation croisée en *leave-one-out*, la capacité de chaque théorie à classer les gènes responsables de DI (*i.e.*, les exemples positifs) comme tels. La Table 3.10 présente les résultats de cette validation croisée pour les expériences d'apprentissage $no-GO$, et $G1$–4. Les résultats montrent que sans utiliser d'annotations GO ($no-GO$), la précision de la classification est plutôt faible (59,6%), avec une haute spécificité mais un faible rappel. Utiliser les termes GO et autoriser la généralisation augmente la précision jusqu'à 69,8%, sans diminuer le rappel. Nous observons donc qu'au-delà des simples annotations GO, les connaissances de domaine proposées par l'ontologie GO permettent une meilleure caractérisation des gènes responsables de DI.

Discussion et conclusion

Une limite assez évidente est le passage à l'échelle dans la prise en considération du LOD et de l'ontologie. En effet, nous nous sommes restreints à n'utiliser qu'un ordinateur personnel lors de ces expérimentations et avec cet équipement nous avons été rapidement limités par les besoins en mémoire d'Aleph (l'implémentation de la PLI utilisée). Cela nous a poussé à faire une sélection manuelle des propriétés du LOD qui nous semblaient pouvoir être discriminantes pour nos gènes (sans d'ailleurs que la pertinence des propriétés choisies n'ait été testée) et nous a forcé à arrêter nos expériences à 4 niveaux de généralisation. En effet, ajouter un niveau de généralisation, enrichit le nombre de descriptions des exemples associés à un concept GO (ou plus) créant un ensemble de descriptions trop grand pour une machine classique au delà de 4.

Ces limitations ne nous ont cependant pas empêché d'observer que de façon générale la PLI se révèle appropriée à la manipulation des LOD et à la prise en compte des connaissances de domaine représentées dans des ontologies. Un avantage de la PLI, notamment par rapport à une distance sémantique qui considère plusieurs ontologies comme dans [Personeni *et al.*, 2018], est la prise en compte des différents aspects de GO uniquement lorsque nécessaire. Nous observons 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, garantissant un bon ratio de vrais négatifs, 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 dans le LOD. 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 utilisées pour considérer un ensemble de gènes tests qui pourraient être impliqués dans les DI. La classification des ces gènes tests par la meilleure théorie permettrait d'en isoler certains que nous pourrions proposer ensuite comme gènes candidats aux experts.

3.3.2 Complétion d'ontologies à partir des données du LOD

Les LOD et les ontologies associées ont des qualités et des niveaux de complétude très variés. Cela amène assez naturellement à considérer les méthodes de fouille pour compléter les LOD avec des méthodes de prédiction de liens comme ceux présentés dans la section 3.2. Mais fouiller les LOD peut également permettre de découvrir des régularités qui peuvent venir enrichir les éléments de connaissance définis dans une ontologie du domaine.

Nous avons proposé dans [Monnin *et al.*, 2017b, Monnin *et al.*, 2017c] de découvrir des axiomes de subsomption (de la forme $C \sqsubseteq D$) entre concepts d'ontologie et cela à partir de régularités dans les données du LOD. Une particularité de notre approche est de considérer les données mais aussi leur façon d'être annotées avec des ontologies. Nous avons pour cela utilisé (encore !) l'analyse formelle de concept (AFC) présentée dans le Chapitre 1. En particulier nous proposons d'utiliser la structure hiérarchique des treillis de concepts pour grouper des entités RDF en observant les *prédicats* dont ces entités sont le *sujet* dans un graphe de connaissances. Ici les rôles de sujet et prédicat sont utilisés dans le sens des triplets RDF qui sont de la forme *(sujet, prédicat, objet)*. Nous définissons la notion d'*annotation de concept* (ou annotation de concept formel pour être précis) qui revient à associer un ensemble de concepts ontologiques à un concept formel d'AFC. L'enchaînement des deux étapes que sont la construction du treillis et l'annotation de ses concepts permet de suggérer de nouveaux axiomes de subsomption pour compléter une ontologie associée aux données.

Exemple jouet et notations

Pour illustrer notre approche nous considérons dans cette section un ensemble de données RDF et une ontologie jouets représentés respectivement dans le Tableau 3.11 et la Figure 3.6.

L'élément atomique d'un graphe RDF est un triplet noté $\langle sujet,\ prédicat, objet\rangle \in (U \cup B) \times (U \cup B) \times (U \cup B \cup L)$ où U est l'ensemble des URIs, L celui des littéraux et B celui des nœuds blancs. Notons $\mathcal{C}_{\mathcal{O}}$ l'ensemble des concepts de $\mathcal{O}$. Nous cherchons à découvrir des axiomes de subsomption, *i.e.*, une relation transitive notée $\sqsubseteq$, où $C \sqsubseteq D$ représente que toutes les instances de C sont également instances de D. Pour ne pas réduire nos ontologies à celles représentées en OWL ou RDF ou SKOS ou autre, nous représentons de façon générique la relation de subsomption dans notre ontologie $\mathcal{O}$ avec le prédicat `abstract:subClassOf` et la relation d'instanciation avec `abstract:type`. La différence entre ces formalismes n'impacte pas notre approche, mais nous permet de rester générique vis à vis du format d'encodage des connaissances (*i.e.*, SKOS, RDF, OWL, OBO).

Nous notons également le TYPE d'une entité `r` comme l'ensemble des concepts de $\mathcal{O}$ que `r` instancie. Formellement,

$$\text{TYPE}(\texttt{r}) = \{\texttt{c} \in \mathcal{C}_{\mathcal{O}} \mid \langle \texttt{r}, \texttt{abstract:type}, \texttt{c}\rangle\}.$$

Suivant la définition de la subsomption, nous définissons le *type étendu* ou X-TYPE (pour *extended type* en anglais) de `r` comme l'ensemble des concepts de TYPE(`r`) et de leurs parents.

TABLE 3.11 – Ensemble jouet de triplets RDF, écrits avec la syntaxe Turtle.

$\mathtt{r}_1$	abstract:type	$\mathtt{k}_1, \mathtt{k}_2$.	$\mathtt{r}_2$	$\mathtt{pred}_3$	$\mathtt{o}_5$.
$\mathtt{r}_1$	$\mathtt{pred}_1$	$\mathtt{o}_1$.	$\mathtt{r}_3$	abstract:type	$\mathtt{k}_1, \mathtt{k}_2$.
$\mathtt{r}_1$	$\mathtt{pred}_2$	$\mathtt{o}_2$.	$\mathtt{r}_3$	$\mathtt{pred}_1$	$\mathtt{o}_6$.
$\mathtt{r}_2$	abstract:type	$\mathtt{k}_1, \mathtt{k}_2, \mathtt{k}_4, \mathtt{k}_5$.	$\mathtt{r}_4$	abstract:type	$\mathtt{k}_1, \mathtt{k}_2, \mathtt{k}_5$.
$\mathtt{r}_2$	$\mathtt{pred}_1$	$\mathtt{o}_3$.	$\mathtt{r}_4$	$\mathtt{pred}_2$	$\mathtt{o}_7$.
$\mathtt{r}_2$	$\mathtt{pred}_2$	$\mathtt{o}_4$.	$\mathtt{r}_4$	$\mathtt{pred}_3$	$\mathtt{o}_8$.

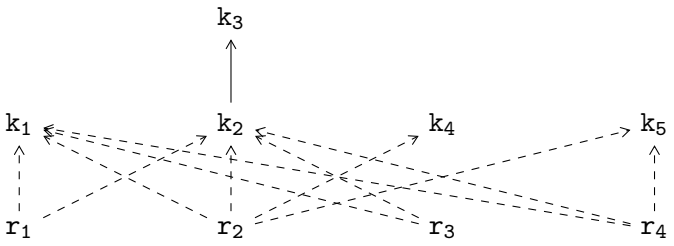


FIGURE 3.6 – Exemple jouet de concepts d'ontologie instanciés par des entités RDF. Les instanciations sont représentées par des flèches pointillées et la subsomption par une flèche continue. Par exemple, $\mathtt{r}_1$ instancie $\mathtt{k}_1$ et $\mathtt{k}_2$, et $\mathtt{k}_2$ est subsumé par $\mathtt{k}_3$.

Formellement,

$$\text{X-TYPE}(\mathtt{r}) = \text{TYPE}(\mathtt{r}) \cup \{\mathtt{d} \in \mathcal{C}_\mathcal{O} \mid \exists \mathtt{c} \in \text{TYPE}(\mathtt{r}),\ \langle \mathtt{c}, \mathtt{abstract:subClassOf}, \mathtt{d} \rangle\}.$$

Nous définissons un contexte formel (G, M, I) où l'ensemble des objets G est l'ensemble des entités sujets d'un triplet dans notre graphe, et l'ensemble de attributs M est l'ensemble des prédicats des triplets de notre graphe. La relation d'incidence $\mathtt{r}\ I\ \mathtt{pred}, \mathtt{r} \in G, \mathtt{pred} \in M$ représente que $\mathtt{r}$ est le sujet d'un triplet dont le prédicat est $\mathtt{pred}$ dans notre graphe. Suivant cette définition, nous représentons dans le Tableau 3.12 les données de notre exemple sous la forme d'un contexte formel de cette forme. En utilisant l'AFC, nous pouvons alors construire une treillis de concepts, comme celui représenté Figure 3.7. Précisons que seuls sujets et prédicats sont considérés pour construire le contexte (et le treillis), les objets des triplets ne sont pas considérés dans ce travail, sauf s'ils sont également sujet. A ce stade le treillis est composé de concepts formels de la forme (A, B) où l'extension A est un ensemble de sujets et l'intention B est un ensemble de prédicats.

Nous définissons ici l'*annotation de concept* qui associe aux concepts formels un ensemble de concepts d'ontologie. Pour un concept formel (A, B), son annotation est définie comme

$$A^\diamond = \bigcap_{\mathtt{r} \in A} \text{X-TYPE}(\mathtt{r})$$

Cette annotation représente l'ensemble des concepts ontologiques communs aux types étendus des sujets de A. Soit deux concepts formels $(A_1, B_1) \leqslant (A_2, B_2)$, tels que $A_1 \subseteq A_2$, nous avons $A_2^\diamond \subseteq A_1^\diamond$. Alors l'annotation de concept peut être écrite de façon réduite en adaptant à nos annotations la notation réduite des treillis [Ganter and Wille, 1999]. Nous nommons l'*annotation réduite* d'un concept, son annotation qui exclut les concepts ontologiques qui apparaissent dans l'annotation des concepts formels plus haut placés dans le treillis. La Figure 3.7 présente notre treillis exemple et en gras l'annotation $(\{\cdot\}_A)$ de ses concepts en notation réduite.

TABLE 3.12 – Contexte formel construit à partir du graphe RDF du Tableau 3.11. Une croix dans le tableau relie le sujet et le prédicat seulement si un triplet existe avec le sujet comme sujet et le prédicat comme prédicat.

	abstract:type	$\texttt{pred}_1$	$\texttt{pred}_2$	$\texttt{pred}_3$
$\texttt{r}_1$	×	×	×	
$\texttt{r}_2$	×	×	×	×
$\texttt{r}_3$	×	×		
$\texttt{r}_4$	×		×	×

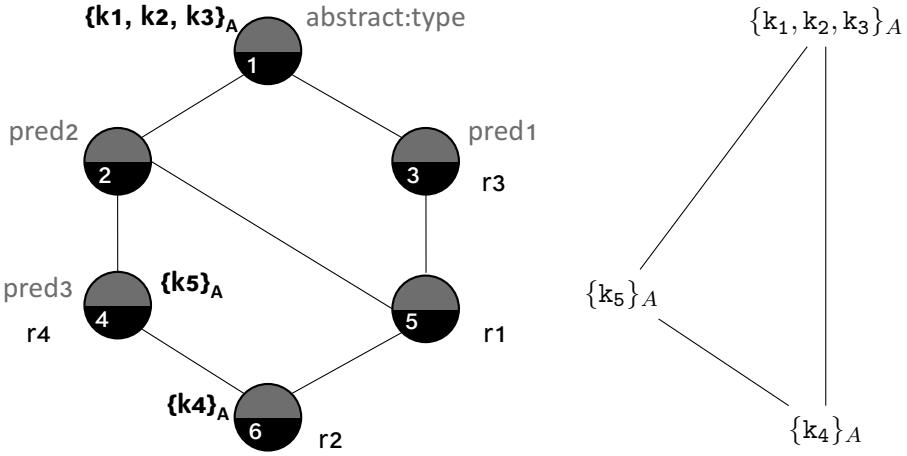


FIGURE 3.7 – A gauche, le treillis de concept *annoté* construit à partir du contexte du Tableau 3.12 et annoté avec les concepts de l'ontologie de la Figure 3.6. A droite, l'ordre induit des annotations. Le treillis, ses annotations et leur ordre sont représentés selon la notation réduite. Les sujets (*i.e.*, extensions) sont en noir, les prédicats (*i.e.*, intentions) sont en gris et les annotations sont entre crochés avec la notation $\{\cdot\}_A$. Les concepts formels sont arbitrairement annotés de 1 à 6.

A partir d'un treillis annoté, des axiomes de subsomption peuvent être découverts de la façon suivante.

Considérons un concept (A, B), *e.g.* le concept 6, et un de ses concepts recouvrant (E, F), *i.e.*, un des concepts directement plus haut dans le treillis comme le concept 4. De la même façon que dans la Figure 3.7, nous notons $A^{\diamond}_{A} = \{\mathtt{x_1}, \mathtt{x_2}, \ldots, \mathtt{x_p}\}_A$ (ici, $\{\mathtt{k_4}\}_A$) et $E^{\diamond}_{A} = \{\mathtt{y_1}, \mathtt{y_2}, \ldots, \mathtt{y_q}\}_A$ (ici, $\{\mathtt{k_5}\}_A$) l'annotation réduite des deux concepts. Alors, considérons deux concepts de l'ontologie $\mathtt{x_i} \in A^{\diamond}_{A}$ et $\mathtt{y_j} \in E^{\diamond}_{A}$. Par définition, $\mathtt{x_i}$ apparaît dans le type étendu de tous les sujets de A et $\mathtt{y_j}$ apparaît dans le type étendu de tous les sujets de E. Comme $A \subseteq E$, $\mathtt{y_j}$ apparaît dans le type étendu de tous les sujets où $\mathtt{x_i}$ apparaît également. De plus, $\mathtt{y_j}$ peut apparaître également dans le type étendu d'autres sujets. Pour cette raison, nous considérons que ceci permet de décrire un axiome de subsomption tel que $\mathtt{x_i}$ est subsumé par $\mathtt{y_j}$ (suivant l'exemple $\mathtt{k_4}$ devrait être subsumé par $\mathtt{k_5}$). Cet axiome est alors comparé aux axiomes déjà définis dans l'ontologie :

- *(i)* si $\mathtt{x_i} \sqsubseteq \mathtt{y_j}$ est déjà explicitement représenté, nous le qualifions d'*axiome redondant*,
- *(ii)* si $\mathtt{x_i} \sqsubseteq \mathtt{y_j}$ n'est pas explicitement représenté, mais peut être inféré, nous le qualifions d'*axiome inférable* et
- *(iii)* si $\mathtt{x_i} \sqsubseteq \mathtt{y_j}$ n'est ni représenté explicitement, ni inférable, nous le classons comme une qualifions de *nouvel axiome de subsomption* découvert.

Suivant notre exemple l'axiome $\mathtt{k_4} \sqsubseteq \mathtt{k_5}$ est nouveau.

Si nous ne considérions que les concepts recouvrants (*i.e.*, directement parent) durant le processus, cela ferait manquer de nouveaux axiomes. Par exemple dans la Figure 3.7, comme l'annotation réduite du concept 2 est vide, nous ne pourrions pas découvrir d'axiome quand nous considérant le concept 4 et seulement son parent direct. Pour cette raison, nous proposons de rechercher des axiomes à partir de l'ordre des annotations induit du treillis. Cet ordre est obtenu en ne considérant que les annotations et en les ordonnant suivant l'ordre défini par l'inclusion *i.e.*, l'ordre défini par les concepts formels. Cet ordre est représenté en notation réduite sur la partie droite de la Figure 3.7. L'inclusion des annotations se lit de haut en bas. Alors les axiomes de subsomption peuvent être lus directement sur la notation réduite de bas en haut. Dans notre exemple jouet nous trouvons les axiomes :

$$\mathtt{k_5} \sqsubseteq \mathtt{k_1}\ \mathit{and}\ \mathtt{k_2}\ \mathit{and}\ \mathtt{k_3}$$
$$\mathtt{k_4} \sqsubseteq \mathtt{k_1}\ \mathit{and}\ \mathtt{k_2}\ \mathit{and}\ \mathtt{k_3}.$$

Ces axiomes ne peuvent pas être découverts en considérant uniquement les concepts recouvrants.

Expérimentation préliminaire

Nous avons mené une expérimentation préliminaire sur DBpedia [Lehmann and *et al.*, 2015], des données liées assemblées à partir de Wikipedia. Nous avons classé 904 pages de DBPedia (version 2014) qui concernent des personnes décédées enter le 01/01/2000 et 07/01/2000. Le treillis associé est constitué de 15 234 concepts que nous avons annotés avec deux ontologies associées à DBPedia : DBPedia Ontology et YAGO. Nous avons retrouvé 11 axiomes redondants pour DBPedia Ontology et 199 pour YAGO. Par exemple nous avons retrouvé l'axiome `Boxer` $\sqsubseteq$ `Athlete` qui est déjà défini dans DBPedia Ontology. Nous n'avons pas trouvé d'autres axiomes (inférables ou nouveaux) pour DBPedia Ontology, probablement à cause de sa petite taille (683 classes). Par contre nous avons trouvé 2 250 axiomes inférables et 1 372 nouveaux pour YAGO. Parmi eux le nouvel axiome `FilipinoChildActors` $\sqsubseteq$ `FilipinoActors` qui semble valide.

3.3.3 Discussion et conclusion

Nous avons sélectionné un sous-ensemble de DBPedia et produit à partir de celui-ci un nombre d'axiomes élevé, nécessitant une étape de post-processing pour en sélectionner les plus intéressant. Une limite de notre évaluation est liée au fait de ne considérer qu'une portion de DBPedia et que par conséquent nos axiomes ne sont valides que sur ce sous-ensemble qui omet une partie des données. En effet les axiomes peuvent ne plus tenir avec plus de données. Il demeure que nous trouvons des axiomes qui semblent valides mais la question de leur trivialité se pose : est-ce que les axiomes nouveaux sont tous aussi intéressants à ajouter dans l'ontologie ? Est-ce que certain n'alourdissent pas la hiérarchie sans ajouter à sa qualité ? Une façon d'évaluer cela serait de considérer les différentes versions de DBPedia et des ontologies associées et de voir si nous trouvons à partir des versions anciennes des axiomes ajoutés dans les version plus récentes. Cette expérience nous semble d'autant plus intéressante que l'axiome trouvé `Filipino_child_singers` $\sqsubseteq$ `Child_singers` était absent de notre version, mais est dans la version courante de l'ontologie. D'un point de vue méthodologique, il serait intéressant de comparer nos résultats avec ceux obtenus avec d'autres méthodes. En particulier, nos annotations de concepts formels peuvent être vus comme une troisième dimension associée aux concepts formels (dont les deux premières dimensions sont l'extension et l'intention). Pour cette raison il nous semble intéressant d'explorer l'*analyse de concepts triadiques* [Lehmann and Wille, 1995], une extension de la FCA qui introduit une troisième dimension dans la description des données.

Une autre perspective intéressante de l'annotation de treillis apparaît dans le cas où, comme dans notre expérimentation, plusieurs ontologies sont associées au même jeu de données. Alors il est possible d'ajouter autant d'annotations aux concepts formels du treillis qu'il y a d'ontologies associées aux données. Dans ce cas le treillis peut être vu comme une structure "pivot" capable de proposer des alignements entre les ontologies. Nous avons poussé ce raisonnement et adapté le formalisme des structures de patrons aux treillis annotés en partant de l'observation que le treillis offre un ordre sur les annotations et que cela nous permet de définir un opérateur de similarité, donc une structure de patron, donc un nouveau treillis [Monnin *et al.*, 2018a]. Ce treillis nous permet de mettre en évidence des concepts équivalents entre deux ontologies (ou plus) et quand cela n'est pas possible de suggérer des relations de généralisation entre concepts d'ontologies différentes. Cette fois encore l'étape suivante qui semble la plus naturelle est de considérer l'analyse de concepts triadiques pour disposer d'un formalisme naturellement défini pour manipuler des descriptions de données à trois dimensions, ou plus.

3.4 Discussion générale

Et de trois ! Après avoir construit une ontologie pour la pharmacogénomique pendant ma thèse (SO-Pharm, [Coulet *et al.*, 2006]), j'en ai construit deux autres (PHARE et PGxO), ce qui fait trois au total ! A première vue cela peut sembler assez incohérent, car il pourrait sembler plus rationnel de réutiliser et éventuellement compléter la première ontologie plutôt que d'en ajouter deux au même domaine dans un panorama des ontologies biomédicales déjà bien riche. La réalité est que ces trois ontologies sont très différentes, que leur construction a été motivée par des applications et des besoins distincts auxquels les ontologies précédentes ne répondaient pas. En effet, SO-Pharm que j'ai développé pendant ma thèse avait un but d'intégration de données, il s'agit pour cette raison d'un modèle composite qui reprend et connecte des morceaux de schéma et d'ontologies nécessaires à structurer les données. La seconde, PHARE, est une ontologie construite semi-automatiquement à partir de texte pour la normalisation de relations extraites également à partir de texte. Elle modélise notamment les dépendances grammaticales qui peuvent exister entre les mots qui servent fréquemment à décrire dans la littérature des entités d'intérêt en pharmacogénomique ; un contenu particulièrement difficile à utiliser dans un contexte différent de celui-ci [Coulet *et al.*, 2010]. Et la dernière ontologie, PGxO, présentée dans ce chapitre a pour objectif de grouper un ensemble de concepts minimal pour représenter les unités de connaissances que sont les relations pharmacogénomiques. Je trouve intéressant de faire cette précision pour illustrer qu'il peut être utile en informatique d'avoir plusieurs ontologies concurrentes pour un même domaine et donc plusieurs définitions des mêmes concepts dans des contextes différents – ce qui n'est pas consensuel [Ghazvinian *et al.*, 2011]. Selon nous l'intérêt n'est en général pas l'ontologie en elle même, mais ce qu'elle permet de faire. Donc aucune raison de s'arrêter à trois !

Les méthodes hybrides Dans ce dernier chapitre, comme dans l'ensemble de ce mémoire nous avons décrit des méthodes symboliques, *i.e.*, adaptées à la manipulation de données nominales ou catégorielles comme le *pattern mining* [Aggarwal and Han, 2014] ou l'AFC, et d'autres numériques *i.e.*, adaptées à la manipulation de données numériques comme les SVM ou les réseaux de neurones. Même si la frontière entre les deux mondes n'est pas stricte, classiquement les systèmes experts qui manipulent des connaissances formelles sont dits symboliques et les méthodes d'apprentissage supervisées performantes sur de gros volumes de données sont dites numériques. Lorsque l'objectif est de *(i)* fouiller un grand volume de données symboliques comme les graphes de connaissances ou *(ii)* de contraindre la fouille de grands volumes de données numériques à l'aide de connaissances formelles, il semble intéressant d'étudier comment combiner les deux types d'approches pour bénéficier du meilleur des deux mondes. L'équipe Orpailleur a déjà obtenu de premiers résultats en proposant des approches hybrides pour la découverte de connaissances [Bosc *et al.*, 2018, Grissa *et al.*, 2016]. Une des directions qui nous semble la plus intéressante est le développement d'approches dites *neuro-symboliques* qui cherchent à réconcilier apprentissage et raisonnement [d'Avila Garcez *et al.*, 2019, van Krieken *et al.*, 2019]. Une manière concrète de tendre vers cet objectif est d'intégrer des éléments de connaissance dans les représentations numériques (par exemple les *embeddings*) et de faire en sorte que lorsque celle-ci sont considérées, comme par exemple dans la fonction de perte d'un réseau de neurone, la fonction de perte puisse reconnaître les représentations de connaissances encodées numériquement et effectuer des tâches de raisonnement sur celle-ci. Ceci n'est pas trivial et cette combinaison apprentissage, raisonnement est un problème ouvert [Valiant, 2003].

Une motivation particulière pour le développement de systèmes hybrides est l'*explicabilité* des décisions. Nous avons déjà discuté ce thème à la fin du Chapitre 2, à propos de l'utilisabilité des

systèmes d'aide à la décision clinique, mais c'est un besoin général des méthodes d'apprentissage. Dans le cas de la complétion de connaissances, que ce soit par la prédiction de lien, la proposition de nouvelles connaissances de domaines, il est important et intéressant de pouvoir fournir à l'expert avec qui l'on travail des éléments interprétables qui puissent lui faire comprendre pourquoi ces nouveaux éléments de connaissance sont proposés. Nous avons vu à plusieurs reprises que les méthodes symboliques utilisées comme l'ILP ou la FCA pouvaient avoir du mal à passer à l'échelle quand elles sont appliquées à de grands volumes de données comme les LOD. Elles ont en revanche un avantage qui est de produire des résultats directement lisibles et interprétables. Les règles des théories d'ILP, ou la structure du treillis qui permet d'inférer de nouveaux axiomes en sont des exemples.

Perspectives de recherche

Mes perspectives de recherches visent à continuer d'étudier les méthodes d'intelligence artificielle symboliques, numériques et hybrides et leur application pour le développement d'une médecine de précision. J'entends par là développer des méthodes de détection, prédiction et quantification de phénomènes biomédicaux à partir de données cliniques. Le premier cas d'application que je propose de considérer est la variabilité inter-individuelle de réponse aux médicaments. En particulier, j'aimerais travailler sur le développement de modèles prédictifs et utiliser ces modèles dans une fonction contrafactuelle pour identifier et quantifier l'hétérogénéité de la réponse aux traitements à partir de données cliniques observationelles.

Contexte applicatif et méthodologique

La **médecine de précision** a pour objectif d'améliorer la qualité des soins en prenant en considération des informations individuelles propres à chaque patient comme son génome, son environnement ou son mode de vie [Jameson and Longo, 2015]. La considération de ces informations doit aider à prescrire le médicament le plus adapté, à la dose adaptée ; et ainsi réduire les effets indésirables médicamenteux (EIM), mais aussi le manque d'efficacité de certaines prescriptions. La médecine de précision peut ainsi avoir un impact sociétal direct, en réduisant les EIMs qui sont la cause de nombreuses hospitalisations, 280 000 par an aux États-unis selon [U.S. Department of Health and Human Services, 2014]. D'autre part, elle a le potentiel, en identifiant des sous groupes qui bénéficient d'un traitement, potentiellement inutile ou dangereux pour d'autres, et ainsi d'aider au développement et à l'accréditation pour mise sur le marché de ces traitements [Stewart *et al.*, 2007].

La variabilité des réponses aux médicaments est cependant complexe à étudier, tout d'abord à cause de la grande variabilité de ses causes (l'état de santé du patient, les interactions avec d'autres médicaments, avec des aliments, etc.) et de notre connaissance incomplète de la liste des facteurs à considérer. De plus, la rareté de certains facteurs de variabilité au sein d'une population peut rendre difficile leur prise en considération.

Les entrepôts de **Dossiers Patients Électronique (DPE)** constituent une source de données prometteuse et sous-exploitée pour l'étude de cette variabilité. En effet, le grand nombre et le large spectre des variables enregistrés dans ces entrepôts permettent de considérer une grande partie des variables qui pourraient impacter la réponse à un médicament [Stewart *et al.*, 2007]. D'autre part, l'émergence de consoritums, comme OHDSI (Observational Health Data Sciences and Informatics) [Hripcsak *et al.*, 2015], facilite le partage entre sites de schémas de données communs pour les DPE et d'outils d'analyse communs compatibles avec ces schéma. Les schéma et outils ainsi partagés permettent par exemple de considérer dans une analyse multi-sites des variables peu fréquentes à l'échelle d'un seul site, ou de confirmer une observation faite sur un site, chez un ensemble de partenaires du consortium [Hripcsak *et al.*, 2016].

Classiquement, l'étude de l'effet d'un médicament se fait dans le cadre d'essais cliniques randomisés contrôlés (ECR), où une portion de la population est traitée par le médicament étudié et l'autre par un placebo. L'effet du médicament est quantifié par des approches statistiques comme la mesure de l'**effet moyen du traitement** [Rubin, 2005, Pearl, 2010] qui consiste à calculer la différence des moyennes des réponses individuelles observées. Ces méthodes sont adaptées aux ECR mais ne le sont pas pour les études dîtes observationnelles, *i.e.,* à partir de données récoltées indépendamment de l'étude en cours comme c'est le cas pour les DPE. De plus, les méthodes classiques permettent de mesuder un effet moyen, mais pas les effets à l'échelle de sous-groupes particuliers qui pourraient répondre différement au traitement et qu'il est nécessaire d'étudier si l'on souhaite étudier l'hétérogénéité des réponses aux traitements [Ballarini *et al.*, 2018].

Les méthodes d'évaluation de l'**hétérogénéité de l'effet d'un traitement** à partir de données observationnelles s'appuient en général sur des approches contrafactuelles qui utilisent *(i)* des systèmes de prédiction de l'effet du traitement et *(ii)* des méthodes de *patient matching* qui permettent de trouver des groupes de patients similaires. Jusqu'ici les méthodes décrites utilisaient des modèles statistiques linéaires et étaient entraînées sur les données d'ECR uniquement (par exemple [Burke *et al.*, 2014, Tian *et al.*, 2014]). Des résultats récents montrent *(1)* que des modèles d'apprentissage plus sophistiqués caractérisent mieux l'hétérogénéité des effets des traitement à partir les données d'ECR [Rigdon *et al.*, 2018, Wager and Athey, 2018] et *(2)* que l'adaptation de ces méthodes à des études observationnelles, à partir de DPE par exemple, présentent de nombreux challenges méthodologiques et applicatifs [Alaa and van der Schaar, 2018].

Objectifs et défis scientifiques

D'un point de vue informatique, je me donne trois objectifs : étudier l'amélioration de la **détection**, la **prédiction** et la **quantification** de conditions particulières à partir de données observationnelles, et en particulier de la variabilité de réponse aux médicaments. Il s'agit donc essentiellement d'adapter des méthodes au contexte de ces données particulières, et de démontrer leur validité en les appliquant à des cas réels, sur un ou plusieurs vrais jeux de données. Ces trois objectifs sont dépendants les uns des autres, puisque la détection des réponses et de leur caractère variable est indispensable à leur prédiction dans une tâche d'apprentissage supervisé. Ensuite, la structure (*e.g.*, un arbre, une forêt, un réseau de neurones) du modèle associé à la méthode choisie guidera le choix de la méthode d'explication de la prédiction.

Notons que **les données** que je projette d'utiliser sont celles des hôpitaux et des assurances santé. Un des objectifs est de travailler avec les données de collaborateurs, en particulier avec les responsables du Système de Données de Santé Français (SNDS) ou les Départements d'Informatique Médicale (DIM) d'hôpitaux universitaires, comme le CHRU de Nancy, l'Hôpital Européen Georges Pompidou et l'hôpital de Stanford avec lesquels je collabore déjà.

La détection de l'hétérogénéité de la réponse à un médicament dans des DPE. Dans le cas simple et idéal, l'effet d'un médicament est mesurable par une variable systématiquement mesurée après la prise du médicament. C'est le cas par exemple pour les anticoagulants dont l'effet est mesuré par un test de laboratoire appelé l'INR (International Normalized Ratio) qui retourne une valeur indicative de la capacité du sang à coaguler. Ce test est réalisé de façon routinière et à intervalles réguliers après la prescription d'un anticoagulant. Mais l'effet d'un médicament est rarement aussi facile à observer et nécessite de définir, en collaboration avec des experts médecins ou pharmaciens, un ensemble de règles ou un algorithme qui permettent de décrire et caractériser la réponse à un médicament particulier. Ces règles peuvent nécessiter par

exemple de considérer le contenu textuel des DPE, ou la hiérarchie d'un vocabulaire contrôlé utilisé pour encoder les actes médicaux. Le processus de formalisation de ces algorithmes est appelé *phenotyping* et le portail PheKB [Kirby *et al.*, 2016] tâche de les regrouper et de les partager, mais sans pour autant qu'il existe de standard de représentation pour ces algorithmes.

J'aimerais utiliser des méthodes de fouille de texte, d'annotation, les ontologies, ainsi que les standards du consortium OHDSI, pour développer, en collaboration avec des experts, des algorithmes de *phenotyping* qui permettent de capturer de façon précise et reproductible les réponses aux médicaments.

Un élément important à considérer dans ce processus est l'importance de disposer d'un ordre, au moins partiel, entre les descriptions des réponses à un médicament, de manière à pouvoir comparer les réponses, et plus largement les patients.

La prédiction de la réponse. Il existe actuellement deux façons de prédire des conditions. La prédiction vise à utiliser des algorithmes d'apprentissage supervisés, entraînés sur l'histoire des individus, *i.e.*, des variables mesurées préalablement à l'évènement que l'on souhaite prédire, pour constituer un jeux d'apprentissage qui permettra pour un nouvel individu, de prédire la survenue de l'évènement en question, au regard de sa propre histoire.

Un premier défi est la nécessité pour des médicaments rarement prescrits, ou pour des variables responsables de variabilité rarement observée, de considérer les données de plusieurs sites. Pour cela, je prendrai part activement aux groupes de travail du consortium OHDSI, et notamment à celui qui se concentre sur les prédictions au niveau du patient (ou *Patient Level Prediction* en anglais) [OHDSI Consortium Wiki, 2018]. L'objet de cette participation est de permettre de valider une observation faite sur des données locales sur les données d'autres partenaires du consortium OHDSI ou de combiner les résultats obtenus sur plusieurs sites dans une "meta-analyse".

Un second défi est de garantir que les modèles prédictifs que je pourrais développer soient équitables [Barocas *et al.*, 2018, Char *et al.*, 2018]. En effet, dans le contexte d'une étude observationnelle, il est possible qu'un modèle ait des performances différentes pour les sous-groupes identifiés dans une population. Une approche qui nous semble prometteuse pour répondre à ce défi est la méthode proposée par [Hebert-Johnson *et al.*, 2018] qui a l'inconvénient d'être complexe algorithmiquement, mais qui a l'avantage de ne pas faire d'a priori sur la définition des sous-groupes pour lesquels les performances du modèle sont comparées. Un autre avantage est qu'elle s'applique à ne pas diminuer les performances du modèle pour les groupes pour lesquels le modèle est initialement le plus efficace. Un troisième défi est de développer des systèmes prédicifs avec une capacité à fournir des éléments explicatifs aux experts, qui les aideront à prendre la décision finale dans un contexte d'aide à la décision [Ribeiro *et al.*, 2016]. Pour cela je propose d'étudier la combinaison de méthodes d'apprentissage numériques et symboliques, comme par exemple les combinaisons décrites dans [van Krieken *et al.*, 2019] ou [Bosc *et al.*, 2018].

La quantification. Il s'agit ici d'identifier l'hétérogénité dans la réponse au traitement et de quantifier cette hétérogénéité. Nous proposons pour cela d'utiliser des méthodes d'inférence causale [Imbens and Rubin, 2015, Hernán and Robins, 2019], qui cherchent à déterminer si un traitement administré (*i.e.*, une modification quelconque) cause une réponse (*i.e.*, des valeurs particulières pour un ensemble de variables observées) et cela dans un cadre observationnel. Un premier défi pour les méthodes d'inférence causale est de contrôler l'identifiabilité, même partielle de nos modèles, c'est-à-dire si leurs paramètres sont calculables de façon exacte avec un nombre infini d'observations. Les méthodes d'évaluation de l'*effet moyen d'un traitement*

[Rubin, 2005] ne sont identifiables que sous certaines hypothèses, la positivité, la stabilité des unités de traitement et l'ignorabilité, qui sont usuellement vérifiées dans le cadre des ECR, mais pas nécessairement avec des DPE. Il est donc important de considérer ces restrictions dans la conception d'une étude contrafactuelle.

Annexe A

Annexe du Chapitre 1

DO Concept	Deaths	Cause of Mortality Datasheet Label	Datasheet Source
Heart disease	616067	Diseases of the heart	CDC
Cancer	562875	Malignant neoplasms	CDC
Cerebrovascular disorder	135952	Cerebrovascular diseases	CDC
Lower respiratory tract disease	127924	Chronic lower respiratory diseases	CDC
Alzheimer's disease	74632	Alzheimer's disease	CDC
Siabetes mellitus	71382	Diabetes mellitus	CDC
Pneumonia	52717	Influenza and pneumonia	CDC
Nephritis, nephrosis, and nephrotic syndrome	46448	Nephritis, nephrotic syndrome and nephrosis	CDC
Bacterial infectious disease	34828	Septicemia	CDC
Liver cirrhosis	29165	Chronic liver disease and cirhosis	CDC
Hypertension	23965	Essential hypertension and hypertensive renal disease	CDC
Parkinson disease	20058	Parkinson's disease	CDC
Cardiovascular system disease	17072898	Cardiovascular diseases	WHO
Respiratory system disease	8294293	Respiratory infections; Respiratory diseases	WHO
Cancer	7424123	Malignant neoplasms	WHO
Myocardial ischemia	7198257	Ischaemic heart disease	WHO
Cerebrovascular disorder	5712241	Cerebrovascular disease	WHO
Pregnancy complication	3707046	Maternal conditions; Perinatal conditions	WHO
Diarrhea	2163283	Diarrheal diseases	WHO
Acquired immunodeficiency syndrome	2039727	HIV/AIDS	WHO
Tuberculosis	1463792	Tuberculosis	WHO
Diabetes mellitus	1140881	Diabetes mellitus	WHO
Hypertension	986560	Hypertensive heart disease	WHO
Urinary system disease	927591	Diseases of the genitourinary system	WHO
Malaria	889186	Malaria	WHO
Nephritis, nephrosis, and nephrotic syndrome	738908	Nephritis/nephrosis	WHO
Dementia	492390	Alzheimer and other dementias	WHO
Deficiency disease	486762	Nutritional deficiencies	WHO
Measles	423710	Measles	WHO
Meningitis	339945	Meningitis	WHO

FIGURE A.1 – Correspondances entre la *Disease Ontology* et les sources des taux de mortalité de l'OMS et du CDC.

Annexe B

Annexes du Chapitre 2

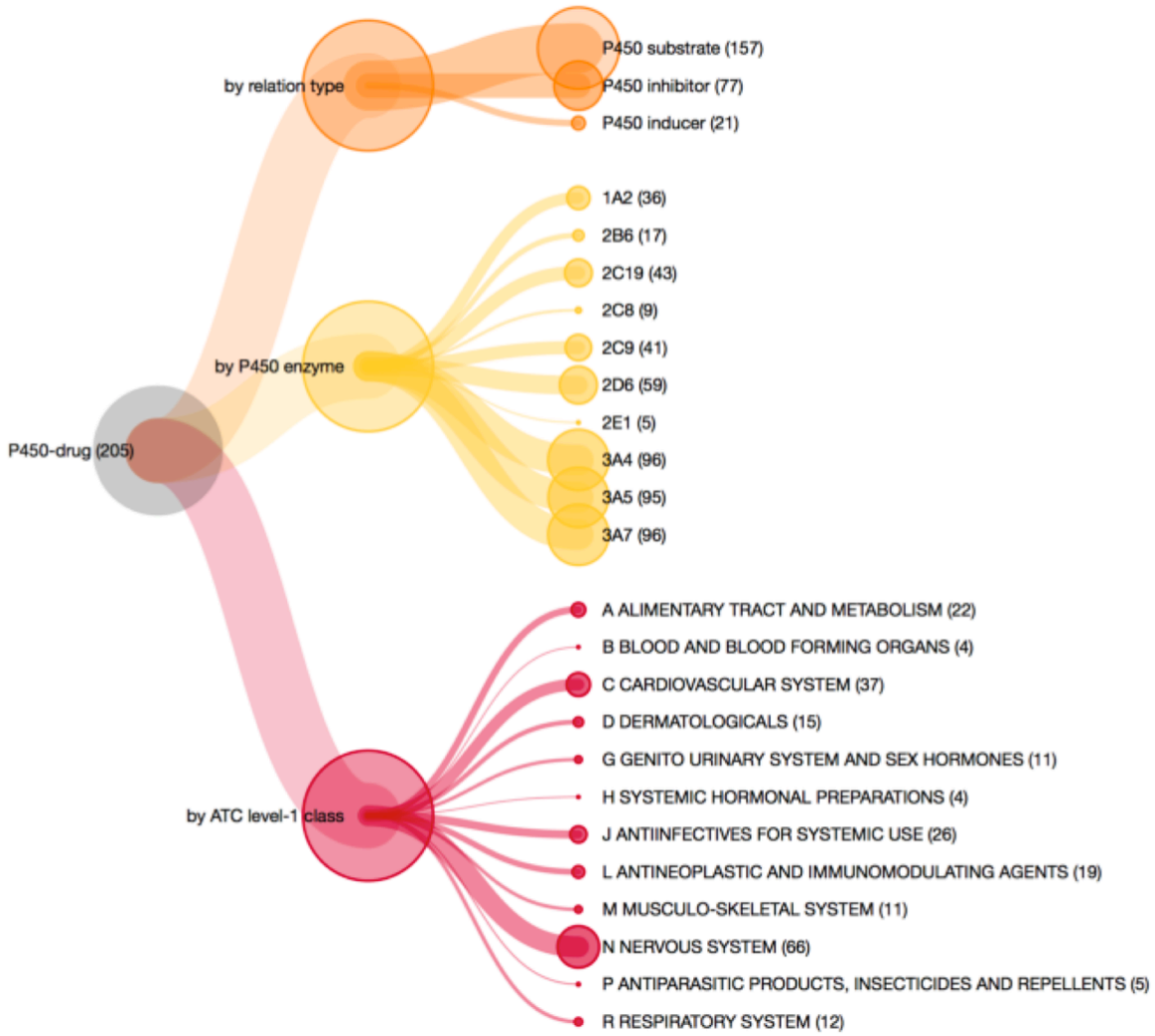


FIGURE B.1 – 25 groupes de médicaments associés aux enzymes P450, organisés selon 3 critères : l'enzyme P450 avec laquelle ils interagissent, le type de relation avec cet enzyme P450 et leur classe ATC (1er niveau seulement). Un médicament peut appartenir à plusieurs groupes. La taille de chaque noeud est proportionnelle à la taille du groupe qui est donné en parenthèse.

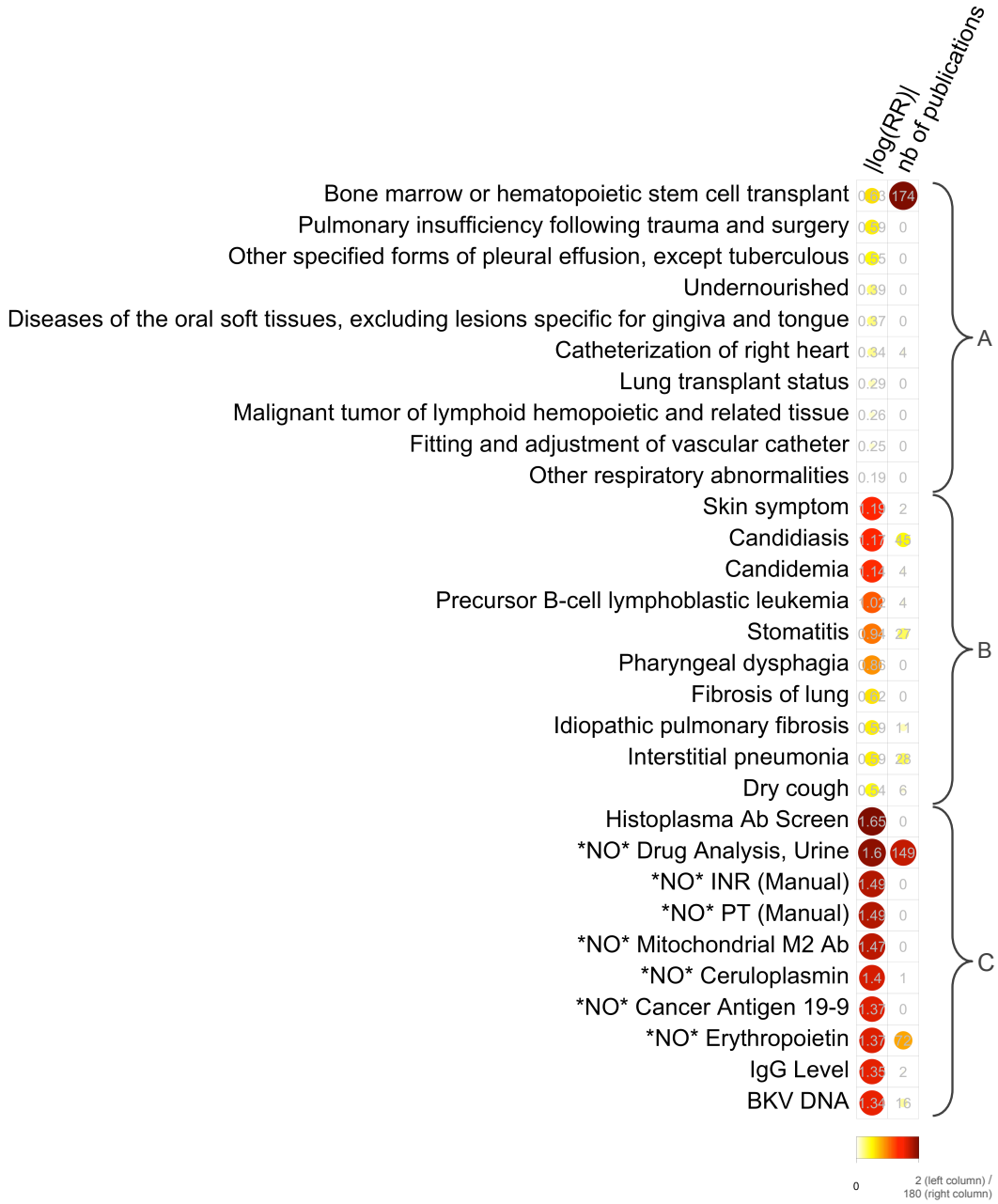


FIGURE B.2 – Exemple de profil phénotypique. L'exemple est réduit aux 10 premiers diagnostics (A), conditions (B), et examens biologiques (C) qui sont observés avant une prescription de *tacrolimus* chez les patients qui nécessitent ensuite une réduction de dose. Chaque phénotype est associé à une valeur p significative (test hypergéométrique, p<0.05, correction de Bonferroni) et est trié selon la valeur des valeurs absolues du logarithme du risque relatif (RR) (représentées dans la première colonne sur une échelle de 0 à 2). Pour aider l'interprétation de ces profils, le nombre d'articles de PubMed qui mentionnent à la fois le médicament (tacrolimus) et le phénotype sont fournis (seconde colonne, sur une échelle de 0 à 180). Par exemple 45 articles mentionnent à la fois le tacrolimus et la candidose (*candidiasis*). (A) Les diagnostics sont les codes ICD-9-CM associés avec les visites de patients ; (B) les conditions sont les phénotypes mentionnés dans le texte de notes cliniques ; (C) les examens de laboratoires commandés. Certains noms d'examens sont préfixés avec "*NO*" pour indiquer une relation négative (RR<1) entre l'examen de laboratoire est la réduction de dose de tacrolimus.

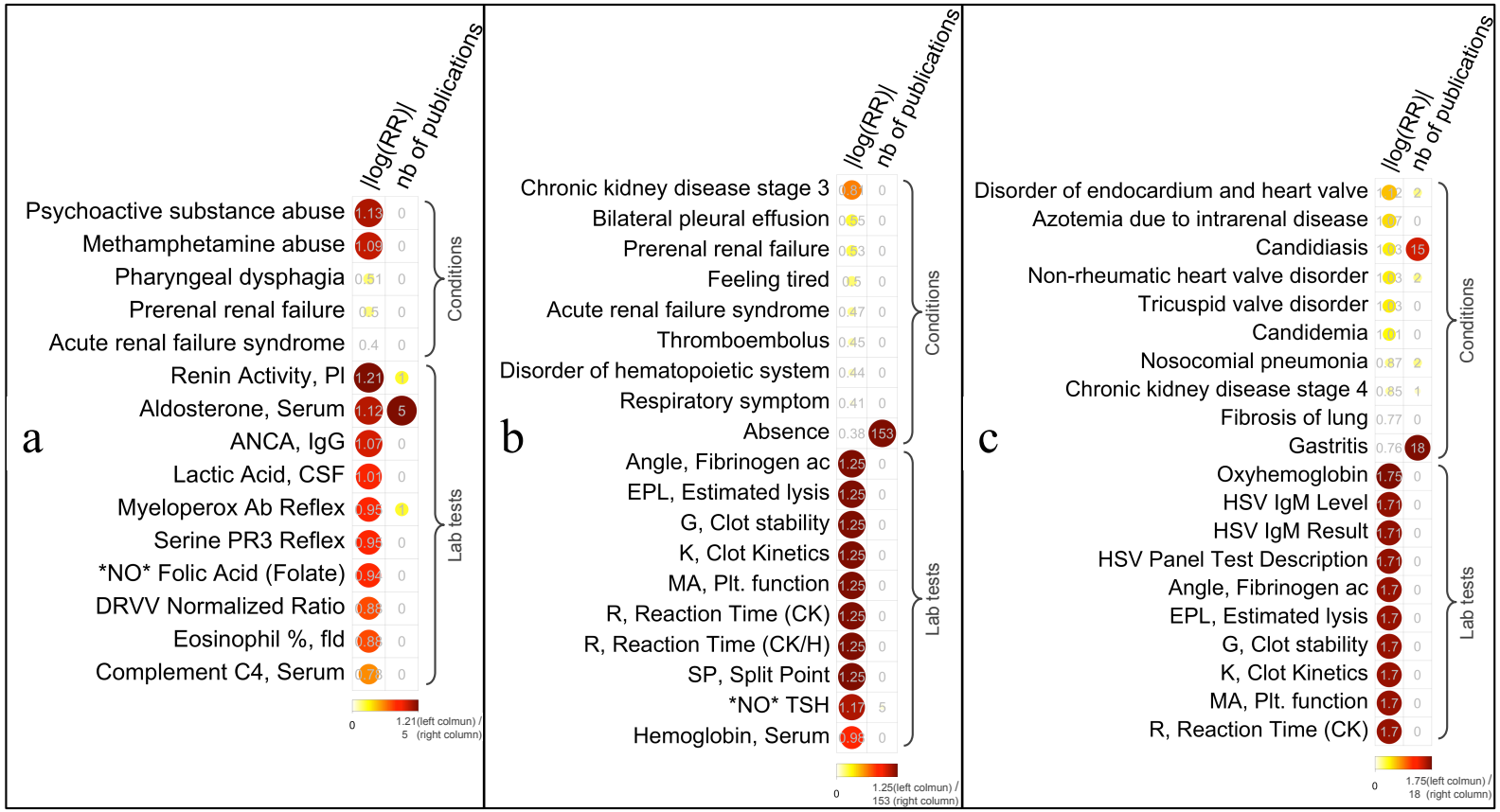


FIGURE B.3 – Trois exemples de profils phénotypiques associés avec les réductions de doses des *labetalol* (a), *sildenafil* (b) et *warfarin* (c). Voir la Figure B.2 pour plus d'indications sur la lecture des profils phénotypiques.

Annexe C

Annexes du Chapitre 3

C.1 Définitions des règles de réconciliation définies sur PGxLOD

C.1.1 Notations

Les instances du concept `PharmacogenomicRelationship` associent trois types de composants : des instances de `Drugs`, `GeneticFactors` et `Phenotypes`.

Soit `r`, une instance de `PharmacogenomicRelationship` dans une base de connaissance $\mathcal{KB}$, nous définissons les ensembles d'individus et de concepts associés à `r` comme suit :

Notation 4 *Soit D, l'ensemble des instances de* `Drug` *qui causent* `r`*, défini comme*

$$D = \{\mathtt{d} \mid \mathcal{KB} \vDash \mathtt{Drug(d)} \text{ and } \mathcal{KB} \vDash \mathtt{causes(d,r)}\}$$

Notation 5 *Soit G, l'ensemble des instances de*`GeneticFactor` *qui causent* `r`*, défini comme*

$$G = \{\mathtt{g} \mid \mathcal{KB} \vDash \mathtt{GeneticFactor(g)} \text{ and } \mathcal{KB} \vDash \mathtt{causes(g,r)}\}$$

Notation 6 *Soit P, l'ensemble des instances de* `Phenotype` *causés par* `r`*, défini comme*

$$P = \{\mathtt{p} \mid \mathcal{KB} \vDash \mathtt{Phenotype(p)} \text{ and } \mathcal{KB} \vDash \mathtt{causes(r,p)}\}$$

Notation 7 *Notons DC, l'ensemble des concepts instanciés par les membres de D. DC est défini comme suit :*

$$DC = \{\mathcal{C} \mid \forall\ \mathtt{d} \in D,\ \mathcal{KB} \vDash \mathcal{C}(\mathtt{d})\}$$

Notation 8 *Notons PC, l'ensemble des concepts instanciés par les membres de P. PC est défini comme suit :*

$$PC = \{\mathcal{C} \mid \forall\ \mathtt{p} \in P,\ \mathcal{KB} \vDash \mathcal{C}(\mathtt{p})\}$$

Notation 9 *Notons GHP, l'ensemble des instances de* `GeneticFactor` *associées par le prédicat* `hasPart` *aux individus de G. GHP est défini comme suit :*

$$GHP = \{\mathtt{g} \mid \mathcal{KB} \vDash \mathtt{GeneticFactor(g)} \text{ and } \exists\ v \in G,\ \mathcal{KB} \vDash \mathtt{hasPart(g,v)}\}$$

Intuitivement, GHP contient les gènes dont les variants sont dans G.

Notation 10 *Notons DOP, l'enemsble des individus associés par le prédicat* $\texttt{dependsOn}^-$ *aux individuas de P. DOP est défini comme suit :*

$$DOP = \{\texttt{e} \mid \exists\ p \in P,\ \mathcal{KB} \vDash \texttt{dependsOn}(\texttt{p}, \texttt{e})\}$$

C.1.2 Définitions

Nous considérons $\texttt{r}_1$ et $\texttt{r}_2$ deux instances du concept `PharmacogenomicRelationship`. Suivant les notations définie dans la section précédente, nous considérons les ensembles d'individus et de classes qui leur sont associé D_1, G_1, P_1, DC_1, PC_1, GHP_1 et DOP_1 (respectivement D_2, G_2, P_2, DC_2, PC_2, GHP_2 et DOP_2).

Quand $\texttt{r}_1$ et $\texttt{r}_2$ sont équivalents.
Règle 1

$$D_1 = D_2 \quad G_1 = G_2 \quad P_1 = P_2 \Rightarrow \texttt{owl:sameAs}(\texttt{r}_1, \texttt{r}_2)$$

Quand $\texttt{r}_1$ est plus spécifique que $\texttt{r}_2$
Règle 2 (Quand les trois types de composants existent)

$$\begin{array}{r} [D_1 \neq \varnothing \quad G_1 \neq \varnothing \quad P_1 \neq \varnothing] \\ [D_1 \subseteq D_2 \quad (DC_2 \neq \varnothing \quad DC_2 \subseteq DC_1)] \\ [G_1 \subseteq G_2] \\ [P_1 \subseteq P_2 \quad (PC_2 \neq \varnothing \quad PC_2 \subseteq PC_1)] \end{array} \Rightarrow \texttt{skos:broadMatch}(\texttt{r}_1, \texttt{r}_2)$$

Règle 3 (Quand un type de composant manque)

$$\begin{array}{r} \Big([D_2 = \varnothing \quad G_1 \neq \varnothing \quad P_1 \neq \varnothing] \\ [G_1 \subseteq G_2] \\ [P_1 \subseteq P_2 \quad (PC_2 \neq \varnothing \quad PC_2 \subseteq PC_1)] \Big) \\ \Big([D_1 \neq \varnothing \quad G_2 = \varnothing \quad P_1 \neq \varnothing] \\ [D_1 \subseteq D_2 \quad (DC_2 \neq \varnothing \quad DC_2 \subseteq DC_1)] \\ [P_1 \subseteq P_2 \quad (PC_2 \neq \varnothing \quad PC_2 \subseteq PC_1)] \Big) \\ \Big([D_1 \neq \varnothing \quad G_1 \neq \varnothing \quad P_2 = \varnothing] \\ [D_1 \subseteq D_2 \quad (DC_2 \neq \varnothing \quad DC_2 \subseteq DC_1)] \\ [G_1 \subseteq G_2] \Big) \end{array} \Rightarrow \texttt{skos:broadMatch}(\texttt{r}_1, \texttt{r}_2)$$

Règle 4 (Quand $\mathtt{r_1}$ est décrit au niveau du variant et $\mathtt{r_2}$ au niveau du gène)

$$\Big[GHP_1 \neq \varnothing \;\; GHP_1 \subseteq G_2\Big]$$
$$\Big[\Big(\;[D_2 = \varnothing \;\; P_1 \neq \varnothing]$$
$$[P_1 \subseteq P_2 \;\; (PC_2 \neq \varnothing \;\; PC_2 \subseteq PC_1)]\;\Big)$$
$$\Big(\;[D_1 \neq \varnothing \;\; P_2 = \varnothing]$$
$$[D_1 \subseteq D_2 \;\; (DC_2 \neq \varnothing \;\; DC_2 \subseteq DC_1)]\;\Big)$$
$$\Big(\;[D_1 \neq \varnothing \;\; P_1 \neq \varnothing]$$
$$[D_1 \subseteq D_2 \;\; (DC_2 \neq \varnothing \;\; DC_2 \subseteq DC_1)]$$
$$[P_1 \subseteq P_2 \;\; (PC_2 \neq \varnothing \;\; PC_2 \subseteq PC_1)]\;\Big)\Big] \Rightarrow \mathtt{skos:broadMatch(r_1, r_2)}$$

Quand $\mathtt{r_1}$ et $\mathtt{r_2}$ sont reliés

Règle 5

$$DOP_1 \neq \varnothing$$
$$[DOP_1 = DOP_2 \;\; DOP_1 = D_2 \;\; DOP_1 = G_2]$$
$$[(G_1 \neq \varnothing \;\; G_1 = G_2) \;\; (D_1 \neq \varnothing \;\; D_1 = D_2)] \;\Rightarrow \mathtt{skos:relatedMatch(r_1, r_2)}$$

C.2 Exemples d'application des règles

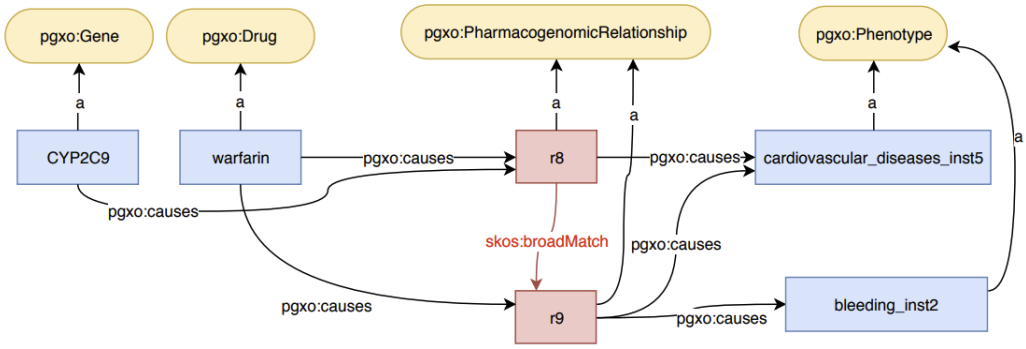


FIGURE C.1 – Exemple d'application de la règle 3.

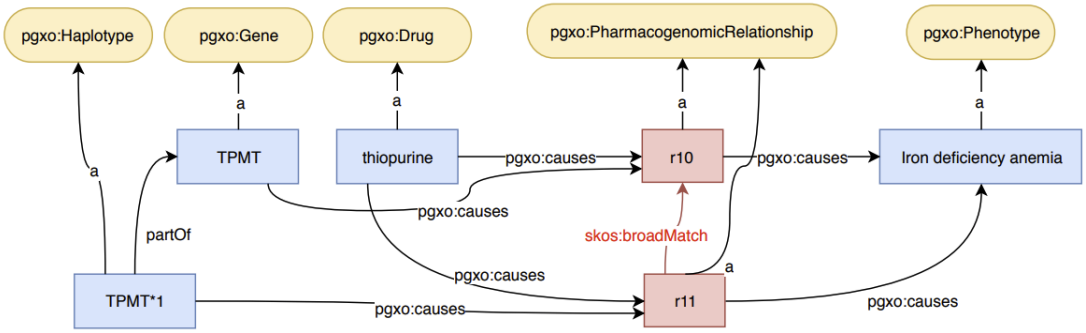


FIGURE C.2 – Exemple d'application de la règle 4.

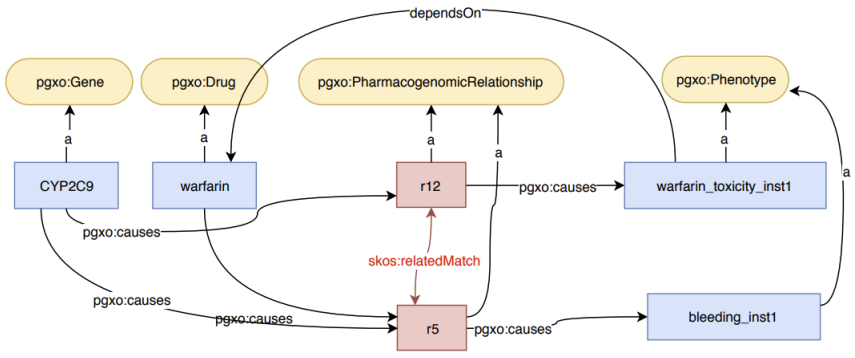


FIGURE C.3 – Exemple d'application de la règle 5.

Bibliographie

[Aggarwal and Han, 2014] Charu C. Aggarwal and Jiawei Han, editors. *Frequent Pattern Mining.* Springer, 2014.

[Alaa and van der Schaar, 2018] Ahmed Alaa and Mihaela van der Schaar. Limits of estimating heterogeneous treatment effects : Guidelines for practical algorithm design. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 129–138, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[Alam *et al.*, 2015] Mehwish Alam, Aleksey Buzmakov, Víctor Codocedo, and Amedeo Napoli. Mining definitions from RDF annotations using formal concept analysis. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pages 823–829, 2015.

[Alam, 2015] Mehwish Alam. *Interactive Knowledge Discovery over Web of Data. (Découverte interactive de connaissances dans le web des données).* PhD thesis, University of Lorraine, Nancy, France, 2015.

[Alexa *et al.*, 2006] Adrian Alexa, Jörg Rahnenführer, and Thomas Lengauer. Improved scoring of functional groups from gene expression data by decorrelating go graph structure. *Bioinformatics*, 22(13) :1600–1607, 2006.

[Amores, 2013] Jaume Amores. Multiple instance classification : Review, taxonomy and comparative study. *Artif. Intell.*, 201 :81–105, 2013.

[Andreasen *et al.*, 2011] Troels Andreasen, Henrik Bulskov, Sine Zambach, Tine Lassen, Bodil Nistrup Madsen, Per Anker Jensen, Hanne Erdman Thomsen, and Jørgen Fischer Nilsson. A semantics-based approach to retrieving biomedical information. In *Flexible Query Answering Systems - 9th International Conference, FQAS 2011, Ghent, Belgium, October 26-28, 2011 Proceedings*, pages 108–118, 2011.

[Antezana *et al.*, 2009] Erick Antezana, Martin Kuiper, and Vladimir Mironov. Biological knowledge management : the emerging role of the Semantic Web technologies. *Briefings in Bioinformatics*, 2009.

[Ashburner *et al.*, 2000] M Ashburner, C A Ball, J A Blake, D Botstein, H Butler, J M Cherry, A P Davis, K Dolinski, S S Dwight, J T Eppig, M A Harris, D P Hill, L Issel-Tarver, A Kasarskis, S Lewis, J C Matese, J E Richardson, M Ringwald, G M Rubin, and G Sherlock. Gene ontology : tool for the unification of biology. the gene ontology consortium. *Nature Genetics*, 25(1) :25–29, May 2000.

[Atencia *et al.*, 2019] Manuel Atencia, Jérôme David, Jérôme Euzenat, Amedeo Napoli, and Jérémy Vizzini. Link key candidate extraction with relational concept analysis. *Discrete Applied Mathematics*, 2019.

[Baader *et al.*, 1999] Franz Baader, Ralf Küsters, and Ralf Molitor. Computing least common subsumers in description logics with existential restrictions. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, IJCAI 99, Stockholm, Sweden, July 31 - August 6, 1999. 2 Volumes, 1450 pages*, pages 96–103, 1999.

[Baader *et al.*, 2010] Franz Baader, Diego Calvanese, Deborah L. McGuinness, Daniele Nardi, and Peter F. Patel-Schneider. *The Description Logic Handbook : Theory, Implementation and Applications*. Cambridge University Press, New York, NY, USA, 2nd edition, 2010.

[Ballarini *et al.*, 2018] Nicolas M. Ballarini, Gerd K. Rosenkranz, Thomas Jaki, Franz Konig, and Martin Posch. Subgroup identification in clinical trials via the predicted individual treatment effect. *PLOS ONE*, 13(10) :1–22, 10 2018.

[Banda *et al.*, 2016] Juan Banda, Lee Evans, Rami Vanguri, Patrick Ryan, and Nigam Shah. A curated and standardized adverse drug event resource to accelerate drug safety research. *Scientific Data*, 3 :160026, 05 2016.

[Barbut and Monjardet, 1970] M. Barbut and B. Monjardet, editors. *Ordres et classification : Algèbre et combinatoire (tome II)*. Hachette, Paris, 1970.

[Barocas *et al.*, 2018] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2018. `http://www.fairmlbook.org`.

[Barocas *et al.*, 2019] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. `http://www.fairmlbook.org`.

[Bates *et al.*, 2003] David Bates, Gilad Kuperman, Samuel Wang, Tejal Gandhi, Anne Kittler, Lynn Volk, Cynthia Spurr, Ramin Khorasani, Milenko Tanasijevic, and Blackford Middleton. Ten commandments for effective clinical decision support : Making the practice of evidence-based medicine a reality. *Journal of the American Medical Informatics Association : JAMIA*, 10 :523–30, 11 2003.

[Bengio *et al.*, 2003] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3 :1137–1155, March 2003.

[Berners-Lee *et al.*, 2001] Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web : A new form of web content that is meaningful to computers will unleash a revolution of new possibilities. *ScientificAmerican.com*, 05 2001.

[Bizer *et al.*, 2009] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data - the story so far. *Int. J. Semantic Web Inf. Syst.*, 5(3) :1–22, 2009.

[Bodenreider, 2004] Olivier Bodenreider. The Unified Medical Language System (UMLS) : integrating biomedical terminology. *Nucleic Acids Research*, 32(Database-Issue) :267–270, 2004.

[Bordes *et al.*, 2011] Antoine Bordes, Jason Weston, Ronan Collobert, and Yoshua Bengio. Learning structured embeddings of knowledge bases. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2011, San Francisco, California, USA, August 7-11, 2011*, 2011.

[Bosc *et al.*, 2018] Guillaume Bosc, Jean-François Boulicaut, Chedy Raïssi, and Mehdi Kaytoue. Anytime discovery of a diverse set of patterns with monte carlo tree search. *Data Min. Knowl. Discov.*, 32(3) :604–650, 2018.

[Breiman, 2001] Leo Breiman. Random forests. *Machine Learning*, 45(1) :5–32, October 2001.

[Burke *et al.*, 2014] James F. Burke, Rodney A. Hayward, Jason P. Nelson, and David M. Kent. Using internally developed risk models to assess heterogeneity in treatment effects in clinical trials. *Circulation : Cardiovascular Quality and Outcomes*, 7(1) :163–169, 2014.

[Callahan *et al.*, 2013] Alison Callahan, Jose Cruz-Toledo, Peter Ansell, and Michel Dumontier. Bio2rdf release 2 : Improved coverage, interoperability and provenance of life science linked data. In *ESWC 2013*, pages 200–12, 2013.

[Canuel *et al.*, 2015] Vincent Canuel, Bastien Rance, Paul Avillach, Patrice Degoulet, and Anita Burgun. Translational research platforms integrating clinical and omics data : a review of publicly available solutions. *Briefings in Bioinformatics*, 16(2) :280–290, 2015.

[Cawley and Talbot, 2010] Gavin C Cawley and Nicola LC Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research*, 11 :2079–2107, 2010.

[Char *et al.*, 2018] Danton S. Char, Nigam H. Shah, and David Magnus. Implementing Machine Learning in Health Care ? Addressing Ethical Challenges. *New England Journal of Medicine*, 378(11) :981–983, mar 2018.

[Cheatham *et al.*, 2017] Michelle Cheatham, Isabel F. Cruz, Jérôme Euzenat, and Catia Pesquita. Special issue on ontology and linked data matching. *Semantic Web*, 8(2) :183–184, 2017.

[Cimiano *et al.*, 2004] Philipp Cimiano, Andreas Hotho, Gerd Stumme, and Julien Tane. Conceptual knowledge processing with formal concept analysis and ontologies. In Peter Eklund, editor, *Concept Lattices*, pages 189–207, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg.

[Collobert *et al.*, 2011] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12 :2493–2537, November 2011.

[Consortium, 2018] The UniProt Consortium. UniProt : a worldwide hub of protein knowledge. *Nucleic Acids Research*, 47(D1) :D506–D515, 11 2018.

[Coulet and Smaïl-Tabbone, 2016] Adrien Coulet and Malika Smaïl-Tabbone. Mining Electronic Health Records to Validate Knowledge in Pharmacogenomics. *ERCIM News*, (104) :56, 2016.

[Coulet *et al.*, 2006] Adrien Coulet, Malika Smaïl-Tabbone, Amedeo Napoli, and Marie-Dominique Devignes. Suggested ontology for pharmacogenomics (so-pharm) : Modular construction and preliminary testing. In *OTM Workshops (1)*, volume 4277 of *LNCS series*, pages 648–657. Springer, 2006.

[Coulet *et al.*, 2010] Adrien Coulet, Nigam H. Shah, Yael Garten, Mark A. Musen, and Russ B. Altman. Using text to build semantic networks for pharmacogenomics. *Journal of Biomedical Informatics*, 43(6) :1009–1019, 2010.

[Coulet *et al.*, 2011] Adrien Coulet, Yael Garten, Michel Dumontier, Russ B Altman, Mark A Musen, and Nigam H Shah. Integration and publication of heterogeneous text-mined relationships on the Semantic Web. *Journal of Biomedical Semantics*, 2(Suppl 2) :S10, 2011.

[Coulet *et al.*, 2013] Adrien Coulet, Florent Domenach, Mehdi Kaytoue, and Amedeo Napoli. Using Pattern Structures for Analyzing Ontology-Based Annotations of Biomedical Data. In *International Conference on Formal Concept Analysis*, LNCS/LNAI series, Dresden, Germany, May 2013. Springer.

[Coulet *et al.*, 2018] Adrien Coulet, Nigam H. Shah, Maxime Wack, Mohammad Chawki, Nicolas Jay, and Michel Dumontier. Predicting the need for a reduced drug dose, at first prescription. *Scientific Reports*, 8(1), 2018.

[Coulet, 2008] Adrien Coulet. *Construction et utilisation d'une base de connaissances pharmacogénomique pour l'intégration de données et la découverte de connaissances. (Construction and use of a pharmacogenomic knowledge base for data integration and knowledge discovery).* PhD thesis, Henri Poincaré University, Nancy, France, 2008.

[Dai *et al.*, 2008] Manhong Dai, Nigam H Shah, Wei Xuan, Mark A Musen, Stanley J Watson, Brian D Athey, Fan Meng, et al. An efficient solution for mapping free text to ontology terms. *AMIA summit on translational bioinformatics*, 21, 2008.

[Dalleau *et al.*, 2015] Kevin Dalleau, Ndeye Coumba Ndiaye, and Adrien Coulet. Suggesting valid pharmacogenes by mining linked data. In *International Conference on the Semantic Web Applications and Tools for Life Sciences (SWAT4LS) 2015*, Cambridge, United Kingdom, 2015.

[Dalleau *et al.*, 2017] Kevin Dalleau, Yassine Marzougui, Sébastien Da Silva, Patrice Ringot, Ndeye Coumba Ndiaye, and Adrien Coulet. Learning from biomedical linked data to suggest valid pharmacogenes. *Journal of Biomedical Semantics*, 8(29), 2017.

[Dao *et al.*, 2018] Tri Dao, Albert Gu, Alexander J. Ratner, Virginia Smith, Christopher De Sa, and Christopher Ré. A kernel theory of modern data augmentation. *CoRR*, abs/1803.06084, 2018.

[d'Avila Garcez *et al.*, 2019] A. S. d'Avila Garcez, M. Gori, L. C. Lamb, L. Serafini, M. Spranger, and S. N. Tran. Neural-symbolic computing : An effective methodology for principled integration of machine learning and reasoning. *FLAP*, 6(4) :611–632, 2019.

[de Vries and de Rooij, 2015] Gerben Klaas Dirk de Vries and Steven de Rooij. Substructure counting graph kernels for machine learning from RDF data. *J. Web Sem.*, 35 :71–84, 2015.

[Denny *et al.*, 2017] Joshua Denny, Sara Van Driest, Wei-Qi Wei, and Dan Roden. The influence of big (clinical) data and genomics on precision medicine and drug development. *Clinical Pharmacology & Therapeutics*, 103(3) :409–418, 2017.

[Domenach and Portides, 2014] Florent Domenach and George Portides. Similarity measures on concept lattices. In *Analysis of Large and Complex Data - Second European Conference on Data Analysis, ECDA 2014, Bremen, Germany, July 2-4, 2014*, pages 159–169, 2014.

[Domenach, 2013] Florent Domenach. Similarity measures of concept lattices. In *Data Science, Learning by Latent Structures, and Knowledge Discovery [revised versions of selected papers presented during the European Conference on Data Analysis (ECDA 2013), Luxembourg, July 2013]*., pages 89–99, 2013.

[Domenach, 2017] Florent Domenach. Mesure de similarité entre treillis basée sur des correspondances explicites. In *17ème Journées Francophones Extraction et Gestion des Connaissances, EGC 2017, 24-27 Janvier 2017, Grenoble, France*, pages 363–368, 2017.

[Draghici *et al.*, 2003] Sorin Draghici, Purvesh Khatri, Pratik Bhavsar, Abhik Shah, Stephen A. Krawetz, and Michael A. Tainsky. Onto-Tools, the toolkit of the modern biologist : Onto-Express, Onto-Compare, Onto-Design and Onto-Translate. *Nucleic Acids Research*, 31(13) :3775–3781, 07 2003.

[Ehrlinger and Wöß, 2016] Lisa Ehrlinger and Wolfram Wöß. Towards a definition of knowledge graphs. In *Joint Proceedings of the Posters and Demos Track of the 12th International Conference on Semantic Systems - SEMANTiCS2016 and the 1st International Workshop on Semantic Change & Evolving Semantics (SuCCESS'16) co-located with the 12th International Conference on Semantic Systems (SEMANTiCS 2016), Leipzig, Germany, September 12-15, 2016.*, 2016.

[Euzenat and Shvaiko, 2013] Jérôme Euzenat and Pavel Shvaiko. *Ontology Matching, Second Edition*. Springer, 2013.

[Fayyad *et al.*, 1996] Usama M. Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery : An overview. In *Advances in Knowledge Discovery and Data Mining*, pages 1–34. 1996.

[Fernandez-Lopez and Corcho, 2010] Mariano Fernandez-Lopez and Oscar Corcho. *Ontological Engineering : With Examples from the Areas of Knowledge Management, e-Commerce and the Semantic Web. First Edition*. Springer Publishing Company, Incorporated, 2010.

[Flockart, 2007] DA Flockart. Drug interactions : Cytochrome p450 drug interaction table, 2007. (visitée le 18 octobre 2019.).

[Frey *et al.*, 2016] Lewis J Frey, Elmer V Bernstam, and Joshua C Denny. Precision medicine informatics. *Journal of the American Medical Informatics Association*, 23(4) :668–670, 06 2016.

[Ganter and Kuznetsov, 2001] Bernhard Ganter and Sergei O. Kuznetsov. Pattern structures and their projections. In *Conceptual Structures : Broadening the Base, 9th International Conference on Conceptual Structures, ICCS 2001, Stanford, CA, USA, July 30-August 3, 2001, Proceedings*, pages 129–142, 2001.

[Ganter and Wille, 1999] Bernhard Ganter and Rudolf Wille. *Formal Concept Analysis - Mathematical Foundations*. Springer, 1999.

[Garten, 2010] Yael Garten. *Text mining of the scientific literature to identify pharmacogenomic interactions*. PhD thesis, Stanford University, California, 2010.

[Ghazvinian *et al.*, 2009] Amir Ghazvinian, Natalya Fridman Noy, Clément Jonquet, Nigam H. Shah, and Mark A. Musen. What four million mappings can tell you about two hundred ontologies. In *The Semantic Web - ISWC 2009, 8th International Semantic Web Conference, ISWC 2009, Chantilly, VA, USA, October 25-29, 2009. Proceedings*, pages 229–242, 2009.

[Ghazvinian *et al.*, 2011] Amir Ghazvinian, Natasha Noy, and Mark Musen. How orthogonal are the obo foundry ontologies ? *Journal of biomedical semantics*, 2 Suppl 2 :S2, 05 2011.

[Goldman and Goldman, 2019] Lee Goldman and Jill S. Goldman. Precision Medicine for Clinicians : The Future Begins NowPrecision Medicine for Clinicians : The Future Begins Now. *Annals of Internal Medicine*, 170(9) :660–661, 05 2019.

[Gombar *et al.*, 2019] Saurabh Gombar, Alison Callahan, Robert Califf, Robert Harrington, and Nigam H. Shah. It is time to learn from patients like mine. *npj Digital Medicine*, 2(16), 2019.

[Grissa *et al.*, 2016] Dhouha Grissa, Blandine Comte, Estelle Pujos-Guillot, and Amedeo Napoli. A hybrid knowledge discovery approach for mining predictive biomarkers in metabolomic data. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I*, pages 572–587, 2016.

[Hassan *et al.*, 2015] Mohsen Hassan, Olfa Makkaoui, Adrien Coulet, and Yannick Toussain. Extracting disease-symptom relationships by learning syntactic patterns from dependency graphs. In *Proceedings of BioNLP 15*, pages 71–80, Beijing, China, 2015. Association for Computational Linguistics.

[Hassan, 2017] Mohsen Hassan. *Knowledge Discovery Considering Domain Literature and Ontologies : Application to Rare Diseases. (Découverte de connaissances considérant la littérature et les ontologies de domaine : application aux maladies rares)*. PhD thesis, University of Lorraine, Nancy, France, 2017.

[Hebert-Johnson *et al.*, 2018] Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration : Calibration for the (Computationally-identifiable) masses. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1939–1948, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[Hernán and Robins, 2019] Miguel A. Hernán and James M. Robins. *Causal Inference*. Chapman & Hall/CRC, Boca Raton, FL, USA, 2019.

[Holm, 1979] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

[Hripcsak *et al.*, 2015] George Hripcsak, Jon D Duke, Nigam Shah, Christian G Reich, Vojtech Huser, Martijn J Schuemie, Marc A Suchard, Rae Woong Park, Ian Chi Kei Wong, Peter Rijnbeek, Johan Lei, Nicole Pratt, Niklas NorÃ©n, Yu-Chuan Li, Paul E Stang, David Madigan, and Patrick B Ryan. Observational health data sciences and informatics (ohdsi) : Opportunities for observational researchers. *Studies in health technology and informatics*, 216 :574–8, 08 2015.

[Hripcsak *et al.*, 2016] G. Hripcsak, P.B. Ryan, J. D. Duke, N. H. Shah, R. W. Park, V. Huser, M. A. Suchard, M. J. Schuemie, F. J. DeFalco, A. Perotte, J. M. Banda, C. G. Reich, L. M.Schilling, M.E. Matheny, D. Meeker, N. Pratt, and D. Madigan. Characterizing treatment pathways at scale using the ohdsi network. *Proceedings of the National Academy of Sciences*, 113(27) :7329–7336, 2016.

[Imanishi and Nakaoka, 2009] Tadashi Imanishi and Hajime Nakaoka. Hyperlink Management System and ID Converter System : enabling maintenance-free hyperlinks among major biological databases. *Nucleic Acids Research*, 37(Web Server issue) :W17–W22, July 2009.

[Imbens and Rubin, 2015] Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences : An Introduction*. Cambridge University Press, New York, NY, USA, 2015.

[Inlow and Restifo, 2004] Jennifer K Inlow and Linda L Restifo. Molecular and comparative genetics of mental retardation. *Genetics*, 166(2) :835–881, 2004.

[Ioannidis, 2013] John P.A. Ioannidis. To replicate or not to replicate : The case of pharmacogenetic studies. *Circulation : Cardiovascular Genetics*, 6 :413–8, 2013.

[Jameson and Longo, 2015] J. Larry Jameson and Dan L. Longo. Precision medicine — personalized, problematic, and promising. *New England Journal of Medicine*, 372(23) :2229–2234, 2015.

[Jensen *et al.*, 2012] Peter B. Jensen, Lars J. Jensen, and Soren Brunak. Mining electronic health records : towards better research applications and clinical care. *Nat Rev Genet.*, 13(6) :395–405, 2012.

[Jonquet *et al.*, 2009] Clement Jonquet, Nigam H. Shah, and Mark A. Musen. The Open Biomedical Annotator. In *American Medical Informatics Association Symposium on Translational BioInformatics, AMIA-TBI'09*, pages 56–60, San Francisco, CA, USA, March 2009.

[Jonquet *et al.*, 2010] Clement Jonquet, Adrien Coulet, Nigam H. Shah, and Mark A. Musen. Indexation et intégration de ressources textuelles á l'aide d'ontologies : application au domaine biomédical. In *21èmes Journées Francophones d'Ingénierie des Connaissances, IC'10*, pages 271–282, 2010.

[Jonquet *et al.*, 2011] Clement Jonquet, Paea LePendu, Sean M. Falconer, Adrien Coulet, Natalya Fridman Noy, Mark A. Musen, and Nigam H. Shah. Ncbo resource index : Ontology-based search and mining of biomedical resources. *Journal of Web Semantics*, 9(3) :316–324, 2011.

[Jung and Shah, 2015] Kenneth Jung and Nigam H Shah. Implications of non-stationarity on predictive modeling using ehrs. *Journal of biomedical informatics*, 58 :168–174, 2015.

[Jung *et al.*, 2014] Kenneth Jung, Paea LePendu, William S. Chen, Srinivasan V. Iyer, Ben Readhead, Joel T. Dudley, and Nigam H. Shah. Automated detection of off-label drug use. *PLOS ONE*, 9(2) :1–9, 02 2014.

[Jupp *et al.*, 2008] Simon Jupp, Sean Bechhofer, and Robert Stevens. SKOS with OWL : don't be full-ish ! In *Proceedings of the Fifth OWLED Workshop on OWL : Experiences and Directions, collocated with the 7th International Semantic Web Conference (ISWC-2008), Karlsruhe, Germany, October 26-27, 2008*, 2008.

[Jupp *et al.*, 2014] Simon Jupp, James Malone, Jerven Bolleman, Marco Brandizi, Mark Davies, Leyla J. Garcia, Anna Gaulton, Sebastien Gehant, Camille Laibe, Nicole Redaschi, Sarala M. Wimalaratne, Maria Jesus Martin, Nicolas Le Novère, Helen E. Parkinson, Ewan Birney, and Andrew M. Jenkinson. The EBI RDF platform : linked open data for the life sciences. *Bioinformatics*, 30(9) :1338–1339, 2014.

[Kawai *et al.*, 2014] V. K. Kawai, A. Cunningham, S. I. Vear, S. L. Van Driest, A. Oginni, H. Xu, M. Jiang, C. Li, J. C. Denny, C. Shaffer, E. Bowton, B. F. Gage, W. A. Ray, D. M. Roden, and C. M. Stein. Genotype and risk of major bleeding during warfarin treatment. *Pharmacogenomics*, 15(16) :1973–1983, Dec 2014.

[Kayser, 1997] Daniel Kayser. *La représentation des connaissances*. Hermès, 1997.

[Kaytoue *et al.*, 2011a] M. Kaytoue, S.O. Kuznetsov, A. Napoli, and S. Duplessis. Mining Gene Expression Data with Pattern Structures in Formal Concept Analysis. *Information Science*, 181(10) :1989–2001, 2011.

[Kaytoue *et al.*, 2011b] Mehdi Kaytoue, Sergei O. Kuznetsov, and Amedeo Napoli. Revisiting numerical pattern mining with formal concept analysis. In *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011*, pages 1342–1347, 2011.

[Kinjo *et al.*, 2012] Akira R. Kinjo *et al.* Protein Data Bank Japan (PDBj) : maintaining a structural data archive and resource description framework format. *Nucleic Acids Research*, 2012.

[Kirby *et al.*, 2016] Jacqueline C Kirby, Peter Speltz, Luke V Rasmussen, Melissa Basford, Omri Gottesman, Peggy L Peissig, Jennifer A Pacheco, Gerard Tromp, Jyotishman Pathak, David S Carrell, Stephen B Ellis, Todd Lingren, Will K Thompson, Guergana Savova, Jonathan Haines, Dan M Roden, Paul A Harris, and Joshua C Denny. Phekb : a catalog and workflow for creating electronic phenotype algorithms for transportability. *Journal of the American Medical Informatics Association*, 23(6) :1046–1052, 2016.

[Kuhn *et al.*, 2016] Michael Kuhn, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. The SIDER database of drugs and side effects. *Nucleic Acids Research*, 44(Database-Issue) :1075–1079, 2016.

[Kuznetsov and Samokhin, 2005] Sergei O. Kuznetsov and Mikhail V. Samokhin. Learning closed sets of labeled graphs for chemical applications. In *Inductive Logic Programming, 15th International Conference, ILP 2005, Bonn, Germany, August 10-13, 2005, Proceedings*, pages 190–208, 2005.

[Landrum *et al.*, 2014] Melissa J. Landrum, Jennifer M. Lee, George R. Riley, Wonhee Jang, Wendy S. Rubinstein, Deanna M. Church, and Donna R. Maglott. Clinvar : public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research*, 42(Database-Issue) :980–985, 2014.

[Lebo *et al.*, 2013] Timothy Lebo, Satya Sahoo, Deborah McGuinness, Khalid Belhajjame, James Cheney, David Corsar, Daniel Garijo, Stian Soiland-Reyes, Stephan Zednik, and Jun Zhao. PROV-O : The PROV Ontology. *W3C recommendation*, 30, 2013.

[Legrand *et al.*, 2018] Joël Legrand, Yannick Toussaint, Chedy Raïssi, and Adrien Coulet. Syntax-based Transfer Learning for the Task of Biomedical Relation Extraction. In *LOUHI 2018 - The Ninth International Workshop on Health Text Mining and Information Analysis*, Proceedings of LOUHI 2018 : The Ninth International Workshop on Health Text Mining and Information Analysis, Brussels, Belgium, October 2018.

[Legrand *et al.*, 2019] Joël Legrand, Romain Gogdemir, Cédric Bousquet, Kevin Dalleau, Marie-Dominique Devignes, William Digan, Chia-Ju Lee, Ndeye-Coumba Ndiaye, Nadine Petitpain, Patrice Ringot, Malika Smaïl-Tabbone, Yannick Toussaint, and Adrien Coulet. Pgxcorpus : a manually annotated corpus for pharmacogenomics. *bioRxiv*, 2019.

[Lehmann and *et al.*, 2015] Jens Lehmann and *et al.* DBpedia - A large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2) :167–195, 2015.

[Lehmann and Wille, 1995] Fritz Lehmann and Rudolf Wille. A triadic approach to formal concept analysis. *Conceptual structures : applications, implementation and theory*, pages 32–43, 1995.

[LePendu *et al.*, 2011] Paea LePendu, Mark A Musen, and Nigam H Shah. Enabling enrichment analysis with the human disease ontology. *Journal of biomedical informatics*, 44 :S31–S38, 2011.

[LePendu *et al.*, 2013] P LePendu, S V Iyer, A Bauer-Mehren, R Harpaz, J M Mortensen, T Podchiyska, T A Ferris, and N H Shah. Pharmacovigilance using clinical notes. *Clinical Pharmacology & Therapeutics*, 93(6) :547–555, 2013.

[Lieber *et al.*, 2006] Jean Lieber, Amedeo Napoli, Laszlo Szathmary, and Yannick Toussaint. First elements on knowledge discovery guided by domain knowledge (KDDK). In *Concept Lattices and Their Applications, Fourth International Conference, CLA 2006, Tunis, Tunisia, October 30 - November 1, 2006, Selected Papers*, pages 22–41, 2006.

[Lin and Chung, 2019] Bohan Lin and Wendy K. Chung. Cases in Precision Medicine : The Role of Pharmacogenetics in Precision Prescribing. *Annals of Internal Medicine*, 05 2019.

[Liu *et al.*, 2012] Yi Liu, Adrien Coulet, Paea LePendu, and Nigam H. Shah. Using ontology-based annotation to profile disease research. *J Am Med Inform Assoc.*, 19(e1), 2012.

[Lowe *et al.*, 2009] Henry J. Lowe, Todd A. Ferris, Penni M. Hernandez, and Susan C. Weber. STRIDE – an integrated standards-based translational research informatics platform. *AMIA Annu Symp Proc*, 2009 :391–395, 2009.

[Mandel *et al.*, 2016] JC Mandel, DA Kreda, K D Mandl, IS Kohane, and RB Ramoni. SMART on FHIR : a standards-based, interoperable apps platform for electronic health records. *JAMIA*, 23(5) :899–908, 2016.

[Masini *et al.*, 1989] G. Masini, A. Napoli, D. Colnet, D. Léonard, and K. Tombre. *Les langages à objets*. InterEditions, Paris, France, 1989.

[Mathers *et al.*, 2009] C.D. Mathers, G. Stevens, and M. Mascarenhas. Global health risks : Mortality and burden of disease attributable to selected major risks, 2009.

[McCray, 2003] Alexa T. McCray. An upper level ontology for the biomedical domain. *Comp. Funct. Genom.*, 4 :80–84, 2003.

[McCusker *et al.*, 2018] James P McCusker, John S. Erickson, Katherine Chastain, Sabbir Rashid, Rukmal Weerawarana, and Deborah L McGuinness. What is a knowledge graph ? *Pre-Print at http://www.semantic-web-journal.net/content/what-knowledge-graph*, 2018.

[Monnin *et al.*, 2017a] Pierre Monnin, Clement Jonquet, Joël Legrand, Amedeo Napoli, and Adrien Coulet. PGxO : A very lite ontology to reconcile pharmacogenomic knowledge units. In *Methods, tools & platforms for Personalized Medicine in the Big Data Era*, NETTAB 2017 Workshop Collection, Palermo, Italy, Oct 2017.

[Monnin *et al.*, 2017b] Pierre Monnin, Mario Lezoche, Amedeo Napoli, and Adrien Coulet. Using formal concept analysis for checking the structure of an ontology in LOD : the example of DBpedia. In *23rd International Symposium on Methodologies for Intelligent Systems, ISMIS 2017*, Warsaw, Poland, June 2017.

[Monnin *et al.*, 2017c] Pierre Monnin, Amedeo Napoli, and Adrien Coulet. Discovering Subsumption Axioms with Concept Annotation. Gestion de Données - Principes, Technologies et Applications (BDA 2017), November 2017. Poster.

[Monnin *et al.*, 2018a] Pierre Monnin, Amedeo Napoli, and Adrien Coulet. Combining Concept Annotation and Pattern Structures for Guiding Ontology Mapping. In *FCA4AI@IJCAI2018 - 6th International Workshop "What can FCA do for Artificial Intelligence ?"*, volume CEUR Workshop Proceedings of *Proceedings of the 6th International Workshop "What can FCA do for Artificial Intelligence" ? co-located with International Joint Conference on Artificial Intelligence and European Conference on Artificial Intelligence (IJCAI/ECAI 2018), Stockholm, Sweden, July 13, 2018*, Stockholm, Sweden, July 2018.

[Monnin *et al.*, 2018b] Pierre Monnin, Amedeo Napoli, and Adrien Coulet. Data-Interlinking : the Seed of Knowledge Reconciliation in Pharmacogenomics. working paper or preprint, 2018.

[Monnin *et al.*, 2019a] Pierre Monnin, Joël Legrand, Graziella Husson, Patrice Ringot, Andon Tchechmedjiev, Clément Jonquet, Amedeo Napoli, and Adrien Coulet. Pgxo and pgxlod : a reconciliation of pharmacogenomic knowledge of various provenances, enabling further comparison. *BMC Bioinformatics*, 20(S4) :139, 2019.

[Monnin *et al.*, 2019b] Pierre Monnin, Chedy Raïssi, Amedeo Napoli, and Adrien Coulet. Knowledge Reconciliation with Graph Convolutional Networks : Preliminary Results. In *DL4KG2019 - Workshop on Deep Learning for Knowledge Graphs*, volume CEUR Workshop Proceedings of *Proceedings of the Workshop on Deep Learning for Knowledge Graphs (DL4KG2019) Co-located with the 16th Extended Semantic Web Conference 2019 (ESWC 2019)*, Portoroz, Slovenia, June 2019.

[Muggleton, 1991] Stephen Muggleton. Inductive Logic Programming. *New Generation Computing*, 8(4) :295–318, 1991.

[Neuraz *et al.*, 2013] A Neuraz, L Chouchana, et al. Phenome-wide association studies on a quantitative trait : Application to TPMT enzyme activity and thiopurine therapy in pharmacogenomics. *PLoS Comp Bio.*, 9(12), 2013.

[Noy *et al.*, 2009] Natalya Fridman Noy, Nigam H. Shah, Patricia L. Whetzel, Benjamin Dai, Michael Dorf, Nicholas Griffith, and Clement Jonquet *et al.* Bioportal : ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research*, 37(Web-Server-Issue) :170–173, 2009.

[OHDSI Consortium Wiki, 2018] OHDSI Consortium Wiki. Patient-level prediction workgroup, 2018.

[Osborne *et al.*, 2009] John D Osborne, Jared Flatow, Michelle Holko, Simon M Lin, Warren A Kibbe, Lihua Julie Zhu, Maria I Danila, Gang Feng, and Rex L Chisholm. Annotating the human genome with disease ontology. *BMC genomics*, 10(1) :S6, 2009.

[Pearl, 2010] Judea Pearl. Causal inference. In *Causality : Objectives and Assessment (NIPS 2008 Workshop), Whistler, Canada, December 12, 2008*, pages 39–58, 2010.

[Percha and Altman, 2013] Bethany Percha and Russ B Altman. Inferring the semantic relationships of words within an ontology using random indexing : applications to pharmacogenomics. In *AMIA Annual Symposium Proceedings*, volume 2013, page 1123. American Medical Informatics Association, 2013.

[Percha *et al.*, 2012] B Percha, Y Garten, and RB. Altman. Discovery and explanation of drug-drug interactions via text mining. In *PSB 2012, Pacific Symposium on Biocomputing*, pages 16–31, 2012.

[Personeni *et al.*, 2014] Gabin Personeni, Simon Daget, Céline Bonnet, Philippe Jonveaux, Marie-Dominique Devignes, Malika Smaïl-Tabbone, and Adrien Coulet. Mining Linked Open Data : A Case Study with Genes Responsible for Intellectual Disability. In *10th International Conference on Data Integration in the Life Sciences, DILS 2014*, volume 8574 of *LNCS series*, pages 16 – 31. Springer, July 2014.

[Personeni *et al.*, 2017] Gabin Personeni, Emmanuel Bresso, Marie-Dominique Devignes, Michel Dumontier, Malika Smaïl-Tabbone, and Adrien Coulet. Discovering associations between adverse drug events using pattern structures and ontologies. *Journal of Biomedical Semantics*, 8(16), 2017.

[Personeni *et al.*, 2018] Gabin Personeni, Marie-Dominique Devignes, Malika Smaïl-Tabbone, Philippe Jonveaux, Céline Bonnet, and Adrien Coulet. Cooperation of bio-ontologies for the classification of genetic intellectual disabilities : a diseasome approach. In *SWAT4HCLS 2018 - 11th International SWAT4HCLS Conference Semantic Web Applications and Tools for Healthcare and Life Sciences*, Antwerp, Belgium, December 2018.

[Personeni, 2018] Gabin Personeni. *Apport des ontologies de domaine pour l'extraction de connaissances à partir de données biomédicales. (Contribution of domain ontologies for knowledge discovery in biomedical data)*. PhD thesis, University of Lorraine, Nancy, France, 2018.

[PharmGKB web page, 2019] PharmGKB web page. Levels of evidence of annotations [visited sept 24, 2019] : `https://www.pharmgkb.org/page/clinAnnLevels`. 2019.

[Piñero *et al.*, 2015] Janet Piñero, Núria Queralt-Rosinach, Àlex Bravo, Jordi Deu-Pons, Anna Bauer-Mehren, Martin Baron, Ferran Sanz, and Laura Inés Furlong. Disgenet : a discovery platform for the dynamical exploration of human diseases and their genes. *Database*, 2015, 2015.

[Ratner *et al.*, 2017] Alexander Ratner, Stephen H. Bach, Henry R. Ehrenberg, Jason Alan Fries, Sen Wu, and Christopher Ré. Snorkel : Rapid training data creation with weak supervision. *CoRR*, abs/1711.10160, 2017.

[Relling and Klein, 2011] MV Relling and TE Klein. Cpic : Clinical pharmacogenetics implementation consortium of the pharmacogenomics research network. *Clinical Pharmacology & Therapeutics*, 89(3) :464–467, 2011.

[Ribeiro *et al.*, 2016] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you ?" : Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144, 2016.

[Rigdon *et al.*, 2018] Joseph Rigdon, Michael Baiocchi, and Sanjay Basu. Preventing false discovery of heterogeneous treatment effect subgroups in randomized trials. *Trials*, 19(1) :382, Jul 2018.

[Roden *et al.*, 2018] Dan M. Roden, Sara L Van Driest, Jonathan D. Mosley, Quinn S. Wells, Jamie R. Robinson, Joshua C. Denny, and Josh F. Peterson. Benefit of preemptive pharmacogenetic information on clinical outcome. *Clinical Pharmacology & Therapeutics*, 103(5), 2018.

[Rubin, 2005] Donald B. Rubin. Causal inference using potential outcomes : Design, modeling, decisions. *Journal of the American Statistical Association*, 100 :322–331, 2005.

[Saigo *et al.*, 2009] Hiroto Saigo, Sebastian Nowozin, Tadashi Kadowaki, Taku Kudo, and Koji Tsuda. gboost : a mathematical programming approach to graph classification and regression. *Machine Learning*, 75(1) :69–89, 2009.

[Samwald *et al.*, 2011] Matthias Samwald, Anja Jentzsch, Christopher Bouton, Claus Stie Kallesoe, Egon Willighagen, Janos Hajagos, M Scott Marshall, Eric Prud'hommeaux, Oktie Hassenzadeh, Elgar Pichler, and Susie Stephens. Linked open drug data for pharmaceutical research and development. *Journal of Cheminformatics*, 3 :19, May 2011.

[Saïs, 2007] Fatiha Saïs. *Intégration sémantique de données guidée par un ontologie*. PhD thesis, Université Paris-Sud, Orsay, France, 2007.

[Schaffert *et al.*, 2009] Sebastian Schaffert, Julia Eder, Szaby Grünwald, Thomas Kurz, and Mihai Radulescu. Kiwi - A platform for semantic social software (demonstration). In *The Semantic Web : Research and Applications, 6th European Semantic Web Conference, ESWC 2009, Heraklion, Crete, Greece, May 31-June 4, 2009, Proceedings*, pages 888–892, 2009.

[Schreiber and Akkermans, 2000] Guus T. Schreiber and Hans Akkermans. *Knowledge Engineering and Management : The CommonKADS Methodology*. MIT Press, Cambridge, MA, USA, 2000.

[Shah *et al.*, 2019] Nigam H. Shah, Arnold Milstein, and Steven C. Bagley, PhD. Making Machine Learning Models Clinically Useful. *JAMA*, 322(14) :1351–1352, 10 2019.

[Shi *et al.*, 2011] Lian Shi, Yannick Toussaint, Amedeo Napoli, and Alexandre Blansché. Mining for reengineering : An application to semantic wikis using formal and relational concept analysis. In Grigoris Antoniou, Marko Grobelnik, Elena Simperl, Bijan Parsia, Dimitris Plexousakis, Pieter De Leenheer, and Jeff Pan, editors, *The Semanic Web : Research and Applications*, pages 421–435, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[Srinivasan, 2007] Ashwin Srinivasan. The Aleph Manual. Available at http ://www.comlab.ox.ac.uk/oucl/research/areas/machlearn/Aleph/, 2007.

[Steinhaus, 1956] Hugo Steinhaus. Sur la division des corp materiels en parties. *Bull. Acad. Polon. Sci*, 1(804) :801, 1956.

[Stewart *et al.*, 2007] Walter F. Stewart, Nirav R. Shah, Mark J. Selna, Ronald A. Paulus, and James M. Walker. Bridging the inferential gap : The electronic health record and clinical evidence. *Health Affairs*, 26(2) :w181–w191, 2007.

[Subramanian *et al.*, 2005] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set enrichment analysis : A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43) :15545–15550, 2005.

[Tchechmedjiev *et al.*, 2018] Andon Tchechmedjiev, Amine Abdaoui, Vincent Emonet, Soumia Melzi, Jitendra Jonnagaddala, and Clement Jonquet. Enhanced Functionalities for Annotating and Indexing Clinical Text with the NCBO Annotator+. *Bioinformatics*, 34(11) :1962–1965, June 2018.

[Tian *et al.*, 2014] Lu Tian, Ash A. Alizadeh, Andrew J. Gentles, and Robert Tibshirani. A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, 109(508) :1517–1532, 2014. PMID : 25729117.

[U.S. Department of Health and Human Services, 2014] U.S. Department of Health and Human Services. National action plan for adverse drug event prevention. 2014.

[U.S. Food & Drug Administration, 2018 visitée le 17102019] U.S. Food & Drug Administration. Fda adverse event reporting system, 2018, (visitée le 17/10/2019).

[U.S. National Library of Medicine, 2018] U.S. National Library of Medicine. Introduction to mesh, 2018.

[Valiant, 2003] Leslie G. Valiant. Three problems in computer science. *J. ACM*, 50(1) :96–99, 2003.

[van Krieken *et al.*, 2019] E. van Krieken, E. Acar, and F. van Harmelen. Semi-supervised learning using differentiable reasoning. *FLAP*, 6(4) :633–652, 2019.

[Vasudevan and Ginzler, 2009] Archana R. Vasudevan and Ellen M. Ginzler. Established and novel treatments for lupus. *The Journal of Musculoskeletal Medicine*, 26, 2009.

[Volpi *et al.*, 2018] Simona Volpi, Carol Bult, Rex Chisholm, Patricia Deverka, Geoffrey Ginsburg, Howard Jacob, Melpomeni Kasapi, Howard McLeod, Dan Roden, Marc Williams, Eric Green, Laura Rodriguez, Samuel Aronson, Larisa Cavallari, Joshua Denny, Lynn Dressler, Julie Johnson, Teri Klein, J Steven Leeder, and Mary Relling. Research directions in the clinical implementation of pharmacogenomics : An overview of us programs and projects. 103(5) :778–786, 2018.

[Voss *et al.*, 2015] EA Voss, R Makadia, A Matcho, Q Ma, C Knoll, et al. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. *JAMIA*, 22(3) :553–564, 2015.

[Wager and Athey, 2018] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523) :1228–1242, 2018.

[Wagner *et al.*, 2016] Alex H. Wagner, Adam C. Coffman, Benjamin J. Ainscough, Nicholas C. Spies, Zachary L. Skidmore, Katie M. Campbell, Kilannin Krysiak, Deng Pan, Joshua F. McMichael, James M. Eldred, Jason R. Walker, Richard K. Wilson, Elaine R. Mardis, Malachi Griffith, and Obi L. Griffith. Dgidb 2.0 : mining clinically relevant drug-gene interactions. *Nucleic Acids Research*, 44(Database-Issue) :1036–1044, 2016.

[Weinshilboum and Wang, 2017] Richard M Weinshilboum and Liewei Wang. Pharmacogenomics : Precision medicine and drug response. *Mayo Clinic proceedings*, 92(11) :1711–1722, 2017.

[Whirl-Carrillo *et al.*, 2012] M. Whirl-Carrillo, E.M. McDonagh, J. M. Hebert, L. Gong, K. Sangkuhl, C.F. Thorn, R.B. Altman, and T.E. Klein. Pharmacogenomics knowledge for personalized medicine. *Clinical Pharmacology & Therapeutics*, 92(4) :414–7, 2012.

[WHO Collaborating Centre for Drug Statistics Methodology, 2018] WHO Collaborating Centre for Drug Statistics Methodology. Atc structure and principles, 2018.

[Wille, 2002] Rudolf Wille. Why can concept lattices support knowledge discovery in databases ? *Journal of Experimental & Theoretical Artificial Intelligence*, 14(2-3) :81–92, 2002.

[Wishart *et al.*, 2008] David S. Wishart, Craig Knox, Anchi Guo, Dean Cheng, Savita Shrivastava, Dan Tzur, Bijaya Gautam, and Murtaza Hassanali. Drugbank : a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Research*, 36(Database-Issue) :901–906, 2008.

[Xu *et al.*, 2007] Rong Xu, Yael Garten, Kaustubh S. Supekar, Amar K. Das, Russ B. Altman, and Alan M. Garber. Extracting subject demographic information from abstracts of randomized clinical trial reports. In *MedInfo*, volume 129 of *Studies in Health Technology and Informatics*, pages 550–4. IOS Press, 2007.

[Yan and Han, 2002] Xifeng Yan and Jiawei Han. gspan : Graph-based substructure pattern mining. In *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*, pages 721–724. IEEE, 2002.

[Zeng *et al.*, 2007] Kelly Zeng, Olivier Bodenreider, John Kilbourne, and Stuart Nelson. Rxnav : Towards an integrated view on drug information. In *Medinfo*, page 386, 2007.

[Zineh *et al.*, 2013] Issam Zineh, Michael Pacanowski, and Janet Woodcock. Pharmacogenetics and coumarin dosing ? recalibrating expectations. *N Engl J Med*, 369 :2273–5, 2013.